



Semantic encoder-decoder nonnegative matrix factorization with kullback-leibler divergence

Sayvan Soleymanbaigi¹ · Amjad Seyedi² · Fatemeh Daneshfar¹ · Fardin Akhlaghian Tab¹

Received: 26 January 2025 / Accepted: 11 November 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract

Nonnegative Matrix Factorization (NMF), as a group representation learning model, produces part-based representation with interpretable features and can be applied to various problems, such as text clustering. The findings indicate that the NMF model with Kullback-Leibler divergence (NMFk) exhibits promising performance in the task of text clustering. However, existing NMF-based text clustering methods are defined within a latent decoder model, lacking a verification mechanism. Recently, self-representation techniques have been applied to a wide range of tasks, empowering models to learn and verify representations of their input data autonomously. This paper proposes a self-representation factorization model for text clustering that incorporates semantic information into its learning process. The Semantic-aware Encoder-Decoder NMF model based on Kullback-Liebler divergence (SEDNMFk), integrates encoder and decoder factorizations into a Kullback-Liebler cost function that mutually verify and refine each other, resulting in the formation of more distinct clusters. To further enhance the semantic properties of the method, we add a tailored semantic regularization to the model. Due to its autoencoder-like architecture, SEDNMFk, and utilization of contextual information, produces more informative word embeddings with generalization abilities that are applicable to out-of-sample data. We present an efficient and effective optimization algorithm based on multiplicative update rules to solve the proposed unified model. The experimental results on the seven well-known datasets show that the proposed SEDNMFk model outperforms other state-of-the-art text clustering methods in both fully observed and out-of-sample settings.

Keywords Nonnegative matrix factorization · Encoder-decoder structure · Kullback-Leibler divergence · Self-representation · Text clustering

1 Introduction

Text clustering is a process of extracting useful knowledge or patterns from text collections that are not well-formed or partially formed, by grouping texts into different clusters based on their content similarity. This task can assist in organizing, summarizing, and visualizing large amounts of text data, such as news articles, social messages, scientific papers, etc [1, 2]. Text clustering is a difficult task because text data are often noisy, high-dimensional, and ambiguous [3], and it additionally requires choosing suitable data representation methods for different types of text data and applications [4, 5]. Representation learning is a fundamental topic in unsupervised learning and an essential step for various applications [6–8]. Its main objective is to overcome the limitations of hand-crafted features and reduce the data dimensionality and complexity. This is particularly crucial for high-dimensional data commonly used in domains such

✉ Fardin Akhlaghian Tab
f.akhlaghian@uok.ac.ir

Sayvan Soleymanbaigi
s.soleymanbaigi@uok.ac.ir

Amjad Seyedi
seyedamjad.seyedi@umons.ac.be

Fatemeh Daneshfar
f.daneshfar@uok.ac.ir

¹ Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

² Department of Mathematics and Operational Research, University of Mons, Mons, Belgium

as computer vision, clustering, and information retrieval, where direct use for model learning is not feasible [9]. Furthermore, effective data representation can also contribute to improved performance and generalization in downstream tasks [10].

Nonnegative Matrix Factorization (NMF) is a dimensionality reduction and data representation technique proposed by Lee and Seung [11], and has demonstrated its effectiveness in various applications such as data clustering [12, 13], community detection [14–17], text clustering [18, 19], topic modeling [20], face recognition [21, 22], matrix completion [23–25], link prediction [26, 27], cross-modal retrieval [28], and hyperspectral image unmixing [29]. NMF incorporates the nonnegativity constraint and thus obtains the parts-based representation as well as enhancing the interpretability of the issue correspondingly. Text clustering using NMF usually decomposes a term-document matrix into a term-cluster matrix and a cluster-document matrix [30]. The term-cluster matrix can be interpreted as a representation of how terms are associated with different clusters or topics, while the cluster-document matrix reveals how documents are related to these clusters.

Different applications and data types use various cost functions to measure NMF approximation quality. Square error distance (SED) or Frobenius norm is the most common cost function for a wide range of applications, including text clustering [31]. Various methods have been developed to perform text clustering and topic modeling using this cost function. For example, semantic-assisted NMF (SeaNMF) [32] employs the basic NMF to decompose the term-document matrix and semantic correlation matrix into the shared latent space of the term-topic matrix. Deep NMF topic modeling [33] explores the unsupervised deep NMF framework which contains two parts, the first part is the unsupervised deep neural network for representing data and the second part is the basic NMF to learn the term-topic matrix. Nonnegative Matrix Tri-Factorization (NMTF) has proven to be useful for data co-clustering [34]. In the text clustering task, NMTF decomposes the term-document matrix into three nonnegative latent factor matrices. Parallel Nonnegative Matrix Tri-Factorization (PNMTF) [35] proposes a scalable method that is capable of updating matrix factors in parallel for text co-clustering tasks. The Kullback-Leibler divergence, commonly referred to as KL-divergence, functions as a cost metric within the context of NMF [36], aligning with the concept of additive Poisson noise [37]. It is proved that NMF with KL divergence (NMFk) is equivalent to Probabilistic Latent Semantic Indexing (PLSI) [38] and it has been used for text clustering. Regularized Asymmetric NMF (RANMF) [39] uses NMFk as the main objective function and the similarity between documents as regularized constraints. Similar to the SeaNMF model, Semantic

NMFk (SeNMFk) [40] decomposes the term-document matrix and word-word co-occurrence matrix to the common space of a term-topic matrix. In addition, this model ensembles multiple NMFk to determine an accurate latent number of topics in text corpora. More recently, Yuan et al. [41] introduced Biorthogonal- β NMF, employing generalized β -divergence as a dissimilarity metric instead of the Euclidean distance for matrix comparison. Moreover, they integrated biorthogonal constraints into the factorization, guaranteeing both feature orthogonality and data sparsity in representation. Recently, Concept Factorization (CF), a matrix factorization technique, has gained attention for its applications in image clustering, data representation, and combined tasks [42]. It works by representing concepts as linear combinations of data points and approximating data points as linear combinations of these concepts [43].

Existing NMF-based text clustering methods, such as NMFk and SeaNMF, typically operate within a decoder-only framework, decomposing the term-document matrix into latent factors without a mechanism to verify or refine these representations against the input data. This lack of verification stems from their reliance on a single-directional factorization, which assumes the latent space accurately captures the data structure without explicitly checking its consistency. In real-world applications, such as news article clustering or social media analysis, this limitation can lead to suboptimal performance due to the inability to adapt to noisy, high-dimensional, or out-of-sample data. Without a feedback loop to validate the learned clusters, these models may produce less distinct or interpretable groupings, reducing their effectiveness in dynamic, large-scale text processing tasks where robustness and generalization are critical.

Despite the interpretable part-based representation offered by the aforementioned methods and the integration of semantic information, there exists a notable absence of research that systematically investigates the synergistic interplay between self-representation and semantic information integration within these methods. This research gap specifically concerns the investigation into the comprehensive effects of combining self-representation and semantic information on key aspects such as model performance, interpretability, and the ability to generalize to out-of-sample data. This research aims to address this research gap by developing an innovative self-representation NMFk model.

The proposed SEDNMFk differs significantly from existing NMF-based methods, such as NMFk, SeaNMF, and BO- β NMF, by integrating an encoder-decoder framework with semantic regularization into a unified KL-divergence-based cost function. Unlike NMFk and SeaNMF, which rely on decoder-only factorization and lack a verification mechanism, SEDNMFk employs a self-representation approach where the encoder (transforming data into latent

space) and decoder (reconstructing the input) mutually refine each other, enhancing cluster accuracy and generalization, particularly for out-of-sample data. While SeaNMF incorporates semantic correlations and BO- β NMF uses β -divergence with biorthogonal constraints for noise robustness, SEDNMFk uniquely combines word-word co-occurrence information (via SPPMI) as a tailored regularization term with the encoder-decoder structure, improving both interpretability and clustering performance. This novel synergy enables SEDNMFk to outperform state-of-the-art methods by capturing both local semantic relationships and global data structure, addressing limitations in robustness and adaptability found in prior works. Distinct from prior NMF-based approaches that focus solely on decoder factorization, our work introduces SEDNMFk, which combines an encoder-decoder structure with semantic regularization using word-word co-occurrence information. This novel integration not only improves clustering accuracy but also provides new insights into achieving interpretable and generalizable representations for text data, surpassing the limitations of existing methods. This model not only achieves high text clustering performance with superior generalization ability but also attempts to enhance the interpretability of NMF, thereby bridging the existing gap between model performance and interpretability within the context of NMF.

Recently, the Encoder-Decoder NMF model has been introduced for community detection tasks, which consists of decoder and encoder factorization modules [44]. The Decoder part is a basic NMF that reconstructs the original data matrix from a multiplicative of factors matrix ($\mathbf{X} \approx \mathbf{WH}$), and the encoder part transforms the original data matrix to latent representation space using basis matrix ($\mathbf{H} \approx \mathbf{W}^T \mathbf{X}$). This self-representation model integrates both decoder and encoder terms into a unified cost function. In this paper, we propose a novel Semantic Encoder-Decoder NMF with KL-divergence (SEDNMFk) for text clustering. This model integrates encoder and decoder NMF parts based on the KL-divergence loss function compatible with the text clustering problem. In this self-representation model, encoder and decoder NMFk modules by refining and verifying each other can form more precise clusters and learn global topic modeling information. Consequently, the extracted term-cluster matrix by this model has generalization properties and can handle out-of-sample documents more efficiently. In addition, to induce local semantic information and to enhance interpretability, word-word co-occurrence information is incorporated into this model by defining a tailored regularization term. This paper also develops an optimization scheme to solve the cost function by multiplicative updating rules. The key contributions of this paper are summarized as follows:

- We introduce a novel Semantic Encoder-Decoder NMF with KL-divergence that leverages a self-representation mechanism to enhance document clustering, topic modeling, and generalization for both in-sample and out-of-sample data.
- The proposed model incorporates word-word co-occurrence information as semantic regularization, improving clustering performance and interpretability.
- We develop an efficient optimization scheme with tailored multiplicative updating rules to solve the unified cost function, ensuring rapid and effective convergence.

This paper has the following structure: First, the backgrounds of basic NMF, Encoder-Decoder NMF, and NMF with KL divergence are introduced in Section 2. In Section 3, the proposed models and their numerical solutions are presented. The experimental results that show the effectiveness of our method are provided in Section 4. Finally, the conclusion will be provided in Section 5.

2 Background

This section introduces some preliminaries including, basic NMF, Encoder-Decoder NMF, and NMF with Kullback-Leibler Divergence models. In this paper, we use capital bold letters (like \mathbf{X}) for matrices, lowercase bold letters (like \mathbf{x}) for vectors, and regular letters (like a) for scalars. We also use \mathbf{x}_i , $\mathbf{x}^{(j)}$, and X_{ij} to mean the i -th column vector, the j -th row vector, and the element in the i -th row and j -th column of matrix \mathbf{X} , respectively. In addition, KL divergence and Frobenius norm are denoted by $D(\cdot|\cdot)$ and $\|\cdot\|_F$, respectively.

2.1 Basic nonnegative matrix factorization

Given nonnegative data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ consists of n samples where each sample has d features. NMF is a technique that decomposes the matrix \mathbf{X} into two nonnegative matrices \mathbf{W} and \mathbf{H} , such that $\mathbf{X} \approx \mathbf{WH}$. This factorization provides a part-based representation of the data, where each element is approximated as a nonnegative combination of the basis vectors, $\mathbf{x}_i \approx \mathbf{W}\mathbf{h}_i$. Thus, the objective function of NMF with Frobenius norm is defined as follows:

$$\min_{\mathbf{W}, \mathbf{H}} \mathcal{L}_f = \|\mathbf{X} - \mathbf{WH}\|_F^2 \quad \text{s.t.} \quad \mathbf{W}, \mathbf{H} \geq 0, \quad (1)$$

where $\|\cdot\|_F$ indicates the Frobenius norm of a matrix, $\mathbf{W} \in \mathbb{R}^{d \times k}$ is the basis matrix, $\mathbf{H} \in \mathbb{R}^{k \times n}$ is the low dimensional representation matrix, k is dimension of the

low-dimensional representation [36]. The update rules associated with the objective function (1) are

$$W \leftarrow W \odot \frac{XH^T}{WHH^T}, \tag{2}$$

$$H \leftarrow H \odot \frac{W^T X}{W^T W H}. \tag{3}$$

where \odot indicates the Hadamard product.

2.2 Encoder–decoder nonnegative matrix factorization

From the previous model, we can infer that the NMF model depends only on the decoder loss function that reconstructs the original input X from the latent space. Furthermore, having only a decoder structure requires executing the whole procedure for every out-of-sample as input for the factorization process. To address these issues and utilize the autoencoder’s representation learning potential, one of the most effective approaches is to modify the loss function to include the encoder stage in the factorization process [45]. This modification enables the unification of the loss functions between the encoder and decoder stages. The loss function of the Encoder NMF with Frobenius norm can be considered as follows:

$$\min_{W,H} \mathcal{L}_{Ef} = \|H - W^T X\|_F^2 \quad \text{s.t.} \quad W, H \geq 0, \tag{4}$$

The objective function of encoder-decoder NMF is:

$$\min_{W,H} \mathcal{L}_{EDf} = \|X - WH\|_F^2 + \|H - W^T X\|_F^2 \quad \text{s.t.} \quad W, H \geq 0, \tag{5}$$

Bing-Jie et al. [46] proposed multiplicative updating rules to update W and H :

$$W \leftarrow W \odot \frac{2XH^T}{WHH^T + XX^T W} \tag{6}$$

$$H \leftarrow H \odot \frac{2W^T X}{W^T W H + H} \tag{7}$$

2.3 Nonnegative matrix factorization with kullback-leibler divergence

Kullback–Leibler (KL) measures divergence between two distributions which is used as a common cost function to quantify the approximation. In the context of NMF, KL

divergence represents an asymmetric measure of divergence between two matrices [36] as follows,

$$\min_{W,H} \mathcal{L}_k = \mathcal{D}(X \| WH) = \sum_{i=1}^d \sum_{j=1}^n X_{ij} \log \frac{X_{ij}}{[WH]_{ij}} - X_{ij} + [WH]_{ij} \quad \text{s.t.} \quad W, H \geq 0. \tag{8}$$

This divergence quantifies how different the distribution X is from the distribution WH , and it is always greater than or equal to zero. It only becomes zero when $X = WH$. However, it cannot be considered as a distance due to its absence of symmetry, thus violating the properties of a metric, and is often referred to as a divergence. The multiplicative update rules that were proposed by Lee and Seung for minimizing this cost function are as follows [36]:

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{j=1}^n H_{kj} X_{ij} / [WH]_{ij}}{\sum_{j=1}^n H_{kj}} \tag{9}$$

$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^d W_{ik} X_{ij} / [WH]_{ij}}{\sum_{i=1}^d W_{ik}} \tag{10}$$

3 Proposed model

This section proposes the Semantic Encoder-Decoder NMF model with Kullback-Leibler divergence (SEDNMFk), a specialized factorization for text clustering problem. Its success is mainly due to four factors: (1) It employs a KLD loss to be robust against Poisson noise distribution and to cover the input sparsity; (2) In a KLD-based framework, it adds an encoder NMF term to the Decoder NMF that refines and verifies the factorization process, providing a mechanism to handle the out-of-sample data; (3) It adds the semantic correlation information to the proposed Encoder-Decoder NMFk by a tailored regularization to enhance the semantic properties of factorization; (4) It integrates the above into one joint learning problem and adopts an efficient alternating minimization strategy for optimization. The illustration of the proposed model is presented in Figure 1.

3.1 Pre-processing

We employ the term frequency-inverse document frequency (tf-idf) vector space model to represent documents [47]. Consider the given documents $D = \{d_1, d_2, \dots, d_n\}$, and the terms $V = \{v_1, v_2, \dots, v_p\}$ occur in the documents. The tf-idf for i -th word and j -th document is defined as:

$$\text{tf-idf}(v_i, d_j) = \text{tf}(v_i, d_j) \times \text{idf}(v_i), \tag{11}$$

Fig. 1 The illustration of the semantic Encoder-Decoder NMF with Kullback–Leibler divergence model (SEDNMFk)

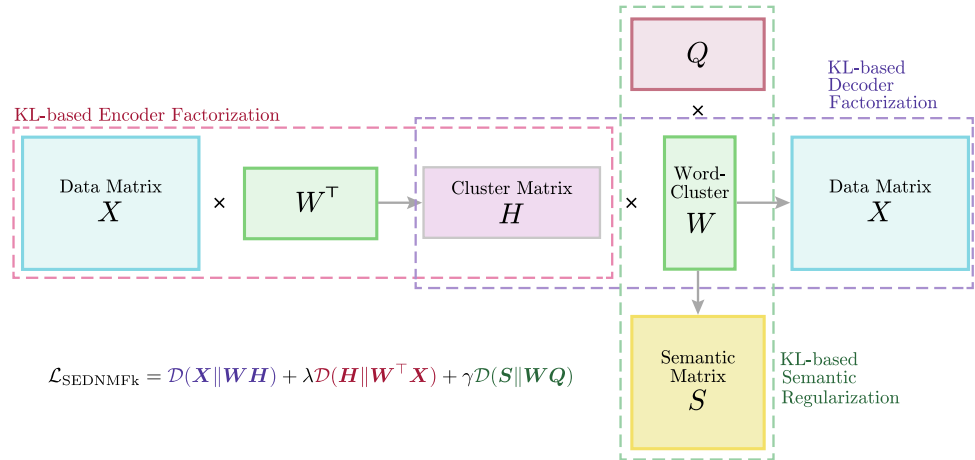


Table 1 Notation used in this paper

Notation	Description
\mathbf{X}	Term-document (word-document) matrix.
\mathbf{S}	Word-word co-occurrence (SPPMI) matrix.
\mathbf{W}	Latent factor matrix of words.
\mathbf{H}	Latent factor matrix of documents.
\mathbf{Q}	Latent factor matrix of contexts.
p	Number of distinct words in the vocabulary.
n	Number of documents in the corpus.
c	Number of clusters.

$$\text{idf}(v_i) = \log \frac{(n + 1)}{\text{df}(v_i) + 1} + 1, \tag{12}$$

where $\text{tf}(v_i, d_j)$ is the number of occurrences of v_i in d_j and $\text{df}(v_i)$ denotes the number of documents in which v_i occurs. The term-document matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$ is created by Eq. (11), where $X_{ij} = \text{tf-idf}(v_i, d_j)$. The important symbols used in this paper are summarized in Table 1.

3.2 Encoder–decoder NMFk

Given a nonnegative term-document matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, each row of input corresponds to a word and each column corresponds to a document. NMFk (8) tries to discover two nonnegative matrices \mathbf{W} and \mathbf{H} where $\mathbf{W} \in \mathbb{R}^{p \times c}$ is a term-cluster matrix and $\mathbf{H} \in \mathbb{R}^{c \times n}$ is a cluster-document matrix. To enhance the quality of representation and handle out-of-sample, EDNMFk introduces encoder part which transform term-document matrix \mathbf{X} into the cluster-document representation \mathbf{H} using term-cluster matrix \mathbf{W} , with KL divergence loss function as follows:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{H} \geq 0} \mathcal{L}_{\text{Ek}} &= \mathcal{D}(\mathbf{H} \|\mathbf{W}^\top \mathbf{X}) \\ &= \sum_{k=1}^c \sum_{j=1}^n H_{kj} \log \frac{H_{kj}}{[\mathbf{W}^\top \mathbf{X}]_{kj}} - H_{kj} + [\mathbf{W}^\top \mathbf{X}]_{kj}, \end{aligned} \tag{13}$$

EDNMFk combines encoder (13) and decoder (8) parts into a unified cost function as follows:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{H} \geq 0} \mathcal{L}_{\text{EDk}} &= \mathcal{D}(\mathbf{X} \|\mathbf{W}\mathbf{H}) + \lambda \mathcal{D}(\mathbf{H} \|\mathbf{W}^\top \mathbf{X}) \\ &= \sum_{i=1}^p \sum_{j=1}^n X_{ij} \log \frac{X_{ij}}{[\mathbf{W}\mathbf{H}]_{ij}} - X_{ij} + [\mathbf{W}\mathbf{H}]_{ij} \\ &\quad + \lambda \sum_{k=1}^c \sum_{j=1}^n H_{kj} \log \frac{H_{kj}}{[\mathbf{W}^\top \mathbf{X}]_{kj}} - H_{kj} + [\mathbf{W}^\top \mathbf{X}]_{kj}, \end{aligned} \tag{14}$$

where the parameter λ controls the importance of self-expression representation.

3.3 Semantic encoder–decoder NMFk

The term-document matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$ provides a global perspective by representing each word solely based on its frequency across the documents. In this study, for capturing local information, we extract the word co-occurrence matrix $\mathbf{S} \in \mathbb{R}^{p \times p}$, which can measure the associations among words [48]. A word-word co-occurrence matrix is a way of representing the relationship between words in a text based on how frequently they appear together in a certain context. The context can be defined as a window of words around the target word. The matrix has words as both rows and columns, with each cell representing the frequency or weight of the co-occurrence between the corresponding row word and column word. We further rely on a non-linear transformation of the word-word co-occurrences, based on the Point-wise Mutual Information (PMI) [49]. The PMI is an information-theoretic measure widely used to quantify the association between pairs of outcomes arising from discrete random variables. PMI quantifies how frequently the two words co-occur, relative to their independent occurrences. The expression for PMI is,

$$PMI(i, l) = \log\left(\frac{c_{il} \times \sum_{i,l \in V} c_{il}}{\sum_{l \in V} c_{il} \times \sum_{i \in V} c_{il}}\right), \tag{15}$$

where co-occurrence c_{il} counts the word pairs i and l , and $\sum_{i,l \in V} c_{il}$ counts the total number of word pairs. Additionally, $\sum_{l \in V} c_{il}$ and $\sum_{i \in V} c_{il}$ give the number of times words i and l appear independently. To ensure non-negativity, the SPPMI word-context matrix S is defined by shifting the PMI using a constant $\tau \geq 0$ and taking the maximum of the resulting value and 0 as follows,

$$S_{i,l} = \max(PMI(i, l) - \log(\tau), 0), \quad \forall i, l \in \{1, 2, \dots, p\}. \tag{16}$$

It has been demonstrated that word and context embeddings can be obtained by factorizing the SPPMI word-context matrix S [50]. According to the sparsity of the word-context matrix S and to define a regularization compatible with the KLD-based objective function (14), we introduce a semantic regularization term as follows:

$$\min_{T, Q \geq 0} \mathcal{R}(S) = \mathcal{D}(S || TQ) = \sum_{i=1}^p \sum_{l=1}^p S_{il} \log \frac{S_{il}}{[TQ]_{il}} - S_{il} + [TQ]_{il}, \tag{17}$$

where T and Q are word embedding and context embedding matrices, respectively. This term improves the distinction of clusters, particularly when dealing with complex datasets involving multiple classes. To learn a more informative word embedding from the frequency information X and the semantic information S , we define a shared factor for self-representation (14) and semantic factorization (17) by assuming $T = W$. Finally, by adding semantic regularization term (17) to the EDNMFk model (14), the objective function of Semantic EDNMFk (SEDNMFk) is expressed as follows:

$$\begin{aligned} & \min_{W, H, Q \geq 0} \mathcal{L}_{SEDNMFk} \\ & = \mathcal{D}(X || WH) + \lambda \mathcal{D}(H || W^T X) + \gamma \mathcal{D}(S || WQ) \\ & = \sum_{i=1}^p \sum_{j=1}^n X_{ij} \log \frac{X_{ij}}{[WH]_{ij}} - X_{ij} + [WH]_{ij} \\ & + \lambda \sum_{k=1}^c \sum_{j=1}^n H_{kj} \log \frac{H_{kj}}{[W^T X]_{kj}} - H_{kj} + [W^T X]_{kj} \\ & + \gamma \sum_{i=1}^p \sum_{l=1}^p S_{il} \log \frac{S_{il}}{[WQ]_{il}} - S_{il} + [WQ]_{il}, \end{aligned} \tag{18}$$

where hyperparameter γ controls the contribution of semantic information.

3.4 Optimization

The objective function (18) is non-convex, and thus quite challenging to solve. To obtain the optimal solution of the SEDNMFk objective function, we apply the multiplicative update rules of $\mathcal{L}_{SEDNMFk}$ to optimize W , H , and Q alternatively and iteratively. The constraint set for the objective function includes $W \geq 0$, $H \geq 0$, and $Q \geq 0$. Therefore, the Lagrange function \mathcal{L} is given by introducing Lagrange multipliers Ψ , Φ , and Γ where $\Psi = [\Psi_{ik}]$, $\Phi = [\Phi_{kj}]$ and $\Gamma = [\Gamma_{il}]$. Finally, the Lagrange function of the SEDNMFk is,

$$\begin{aligned} & L(W, H, Q) \\ & = \sum_{i=1}^p \sum_{j=1}^n X_{ij} \log \frac{X_{ij}}{[WH]_{ij}} - X_{ij} + [WH]_{ij} \\ & + \lambda \sum_{k=1}^c \sum_{j=1}^n H_{kj} \log \frac{H_{kj}}{[W^T X]_{kj}} - H_{kj} + [W^T X]_{kj} \\ & + \gamma \sum_{i=1}^p \sum_{l=1}^p S_{il} \log \frac{S_{il}}{[WQ]_{il}} - S_{il} + [WQ]_{il} \\ & - Tr(\Psi W^T) - Tr(\Phi H^T) - Tr(\Gamma Q^T). \end{aligned} \tag{19}$$

The first-order partial derivatives w.r.t W , H and Q are

$$\begin{aligned} \frac{\partial L}{\partial W_{ik}} & = - \sum_{j=1}^n H_{kj} \frac{X_{ij}}{[WH]_{ij}} + \sum_{j=1}^n H_{kj} \\ & + \lambda \left(- \sum_{j=1}^n X_{ij} \frac{H_{kj}}{[W^T X]_{kj}} + \sum_{j=1}^n X_{ij} \right) \\ & + \gamma \left(- \sum_{l=1}^p Q_{kl} \frac{S_{il}}{[WQ]_{il}} + \sum_{l=1}^p Q_{kl} \right) - \psi_{ik}, \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial L}{\partial H_{kj}} & = - \sum_{i=1}^p W_{ik} \frac{X_{ij}}{[WH]_{ij}} + \sum_{i=1}^p W_{ik} \\ & + \lambda \left(- \log [W^T X]_{kj} + \log H_{kj} \right) - \phi_{kj}, \end{aligned} \tag{21}$$

and

$$\frac{\partial L}{\partial Q_{kl}} = - \sum_{i=1}^p W_{ik} \frac{S_{il}}{[WQ]_{il}} + \sum_{i=1}^p W_{ik} - \gamma_{kl}. \tag{22}$$

By setting the first-order partial derivatives to zero, we have:

$$\begin{aligned} \psi_{ik} = & - \sum_{j=1}^n H_{kj} \frac{X_{ij}}{[WH]_{ij}} + \sum_{j=1}^n H_{kj} \\ & + \lambda \left(- \sum_{j=1}^n X_{ij} \frac{H_{kj}}{[W^T X]_{kj}} + \sum_{j=1}^n X_{ij} \right) \\ & + \gamma \left(- \sum_{i=1}^p Q_{kl} \frac{S_{il}}{[WQ]_{il}} + \sum_{i=1}^p Q_{kl} \right), \end{aligned} \tag{23}$$

$$\begin{aligned} \phi_{kj} = & - \sum_{i=1}^p W_{ik} \frac{X_{ij}}{[WH]_{ij}} + \sum_{i=1}^p W_{ik} \\ & + \lambda (-\log[W^T X]_{kj} + \log H_{kj}), \end{aligned} \tag{24}$$

$$\gamma_{kl} = - \sum_{i=1}^p W_{ik} \frac{S_{il}}{[WQ]_{il}} + \sum_{i=1}^p W_{ik}. \tag{25}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions $\psi_{ik}W_{ik} = 0$, $\phi_{kj}H_{kj} = 0$, and $\gamma_{kl}Q_{kl} = 0$, we obtain all the updating rules, where they are written in matrix form as follows:

$$W \leftarrow W \odot \frac{\frac{X}{WH}H^T + \lambda \frac{H}{W^T X}X^T + \gamma \frac{S}{WQ}Q^T}{1H^T + \lambda 1X^T + \gamma 1Q^T}, \tag{26}$$

$$H \leftarrow H \odot \frac{W^T \frac{X}{WH} + \lambda \log(W^T X)}{W^T 1 + \lambda \log(H)} \tag{27}$$

$$Q \leftarrow Q \odot \frac{W^T \frac{S}{WQ}}{W^T 1} \tag{28}$$

where \odot indicates the Hadamard product and $\mathbf{1}$ is matrix of all ones. The optimization process of SEDNMFk is provided in Algorithm 1.

Input: Term-document matrix X , number of cluster c , scale parameter λ , semantic parameter γ ;

Output: Cluster-document matrix H ;

- 1: Calculate SPPMI matrix S according to (16);
- 2: Initialize W , H , and Q matrices randomly;
- 3: **while** Convergence **do**
- 4: Update word-cluster matrix W according to (26);
- 5: Update cluster-document matrix H according to (27);
- 6: Update word-context matrix Q according to (28);
- 7: **end while**
- 8: **return** W , H , and Q ;

Algorithm 1 Semantic Encoder Decoder NMFk (SEDNMFk)

3.5 Complexity analysis

The proposed Semantic Encoder-Decoder NMF employs an

iterative approach (Algorithm 1) to tackle the non-convex optimization problem. In each iteration, the three factor matrices, W , H , and Q , are updated. The computational complexity of multiplying the matrix $X \in \mathbb{R}^{p \times n}$ with the factor matrices, which are $H \in \mathbb{R}^{c \times n}$ and $W \in \mathbb{R}^{p \times c}$, is $O(pnc)$. Consequently, the time complexity of updating H is $O(pnc)$, aligning with other NMF models. The complexity of updating W and Q depend on the time complexity of calculating the SPPMI for all words, which is $O(p^2)$. As a result, updating W and Q have a time complexity of $O(p^2c)$. Since matrix multiplication is the dominant factor, the overall complexity per iteration of the algorithm is $O(pnc + p^2c)$. In conclusion, the overall complexity of the proposed algorithm is directly proportional to the number of iterations t required for convergence, falling within the range of $O(t(pnc + p^2c))$. This complexity indicates that SEDNMFk scales efficiently with the number of documents (n) and vocabulary size (p), making it practically viable for large-scale text clustering tasks, especially given its rapid convergence demonstrated in experiments.

4 Experimental results

In this section, extensive experiments are conducted to validate the effectiveness of the proposed model in comparison to 10 baseline and state-of-the-art models on seven benchmark datasets. These experiments include performance evaluation, parameter analysis, ablation study, and out-of-sample analysis, which are discussed in detail.

4.1 Datasets

We perform the experiments on seven datasets to show the superior performance and generalization abilities of

SEDNMFk versus the other compared methods. WebKB dataset is the web pages from the computer science departments of some universities that were manually grouped into seven categories including Student, Faculty, Course,

Project, Staff, Department, and Other by the CMU text learning group in 1997. We did not use the Other category because the page contents are very different in this group. The Department and Staff categories are also not used because there were only a few pages from each university in them. The four selected categories include 4199 documents. BBCNews dataset is a collection of news articles from the BBC News website that are categorized into five topics (Business, Entertainment, Politics, Sport, and Tech). This dataset consists of 2225 documents from 2004-2005. The third dataset is Reuters-21578 which contains 21578 texts that appeared on the Reuters news-wire in 1987. We use two versions of Reuters-21578 (i.e. Reuters4c and Reuters10c) with four and ten clusters that contain 7632 and 9979 texts, respectively. AGNews dataset is a collection of news articles from different sources that are categorized into four topics (World, Sports, Business, and Sci/Tech). Finally, the 20Newsgroups dataset has 18821 texts from newsgroups, divided into 20 different topics. Some topics are closely related, while others are distinctly different. The YahooAnswers dataset consists of 60000 user-generated questions and their corresponding answers, organized into 10 distinct categories. A summary of the datasets is provided in Table 2

4.2 Evaluation metrics

To evaluate the proposed method’s performance, we use three quantitative metrics, Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI) to measure the performance of clustering. Based on Mutual Information (MI), NMI is a metric that quantifies the overlap between two distributions in a normalized manner. It is computed as the fraction of the mutual information of the distributions and the mean entropy of the distributions. The Mutual Information (MI) criterion for clusterings C and Y is defined as

$$MI(C, Y) = \sum_{c_i \in C, y_j \in Y} p(c_i, y_j) \log \frac{p(c_i, y_j)}{p(c_i)p(y_j)} \quad (29)$$

where $p(c_i)$ and $p(y_j)$ characterize the likelihood that documents are contained within cluster c_i and cluster y_j ,

Table 2 The detailed of the real-world datasets

Dataset	#document	#word	#class
WebKB	4199	2000	4
BBCNews	2225	2000	5
Reuters4c	7632	500	4
Reuters10c	9979	1000	10
AGNews	7600	1000	4
20Newsgroups	18821	4000	20
YahooAnswers	60000	1000	10

respectively. On the other hand, the probability $p(c_i, y_j)$ signifies the likelihood that the documents belong to both cluster c_i and cluster y_j simultaneously. As an extension of the MI, NMI is expressed as

$$NMI(C, Y) = \frac{MI(C, Y)}{\max(H(C), H(Y))} \quad (30)$$

where $H(C)$ and $H(Y)$ give, respectively, the entropy values of the sets C and true labels Y .

The Clustering Accuracy (ACC) measure shows the proportion of documents that are correctly assigned to their true classes based on the clustering results. Its specific formula is as follows:

$$ACC(C, Y) = \frac{\sum_{i=1}^n \delta(\text{map}(c_i), y_i)}{n}, \quad (31)$$

where n stands for the total count of documents. Each document is assigned an actual label denoted as y_i and the function $\hat{y}_i = \text{map}(c_i)$ represents the optimal mapping function that reorganizes the cluster labels to align as closely as possible with the actual labels. The $\delta(\cdot, \cdot)$ function is applied, yielding a value of 1 when y_i equals \hat{y}_i and 0 otherwise.

The Adjusted Rand Index (ARI) quantifies the similarity between two cluster sets. ARI scores can range from negative values, when the resemblance is lower than random chance, to a maximum score of 1, when the clusters are a perfect match. The equation for calculating ARI is as follows:

$$ARI(C, Y) = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2}}{\sum_j \binom{n_{.j}}{2} / \binom{n}{2}} \quad (32)$$

$$= \frac{1}{2} \left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right] - \sum_j \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}$$

where C refers to the clustering results, and Y denotes the true ground-truth clustering labels. n_{ij} represents the number of documents that are common to both cluster c_i and cluster y_j , while $n_{i.}$ and $n_{.j}$ stands for the document count within cluster c_i and cluster y_j , respectively.

4.3 Compared methods

We compare our proposed method, SEDNMFk with seven other models, including baselines and state-of-the-art NMF

clustering methods. The comparison methods are listed as follows:

- **NMF** model factorizes the input matrix based on square error distance (SED) [36].
- **NMFk** utilizes Kullback-Leibler divergence (KL-divergence) for evaluating the quality of approximation [36].
- **EDNMF** is an NMF extension that integrates encoder and decoder factorization terms [46].
- **SeaNMF** maps the term-document matrix and semantic correlation matrix into a shared term-cluster space for topic modeling [32].
- **WRNMTF** performs word clustering and document clustering, by a regularized Nonnegative Matrix Tri-Factorization [48].
- **RANMF** adds graph regularization to NMFk, which learns the semantic features of documents [39].
- **DGLCF** proposes the dual-graph global and local concept factorization to better reveal the complex inner manifold [51].
- **BO- β NMF** is a novel NMF model that incorporates β -divergence and biorthogonal regularization, allowing it to learn a representation of the data that is robust to noise and outliers [41].
- **OEDFS** (Orthogonal Encoder-Decoder Feature Selection) integrates feature selection into the NMF framework, focusing on discriminative features that enhance clustering performance [52].

- **EDA-TEC** is a self-representation approach designed for text clustering tasks, combining a deep autoencoder, graph regularization, and elastic loss to enhance feature representation and clustering performance [10].

4.4 Results

To evaluate the efficiency of the proposed model, the clustering quality of the SEDNMFk and other methods are compared on the seven real-world datasets. The Results of each method are evaluated by using NMI, ACC, and ARI metrics, which are presented in Tables 3–5. We run each method 10 times and report the mean and standard deviation. In each table, the best results are marked by bold and the second-best ones are marked by underline. From the Tables 3–5, we can derive that:

- Our proposed SEDNMFk method achieves the best performance across all datasets in terms of NMI, ACC, and ARI metrics. BO- β NMF emerges as the second-best-performing method on most datasets after our proposed model. While BO- β NMF provides improvements over traditional NMF variants by accounting for noise and outliers through beta-divergence and biorthogonal regularization, our method gains a significant additional boost by incorporating self-expressive representations and semantic information. On average, SEDNMFk achieves higher NMI, ACC, and ARI scores than the

Table 3 NMI results for seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.3052	0.7697	0.3472	0.4604	0.2368	0.4773	0.0475
	std	0.0153	0.0436	0.0171	0.0360	0.0430	0.0070	0.0095
NMFk	mean	0.3649	0.7621	0.5460	0.5202	0.3607	0.4562	0.1106
	std	0.0300	0.0640	0.0482	0.0201	0.0393	0.0091	0.0321
EDNMF	mean	0.3091	0.7317	0.3777	0.4666	0.2538	0.4402	0.0613
	std	0.0081	0.0601	0.0483	0.0401	0.0120	0.0140	0.0083
WRNMTF	mean	0.3525	0.6752	0.3595	0.4118	0.2762	0.4879	0.0964
	std	0.0091	0.0582	0.0108	0.0431	0.0140	0.0110	0.0092
SeaNMF	mean	0.3559	0.7350	0.3575	0.4818	0.2713	0.4781	0.0907
	std	0.0091	0.0651	0.0021	0.0201	0.0140	0.0081	0.0063
RANMF	mean	0.3463	0.7974	0.5482	0.5249	0.3551	0.5090	0.1235
	std	0.0162	0.0750	0.0302	0.0261	0.0751	0.0293	0.0260
DGLCF	mean	0.3527	0.7816	0.4461	0.5007	0.3152	0.5126	0.1208
	std	0.0150	0.0420	0.0210	0.0220	0.0210	0.0170	0.0124
BO- β NMF	mean	0.3719	<u>0.8203</u>	0.5567	<u>0.5482</u>	<u>0.4007</u>	<u>0.5333</u>	<u>0.1285</u>
	std	0.0125	0.0659	0.0334	0.0182	0.0368	0.0017	0.0253
OEDFS	mean	0.3216	0.7232	0.4430	0.4683	0.2621	0.4813	0.0882
	std	0.0141	0.0924	0.0589	0.0402	0.0210	0.0147	0.0094
EDA-TEC	mean	0.4153	0.7803	<u>0.5632</u>	0.4781	-	0.5256	-
	std	0.0.137	0.0506	0.0589	0.0402	-	0.0108	-
SEDNMFk	mean	<u>0.4026</u>	0.8318	0.6085	0.5571	0.4177	0.5531	0.1339
	std	0.0170	0.0092	0.0630	0.0150	0.0530	0.0211	0.0165

Table 4 ACC results for seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.5311	0.8965	0.5195	0.4561	0.4641	0.4553	0.1856
	std	0.0193	0.0634	0.0116	0.0480	0.0510	0.0110	0.0143
NMFk	mean	0.6473	0.8612	0.7015	0.4766	0.6323	0.4668	0.2749
	std	0.0431	0.0640	0.0538	0.0420	0.0680	0.0330	0.0346
EDNMF	mean	0.5634	0.8810	0.5974	0.4753	0.4537	0.4279	0.2147
	std	0.0380	0.0640	0.0439	0.0340	0.0370	0.0270	0.0232
WRNMTF	mean	0.6040	0.7823	0.5203	0.3695	0.4824	0.4790	0.2514
	std	0.0528	0.0630	0.0377	0.0290	0.0701	0.0170	0.0421
SeaNMF	mean	0.6045	0.8538	0.5245	0.4565	0.4676	0.4597	0.2612
	std	0.0528	0.0940	0.0041	0.0201	0.0450	0.0210	0.0311
RANMF	mean	0.5860	0.8906	0.7247	0.5009	0.6458	0.4973	0.2966
	std	0.0184	0.0980	0.0399	0.0330	0.0980	0.0540	0.0372
DGLCF	mean	0.6340	0.9024	0.4307	0.4861	0.5552	0.4860	0.2871
	std	0.0322	0.0620	0.0210	0.0310	0.0530	0.0220	0.0253
BO- β NMF	mean	0.6428	<u>0.9273</u>	0.7027	<u>0.5080</u>	<u>0.6869</u>	<u>0.5372</u>	<u>0.3052</u>
	std	0.0369	0.0700	0.0372	0.0501	0.0530	0.0145	0.0310
OEDFS	mean	0.5718	0.8435	0.6501	0.4955	0.5642	0.4632	0.2467
	std	0.0201	0.1230	0.0741	0.0286	0.0528	0.0129	0.0115
EDA-TEC	mean	<u>0.6521</u>	0.9045	<u>0.7271</u>	0.4935	-	0.4872	-
	std	0.0245	0.0712	0.0406	0.0360	-	0.0240	-
SEDNMFk	mean	0.6933	0.9388	0.7620	0.5761	0.7255	0.5551	0.3207
	std	0.0210	0.0041	0.0570	0.0460	0.0570	0.0370	0.0366

Table 5 ARI results for seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.2273	0.7891	0.1404	0.2863	0.2252	0.3040	0.0232
	std	0.0181	0.0790	0.0075	0.0601	0.0490	0.0110	0.0077
NMFk	mean	0.3583	0.7615	0.4882	0.3809	0.3670	0.3341	0.0836
	std	0.0349	0.0950	0.0693	0.0430	0.0510	0.0160	0.0129
EDNMF	mean	0.2471	0.7510	0.2257	0.3044	0.2365	0.2135	0.0356
	std	0.0418	0.0940	0.0821	0.0601	0.0220	0.0219	0.0050
WRNMTF	mean	0.3338	0.6337	0.1421	0.1777	0.2584	0.3299	0.0578
	std	0.0470	0.0801	0.0240	0.0610	0.0680	0.0170	0.0102
SeaNMF	mean	0.3066	0.7397	0.1431	0.3087	0.2515	0.3047	0.0519
	std	0.0090	0.1060	0.0051	0.0280	0.0210	0.0120	0.0085
RANMF	mean	0.3306	0.8030	0.5120	0.4030	0.3624	0.3735	0.0979
	std	0.0196	0.0981	0.0457	0.0451	0.0970	0.0250	0.0174
DGLCF	mean	0.3204	0.8019	0.2594	0.3481	0.2715	0.3690	0.0905
	std	0.0280	0.0750	0.0590	0.0230	0.0470	0.0130	0.0166
BO- β NMF	mean	0.3418	<u>0.8451</u>	0.4844	<u>0.4167</u>	<u>0.4292</u>	<u>0.4066</u>	<u>0.1019</u>
	std	0.0430	0.0881	0.0400	0.0565	0.0502	0.0085	0.0113
OEDFS	mean	0.2583	0.7245	0.3218	0.3053	0.2521	0.3142	0.0502
	std	0.0201	0.0595	0.0766	0.0407	0.0753	0.0231	0.0076
EDA-TEC	mean	0.4016	0.7981	<u>0.5542</u>	0.3836	-	0.3655	-
	std	0.0304	0.0745	0.0651	0.0205	-	0.0172	-
SEDNMFk	mean	<u>0.3983</u>	0.8567	0.5952	0.4917	0.4410	0.4262	0.1043
	std	0.0296	0.0090	0.0640	0.0310	0.0660	0.0230	0.0102

second-best methods by 0.025, 0.046, and 0.061, respectively. This means that SEDNMFk can discover more accurate clusters than the other methods.

- The experiments show that the performance of the compared methods varies across different datasets. For

example, NMFk performs well partitioning against the Frobenius-base model in most datasets except BBC-News and 20Newsgroup. The more recent KL-based model (i.e., RANMF) shows low performance on WebKB and 20newsgroup. However, our self-representation

framework outperforms other methods across all the evaluated datasets.

- In terms of NMI criteria, the Frobenius-based models slightly outperform the KL-based models on the 20Newsgroup. However, due to using semantic information the SEDNMFk model achieves the best result among all the models on this complex dataset.
- The results demonstrate that the SEDNMFk method leverages both local and global information from words and documents, consistently achieving stable and strong performance across a wide range of text datasets with diverse characteristics, including vocabulary size, number of clusters, and number of documents.

The superior performance of SEDNMFk over other methods can be attributed to its unique integration of self-representation and semantic information within a KL-divergence framework. Unlike traditional NMF variants (e.g., NMF and NMFk), which rely solely on decoder-based factorization, SEDNMFk’s encoder-decoder structure enables mutual refinement and verification of clusters, enhancing the robustness and precision of the learned representations. The incorporation of semantic regularization via the SPPMI matrix further enriches the model by capturing local contextual relationships, which improves cluster distinctiveness

and interpretability—particularly beneficial for complex datasets like 20Newsgroups. Additionally, the KL-divergence cost function aligns well with the sparse, noisy nature of text data, providing a probabilistic foundation that outperforms Frobenius norm-based methods in generalization, as evidenced by the out-of-sample results. Compared to advanced models like BO- β NMF, which leverages β -divergence and biorthogonal constraints for noise robustness, SEDNMFk’s combination of self-expression and semantic awareness offers a more comprehensive representation, resulting in higher NMI, ACC, and ARI scores across all datasets.

4.5 Parameter analysis

In this section, we analyze the influence of the hyperparameters on the clustering performance, where λ and γ control the contribution of self-expression and semantic information, respectively. In a grid search, we analyze the simultaneous effect of both parameters on the proposed model. Figures 2–4 illustrate the NMI, ACC, and ARI of the proposed method with various λ and γ on six datasets. Note that in these heatmap figures, the two axes correspond to the λ and γ parameters and the measures are represented by color, where darker colors indicate better results. In this

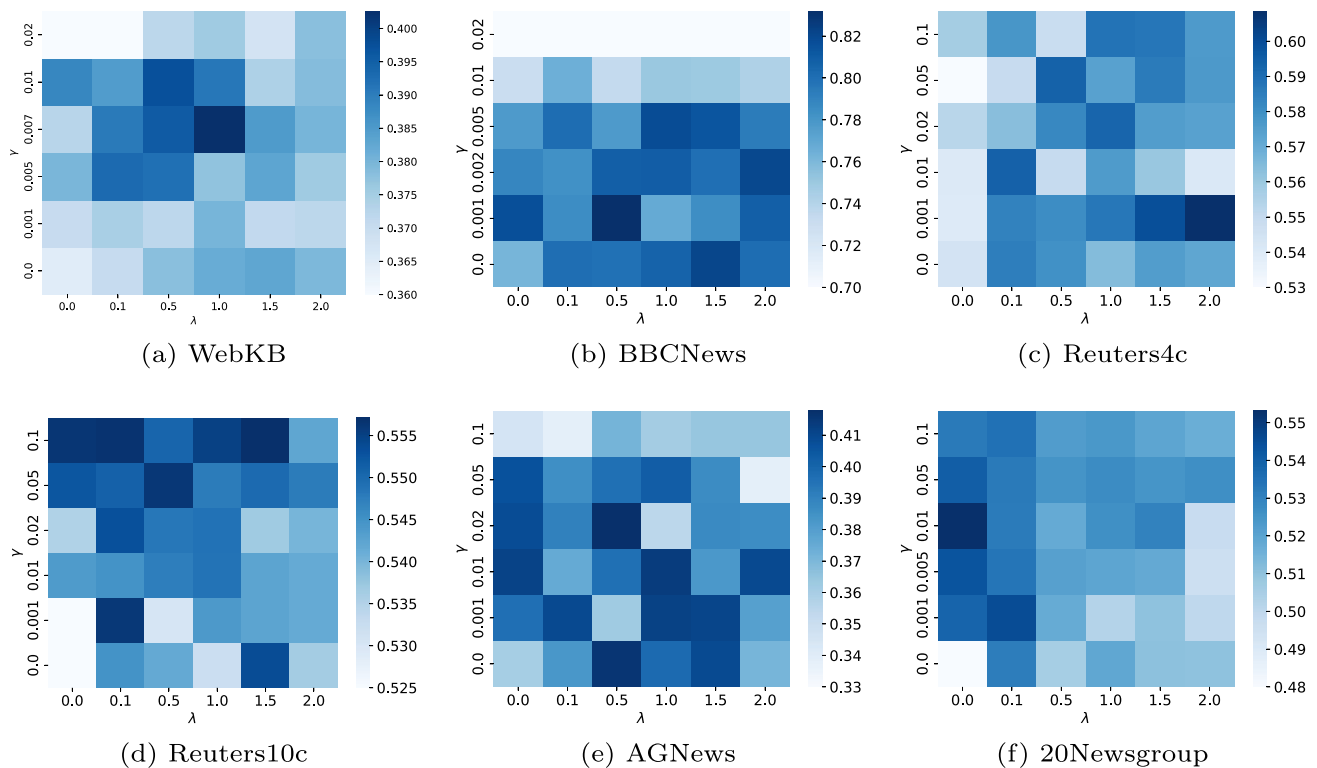


Fig. 2 Parameter analysis (in terms of NMI) on the λ and γ parameters

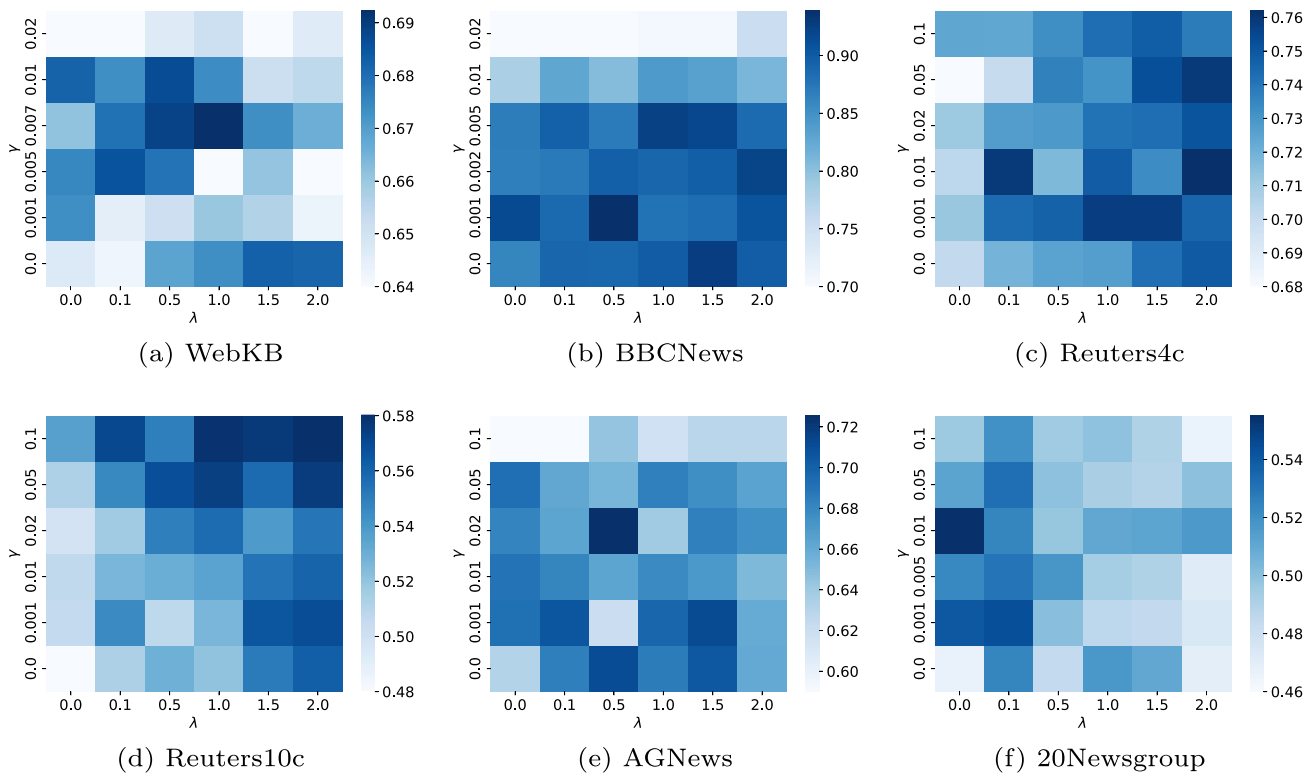


Fig. 3 Parameter analysis (in terms of ACC) on the λ and γ parameters

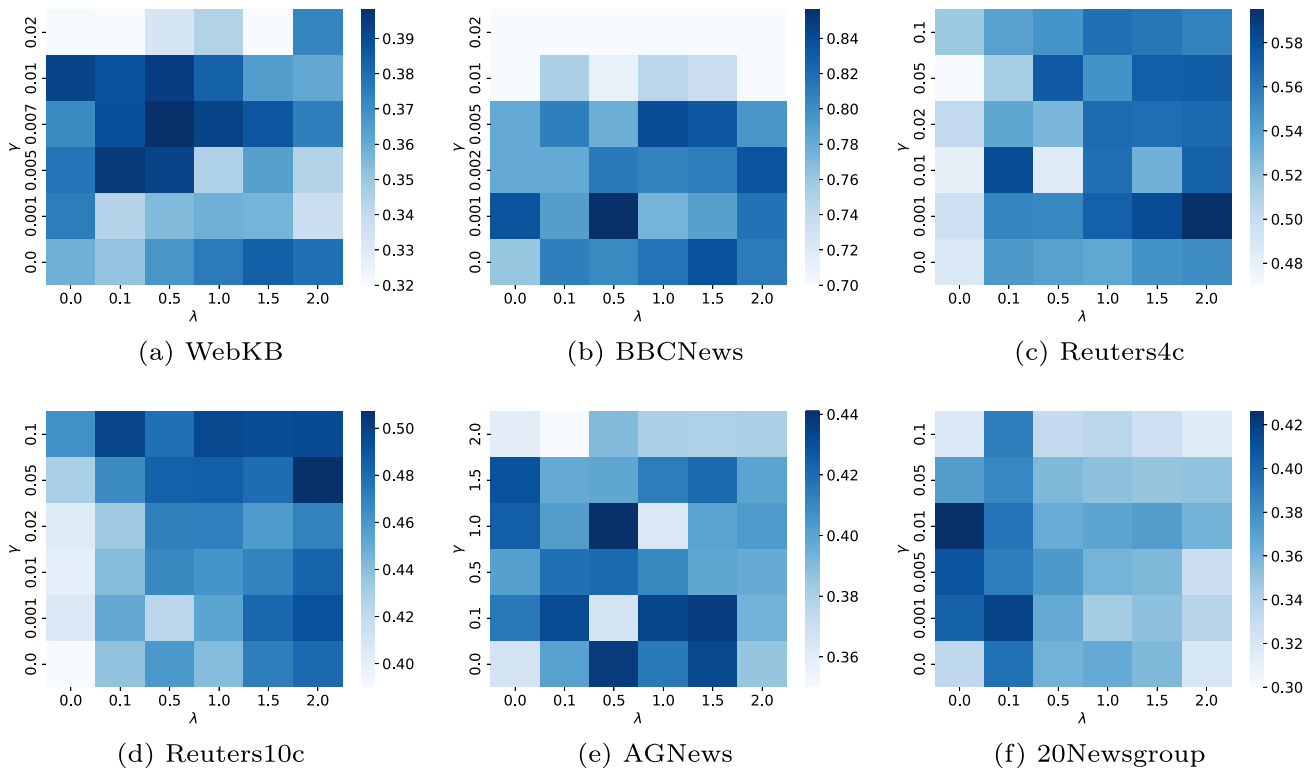


Fig. 4 Parameter analysis (in terms of ARI) on the λ and γ parameters

Fig. 5 Ablation study on the effect of regularization terms of the proposed method based on NMI measure

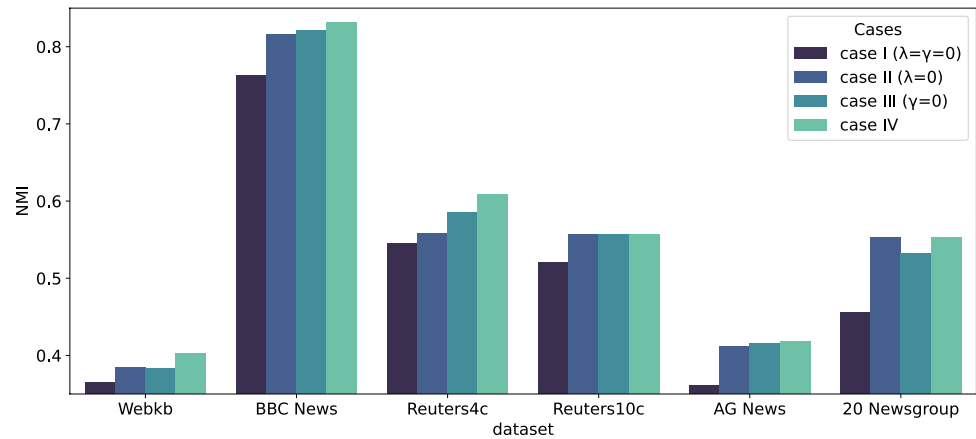


Fig. 6 Ablation study on the effect of regularization terms of the proposed method based on ACC measure

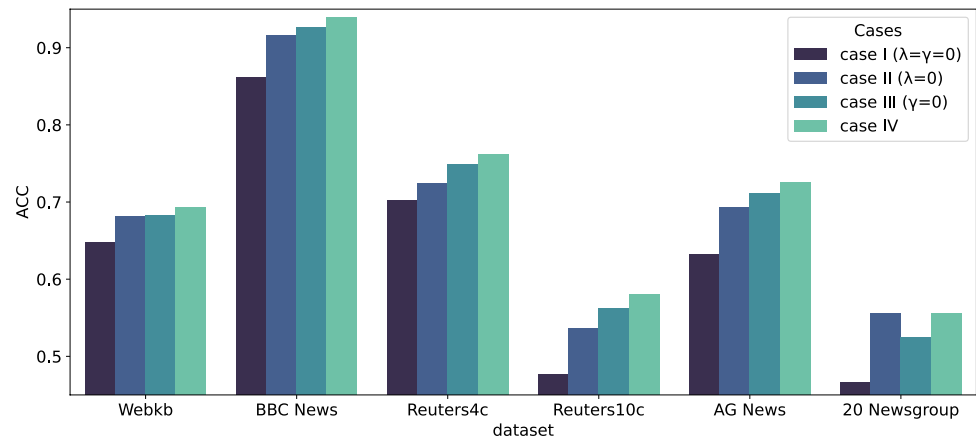
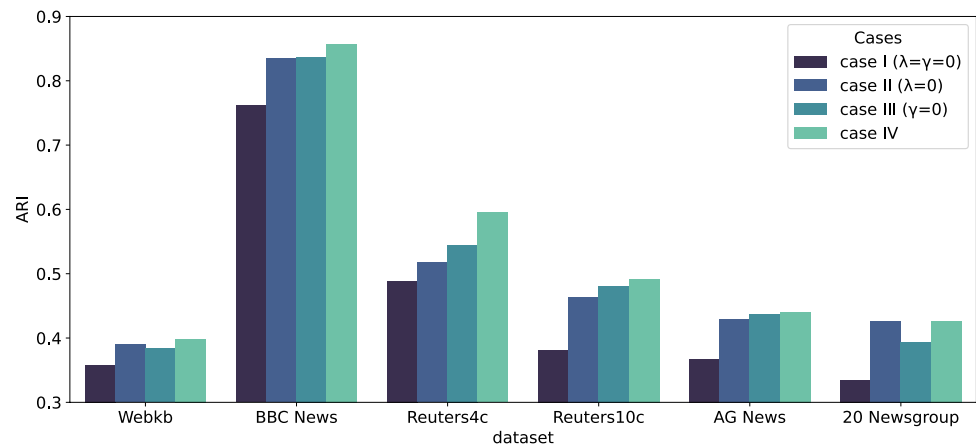


Fig. 7 Ablation study on the effect of regularization terms of the proposed method based on ARI measure



experiment, we chose small values for γ , ranging between 0 and 0.1, since over-emphasizing semantic regularization leads to poor performance when γ is large. Also, we experimented the proposed model with different values of λ from $\{0, 0.1, 0.5, 1, 1.5, 2\}$. From Figures 2–4, we can derive that for Reuters10c, and 20Newsgroup larger γ usually generate better performance, which indicates that the semantic information is more important compared to other datasets.

On the other hand, for BBCNews, Reuters4c, and AGNews, larger λ usually leads to higher performance which implies self-expression is more important. According to the Figures, an acceptable performance can be achieved when the values of the parameters λ and γ are close to 1 and between 0.001 and 0.01, respectively.

4.6 Ablation Study

The proposed model is a basic Decoder NMFk that includes two complementary modules, the Encoder term and the semantic term for realizing self-representation and incorporating semantic information, respectively. In this study, we evaluate how these two modules enhance the clustering performance of the basic NMFk. The NMIs, ACCs, and ARIs of the proposed model in different settings are shown in Figures 5–7. Specifically, case I ($\gamma = \lambda = 0$) means the proposed method equivalent to NMFk, case II ($\lambda = 0$) represents the scenario where the encoder term is exclude, Case III ($\gamma = 0$) represents EDNMFk, and finally, Case IV denotes the SEDNMFk model with both encoder and semantic terms. From Figures 2–4, it is possible to derive the following conclusions:

- On all six datasets, Cases II and III outperform Case I in terms of NMI, ACC, and ARI proving the efficiency of both encoder and semantic term.
- On most datasets, Case III produces better NMI, ACC, and ARI scores than Case II; however, on 20newsgroup, Case II outperforms Case III across these metrics. This observation implies that, in the proposed model, the encoder component often holds more significance than the semantic component.
- To sum up, both terms play a crucial role in significantly enhancing performance. Their individual contributions are not only important on their own but also work in tandem, making them highly effective when combined. Moreover, they are consistent and complementary to each other, reinforcing their collective impact and

ensuring that the overall performance remains robust and reliable across various scenarios.

4.7 Out-of-sample analysis

In this section, we repeat experiment 4.4 to analyze the generalization of the proposed model in the out-of-samples scenario. In an unsupervised manner, we split each dataset into two parts: 80% as the in-sample (train) dataset and 20% as the out-of-sample (test) dataset. The model is trained on the in-sample datasets to find the best projection matrix W . Then, to extract the cluster membership H' for the out-of-sample set X' , projection matrix W maps it to the cluster space by $H' = W^\dagger X'$ where W^\dagger is the pseudo-inverse of matrix W . However, for the encoder-decoder models, we can also use the simpler and faster formula $H' = W^T X'$. We run each method 10 times and report the mean and standard deviation of three performance metrics in Tables 6–8. Based on the results presented in these tables, our proposed SEDNMFk method maintains its top performance across all datasets for NMI, ACC, and ARI metrics. BO- β NMF again emerges as the second best method after our SEDNMFk in some cases. While BO- β NMF demonstrates good generalization ability due to its robust beta-divergence formulation, our proposed SEDNMFk exhibits even better generalization by effectively combining self-expression and semantic terms along with the KL-divergence. The superior out-of-sample performance highlights the enhanced generalization capabilities of our SEDNMFk framework compared to existing NMF variants. In summary, our proposed SEDNMFk method, by unifying self-expressive representations, semantic information, and the KL-divergence base, achieves

Table 6 NMI results for test samples of seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.3014	0.7516	0.3693	0.4960	0.2560	0.4815	0.0430
	std	0.0110	0.0620	0.0390	0.0150	0.0548	0.0060	0.0051
NMFk	mean	0.3347	0.7848	<u>0.5122</u>	0.5263	0.3539	0.5026	0.0704
	std	0.0280	0.0874	0.0201	0.0220	0.0460	0.0150	0.0055
EDNMF	mean	0.3107	0.7195	0.3887	0.5120	0.2645	0.4471	0.0681
	std	0.0150	0.0270	0.0180	0.0270	0.0280	0.0220	0.0077
WRNMTF	mean	0.3445	0.6877	0.3695	0.4592	0.2676	0.4821	0.0512
	std	0.0123	0.0508	0.0170	0.0400	0.0388	0.0170	0.0071
SeaNMF	mean	0.3467	0.7681	0.3963	0.5099	0.3370	0.4857	0.0498
	std	0.0090	0.0421	0.0020	0.0140	0.0580	0.0120	0.0077
RANMF	mean	<u>0.3515</u>	<u>0.7880</u>	0.5022	0.5305	0.3593	0.5035	<u>0.0821</u>
	std	0.0250	0.0980	0.0457	0.0450	0.0490	0.0250	0.0128
DGLCF	mean	0.3375	0.7547	0.3993	0.5227	0.3259	0.5198	0.0731
	std	0.0210	0.0750	0.0590	0.0350	0.0374	0.0130	0.0080
BO- β NMF	mean	0.3174	0.7707	0.4447	<u>0.5323</u>	<u>0.3923</u>	<u>0.5249</u>	0.0805
	std	0.0048	0.0879	0.0292	0.0350	0.0196	0.0024	0.0090
SEDNMFk	mean	0.3751	0.7904	0.5778	0.5413	0.4025	0.5362	0.0928
	std	0.0154	0.0353	0.0392	0.0310	0.0660	0.0230	0.0107

Table 7 ACC results for test samples of seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.5379	0.8751	0.5367	0.4544	0.4709	0.4431	0.1799
	std	0.0270	0.0910	0.0101	0.0150	0.0550	0.0180	0.0097
NMFk	mean	0.6290	<u>0.8869</u>	0.6967	0.5050	0.6531	0.4976	0.2160
	std	0.0270	0.0950	0.0370	0.0201	0.0610	0.0310	0.0074
EDNMF	mean	0.5297	0.8551	0.6075	0.5058	0.4862	0.4215	0.1897
	std	0.0320	0.0130	0.0091	0.0301	0.0070	0.0301	0.0130
WRNMTF	mean	0.6570	0.8187	0.5649	0.4663	0.4616	0.4591	0.1862
	std	0.0123	0.0721	0.0201	0.0418	0.0210	0.0210	0.0051
SeaNMF	mean	<u>0.6722</u>	0.8802	0.5451	0.4782	0.5811	0.4502	0.1867
	std	0.0080	0.0690	0.0030	0.0170	0.0630	0.0160	0.0082
RANMF	mean	0.6489	0.8734	<u>0.7033</u>	<u>0.5179</u>	0.6568	<u>0.4925</u>	<u>0.2274</u>
	std	0.0250	0.1001	0.0510	0.0450	0.0580	0.0290	0.0203
DGLCF	mean	0.6183	0.8665	0.5816	0.4959	0.5417	0.4718	0.1945
	std	0.0250	0.0901	0.0590	0.0401	0.0420	0.0130	0.0104
BO- β NMF	mean	0.5821	0.8775	0.6129	0.5176	<u>0.6638</u>	0.4894	0.2155
	std	0.0317	0.1082	0.0794	0.0506	0.0512	0.0130	0.0124
SEDNMFk	mean	0.6736	0.8943	0.7548	0.6043	0.6716	0.4983	0.2507
	std	0.0270	0.0710	0.0330	0.0320	0.0710	0.0160	0.0117

Table 8 ARI results for test samples of seven datasets. The top-performing result is showcased in **bold** format, while the second-best is indicated with an underline

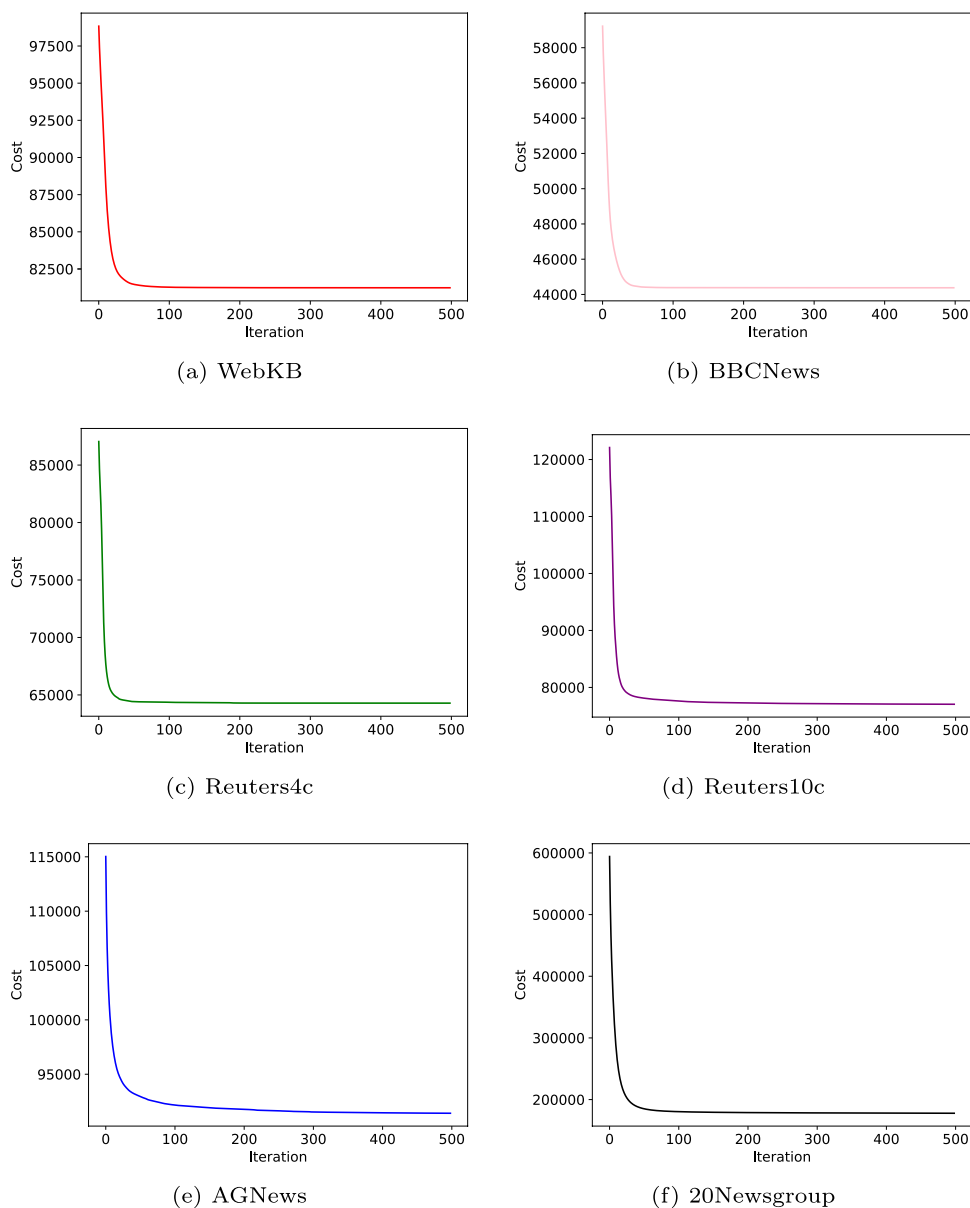
Method		WebKB	BBCNews	Reuters4c	Reuters10c	AGNews	20Newsgroup	YahooAnswers
NMF	mean	0.2372	0.7638	0.1722	0.3094	0.2300	0.2860	0.0191
	std	0.0301	0.1050	0.0090	0.0110	0.0601	0.0090	0.0035
NMFk	mean	0.3282	<u>0.7901</u>	<u>0.4539</u>	0.3976	0.3585	0.3606	0.0357
	std	0.0310	0.0950	0.1102	0.0201	0.0620	0.0230	0.0040
EDNMF	mean	0.2309	0.7349	0.2386	0.3454	0.2413	0.3125	0.0280
	std	0.0401	0.0140	0.0201	0.0540	0.0202	0.0520	0.0075
WRNMTF	mean	0.3245	0.6894	0.1981	0.3105	0.3117	0.2871	0.0243
	std	0.0248	0.0830	0.0101	0.0500	0.0350	0.0210	0.0065
SeaNMF	mean	0.3408	0.7877	0.1967	0.3185	0.3447	0.3009	0.0252
	std	0.0120	0.0780	0.0040	0.0160	0.0701	0.0150	0.0065
RANMF	mean	<u>0.3485</u>	0.7812	0.4406	<u>0.4102</u>	0.3669	0.3625	<u>0.0442</u>
	std	0.0501	0.1200	0.0510	0.0260	0.0580	0.0160	0.0080
DGLCF	mean	0.3057	0.7673	0.1964	0.3268	0.3377	0.3088	0.0344
	std	0.0310	0.0920	0.0590	0.0370	0.0501	0.0110	0.0082
BO- β NMF	mean	0.2654	0.7801	0.3831	0.3947	<u>0.4019</u>	<u>0.3720</u>	0.0427
	std	0.0284	0.1277	0.0374	0.0477	0.0264	0.0092	0.0065
SEDNMFk	mean	0.3638	0.7989	0.5454	0.5236	0.4121	0.3853	0.0601
	std	0.0293	0.0705	0.0370	0.0369	0.0450	0.0120	0.0076

state-of-the-art clustering performance on both in-sample and out-of-sample data across all the benchmark datasets.

4.8 Convergence analysis

In this section, we conduct an empirical analysis of the convergence behavior of the SEDNMFk algorithm. Figure 8 demonstrates how this method consistently minimizes the objective function in accordance with the update rules

explained in Section 3.4. The diagrams within this figure depict the number of iterations on the X-axis and the corresponding objective function values on the Y-axis. Consequently, based on the observed convergence patterns across various datasets, it can be confidently concluded that this method exhibits convergence. The findings from Figure 8 indicate that the model converges rapidly, reaching convergence within a reasonable number of iterations.

Fig. 8 Convergence analysis of the proposed model on six datasets

5 Conclusion

This paper proposed the Semantic Encoder-Decoder Non-negative Matrix Factorization with KL-divergence (SED-NMFk) model. Building upon the foundations of the encoder-decoder NMF model, this novel method has successfully addressed the challenge of accuracy and interpretability in text clustering task. The SEDNMFk model not only achieves high text clustering accuracy but also places emphasis on generalization, in integrating encoder and decoder NMFk modules and formulating a unified cost function based on KL-divergence. This self-representation model refines and verifies clusters while learning interpretable topic modeling information. As a result, the term-cluster matrix extracted by this model demonstrates robust

generalization properties, making it highly effective for handling out-of-sample documents. Incorporating word-word co-occurrence information as a tailored regularization term further enhances the model's ability to capture semantic information, making it more suitable for real-world text clustering scenarios. The optimization scheme proposed in this paper, based on multiplicative updating rules, ensures the efficient and accurate convergence of the cost function. To validate the efficiency and effectiveness of the SEDNMFk model, extensive experiments were conducted on various datasets, including fully observed and out-of-sample settings. The results demonstrated the model's superiority in terms of clustering performance and its capability to handle unseen documents effectively. Furthermore, with a low computational complexity and rapid convergence,

SEDNMFk offers scalable performance suitable for large-scale text clustering applications.

Several directions for future research and development can further enhance its capabilities and applicability. While the KL-divergence is a powerful tool for measuring the similarity between probability distributions, it is just one of several divergence measures. The model could be extended to beta-divergence to cover some other specific distributions or problems. In addition to word-word co-occurrence information, other side information, such as partial label information, could be incorporated into the SEDNMFk model to develop it into a semi-supervised model. The SEDNMFk model was designed for text clustering, but it could be extended to handle other tasks besides clustering, such as recommender systems and link prediction.

Acknowledgements Amjad Seyedi acknowledges the support by the European Union (ERC consolidator, eLinoR, no 101085607).

Author contributions Sayvan Soleymanbaigi: Writing – original draft, Visualization, Implementation, Methodology. Amjad Seyedi: Writing – review & editing, Methodology, Investigation, Conceptualization. Fatemeh Daneshfar: Writing – review & editing, Validation, Supervision. Fardin Akhlaghian Tab: Writing – review & editing, Conceptualization, Supervision, Formal analysis.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interests The authors declare no competing interests.

References

1. Sendhilkumar S et al (2023) Developing a conceptual framework for short text categorization using hybrid cnn-lstm based caledonian crow optimization. *Expert Syst Appl* 212:118517
2. Yamini P, Daneshfar F, Ghorbani A (2024) KurdSM: Transformer-based model for kurdish abstractive text summarization with an annotated corpus. *Iran J Electrical Electron Engineer* 20(4):8–12
3. Sun Y, Qin Y, Li Y, Peng D, Peng X, Hu P (2024) Robust multi-view clustering with noisy correspondence. *IEEE Trans Knowl Data Eng* 36(12):9150–9162
4. Daneshfar F, Soleymanbaigi S, Nafisi A, Yamini P (2024) Elastic deep autoencoder for text embedding clustering by an improved graph regularization. *Expert Syst Appl* 238:121780
5. Seyedi SA, Ghodsi SS, Akhlaghian F, Jalili M, Moradi P (2019) Self-paced multi-label learning with diversity. In: *Proceedings of The Eleventh Asian Conference on Machine Learning* 101:790–805
6. Qin Y, Pu N, Wu H (2023) Elastic multi-view subspace clustering with pairwise and high-order correlations. *IEEE Trans Knowl Data Eng* 36(2):556–568
7. Qin Y, Qian L (2024) Fast elastic-net multi-view clustering: A geometric interpretation perspective. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10164–10172
8. Qin Y, Zhang X, Shen L, Feng G (2022) Maximum block energy guided robust subspace clustering. *IEEE Trans Pattern Anal Mach Intell* 45(2):2652–2659
9. Yao J, Qian Q, Hu J. (2024) Multi-modal proxy learning towards personalized visual multiple clustering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14066–14075
10. Daneshfar F, Saifee BS, Soleymanbaigi S, Amini M (2025) Elastic deep multi-view autoencoder with diversity embedding. *Inf Sci* 689:121482
11. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791
12. Daneshfar F, Soleymanbaigi S, Yamini P, Amini MS (2024) A survey on semi-supervised graph clustering. *Eng Appl Artif Intell* 133:108215
13. Seyedi SA, Moradi P, Tab FA (2017) A weakly-supervised factorization method with dynamic graph embedding. In: *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pp. 213–218
14. Ghodsi S, Seyedi SA, Ntoutsis E (2024) Towards cohesion-fairness harmony: Contrastive regularization in individual fair graph clustering. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 284–296. Springer
15. Mohammadi A, Seyedi SA, Akhlaghian Tab F, Pir Mohammadian R (2024) Diverse joint nonnegative matrix tri-factorization for attributed graph clustering. *Appl Soft Comput* 164:112012
16. Abdollahi R, Amjad Seyedi S, Reza Noorimehr M (2020) Asymmetric semi-nonnegative matrix factorization for directed graph clustering. In: *2020 10th International Conference on Computer and Knowledge Engineering (ICCCKE)*, pp. 323–328
17. Ye F, Chen C, Zheng Z (2018) Deep autoencoder-like nonnegative matrix factorization for community detection. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1393–1402
18. Muthusami R, Mani Kandan N, Saritha K, Narenthiran B, Nagaprasad N, Ramaswamy K (2024) Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Sci Rep* 14(1):12003
19. Soleymanbaigi S, Seyedi A, Akhlaghian Tab F, Daneshfar F (2025) Data clustering by encoder-decoder nonnegative matrix factorization with β -divergence. *Pattern Recognition* 166
20. Gallego V, Lingan J, Freixes A, Juan AA, Osorio C (2024) Applying machine learning in marketing: An analysis using the NMF and k-means algorithms. *Information* 15(7):368
21. Saberi-Movahed F, Biswas B, Tiwari P, Lehmann J, Vahdati S (2024) Deep nonnegative matrix factorization with joint global and local structure preservation. *Expert Syst Appl* 249:123645
22. Barkhoda W, Seyedi A, Gillis N, Akhlaghian Tab, F (2025) Instance-wise distributionally robust nonnegative matrix factorization. *Pattern Recognition* 166
23. Seyedi SA, Tab FA, Lotfi A, Salahian N, Chavoshinejad J (2023) Elastic adversarial deep nonnegative matrix factorization for matrix completion. *Inf Sci* 621:562–579
24. Yahaya F, Puigt M, Delmaire G, Roussel, G (2024) A framework for compressed weighted nonnegative matrix factorization. *IEEE Transactions on Signal Processing*
25. Shajarian Z, Seyedi SA, Moradi P (2017) A clustering-based matrix factorization method to improve the accuracy of recommendation systems. In: *2017 Iranian Conference on Electrical Engineering (ICEE)*, pp. 2241–2246
26. Mahmoodi R, Seyedi SA, Abdollahpour A, Akhlaghian Tab F (2024) Enhancing link prediction through adversarial training in deep nonnegative matrix factorization. *Eng Appl Artif Intell* 133:108641

27. Mahmoodi R, Seyedi SA, Akhlaghian Tab F, Abdollahpouri A (2023) Link prediction by adversarial nonnegative matrix factorization. *Knowl-Based Syst* 280:110998
28. Zhang D, Wu X-J (2022) Robust and discrete matrix factorization hashing for cross-modal retrieval. *Pattern Recogn* 122:108343
29. Feng X-R, Li H-C, Wang R, Du Q, Jia X, Plaza A (2022) Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15:4414–4436
30. Shahnaz F, Berry MW, Pauca VP, Plemmons RJ (2006) Document clustering using nonnegative matrix factorization. *Inf Process Manag* 42(2):373–386
31. Guo Y-T, Li Q-Q, Liang C-S (2024) The rise of nonnegative matrix factorization: algorithms and applications. *Inf Syst* 123:102379
32. Shi T, Kang K, Choo J, Reddy CK (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: *Proceedings of the 2018 World Wide Web Conference*, pp. 1105–1114
33. Wang J, Zhang X-L (2023) Deep NMF topic modeling. *Neurocomputing* 515:157–173
34. Battaglia E, Peiretti F, Pensa RG (2024) Co-clustering: A survey of the main methods, recent trends, and open problems. *ACM Computing Surveys* 57(2)
35. Chen Y, Lei Z, Rao Y, Xie H, Wang FL, Yin J, Li Q (2022) Parallel non-negative matrix tri-factorization for text data co-clustering. *IEEE Trans Knowl Data Eng* 35(5):5132–5146
36. Lee D, Seung HS (2000) Algorithms for non-negative matrix factorization. In: *Adv Neural Inf Process Syst* 13:535–541
37. Devarajan K (2021) A statistical framework for non-negative matrix factorization based on generalized dual divergence. *Neural Netw* 140:309–324
38. Ding C, Li T, Peng W (2006) Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In: *AAAI*, vol. 42, pp. 137–43
39. Aghdam MH, Zanjani MD (2021) A novel regularized asymmetric non-negative matrix factorization for text clustering. *Inf Process Manag* 58(6):102694
40. Vangara R, Skau E, Chennupati G, Djidjev H, Tierney T, Smith JP, Bhattarai M, Stanev VG, Alexandrov BS (2020) Semantic nonnegative matrix factorization with automatic model determination for topic modeling. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 328–335
41. Yuan R, Leng C, Li B, Basu A (2023) β -divergence NMF with biorthogonal regularization for data representation. *Eng Appl Artif Intell* 121:106014
42. Li X, Shen X, Shu Z, Ye Q, Zhao C (2017) Graph regularized multilayer concept factorization for data representation. *Neurocomputing* 238:139–151
43. Shu Z, Zhao C, Huang P (2015) Local regularization concept factorization and its semi-supervised extension for image representation. *Neurocomputing* 158:1–12
44. Luo G, Zhao Z, Liu S, Wu S, Hu A, Zhang N (2024) Integrating topology and content equally in non-negative matrix factorization for community detection. *Expert Syst Appl* 255:124713
45. Salahian N, Tab FA, Seyedi SA, Chavoshinejad J (2023) Deep autoencoder-like NMF with contrastive regularization and feature relationship preservation. *Expert Syst Appl* 214:119051
46. Sun B-J, Shen H, Gao J, Ouyang W, Cheng X (2017) A non-negative symmetric encoder-decoder approach for community detection. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 597–606
47. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 24(5):513–523
48. Salah A, Ailem M, Nadif M (2018) Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In: *Proc AAAI Conf Artif Intell* 32:3992–3999
49. Bullinaria JA, Levy JP (2007) Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav Res Methods* 39(3):510–526
50. Levy O, Goldberg Y (2014) Neural word embedding as implicit matrix factorization. *Adv Neural Inf Process Syst*, 27
51. Li N, Leng C, Cheng I, Basu A, Jiao L (2024) Dual-graph global and local concept factorization for data clustering. *IEEE Trans Neural Netw Learn* 35(1):803–816
52. Mozafari M, Seyedi SA, Mohammadiani RP, Tab FA (2024) Unsupervised feature selection using orthogonal encoder-decoder factorization. *Inf Sci* 663:120277

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.