World Wide Web (2023) 26:3861–3890
https://doi.org/10.1007/s11280-023-01209-4

A semi-supervised framework for concept-based hierarchical document clustering

Seyed Mojtaba Sadjadi[1] · **Hoda Mashayekhi**[1] · **Hamid Hassanpour**[1]

Received: 30 July 2022 / Revised: 21 August 2023 / Accepted: 30 August 2023 /
Published online: 2 October 2023

Abstract

Text clustering is used in various applications of text analysis. In the clustering process, the employed document representation method has a significant impact on the results. Some popular document representation methods cannot effectively maintain the proximity information of the documents or suffer from low interpretability. Although the concept-based representation methods overcome these challenges to some extent, the existing semi-supervised document clustering methods rarely use this type of document representation. In this paper, we propose a concept-based semi-supervised framework for document clustering that uses both labeled and unlabeled data to yield a higher clustering quality. Concepts are composed of a set of semantically similar words. We propose the notion of semi-supervised concepts to benefit from document labels in extracting more relevant concepts. We also propose a new method of clustering documents based on the weights of such concepts. In the first and second steps of the proposed framework, the documents are represented based on the concepts extracted from the set of embedded words in the corpus. The proposed representation is interpretable and preserves the proximity information of documents. In the third step, the semi-supervised hierarchical clustering process utilizes unlabeled data to capture the overall structure of the clusters, and the supervision of a small number of labeled documents to adjust the cluster centroids. The use of concept vectors improves the process of merging clusters in the hierarchical clustering approach. The proposed framework is evaluated using the Reuters, 20-NewsGroups and WebKB text collections, and the results reveal the superiority of the proposed framework compared to several existing semi-supervised and unsupervised clustering approaches.

Keywords Semi-supervised clustering · Document clustering · Word embedding · Concept-based representation · Hierarchical clustering

✉ Hoda Mashayekhi
hmashayekhi@shahroodut.ac.ir

Seyed Mojtaba Sadjadi
sadjadi@shahroodut.ac.ir

Hamid Hassanpour
h.hassanpour@shahroodut.ac.ir

[1] Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

1 Introduction

Nowadays, web and social media have become the main sources of large amounts of text data owing to the widespread surge of such data over the Internet, such as social media feeds, news websites, learning forums, and electronic book libraries [1]. In order to manage and organize a large number of documents, document clustering may be used. Document clustering is a technique for partitioning a set of documents into distinct clusters mainly based on content similarity [2]. Clustering is an essential way to learn without supervision and discover the inherent structure of unlabeled data [3]. This technique facilitates comprehensive analysis in the field of text mining in different applications such as recommender systems [4–6], disease detection and categorization [7], spam detection [8–10], information retrieval [11], topic modeling [12], time series analysis [13], sentiment classification [14] and text summarization [15].

The three principal steps of a document clustering algorithm are text preprocessing, document representation, and clustering. The clustering performance largely relies on the quality of document representation, in which the raw documents are commonly converted into numerical feature vectors to make various clustering algorithms applicable. The document-representation quality can be assessed from two aspects: semantic quality and statistical quality [16]. Semantic quality refers to the interpretability of the feature vector and the degree to which it describes the content of the document. The most popular document representation method, known for its intuitive and simple interpretability, is the BoW model, which represents a document by a feature vector based on its word frequencies [17]. Although BoW is simple to interpret, it suffers from excessive dimensionality, and when the number of unique words increases, cannot retain the proximity information. Statistical quality assessment means how much the feature vector can distinguish documents from each other in different categories. Some methods, such as convolutional neural networks (CNNs) [18], recurrent neural networks (RNNs) [19], and Doc2Vec [20], create low dimensional vectors to represent documents that also preserves the proximity information. Nevertheless, the obtained representations are complicated to interpret because the value of each feature is computed through complicated structures of the neural network weights. To overcome the weakness of previous methods of document representation, another approach is recently introduced, which leverages concepts [21]. After embedding the words, concepts are extracted by clustering the word vectors, and documents are represented based on concepts. Therefore, the document vectors have reasonable dimensions.

The use of labeled data to improve the quality of clustering is increasingly attracting more attention [22]. The main idea is that the unlabeled data capture the overall structure of the clusters and a small number of labeled data adjust the centroid of clusters. This technique utilizes both labeled and unlabeled data to obtain a better clustering model and is referred to as semi-supervised clustering [23]. In general, semi-supervised clustering methods are divided into two categories: similarity-based and search-based approaches. In similarity-based algorithms, the underlying similarity criterion is adapted according to the constraints and labels of the supervised data. The methods proposed by Zhang et al. [24] assume that a mixture model generates the data population. In these methods, the unlabeled data is labeled and used to train the model. However, it is not clear how much data re-labeling is required and how reliable this re-labeling will be. In search-based methods, the semi-supervised clustering algorithm is modified to use labels or restrictions provided by the user to search for the proper partition bias. The TESC semi-supervised clustering [25] uses a search-based clustering approach to classify texts. The primary purpose of this method is to improve the quality of classification, not the clustering. It uses semi-supervised clustering to identify the components of the corpus and leverage these components to predict labels for unlabeled data.

In this paper, we propose a novel semi-supervised document clustering method that, unlike some current popular semi-supervised clustering algorithms [23, 24], uses both unlabeled data and a restricted set of labeled data simultaneously in the clustering process. To the best of our knowledge, none of the current semi-supervised clustering methods uses the conceptual representation of documents. This representation has the benefit of putting documents with common concepts in one cluster, unlike other representation methods that construct the document vectors directly based on the word occurrences. It can capture the underlying characteristics and distinctive features of documents, while preserving their interpretability. It has the advantage of maintaining the proximity information of the documents, and as shown in the experiments, significantly improving the clustering quality.

After embedding the words in the corpus, its components are identified by extracting concepts from the words, and the documents are represented based on the defined concepts. This representation of the documents is used in the semi-supervised clustering process. The proposed method can arbitrarily use the labeled data in the concept extraction phase or the clustering phase. Although document representation based on concepts is previously adopted by some studies outside the context of semi-supervised clustering, in this paper we propose the novel idea of semi-supervised concepts for document representation. The semi-supervised concepts capture both the semantic components of the corpus and their correspondence to the class labels of data. The proposed model is explicable, providing a deeper understanding of the corpus and a more transparent operation logic for reasoning. In addition, as the documents are expressed as a collection of labeled concepts, an intuitive understanding of the comprising semantic components along with their label correspondence is achieved. In the semi-supervised clustering phase, the unlabeled documents are used to identify the overall structure of the clusters and, a small number of labeled documents are utilized to more precisely determine the centroids.

We conduct a comprehensive evaluation of the proposed method using three well-known benchmark datasets and compare the proposed approach from to multiple baseline and state-of-the-art algorithms. The evaluation results show that the proposed method has a significant advantage over previous semi-supervised document clustering methods. The main contributions of this paper are as follows:

- Proposing a novel semi-supervised document clustering framework based on the conceptual representation of the documents.
- Proposing approaches to involve a limited set of labeled documents in either the concept extraction or the document clustering phases.
- Proposing the notion of semi-supervised concepts for document representation.
- Conducting a comprehensive evaluation for analyzing and comparing the proposed methods.

The rest of the paper is organized as follows. In Section 2, we review current studies related to both document representation and semi-supervised clustering. Research Objectives are introduced in Section 3. Our proposed semi-supervised document clustering framework is presented in Section 4. Experimental results, detailed analysis, and discussion are presented in Section 5. Ultimately, our work is concluded in Section 6.

2 Related work

2.1 Document representation

Word embedding [26] is a set of language modeling techniques in natural language processing where semantic relationships of words can be captured in a low-dimensional and dense vector space. One of these techniques is the Word2Vec [27] method which is based on the assumption that words and terms that occur in similar contexts tend to have similar meanings. Le et al. [20] introduced the Doc2Vec model, which uses textual information from words and documents mutually to learn the representation of documents in a continuous vector space. Research has shown that Doc2Vec is more efficient than Word-2Vec in solving clustering and classification problems [28]. In addition, due to the smaller dimensions of the generated document vectors, it is more efficient than BoW. Nevertheless, Doc2Vec has low interpretability, and the logic behind the generation procedure of document vectors is unclear.

In this paper, we represent documents based on concepts. A summary of studies using conceptual factorization for document representation is provided in [29]. Kim et al. [21] proposed the Bag-of-Concepts (BoC), which BoC creates concepts by clustering the word vectors generated by word2vec. It then creates the document vector using the frequencies of concepts in the documents. BoC is an unsupervised method that does not provide a specific solution for clustering and classification applications. Lee et al. [30] introduced concept-based representation which derives the conceptual knowledge from an external knowledge base. They also reduce the concept ambiguity by clustering concepts in an attempt to enhance BoC. Lou et al. [31] proposed a concept-based scheme for clustering and visualization of biomedical documents which the concept embedding is learned through neural networks. Another study [32] presents a decomposition method that generates concept vectors by identifying semantic word communities in a weighted word co-occurrence network extracted from the short text collection. Based on the idea that each entity may have different aspects reflected in different documents, a representation scheme is proposed in [33]. Each entity is modeled through different aspects, where each aspect consists of a mixture of latent topics. The Bag-of-Senses model [34] is based on the assumption that a document is a set of senses of words, and the senses are considered instead of the words. The sense of a word is estimated based on the documents in which it occurs. In the text analysis applications, the performance of feature-based techniques can be significantly hampered by the word mismatch and ambiguity problems. As a solution, in [35] a concept-based approach is proposed using domain-specific ontologies to support automated document classification. In [36], a conceptualization method using a Tagged Bag-of-Concepts (TBoC) is presented to detect sentiment in short texts.

To the best of our knowledge, no previous study aimed at clustering documents using a semi-supervised approach has adopted and analyzed the concept representation of documents. In addition, in contrast to previous studies, we propose a semi-supervised method of concept extraction to generate labeled concepts.

2.2 Semi-supervised clustering

Semi-supervised clustering is developed as an alternative to conventional unsupervised methods where partial domain knowledge is employed in the clustering process to improve its performance. A comprehensive survey of some semi-supervised clustering algorithms is provided by Basu et al. [37].

Inspired by the purpose of label propagation for semi-supervised learning, Zheng et al. [38] developed a method that involves label prediction for unlabeled data. The combination of Naive Bayes and EM algorithms (NBEM) is also adopted for semi-supervised clustering [24]. The model iteratively labels the unlabeled data and employs this newly labeled data to re-train the model. Zhang et al. [25] developed a method called TESC for text classification using semi-supervised clustering. TESC assumes that the data set is composed of different components and uses a clustering process to capture these components.

To improve the clustering performance of documents with supervisory information, a semi-supervised approach to the factorization of concepts is proposed by Lu et al.[39]. They incorporate the pairwise constraints of penalty and rewards in the concept factorization, which can ensure that data points regarding a cluster in the main space are still in the same cluster in the transformed space. In another study, text clustering utilizing automatic generation constraints is used for document classification [40]. The clustering algorithm allows obtaining a set of must-link/cannot-link constraints that can be used in the semi-supervised clustering step. Next, these constraints are used as semi-supervision in the hierarchical clustering algorithm.

Li et al. [41] proposed a distributed semi-supervised clustering method. The clustering process is the same as TESC, except that distributed clustering and collecting the results of the sub-clusters is applied. In [42] a new method of selecting the pairwise constraints from unlabeled data for semi-supervised clustering of documents is presented. A dense data group is selected from each initial cluster, and in these dense groups, the most informative data are identified by the local density estimation method. The identified data are used to form a set of constraint pairs in semi-supervised clustering. In the work of Gan et al. [43], the basic idea is that when the label of a labeled sample is risky, the predictions of the labeled sample and the nearest homogeneous unlabeled samples should be similar. This is performed by unsupervised clustering and then building a local graph to model the relationships between labeled and nearest unlabeled samples.

In [44], a bag of phrases is used to classify texts which incorporate phrases into the vector space model for the document classification task. The Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA) is used to separate phrases from the corpus. In [45], an embedding-based generative framework is used for semi-supervised text categorization based on the integration of labels, metadata, and text. In [46], a text clustering method based on a deep learning approach is proposed that combines (i) a Convolutional Siamese Network (CSN) based on pair constraints for representation learning, and (2) the traditional K-Means algorithm for clustering the learned vectors without supervision. In [47], a semi-supervised approach is proposed to address the costs of labeling unlabeled data by combining a dynamic graph with a self-learning mechanism. In [48], a new selection criterion is proposed using the neighborhood construction algorithm for semi-supervised learning. Unlabeled data are selected close to the decision boundary, and to determine the correct labels for these data points, an agreement is formed between the classifier predictions and the neighborhood construction algorithm. In [49] a semi-supervised method based on deep learning is presented. SDEC learns the features that enhance the clustering tasks. It incorporates pairwise constraints in the feature learning process, such that data samples belonging to the same cluster are close together and data samples belonging to different clusters are far apart in the learned feature space. In [50], a discriminative semi-supervised NMF (DSSNMF) method is proposed which uses the partial label information as a discriminative constraint. The DSSNMF method is investigated with two different cost functions and provides relevant update rules for the optimization problems. A semi-supervised clustering approach based on deep metric learning is presented in [51]. To dynamically update unlabeled data to labeled data, they combine embedding with label propagation strategies and triplet loss with deep metric learning networks.

Although we also make simultaneous use of labeled and unlabeled data, our approach is different from previous approaches. In our proposed method and evaluations, a large amount of the data is unlabeled, and a limited number of labeled data may be used. This difference significantly reduces the cost of labeling data in real-world applications. Furthermore, most of the mentioned semi-supervised clustering methods neglect the document representation issue, which can largely affect the clustering outcome. In this paper, other than using the concept-based document representation in semi-supervised clustering, we propose the notion of labeled concepts generated through semi-supervised concept extraction.

3 Research objectives

This paper proposes a novel semi-supervised clustering method for documents based on their conceptual representation. We assume that the input document collection $D = \left\{D^l, D^u\right\}$ is divided into labeled (D^l) and unlabeled (D^u) documents. A labeled document d has a corresponding label belonging to the set of all labels L ($label(d) \in L$). Each document is composed of a set of words. The union of all words from all documents is denoted as W. We aim to reach a clustering model $C = \left\{C_1, \dots, C_m\right\}$ of the documents, such that $\bigcup_{1 \leq i \leq m} C_i = D$ and $C_i \cap C_j = \emptyset (1 \leq i \neq j \leq m)$.

The main purpose of this study is to present a new framework for semi-supervised clustering of the document collection D. The method of document representation plays an important role in text clustering, and this is further investigated in the experiments. The proposed method is based on conceptual representation of documents, which offers a low-dimensional representation while reflecting the proximity information of documents. Using the concepts extracted from the set of words W, documents are represented as concept vectors. Because tagging unlabeled data is costly, it can be tedious in real-world text analysis applications. One of the principal objectives of this study is to cluster unlabeled textual data using a limited number of labeled data D^l. For this purpose, depending on whether labeled data are used in the concept extraction or document clustering phases, we offer two approaches for semi-supervised clustering, both of which require a small ratio of labeled data. In the former approach, we propose the notion of semi-supervised concepts, which simultaneously captures the intrinsic subjects in the collection, and their label associations. We use an agglomerative hierarchical clustering algorithm to extract the semi-supervised concepts and the document clustering in the former and latter approaches, respectively.

4 The proposed semi-supervised clustering framework

Figure 1 shows the detailed process of the proposed framework for semi-supervised clustering, described in terms of three phases: preprocessing, document representation, and clustering. The representation of documents is based on the concepts extracted from the corpus. The proposed clustering model is semi-supervised, benefiting from labeled documents to improve the clustering quality. Based on whether we want to use the labels in the representation phase

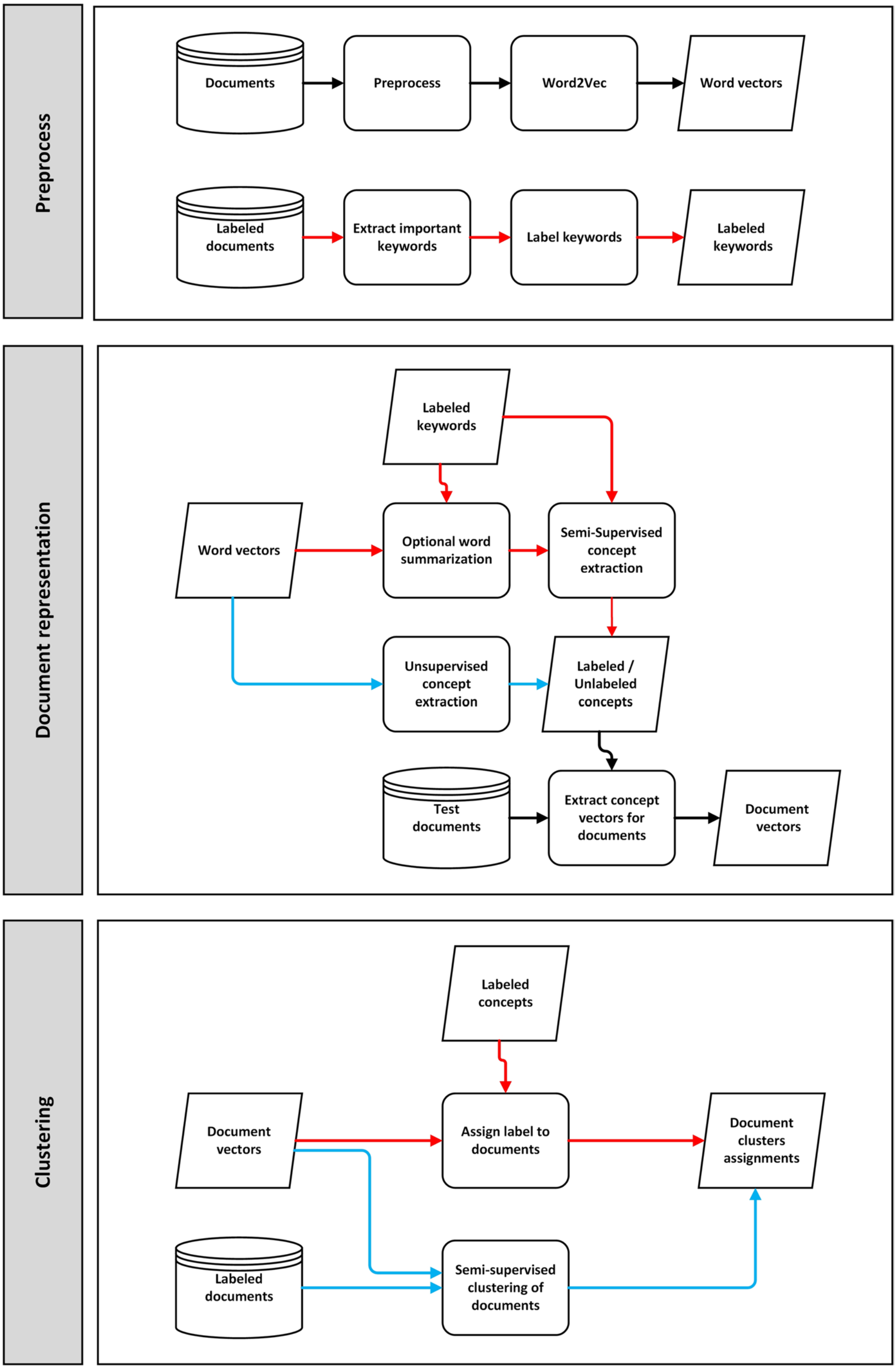

Preprocess
Documents
Preprocess
Word2Vec
Word vectors
Labeled documents
Extract important keywords
Label keywords
Labeled keywords
Document representation
Labeled keywords
Word vectors
Optional word summarization
Semi-Supervised concept extraction
Unsupervised concept extraction
Labeled / Unlabeled concepts
Test documents
Extract concept vectors for documents
Document vectors
Clustering
Labeled concepts
Document vectors
Assign label to documents
Document clusters assignments
Labeled documents
Semi-supervised clustering of documents

Fig. 1 The overall process of the proposed semi-supervised document clustering framework. The two methods proposed in this paper are presented in two different colours: SSConE in red, and SSClusE in blue. The common steps of the two are marked in black

Table 1 Description of the symbols used in the algorithms

Symbol	Description
D, d	The set of documents, a document
D^l, D^u	The set of labeled, unlabeled documents
$\vec{d}, \vec{d}^j$	The concept vector of d_i, the j^{th} component of the vector
L, l	The set of labels, a label
W, w	The set of words in the corpus, a word
W^l	A set of words with labels
T, T_i	The set of concepts, a concept
C, C_i	The set of clusters, a cluster
n, m, z	Number of documents, clusters, concepts
i, j, k, p	Used for indexing

or only in the clustering phase, two approaches are proposed. In the former, the document labels can be utilized in concept modeling, leading to the Semi-Supervised Concept Extraction (SSConE) approach. In the latter, the labels are used in the actual document clustering, leading to the Semi-Supervised Cluster Extraction (SSClusE) approach. Nevertheless, both methods yield a final clustering model of the documents. Table 1 presents the set of symbols used in the algorithm description. In the following subsections, we describe the three phases of the proposed method in detail.

4.1 Preprocessing

Initially, documents are tokenized after eliminating stop-words and preprocessing the texts to obtain a set of words W. The proposed algorithm requires that each word in the set W is represented with a vector in an embedded space. The embedding preserves the semantic relationships between the words. In the experiments, we use two popular word embedding models, namely Word2Vec [52] and BERT [53] to learn word associations from the input corpus.

If the documents labels are to be used in for concept modeling, a set of labeled words W^l are required. To this end, a limited set of more discriminating words W^l are extracted from the set of labeled documents D^l, and labeled as follows. The TF-IDF model is used to assign a weight to each word w in each labeled document as calculated in Eq. 1, where $tf_{w,d}$ denotes the number of occurrences of w in document d, and df_w is the number of labeled documents containing w. A limited set of words with the highest weights are selected, and further labeled according to the document in which they occur. To simplify the algorithm description, the unlabeled words are denoted as W^u, such that $W^l \cup W^u = W$. The labeled words are later used in the representation phase for semi-supervised concept extraction.

$$TF - IDF(w, d) = tf_{w,d} \times \log(\frac{|D^l|}{df_w}) \tag{1}$$

4.2 Representation

The representation of documents is based on the concepts extracted from the corpus. In the representation phase, we first extract a set of concepts $T = \{T_1, \dots, T_z\}$ from the set of words W, such that each concept consists of an exclusive set of words, i.e. $T_i \cap T_j = \emptyset (1 \leq i \neq j \leq z)$ and $\bigcup_{1 \leq i \leq z} T_i = W$. The general procedure of concept extraction involves executing a clustering algorithm on the set of words W to partition it into several clusters, each representing a concept. The concepts may be extracted in a semi-supervised or unsupervised approach, as described next. The former requires the set of labeled words W^l. After constructing the concepts, each document d is represented by a concept vector $\vec{d}$.

4.2.1 Semi-supervised concept extraction

The semi-supervised concept extraction employs a semi-supervised hierarchical clustering algorithm executed on the set of words $W^l \cup W^u$. Each resulting cluster is considered as a concept that may have a label based on the underlying words. The same procedure described here is later used in the semi-supervised clustering of documents in Section 4.3.2. In the clustering procedure, unlabeled data are used to capture the overall structure of data ingredients, while the labeled data are utilized to adjust the centroids of text ingredients.

Figure 2 shows the steps of semi-supervised clustering executed on a set of data points X to obtain a set of clusters P. To use a consistent notation; we introduce a dummy label U assigned to the unlabeled data so that all data points have a label. Initially, A matrix is formed to store the distance between the data. Each data point x_i is considered a primary cluster S_i with the same label of x_i. The main clustering loop executes a maximum of $|X|$ rounds until the primary set of clusters has at most one member. Among all pairs of primary clusters, the two closest clusters, S_i and S_j (two clusters with the smallest cosine distance between their centroids), are selected. Next, based on the labels of primary clusters, either a new cluster is created (mergeable), or the two clusters will be preserved (not mergeable). If neither cluster has the label U and $label(S_i) \neq label(S_j)$, the two clusters are preserved and added to the final partitioning (P). In the other cases, the two clusters are merged into a new cluster S_k which replaces the selected clusters in the primary cluster set. The label of S_k is the same as the label of S_i and S_j if they have similar labels, or is the same as the one whose label is not U. The algorithm iterates by selecting pairs of primary clusters. Final clusters with members below a threshold are removed as outliers (remove noise clusters from P). The final clusters preserve their labels and represent the labeled concepts. The number of clusters (m) is determined by the clustering algorithm and depends on the shape and number of input data. Hence the number of concepts (z) and consequently the length of the document vectors ($\left|\vec{d}\right|$) is not predetermined and is variable.

As the set W is large, and the complexity of the described semi-supervised clustering algorithm is $O(|W|^3)$, the words may be optionally summarized before extracting the concepts to reduce the complexity (Optional word summarization in Fig. 1). To this end, the Spherical K-means [54] clustering is used to cluster W. The resulting clusters are further inspected. Any cluster containing words belonging to more than one label is divided such that the resulting clusters are pure in terms of word labels. The cluster centroids are labeled according to their members and form the new pseudo labeled words $\widehat{W}^l$ as the input of the semi-supervised clustering algorithm.

4.2.2 Unsupervised concept extraction

If labeled data is not used in the concept extraction, the unsupervised spherical K-Means clustering algorithm utilizing the cosine distance is used to cluster W. The

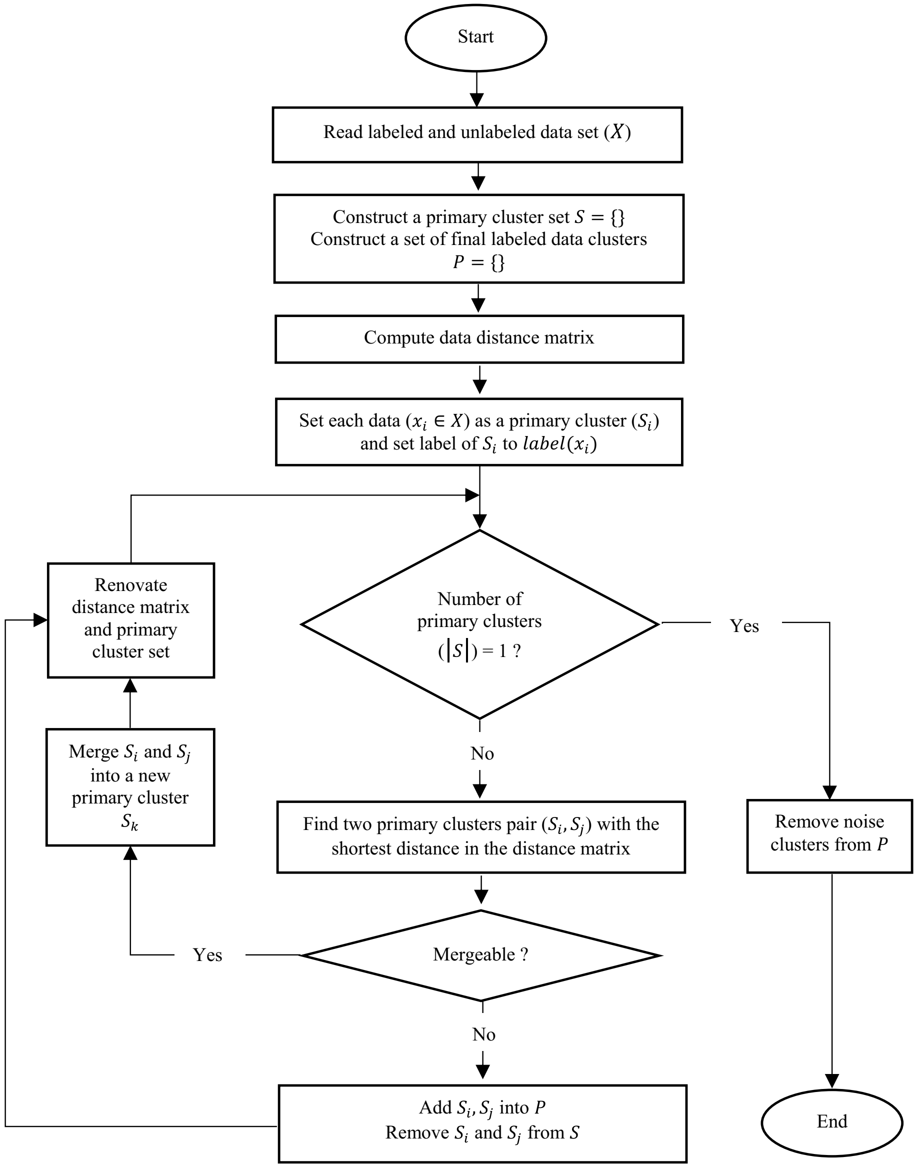

Start
Read labeled and unlabeled data set (X)
Construct a primary cluster set $S = \{\}$
Construct a set of final labeled data clusters $P = \{\}$
Compute data distance matrix
Set each data ($x_i \in X$) as a primary cluster (S_i) and set label of S_i to $label(x_i)$
Number of primary clusters ($|S|$) = 1 ?
Yes
No
Find two primary clusters pair (S_i, S_j) with the shortest distance in the distance matrix
Mergeable ?
Yes
No
Merge S_i and S_j into a new primary cluster S_k
Renovate distance matrix and primary cluster set
Add S_i, S_j into P
Remove S_i and S_j from S
Remove noise clusters from P
End

Fig. 2 Semi-supervised clustering process

resulting clusters are considered as unlabeled concepts extracted from the words. Spherical K-Means is an unsupervised clustering method and requires a predetermined value for the number of clusters for clustering operations. For this reason, the number of concepts extracted by this approach has a fixed and predetermined count.

4.2.3 Document vector extraction

Clusters created by semi-supervised clustering or unsupervised clustering are counted as concepts, and document vectors will be constructed from these concepts. Due to the semantic space trained by embedding method, words with similar meanings or common hypernyms are grouped in a cluster. Consequently, each word in the text will be considered as a member of a concept, and each document can be considered as a collection of concepts. As a concept that occurs in many documents is not a proper discriminator for machine learning applications [55], so Concept Frequency-Inverse Document Frequency (CF-IDF) [17] (Eq. 2) is applied to the generated concepts to eliminate the adverse effects of common concepts between documents. In Eq. 2, $cf_{c,d}$ denotes the number of occurrences of concept c in document d, and df_c is the number of documents containing concepts c.

$$CF - IDF(c,d) = cf_{c,d} \times \log \frac{|D|}{df_c} \tag{2}$$

The length of the document vector may vary based on the extracted concepts. In semi-supervised concept extraction, the number of concepts is automatically determined by the semi-supervised clustering algorithm and depends on the data characteristics and the structure of the labeled data. In unsupervised concept extraction, however, the number of concepts is arbitrary and defined by the user. The effect of changing the number of concepts is evaluated in Section 5.5.1 (A). Changing the number of concepts in the latter approach changes the length of document vectors. Therefore, the storage and computational complexities of the subsequent clustering step is under the control of the user.

4.3 Clustering

Now that the document vectors have been extracted, the document clustering task may be performed. It is expected that two documents containing similar concepts will have similar document vectors. According to whether the concepts are extracted in a semi-supervised or unsupervised manner, the clustering process of the documents is adjusted. In the former case, clustering is performed using the labeled concepts, while in the latter, a semi-supervised clustering of documents is required. The two alternative approaches are described below.

4.3.1 Clustering based on labeled concepts

Each entry in the document vector $\vec{d}$ is associated with an extracted concept. If concepts are extracted from the set of words by a semi-supervised algorithm, each concept T_j has a respective label. The clustering process reduces to the task of identifying the most effective concept in each document vector. The label of this concept determines the label (cluster) of the document, $label(d)$. To determine the most effective concept, different approaches may be employed. In this study, we propose and evaluate three approaches as represented in Section 5.4, and the most accurate of the three is formulated in Eq. 3. In this method the label which has the most aggregate weight in the CF-IDF vector of d is assigned to this document.

$$label(d) = \underset{l}{\operatorname{argmax}} \sum_{label(T_j)=l} \overrightarrow{d^j} \tag{3}$$

4.3.2 Clustering based on unlabeled concepts

If the process of extracting concepts from the set of words is performed without supervision, the set of concepts do not carry labels with them. Therefore, document vectors are clustered with the help of labeled documents through the semi-supervised clustering algorithm previously introduced in Fig. 2. The input of algorithm (X) is set to document vectors $\bigcup_{d\in D} \overrightarrow{d}$, and the output ($P$) is the set of document clusters. As before, the documents without a label are assigned a dummy label U. The algorithm produces clusters of documents, each having an appropriate label.

Both alternative methods of semi-supervised document clustering introduced in this paper cover multiple benefits. Classic methods of text representation fail to maintain the non-linear semantic relationships between words. The document vectors produced by some recent proposals such as Doc2Vec are not intuitive and understandable. The approach adopted in this paper not only preserves the non-linear semantic relationships between words but also has high interpretability and intuition due to the extraction of concepts in the corpus. Since the concepts separate the components of the text, increasing the number of concepts can capture the more petite subcategories of the more fine-grained concepts of the text. As the algorithm requires partially labeled data, the overhead and cost of tagging the documents may be kept low. Accordingly, in applications where large amounts of unlabeled data are to be clustered, such as social media analysis, the method can appear effective. Finally, the logic behind the clustering process is clear, and the resulting clusters are explainable.

5 Experiments

Several experiments are conducted to show the effectiveness of the proposed method in clustering quality. Three common datasets in natural language processing with different number of samples and classes are employed in this paper: Reuters-21578, 20-Newsgroups, and WebKB. The proposed semi-supervised clustering approaches are labeled as SSConE and SSClusE in the experiments. The compared methods are divided into two categories: unsupervised methods and semi-supervised methods, which are described below.

5.1 Compared methods

5.1.1 Unsupervised algorithms

K-means This algorithm divides the sample data into partitions (clusters) by minimizing the sum of squares within the clusters. The number of clusters should be defined before clustering operation. This algorithm has been used in many data clustering applications.

Hierarchical clustering The hierarchical algorithm identifies clusters by merging or breaking them consecutively [56]. The clustering operation is performed in a bottom-up or top-down manner by creating a tree including the root (all data with different classes) at the top, and leaves (individual data samples) at the bottom.

Spectral clustering The basic idea behind spectral clustering is to use the standard clustering approach using eigenvectors of Laplacian matrices of data similarity [57].

Birch Clusters are generated using a tree structure named the clustering feature tree. One of the characteristics of this method is that the nodes have many clustering features [58].

Deep embedded clustering (DEC) This method uses deep neural network features for textual data clustering operations. Kullback-Leibler divergence is used to optimize and adjust the parameters of the model. DEC maps the data to the feature space using a stacked autoencoder [59].

5.1.2 Semi-supervised algorithms

An approach to TExt classification using Semi-supervised Clustering (TESC) It uses an agglomerative hierarchical algorithm for clustering and then classifying documents. The algorithm identifies and labels the test data by merging the clusters using a semi-supervised approach, and then calculating distance of the test instances with each cluster [25].

Doc2Vec This method uses the similar hierarchical clustering algorithm in this paper on the document vectors extracted by Doc2Vec. This comparison can reveal the advantages of the proposed methods over Doc2Vec-based document representation. Doc2Vec converts documents into feature vectors with suitable dimensions. Due to the neural network-based structure of model training, there is no specific logic for the feature identification mechanism used in the document vector [20].

Discriminative semi-supervised non-negative matrix factorization for data clustering (DSSNMF) This method uses the label information of a part of the data as supervision for the clustering process. In this research, two cost functions and new update rules are used for optimization [50].

Semi-supervised deep embedded clustering (SDEC) This model learns features that are beneficial for clustering tasks using a deep learning approach. By considering the pairwise constraints during training, the data belonging to a cluster are located at a close distance to each other in the feature space [49].

5.2 Datasets

In order to show the performance of the proposed model and its applicability, three datasets namely Reuters-21578,[1] 20-Newsgroups[2] and WebKB[3] are used in experiments.

[1] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[2] http://qwone.com/~jason/20Newsgroups/

[3] http://www.cs.cmu.edu/~webkb/

Table 2 Documents distribution in Reuters-21578

Class	Number of documents
Earn	3964
Acq	2369
Money-fx	717
Grain	582
Crude	578
Trade	485
Interest	478
Ship	286
Wheat	283
Corn	237

The Reuters-21578 news dataset includes a collection of news items published on the Reuters website in 1987, which was collected and processed by Reuters' personnel in 1991. Reuters contains 21,578 texts and 135 data classes that are manually labeled. There is an imbalance in the distribution of documents across the classes, so there may be fewer than 10 documents in one class and over 4000 in another. In this study, 9979 documents are used from the top 10 classes. A detailed breakdown of the classes and the number of documents in each class can be found in Table 2.

The second dataset is the 20-Newsgroups documentation, which includes 18,821 newsgroups posts on 20 topics. The distribution of documents in different classes is balanced. There are some classes that are very close to one another and some that are very far apart. This dataset's distribution of documents and classes can be seen in Table 3.

The WebKB dataset contains web pages from the computer science departments at various universities. In total, 8282 web pages are categorized into six imbalanced categories (Students, Faculty, Staff, Department, Course, Project). A miscellaneous category is also included that cannot be compared to the rest so we dumped this category because pages were very different among this group. The Department and Staff classes are also discarded, as there were only a few pages from each university. Table 4 shows the distribution of documents in each class.

5.3 Evaluation metrics

The Normal Mutual Information (NMI) criterion is a clustering validation metric that assesses the quality of the clustering concerning given underlying labeling of the data. NMI measures how closely the clustering algorithm could reconstruct the underlying label distribution [60]. A value of 0 indicates a random cluster assignment, and values closer to one demonstrate that the clustering can recreate the true class membership. The NMI measure is defined in Eq. 4.

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

where $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between the random variables X and Y, $H(X)$ is the Shannon entropy of X, and $H(X|Y)$ is the conditional entropy of X given Y.

Table 3 Documents distribution in 20-Newsgroups

Class	Number of documents
alt.atheism	799
comp.graphics	973
comp.os.ms-windows.misc	966
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	963
comp.windows.x	985
misc.forsale	975
rec.autos	989
rec.motorcycles	996
rec.sport.baseball	994
rec.sport.hockey	999
sci.crypt	991
sci.electronics	984
sci.med	990
sci.space	987
soc.religion.christian	996
talk.politics.guns	909
talk.politics.mideast	940
talk.politics.misc	775
talk.religion.misc	628

Table 4 Documents distribution in WebKB

Class	Number of documents
Student	1641
Faculty	1124
Course	930
Project	504

The Silhouette Coefficient (SC) is a criterion used to calculate the superiority of a clustering technique. This criterion ranges from -1 to 1 and a higher SC score indicates that the clusters are better separated. According to Eq. 7, this criterion is calculated for each data sample in the clustering process:

$$IntraClusterDistance = \frac{1}{\left|C_k\right|\left(\left|C_k\right| - 1\right)} \sum_{w_i, w_j \in C_k, w_i \neq w_j} dist(w_i, w_j) \tag{5}$$

$$InterClusterDistance = \frac{1}{\left|C_k\right|\left|C_p\right|} \sum_{w_i \in C_k, w_i \in C_p} dist\left(w_i, w_j\right) \tag{6}$$

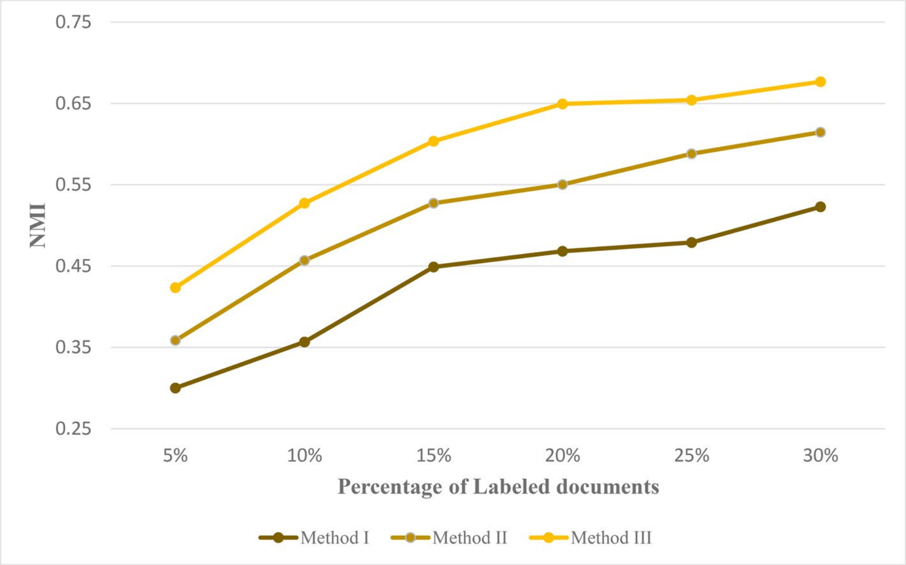


Fig. 3 NMI score of document clustering for the proposed method (SSConE) with the different methods of label assignment for Reuters-21578

where $\left|C_k\right|$, $\left|C_p\right|$ are the number of points in clusters k and$p$.

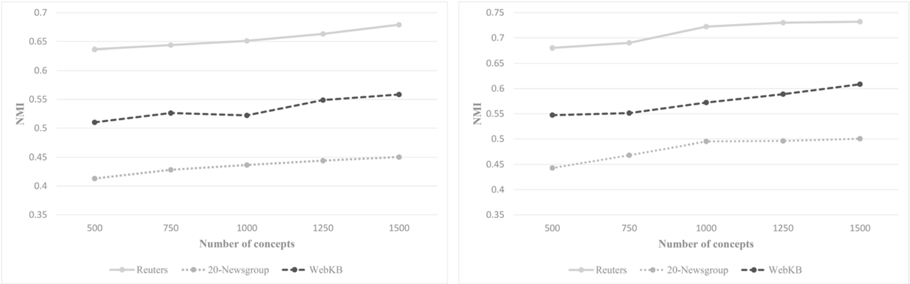
$$Silhouette\ Coefficient = \frac{InterClusterDistance - IntraClusterDistance}{max(IntraClusterDistance, InterClusterDistance)} \tag{7}$$

In this relation, *IntraClusterDistance* is the average distance between a data sample and all other samples in the same cluster, and *InterClusterDistance* is the average distance between a sample and all other samples in the next closest cluster. The average of SC scores for all data points shows the clustering SC. A score of 1 means that the clustering is very dense, while -1 indicates incorrect clustering, and 0 indicates overlapping clusters.

5.4 Evaluation model

In the preprocessing phase, we perform typical tasks such as removing non-alphanumeric characters, stop word elimination, etc. To decrease the complexity of the word embedding, the infrequent words occurring less than five times in the entire dataset are eliminated. In the word summarization task of the semi-supervised concept extraction, according to the available resources, the word vectors are summarized to 2000 pseudo labeled vectors. Although different methods can be used for word embedding, the results are presented with the two popular Word2Vec and BERT embedding models. If unspecified, the default embedding is Word2Vec.

As described in Section 4.3.1, when concepts have labels, the task of assigning a document d to a cluster reduces to finding the most effective label in the document vector $\vec{d}$. Three different methods are proposed and evaluated in this respect, which select the most effective label to be (I) the one having the largest weight in $\vec{d}$, (II) the



(a) SSClusE (Word2Vec) **(b)** SSClusE (BERT)

Fig. 4 NMI score of document clustering for the proposed method (SSClusE) for different datasets when number of concepts varies (**a**) SSClusE (Word2Vec) (**b**) SSClusE (BERT)

Table 5 SC of document clustering for the proposed method (SSClusE) for different datasets when number of concepts varies

Embedding	Dataset	SC				
		500	750	1000	1250	1500
Word2Vec	Reuters-21578	0.1602	0.1614	0.1622	0.1708	0.1768
	20-Newsgroups	0.0501	0.0618	0.0624	0.0806	0.0854
	WebKB	0.0436	0.0521	0.0534	0.0682	0.0765
BERT	Reuters-21578	0.1652	0.1780	0.1865	0.1905	0.2042
	20-Newsgroups	0.0665	0.864	0.1102	0.1126	0.1158
	WebKB	0.0607	0.0689	0.0874	0.0881	0.0907

most frequent label in the concepts associated with $\vec{d}$, and (III) the one having the largest aggregate weight in $\vec{d}$ as described in Section 4.3.1 Eq. 3. All three methods have been tested, and the results of the clustering NMI can be seen in Fig. 3 when the percentage of labeled documents varies. As observed, the third method shows the highest NMI score in clustering the documents. In the rest of the experiments, this method is used for cluster assignment in SSConE.

The selection of initial points in the spherical K-Means algorithm is essential and affects the clustering performance. To address this issue, when spherical K-Means is executed (e.g. when pseudo words are extracted), the algorithm is executed five times with random initial points, and the clustering with minimum within-cluster distance is selected. All the experiments are executed on a system with an Intel Core i5 processor and 6 GB RAM.

Table 6 The performance of the SSClusE when the size of the window varies

	NMI			SC		
	4	8	20	4	8	20
Reuters-21578	0.5986	0.6512	0.7014	0.1455	0.1622	0.1748
20-Newgroup	0.4041	0.4396	0.4400	0.0525	0.0624	0.0792
WebKB	0.5276	0.5428	0.5982	0.0702	0.0787	0.0882

5.5 Results

5.5.1 Parameter study

Number of concepts In the SSClusE model, the number of concepts defines the size of the document vector. Since larger document vectors provide a more fine-grained view into the semantic space, and the distance between documents are obtained by comparing the document vectors, the number of concepts may have a significant effect on the clustering quality. To measure this effect, an experiment is designed in which the number of concepts varies from 500 to 1500, and 25% of the data are labeled. Figure 4 and Table 5 show the values of NMI and SC for the SSClusE model, respectively. According to the results, when the number of concepts increases resulting in larger document vectors, the quality of the created clusters improves in all three datasets.

As an illustration, we opted for the highest number of concepts that could be identified in each dataset and assessed the clustering outcome using the lengthiest document vector. In all documents, words that occurred more than five times (min_freq = 5) are utilized to create concepts. This condition results in 7891, 24,301, and 6324 words remaining for Reuters-21578, 20-Newsgroups, and WebKB datasets, respectively. If we consider each word as a cluster or concept, the longest possible concept vector and document vector are created in the proposed method. We evaluated the document clustering results in this setting, and achieved the NMI values of 0.6867, 0.4796, and 0.5642 for the Reuters-21578, 20-Newsgroups, and WebKB datasets, respectively. Based on the results and Fig. 4, clustering quality improves gradually as the number of concepts increases. Therefore, if the user has sufficient computational and storage resources, incorporating more concepts can be beneficial.

According to Fig. 4(a), the NMI values for the Reuters-21578 dataset when the document vector is created based on 500 and 1500 concepts, are equal to 0.63 and 0.67, respectively. This improving trend can also be seen in the two other datasets. The values in Table 5 show that the SC score of clustering also increases with larger number of concepts. For example, the clustering silhouette for the 20-Newsgroups dataset has increased from 0.05 to 0.08, and from 0.07 to 0.12 with different embeddings. Therefore, the number of extracted concepts shows a direct relationship with the clustering quality. If the number of extracted concepts is small, the concepts will be coarse-grained, each containing a large portion of the words. Such concepts may actually be combined from several, more delicate concepts. Therefore, describing documents in terms of these concepts, results in a less distinguishing capability among the documents covering proximate yet different topics. If the number of concepts is chosen correctly resulting in a suitable granularity of concepts, distinguishing the class of documents will be facilitated.

Table 7 The performance of the SSConE when the size of the window varies

	NMI			SC		
	4	8	20	4	8	20
Reuters-21578	0.5754	0.6393	0.6963	0.1538	0.1821	0.1880
20-Newgroup	0.4566	0.4738	0.4874	0.0642	0.0752	0.0862
WebKB	0.4894	0.5249	0.5306	0.0754	0.0851	0.0894

Table 8 The performance of the SSClusE when the Word2Vec dimension varies

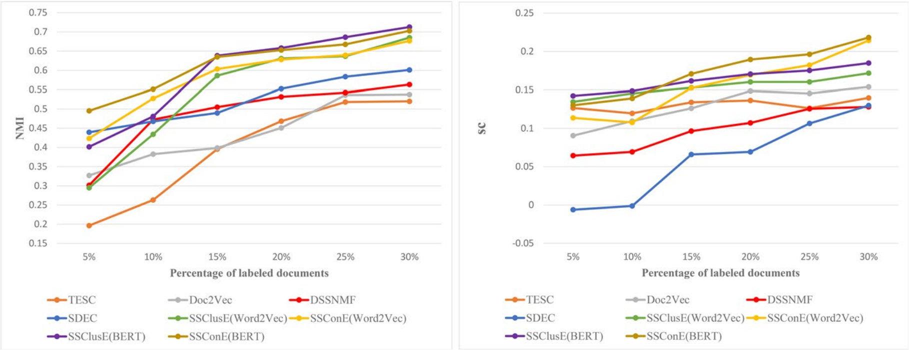
	NMI									
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
Reuters-21578	0.6354	0.6512	**0.6858**	0.6753	0.6648	0.6127	0.6188	0.6287	0.6439	0.6346
20-Newgroup	0.4062	0.4396	0.4337	0.4489	**0.4664**	0.4660	0.4519	0.4507	0.4528	0.4579
WebKB	0.5195	**0.5428**	0. 5344	0.5207	0.5156	0.5105	0.5043	0.5099	0.5080	0.5014

Best NMI value for each dataset is declared in boldface

Embedding window size In the proposed model, Word2Vec embedding is used to maintain the semantic relationships between words. One of the important parameters in embedding is the size of the sliding window. A larger window can capture more semantic relationships among the words, which may have a positive impact on extracting meaningful concepts. We change the window size from four to 20 and observe its effect on the document clustering quality. In this experiment, 25% of the data are labeled and document vector size is 1000. Tables 6 and 7 represent the results. The performance of the proposed model as reported by NMI and SC improves with higher window sizes, with the improvement being larger when changing the window size from four to eight, and marginal when increasing it to 20. For example in Table 6, when the window size changes from four to 20, the NMI value increases from 0.59 to 0.70, indicating a better quality. The SC criterion also shows an increasing trend. This quality improvement in clustering is achieved due to the better conceptual modeling of the corpus. This is explained by the Word2Vec neural network encountering more co-occurrences at each step, and discovering an expanded semantic relation among words. In order to reduce the time complexity, a window size of eight is considered for training the Word2Vec model in the next experiments.

Word vector dimension When it comes to embedding words, the dimensions of the word vectors play a crucial role in determining the outcome of document similarity measurement. Small values may not maintain semantic relationships correctly, while large values may result in a huge computational overhead. Hence, it is crucial to determine the suitable and most efficient value by experimenting with the inherent properties of each dataset. Table 8 shows the results of document clustering with the proposed method when the word vector dimensions change from 500 to 5000. In this experiment, 25% of the data are labeled and window size is 8.

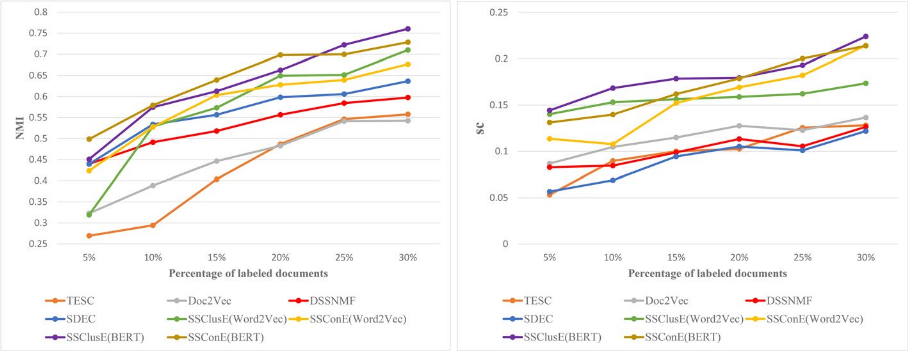
Upon closer examination of Table 8, it becomes apparent that as the dimensions of the word vector increase, the clustering performance and quality improve. However, this increase is not consistent and reaches a point where further changes in word vector dimension are not noticeable for clustering performance. For example, in Reuters-21578, the best



(a) NMI score

(b) Silhouette coefficient

Fig. 5 The performance comparison of different models at the various percentage of labeled documents on Reuters-21578, with 500 concepts (**a**) NMI score (**b**) Silhouette coefficient



(a) NMI score

(b) Silhouette coefficient

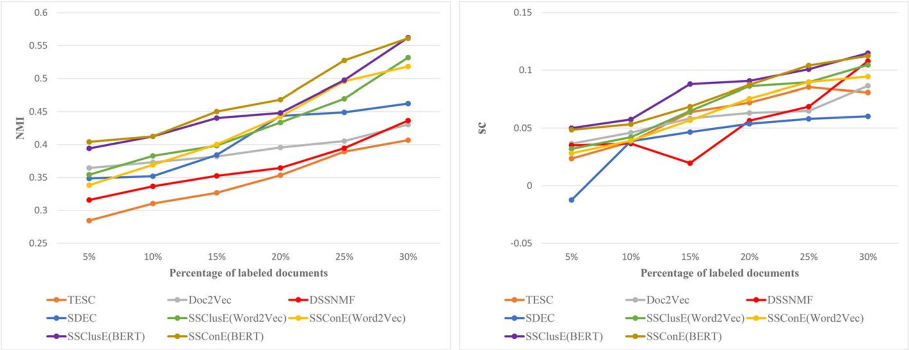
Fig. 6 The performance comparison of different models at the various percentage of labeled documents on Reuters-21578, with 1000 concepts (**a**) NMI score (**b**) Silhouette coefficient

clustering performance is obtained with Word2Vec of length 1500. This value is 2500 and 1000 for the 20-Newgroup and WebKB datasets, respectively.

5.5.2 Effect of the percentage of labeled documents

The number of labeled documents has a literal impact on the quality of clustering performed. To observe this effect, we conducted another experiment in which the size of the labeled data varied. This configuration can assess the performance of the model when encountering different ratios of labeled data. It can reveal in particular whether the model is robust when possessing only a tiny subset of labeled data and if it can benefit from a large number of labeled data to improve its performance.

Figure 5 shows the performance of the proposed methods compared to other semi-supervised methods when the percentage of the labeled data varies. Number of concepts is set to 500 in this experiment. According to Fig. 5(a), the clustering performed



(a) NMI score **(b)** Silhouette coefficient

Fig. 7 The performance comparison of different models at the various percentage of labeled documents on 20-Newsgroups, with 2000 concepts (**a**) NMI score (**b**) Silhouette coefficient

with the proposed method on Reuters-21578 dataset outperforms other methods. When increasing the number of labeled documents, the value of NMI and consequently the clustering quality increases. For example, in the worst case, when 5% of the documents are labeled, the NMI values of SSConE (BERT), SSClusE (BERT), SSConE (Word2Vec), SSClusE (Word2Vec), TESC, Doc2Vec, DSSNMF, and SDEC are 0.4953, 0.4016, 0.42, 0.29, 0.19, 0.32, 0.30, and 0.43, respectively. When a large portion of documents (30%) are labeled, the NMI value of the SSClusE (BERT) reaches 0.7128 which is highest among all methods. For the Silhouette coefficient in Fig. 5(b), with 30% labeled documents, the value of SC for SSConE (BERT) is 0.2182, and at worst, when there are 5% labeled data, SC is 0.13. In evaluating the SC, the proposed methods have maintained their superiority over all other methods. The results of this experiment reveal that the proposed methods can benefit from a small amount of labeled data, while their performance improves when having access to more labeled data.

Figures 6 and 7 show the performance of the proposed methods compared to other semi-supervised document clustering methods with 1000 concepts on Reuters and 2000 concepts on 20-Newsgroups, respectively. From the diagrams in Figs. 5, 6, and 7, it can be seen that the proposed methods are superior to nearly all other methods when a small set of labeled data is available. With more labeled documents, the difference between the NMI of the proposed method and other methods increases. It indicates that access to more labeled data can improve the clustering quality of the proposed methods. In addition, it confirms that the algorithm design is benefiting from a larger labeled set in discovering the clusters in the data. Another interesting observation is that when less labeled data is available, in some cases unsupervised concept extraction (SSClusE) is performing better than semi-supervised concept extraction (SSConE) on 20-Newsgroups. Gradually with the increase in the number of labeled data, SSConE takes over. The results of SC show that the proposed methods have been able to create denser and more distinct clusters than other methods. The proposed methods take an important step towards better clustering of documents by extracting and decomposing text components, compared to other methods. Using labeled data, cluster centers are better tuned, and a more reliable clustering can be expected.

Table 9 The performance comparison of different algorithms

		NMI			SC		
		Reuters-21578	20-Newsgroups	WebKB	Reuters-21578	20-Newsgroups	WebKB
Unsupervised	K-means	0.3898	0.3346	0.3456	0.0786	0.0002	0.0168
	Hierarchical clustering [56]	0.2981	0.2060	0.2432	0.0863	0.0602	0.0552
	Spectral clustering [57]	0.4132	0.1046	0.3105	0.0728	0.0514	0.0148
	Brich [58]	0.4693	0.3242	0.1908	0.0746	-0.0028	0.0123
	DEC [59]	0.3125	0.1584	0.1588	0.0301	-0.0880	-0.0110
Semi-supervised	TESC [25]	0.5464	0.3578	0.2965	0.1256	0.0672	0.0674
	Doc2Vec [20]	0.5416	0.3818	0.3746	0.1229	0.0513	0.0622
	DSSNMF [50]	0.5843	0.3455	0.3641	0.1054	0.0489	0.0514
	SDEC [49]	0.6058	0.4141	0.3891	0.1010	0.0357	0.0586
	SSClusE (Word2Vec)	0.6512	0.4432	0.5428	0.1622	0.0802	0.0787
	SSConE (Word2Vec)	0.6393	0.4874	0.5249	0.1821	0.0820	0.0851
	SSClusE (BERT)	**0.7223**	0.4952	0.5722	0.1865	0.1102	0.0874
	SSConE (BERT)	0.6984	**0.5361**	**0.5846**	**0.2108**	**0.1184**	**0.0968**

Best values for each dataset are declared in boldface

Table 10 Five important words in some extracted concepts of Reuters-21578

Number	Words	True label
1	construction, metals, hut, operatorship, coal	crude
2	economist, industrialists, worker, policymakers, currency, gearing	trade
3	launch, pretax, renewal, pro, CDT	acq
4	financial, Dubai, uneconomic, dominance, minimizing	earn
5	court, prevented, filing, statement, exceed	acq

5.5.3 Comparison with other methods

Table 9 compares the clustering performance of the proposed method with several semi-supervised and unsupervised methods in all three datasets. For this examination, 25% of the labeled data has been used in semi-supervised methods and the evaluation has been done with 1000 concepts or features. Similar to the settings mentioned earlier in Section 5.4 for the spherical K-Means algorithm, K-means is executed five times with random initial points, and the clustering with the minimum within-cluster distance is selected. The Doc2Vec and TESC methods were executed multiple times with varying parameters, and the average results of the top five clustering results were reported in Table 9. The length of documents in all methods is set to 1000 and the window sizes of Word2Vec (for K-means clustering) and Doc2Vec methods are both set to eight. For a comprehensive evaluation of the proposed framework, independent of the embedding method, we have also used the embedded vectors of BERT large language model [53] (SSClusE-BERT and SSConE-BERT in Table 9). In order to substitute Word2Vec vectors, the average of the last four layers from the pre-trained BERT-BASE model is employed for each word, which has a vector length of 768. The proposed SSConE and SSClusE algorithms outperform all other techniques in terms of SC and NMI in all datasets, which confirms that the proposed methods can learn the semantic feature vectors of documents. Using the BERT embeddings produces higher quality clusters.

For the Reuters-21578 dataset, after the proposed methods, the semi-supervised SDEC and DSSNMF methods compete with each other. TESC (BoW) and Doc2Vec are next. Generally, the semi-supervised algorithms are performing better than unsupervised clustering, especially in terms of NMI. Recall that the performance of the SSConE method does not depend on the number of concepts because this method uses a hierarchical algorithm that automatically determines the number of concepts. For the dataset of 20-Newsgroups as well as WebKB, the proposed methods achieve a performance that is superior compared to other models in all configurations. This shows that using conceptual modeling for document representation is effective in document clustering.

5.5.4 Interpretability

Representation of documents using concepts provides high interpretability with a more clear understanding. To illustrate this interpretability, we use the SSConE model, where each final cluster of documents corresponds to an extracted concept. Thus, the important words in each cluster are the same as those in the corresponding concepts. Using the

Table 11 Average values of IntraClusterDistance, InterClusterDistance, and SC for all concepts in Reuters-21578, 20-Newsgroups, and WebKB

Dataset	Intra cluster distance	Inter cluster distance	SC
Reuters-21578	0.7441	0.9399	0.2083
20-Newsgroups	0.8914	0.9585	0.0700
WebKb	0.8937	0.9741	0.0825

Table 12 NMI and SC score of document clustering for the proposed SSConE method when number of words in each concept varies

	Dataset	Original concepts	Step 1	Step 2	Step 3
NMI	Reuters-21578	0.6305	0.6141	0.6141	0.5952
	20-Newsgroups	0.4119	0.3821	0.3857	0.3731
	WebKb	0.5294	0.5123	0.5029	0.4913
SC	Reuters-21578	0.1783	0.1717	0.1632	0.1641
	20-Newsgroups	0.0695	0.0632	0.0540	0.0518
	WebKb	0.0768	0.0746	0.0733	0.0730

TF-IDF weighting of Eq. 1, we can determine the important words in the documents of each cluster. Table 10 shows five important words in some randomly chosen concepts of Reuters-21578. By analyzing the words within each concept, it becomes clear what the concept pertains to and why the related documents are grouped together in their respective clusters. For example, in the first displayed concept, the top five words (construction, metals, hut, operatorship, coal) are all either related to raw materials or construction, which matches with the actual class label, "crude", in the Reuters-21578 dataset. For another example, the top words in the fourth concept (financial, Dubai, uneconomic, dominance, minimizing) are related to the monetary and economic topics, which is consistent with the actual class label, "earn".

The words that occur in the same context are placed close to each other in the embedded space, and the formed concepts maintain the semantic similarities between the words. To demonstrate this more effectively, an experiment was conducted in which the average cosine distance between the words in each concept (Eq. 5) and the words in the nearest neighboring concept (Eq. 6) are measured. Larger *InterClusterDistance* and smaller *IntraClusterDistance* imply better interpretability because the words in a concept k are closer compared to the words in the nearest neighboring concept p. Table 11 shows the average values of *IntraClusterDistance*, *InterClusterDistance*, and SC for all concepts in Reuters-21578, 20-Newsgroups, and WebKB datasets. As can be seen, the proximity of a word to other words within its own concept is greater than its proximity to words in the next nearest concept. This demonstrates the preservation of semantic similarity among words within a concept.

Furthermore, we conducted an experiment to gauge the sensitivity of the clustering methods to subtle changes in the extracted concepts. By removing a percentage of words randomly from each concept, we observed the resulting changes in document clustering. The insightful results of this experiment are presented in Table 12. At each step of the test, 10% of the words in each concept are eliminated (except for the concepts that have less than 5 words) and then the documents are clustered with these new concepts. For this examination, 25% of the labeled data has been used in the semi-supervised method and the evaluation has been done with 400 concepts or features. The tests were conducted 5 times and the average values are reported in Table 12. In the worst case, we observe a 0.0388 and 0.0177 decrease in NMI and SC, respectively.

After analyzing Table 12, it is evident that the decline in clustering performance was minimal, indicating that the proposed clustering method is capable of withstanding variations in concepts.

5.5.5 Discussion

The proposed framework represents documents using dense vectors in the space of constituent concepts of the collection. As observed in the experiments, it achieved better performance compared to methods using other document representation schemes such as TF-IDF (BoW) and Doc2Vec. Unlike neural-networks-based methods [20, 46, 49] which have low interpretability, concept-based algorithms have high interpretability and the logic of creating document vectors is well visible. In neural-network-based document representation methods, it is not clear how each of the document vector elements is calculated and what feature it describes. The BoC [21] representation, although being based on concepts, is unsupervised, does not provide a direct solution for text clustering or classification, and was shown to be less effective compared to the proposed framework. Some other representation methods require an external knowledge base [30] or a domain-specific ontology [44] beside the document collection. Other representation approaches based on model graphs [33], adaptive slider-windows [34], and Latent Dirichlet Allocation [35], may be more complex to extend if labeled data are available and also have not provided a clustering solution. In another method presented for the conceptual representation of short texts [32], the network of semantic communities of words has been used to extract concepts. Such methods may not be suitable for longer texts as used in this paper, due to the computational complexity.

From the algorithm perspective, some semi-supervised algorithms using a small set of labeled data, concentrate on the classification task [45]. Some semi-supervised clustering approaches iteratively label the unlabeled data to use them in model training [24]. This labeling should continue until the model converges, and it is not clear how much labeling is required. Our model in contrast, does not use such labeling, and being based on conceptual representation and hierarchical clustering, is shown to be superior in terms of clustering quality (Figs. 5, 6, and 7). Unlike the methods that cannot maintain the semantic similarity of words in vectors with small dimensions and require model convergence [50, 57], the proposed method has the ability to maintain semantic concepts in vectors with logical length. In TESC [25], a semi-supervised approach is used to categorize documents using a large ratio of labeled data. We use conceptual representation which is able to discover categories of the text in different granularities, and use only a small amount of labeled data. These advantages can be beneficial in various real-world applications where labeled data is scarce. According to the results of Figs. 5, 6, and 7, it is observed that the proposed method outperforms TESC in the face of large amounts of unlabeled data. Unlike some other studies [47, 48], our proposed framework is able to use labeled and unlabeled data simultaneously in the model training phase. It can identify the overall structure of clusters with unlabeled data and tune the centers of each cluster with the help of labeled data. In addition, the use of cumulative hierarchical clustering and the discovery of concepts in the texts allows the proposed framework to benefit from labeled data in assigning unlabeled data to the appropriate cluster. Because the conceptual representation preserves the proximity information of documents as well as the nonlinear relationships between words, it is more convenient to find similar documents compared to the word-based approaches. Overall, the experiment results show that the proposed framework has high efficiency in clustering documents, and can benefit from labeled data in improving the quality of results.

6 Conclusion

In this paper, we presented a concept-based method for semi-supervised document clustering. Our basic assumption was that the type of document representation affects the quality of document clustering and classification tasks. We adopted a concept-based representation based on the concepts extracted using the semantic similarities between the corpus words. This method overcomes the limitations and weaknesses of other methods of text representation based on BoW and document embedding. Additionally, the proposed method has the advantage of describing the documents in a low dimensional space of concepts while offering high interpretability of document vectors. Limited labeled data are used alongside unlabeled data to adjust the cluster centers. In this paper, two methods for engaging labeled data in the clustering process were introduced. In the first method, the labeled documents were employed in the concept extraction task, while in the second method, the labeled data were used in the actual document clustering step. The former method employs supervision in discovering the hidden structure in the embedded space of the words and captures labeled concepts. It allows for a better discovering of the relation between the components in the text with the provided label information. Clustering is a fundamental step of many natural language processing tasks and has numerous practical applications in different diverse areas. Semi-supervised clustering may be beneficial when dealing with tasks that have limited labeled data, such as social network analysis, question answering, topic modeling, concept hierarchy generation, training deep neural networks, and when interpretability is crucial. As shown in the experiments, results on three popular datasets showed the superiority of the proposed methods, especially when a smaller number of labeled data are available. Although the experimental results are satisfying and promising, further extensions to our study can be performed. For example, introducing fuzziness in the produced concepts may yield more realistic modeling of the text components, as each word may participate in more than one concept. In addition, constructing hierarchical overlapping concepts can be investigated in accordance with hierarchical clustering. Keyword expansion may be used to enrich and generalize the set of extracted concepts in dealing with new documents.

Authors' contributions Conceptualization: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi]; Data curation: [Seyed Mojtaba Sadjadi]; Formal analysis: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi], [Hamid Hassanpour]; Investigation: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi]; Methodology: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi]; Project administration: [Hoda Mashayekhi]; Software: [Seyed Mojtaba Sadjadi]; Supervision: [Hoda Mashayekhi], [Hamid Hassanpour]; Validation: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi]; Visualization: [Seyed Mojtaba Sadjadi], [Hoda Mashayekhi], [Hamid Hassanpour]; Writing – original draft: [Seyed Mojtaba Sadjadi]; Writing – review & editing: [Hoda Mashayekhi], [Hamid Hassanpour].

Data availability The datasets used in this study are openly available at the following addresses:

Reuters-21578 at [https://archive.ics.uci.edu/dataset/137/reuters+21578+text+categorization+collection], 20-Newsgroups at [http://qwone.com/~jason/20Newsgroups/] and WebKB at [http://www.cs.cmu.edu/~webkb/].

Declarations

Ethical approval and consent to participate Not applicable.

Human and animal ethics Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

Authors' information Seyed Mojtaba Sadjadi received his M.S. from the Department of Computer Engineering in Shahrood University of Technology in 2021. His main research interests include machine learning, natural language processing, text mining through embedding, and the semantic web.

Hoda Mashayekhi is an Assistant Professor at the faculty of Computer Engineering, Shahrood University of Technology. She received her PhD from Sharif University of Technology in 2013. Her research interests include massive data mining, machine learning, parallel and distributed computing, decision-making, and recommendation systems.

Hamid Hassanpour received his Ph.D. from the Queensland University of Technology, Australia in 2004. He is currently a full professor at the faculty of Computer Engineering, Sharood University of Technololgy, Iran. His research interests include Image Processing, Signal Processing, and Data Mining. He has published over 200 journal and conference papers. He is the Editor-in-Chief for Journal of Artificial Intelligence and Data Mining.

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, pp. 001601–001604, Niagara Falls (2008). https://doi.org/10.1109/CCECE.2008.4564812
2. Janani, R., Vijayarani, S.: Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. Expert Syst. Appl. **134**, 192–200 (2019). https://doi.org/10.1016/J.ESWA.2019.05.030
3. Zhang, W., Yoshida, T., Tang, X., Wang, Q.: Text clustering using frequent itemsets. Knowledge-Based Syst. **23**, 379–388 (2010). https://doi.org/10.1016/j.knosys.2010.01.011
4. Xiao, Y., Liu, B., Yin, J., Hao, Z.: A multiple-instance stream learning framework for adaptive document categorization. Knowledge-Based Syst. **120**, 198–210 (2017). https://doi.org/10.1016/j.knosys.2017.01.001
5. Misztal-Radecka, J., Indurkhya, B.: Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. Inf. Process. Manag. **58**, 102519 (2021). https://doi.org/10.1016/J.IPM.2021.102519
6. Wang, J., Shi, Y., Li, D., Zhang, K., Chen, Z., Li, H.: McHa: a multistage clustering-based hierarchical attention model for knowledge graph-aware recommendation. World Wide Web. **253**(25), 1103–1127 (2022). https://doi.org/10.1007/S11280-022-01022-5
7. Edara, D.C., Vanukuri, L.P., Sistla, V., Kolli, V.K.K.: Sentiment analysis and text categorization of cancer medical records with LSTM. J. Ambient Intell. Humaniz. Comput. 1–17 (2019). https://doi.org/10.1007/s12652-019-01399-8
8. Almeida, T.A., Silva, T.P., Santos, I., Gómez Hidalgo, J.M.: Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering. Knowledge-Based Syst. **108**, 25–32 (2016). https://doi.org/10.1016/j.knosys.2016.05.001
9. Ligthart, A., Catal, C., Tekinerdogan, B.: Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. Appl. Soft Comput. **101**, 107023 (2021). https://doi.org/10.1016/J.ASOC.2020.107023

10 Shakiba, T., Zarifzadeh, S., Derhami, V.: Spam query detection using stream clustering. World Wide Web. **212**(21), 557–572 (2017). https://doi.org/10.1007/S11280-017-0471-Z

11. Djenouri, Y., Belhadi, A., Fournier-Viger, P., Lin, J.C.W.: Fast and effective cluster-based information retrieval using frequent closed itemsets. Inf. Sci. (Ny) **453**, 154–167 (2018). https://doi.org/10.1016/j.ins.2018.04.008

12. Joty, S., Carenini, G., Ng, R.T.: Topic segmentation and labeling in asynchronous conversations. J. Artif. Intell. Res. **47**, 521–573 (2013). https://doi.org/10.1613/jair.3940
13. Paparrizos, J., Gravano, L.: Fast and accurate time-series clustering. ACM Trans. Database Syst. **42**, 1–49 (2017). https://doi.org/10.1145/3044711
14. Li, Y., Guo, H., Zhang, Q., Gu, M., Yang, J.: Imbalanced text sentiment classification using universal and domain-specific knowledge. Knowledge-Based Syst. **160**, 1–15 (2018). https://doi.org/10.1016/j.knosys.2018.06.019
15. Mohd, M., Jan, R., Shah, M.: Text document summarization using word embedding. Expert Syst. Appl. **143**, 112958 (2020). https://doi.org/10.1016/j.eswa.2019.112958
16. Zhang, W., Yoshida, T., Tang, X.: A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Appl. **38**, 2758–2765 (2011). https://doi.org/10.1016/j.eswa.2010.08.066
17. Sayeedunnissa, S.F., Hussain, A.R., Hameed, M.A. Supervised opinion mining of social network data using a bag-of-words approach on the cloud BT. In: Bansal, J.C., Singh, P., Deep, K., Pant, M., Nagar, A. (eds.) Proceedings of seventh international conference on bio-inspired computing: theories and applications (BIC-TA 2012), pp. 299–309. Springer India, India (2013)
18. Jacovi, A., Shalom, O.S., Goldberg, Y. Understanding convolutional neural networks for text classification. In: Proc. 2018 EMNLP Work. BlackboxNLP Anal. Interpret. Neural Networks NLP, pp. 56–65. Association for Computational Linguistics (ACL) (2018). https://doi.org/10.18653/V1/W18-5408
19. N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf., Association for Computational Linguistics (ACL), 2014: pp. 655–665. https://doi.org/10.3115/v1/p14-1062.
20. Le, Q., Mikolov, T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning. PMLR, **32**, 1188–1196. (2014). http://proceedings.mlr.press/v32/le14.html. Accessed 19 March 2021
21. Kim, H.K., Kim, H., Cho, S.: Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. Neurocomputing **266**, 336–352 (2017). https://doi.org/10.1016/j.neucom.2017.05.046
22 van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Mach. Learn. **1092**(109), 373–440 (2019). https://doi.org/10.1007/S10994-019-05855-6
23. 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 Syst. **73**, 149–160 (2015). https://doi.org/10.1016/j.knosys.2014.09.014
24. Zhang, W., Yang, Y., Wang, Q.: Using Bayesian regression and EM algorithm with missing handling for software effort prediction. Inf. Softw. Technol. **58**, 58–70 (2015). https://doi.org/10.1016/j.infsof.2014.10.005
25. Zhang, W., Tang, X., Yoshida, T.: TESC: An approach to TExt classification using Semi-supervised Clustering. Knowledge-Based Syst. **75**, 152–160 (2015). https://doi.org/10.1016/j.knosys.2014.11.028
26. Lai, S., Liu, K., He, S., Zhao, J.: How to generate a good word embedding. IEEE Intell. Syst. **31**, 5–14 (2016). https://doi.org/10.1109/MIS.2016.45
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. (2013). http://arxiv.org/abs/1301.3781. Accessed 16 Sept 2023
28. Dai, A.M., Olah, C., Le, Q.V.: Document embedding with paragraph vectors. 1–8 (2015). http://arxiv.org/abs/1507.07998. Accessed 16 Sept 2023
29. Zhang, Z., Zhang, Y., Xu, M., Zhang, L., Yang, Y., Yan, S.: A survey on concept factorization: from shallow to deep representation learning. Inf. Process. Manag. **58**, 102534 (2021). https://doi.org/10.1016/J.IPM.2021.102534
30. 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 Syst. **193**, 105436 (2020). https://doi.org/10.1016/j.knosys.2019.105436
31. Luo, X., Shah, S.: Concept embedding-based weighting scheme for biomedical text clustering and visualization. Appl. Informatics. **5**, 1–19 (2018). https://doi.org/10.1186/s40535-018-0055-8
32. Jia, C., Carson, M.B., Wang, X., Yu, J.: Concept decompositions for short text clustering by identifying word communities. Pattern Recognit. **76**, 691–703 (2018). https://doi.org/10.1016/j.patcog.2017.09.045
33. Wu, C., Kanoulas, E., de Rijke, M.: Learning entity-centric document representations using an entity facet topic model. Inf. Process. Manag. **57**, 102216 (2020). https://doi.org/10.1016/J.IPM.2020.102216
34. Li, W., Suzuki, E.: Adaptive and hybrid context-aware fine-grained word sense disambiguation in topic modeling based document representation. Inf. Process. Manag. **58**, 102592 (2021). https://doi.org/10.1016/J.IPM.2021.102592

35. Lee, Y.H., Hu, P.J.H., Tsao, W.J., Li, L.: Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. Expert Syst. Appl. **174**, 114681 (2021). https://doi.org/10.1016/J.ESWA.2021.114681

36 Mehanna, Y.S., Bin Mahmuddin, M.: A semantic conceptualization using tagged bag-of-concepts for sentiment analysis. IEEE Access. **9**, 118736–118756 (2021). https://doi.org/10.1109/ACCESS.2021.3107237

37. Basu, S., Davidson, I., Wagstaff, K.: Constrained clustering: advances in algorithms, theory, and applications, 1st. edn. Chapman & Hall/CRC (2008)

38. Zhang, Z., Zhang, Y., Liu, G., Tang, J., Yan, S., Wang, M.: Joint label prediction based semi-supervised adaptive concept factorization for robust data representation. IEEE Trans. Knowl. Data Eng. **32**, 952–970 (2020). https://doi.org/10.1109/TKDE.2019.2893956

39. Lu, M., Zhao, X.J., Zhang, L., Li, F.Z.: Semi-supervised concept factorization for document clustering. Inf. Sci. (Ny) **331**, 86–98 (2016). https://doi.org/10.1016/j.ins.2015.10.038

40. 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 Comput. **20**, 2329–2339 (2016). https://doi.org/10.1007/s00500-015-1643-3

41 Li, P., Deng, Z.: Use of distributed semi-supervised clustering for text classification. J. Circuits Syst. Comput. **28**, 1–13 (2019). https://doi.org/10.1142/S0218126619501275

42. Masud, M.A., Huang, J.Z., Zhong, M., Fu, X.: Generate pairwise constraints from unlabeled data for semi-supervised clustering. Data Knowl. Eng. **123**, 101715 (2019). https://doi.org/10.1016/j.datak.2019.101715

43. Gan, H., Fan, Y., Luo, Z., Zhang, Q.: Local homogeneous consistent safe semi-supervised clustering. Expert Syst. Appl. **97**, 384–393 (2018). https://doi.org/10.1016/j.eswa.2017.12.046

44. Agarwal, R.: Phrases based document classification from semi supervised hierarchical LDA. In: 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 332–337. Dubai (2021). https://doi.org/10.1109/ICCAKM50778.2021.9357720

45. Zhang, Y., Chen, X., Meng, Y., Han, J.: Hierarchical metadata-aware document categorization under weak supervision. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21), pp. 770–778. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3437963.3441730

46. Vilhagra, L.A., Fernandes, E.R., Nogueira, B.M. TextCSN: A semi-supervised approach for text clustering using pairwise constraints and convolutional siamese network. Proc. ACM Symp. Appl. Comput. 1135–1142 (2020). https://doi.org/10.1145/3341105.3374018

47. Li, L., Zhao, K., Gan, J., Cai, S., Liu, T., Mu, H., Sun, R.: Robust adaptive semi-supervised classification method based on dynamic graph and self-paced learning. Inf. Process. Manag. **58**, 102433 (2021). https://doi.org/10.1016/J.IPM.2020.102433

48. Emadi, M., Tanha, J., Shiri, M.E., Aghdam, M.H.: A Selection Metric for semi-supervised learning based on neighborhood construction. Inf. Process. Manag. **58**, 102444 (2021). https://doi.org/10.1016/J.IPM.2020.102444

49. Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S.C.H., Xu, Z.: Semi-supervised deep embedded clustering. Neurocomputing **325**, 121–130 (2019). https://doi.org/10.1016/J.NEUCOM.2018.10.016

50. Xing, Z., Wen, M., Peng, J., Feng, J.: Discriminative semi-supervised non-negative matrix factorization for data clustering. Eng. Appl. Artif. Intell. **103**, 104289 (2021). https://doi.org/10.1016/J.ENGAPPAI.2021.104289

51 Li, X., Yin, H., Zhou, K., Zhou, X.: Semi-supervised clustering with deep metric learning and graph embedding. World Wide Web. **232**(23), 781–798 (2019). https://doi.org/10.1007/S11280-019-00723-8

52. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119 (2013). http://arxiv.org/abs/1310.4546. Accessed 11 May 2021

53. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 1–2 (2019)

54. Hornik, K., Feinerer, I., Wu, M.K., Wien, W., Buchta, C.: Spherical k-means clustering. J. Stat. Softw. **50**, 1–22 (2012)

55. Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. J. Doc. **60**, 503–520 (2004). https://doi.org/10.1108/00220410410560582

56. Li, C., Bai, J., Wenjun, Z., Xihao, Y.: Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment. Inf. Process. Manag. **56**, 91–109 (2019). https://doi.org/10.1016/j.ipm.2018.10.004

57. Semertzidis, T., Rafailidis, D., Strintzis, M.G., Daras, P.: Large-scale spectral clustering based on pairwise constraints. Inf. Process. Manag. **51**, 616–624 (2015). https://doi.org/10.1016/J.IPM.2015.05.007
58. Zhang, T., Ramakrishnan, R., Livny, M. BIRCH: an efficient data clustering method for very large databases. In: SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data), pp. 103–114. ACM PUB27, New York (1996). https://doi.org/10.1145/235968.233324
59. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on Machine Learning, pp. 478–487. PMLR (2016)
60. Strehl, A., Ghosh, J., Mooney, R. Impact of similarity measures on web-page clustering. Work. Artif. Intell. Web Search (AAAI 2000). **58**, 64 (2000)

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