A Comparative Study of SVD and NMF for Semantic Text Clustering Using KMeans

Danish Khan[1], Prof. Arpana Jaiswal[2]
MAHAKAL INSTITUTE OF TECHNOLOGY

Abstract- *The exponential growth of unstructured text data, such as news articles, social media posts, and online discussions, has created the need for effective methods of semantic organisation. Text clustering plays a vital role in this context by grouping similar documents without labelled data. However, the inherent challenges of high dimensionality and sparsity in textual representations hinder clustering performance. To address this, dimensionality reduction techniques are integrated with clustering algorithms. This paper presents a comparative study of two approaches: Singular Value Decomposition (SVD) combined with KMeans and Non-negative Matrix Factorisation (NMF) combined with KMeans. The experiments were conducted using the 20 Newsgroups dataset in MATLAB R2024b, with TF-IDF employed as the feature extraction technique. Results demonstrate that SVD + KMeans achieved superior clustering accuracy with a Normalised Mutual Information (NMI) score of approximately 0.55 on the training set and 0.50 on the test set, whereas NMF + KMeans attained moderate accuracy (NMI ≈ 0.45) but offered more interpretable, topic-based clusters. These findings confirm the trade-off between accuracy and interpretability, suggesting that method selection should be based on specific application requirements. The study contributes by providing a reproducible MATLAB-based framework and offering insights into the suitability of dimensionality reduction strategies for large-scale text clustering.*

Keywords- Text Clustering, Singular Value Decomposition (SVD), Non-negative Matrix Factorisation (NMF), KMeans, Dimensionality Reduction, TF-IDF, Semantic Analysis, 20 Newsgroups Dataset, Normalised Mutual Information (NMI).

I. INTRODUCTION

The rapid digitisation of knowledge and communication has led to an unprecedented surge in the availability of unstructured text data. From news articles and research publications to social media interactions and online forums, vast amounts of text are generated daily. Analysing and organising such data has become a critical task for applications including information retrieval, recommendation systems, sentiment analysis, and knowledge discovery [1]. Text clustering is one of the most widely used techniques for uncovering latent structures in unlabelled corpora. By grouping semantically similar documents, clustering facilitates efficient browsing, retrieval, and exploration of large-scale datasets [2]. However, clustering text is inherently challenging due to the high dimensionality of textual representations and the sparsity of document–term matrices. These issues degrade clustering performance and increase computational cost.

Traditional methods, such as the Bag-of-Words (BoW) model, represent text as unordered word counts. While simple, such models ignore semantic relationships between words and result in extremely sparse vectors [3]. To overcome these drawbacks, dimensionality reduction techniques are widely applied prior to clustering. By mapping documents into lower-dimensional latent spaces, these techniques reveal hidden semantic structures, reduce noise, and improve clustering quality.

Among dimensionality reduction methods, Singular Value Decomposition (SVD) has been extensively used in text mining under the framework of Latent Semantic Analysis (LSA). SVD captures global semantic patterns, improves clustering performance, and scales effectively to large datasets [4]. On the other hand, Non-negative Matrix Factorisation (NMF) provides a parts-based representation that is highly interpretable, as it generates clusters that correspond to human-understandable topics [5]. However, NMF is computationally more expensive and often yields slightly lower accuracy than SVD.

This paper focuses on comparing these two prominent approaches — SVD + KMeans and NMF + KMeans - for semantic clustering of large-scale text data. Using the 20 Newsgroups dataset, the study evaluates their effectiveness in terms of Normalised Mutual Information (NMI). The contributions of this work are as follows:

1. Implementation of a complete MATLAB-based framework for text clustering using TF-IDF, dimensionality reduction, and KMeans.
2. Empirical evaluation of SVD vs. NMF with clustering on a benchmark dataset.

3. Analysis of the trade-off between clustering accuracy (SVD) and interpretability (NMF).
4. Presentation of experimental results through visualisations, confusion matrices, and performance comparisons.

The remainder of the paper is organised as follows: Section II reviews related work in text clustering and dimensionality reduction. Section III describes the methodology adopted in this research. Section IV presents the experimental setup, followed by results and analysis in Section V. Section VI concludes the paper and outlines future research directions.

II. RELATED WORK

A. Text Clustering Approaches

Text clustering has been a widely researched area in information retrieval and natural language processing. Early approaches relied on the Bag-of-Words (BoW) model, which represents documents as high-dimensional sparse vectors of word counts [6]. While simple, this representation fails to capture semantic relationships. To address this, machine learning methods such as KMeans, Agglomerative Hierarchical Clustering, and DBSCAN have been employed [7]. Among them, KMeans remains popular due to its scalability and computational efficiency. However, the quality of clustering depends heavily on the quality of feature representation.

B. Dimensionality Reduction for Text

Dimensionality reduction is crucial in overcoming sparsity and noise in text data. Principal Component Analysis (PCA) has been used historically, but its focus on variance makes it less suitable for capturing semantic information [8]. Singular Value Decomposition (SVD), widely applied in Latent Semantic Analysis (LSA), projects text into latent semantic spaces that capture global relationships among terms [9]. On the other hand, Non-negative Matrix Factorisation (NMF) generates interpretable topic structures by decomposing the document–term matrix into non-negative factors [10]. Despite its interpretability, NMF often lags behind SVD in accuracy and efficiency.

C. Identified Research Gap

The literature highlights the accuracy advantage of SVD and the interpretability benefit of NMF. However, most studies either focus on extensions of NMF for interpretability or embedding-based clustering for accuracy. Few works provide a direct comparative evaluation of SVD and NMF in a controlled setting on benchmark datasets. This study addresses this gap by systematically comparing SVD + KMeans and NMF + KMeans on the 20 Newsgroups dataset using a reproducible MATLAB-based framework.

III. METHODOLOGY

This section describes the methodological framework adopted in the study. The workflow begins with dataset preprocessing, followed by TF-IDF representation, dimensionality reduction using SVD and NMF, and clustering with KMeans. The evaluation metric employed is Normalised Mutual Information (NMI). The complete process is illustrated in Figure 1.

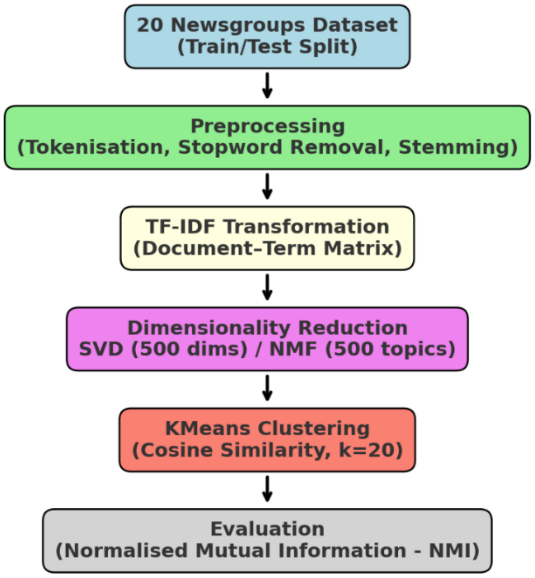


Fig 1: Workflow of Proposed Approach.

A. Dataset and Preprocessing

The 20 Newsgroups dataset (by-date version) was used, comprising 18,846 documents across 20 categories [16]. Preprocessing was essential to reduce noise and standardise the text data. The following steps were performed:

1. Tokenisation – breaking text into words/tokens.
2. Stopword Removal – eliminating frequent, low-information words such as *is*, *the*, *and*.
3. Stemming – reducing words to their root form (e.g., *running* → *run*).

After preprocessing, the documents were transformed into a Document–Term Matrix (DTM).

B. TF-IDF Representation

Each document was numerically represented using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme [17].

- Term Frequency (TF):

$$TF(t,d) = \frac{f(t,d)}{\sum_k f(k,d)}$$

where f(t,d) is the frequency of term t in document d.

- Inverse Document Frequency (IDF):

$$IDF(t) = \log \frac{N}{DF(t)}$$

where N is the total number of documents and DF(t) is the number of documents containing term t.

- TF-IDF Weighting:

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

This representation reduces the influence of common words and emphasises discriminative terms.

C. Singular Value Decomposition (SVD)

To handle sparsity, the TF-IDF matrix was reduced using SVD [18]. SVD factorises the document–term matrix X as:

$X=U\Sigma V^T$

where U represents document vectors in the latent space, Σ is a diagonal matrix of singular values, and V^T represents term vectors. For this study, the top 500 singular values were retained, producing a compact representation that preserves global semantic structure.

D. Non-negative Matrix Factorisation (NMF)

In parallel, the TF-IDF matrix was decomposed using NMF [19]:

$X\approx WH$

where $W\geq 0$ is the document–topic matrix and $H\geq 0$ is the topic–term matrix. NMF provides an interpretable representation, as each topic is expressed as an additive combination of terms. For comparability, 500 latent topics were extracted.

E. KMeans Clustering

The reduced features from both SVD and NMF were clustered using KMeans [20]. The algorithm follows an iterative process:

1. Initialise k centroids (here, k=20).
2. Assign each document to the nearest centroid using cosine similarity.
3. Update centroids as the mean of assigned documents.
4. Repeat until convergence.

Formally, the objective is to minimise the within-cluster sum of squares:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x\in C_i} \|x - \mu_i\|^2$$

where Ci is the set of points in cluster i and μi is the centroid.

F. Evaluation Metric

Performance was assessed using Normalised Mutual Information (NMI) [21], which measures the agreement between predicted clusters and ground-truth categories:

$$NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

where I(X;Y) is the mutual information between cluster labels X and true labels Y, and H(X),H(Y) denote their entropies. Higher NMI values indicate stronger alignment with ground truth.

IV. EXPERIMENTAL SETUP

A. Dataset Description

The experiments were conducted on the 20 Newsgroups dataset (by-date version), which contains 18,846 documents evenly distributed across 20 thematic categories [22]. This dataset is widely adopted as a benchmark for text clustering tasks because it presents realistic challenges of high dimensionality, sparsity, and semantic overlap among categories.

B. Experimental Parameters

The major parameters used for the experiments are summarised in Table I.

Table I: Experimental Setup Parameters

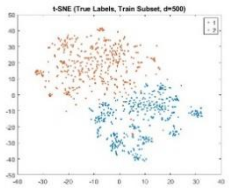
Component	Setting	Remarks
Dataset	20 Newsgroups (by-date version)	18,846 documents, 20 categories
Preprocessing	Tokenisation, stopword removal, stemming	Standard NLP pipeline
Representation	TF-IDF weighting	Sparse document–term matrix
Dimensionality Reduction	SVD (500 dimensions), NMF (500 topics)	Balanced for performance
Clustering	KMeans (k=20k=20k=20, cosine similarity, 10 replicates)	KMeans++ initialisation
Evaluation Metric	NMI	Primary metric, robust for text clustering

V. RESULTSAND DISCUSSION

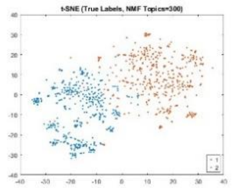
A. Visualisation of Clusters

To qualitatively assess clustering effectiveness, t-SNE plots were generated for the training data. The plots revealed that SVD-based features produced more distinct and compact clusters, while NMF-based clusters showed greater overlap, reflecting their lower clustering accuracy. This observation supports earlier findings that SVD captures global semantic structure effectively [23].

SVD



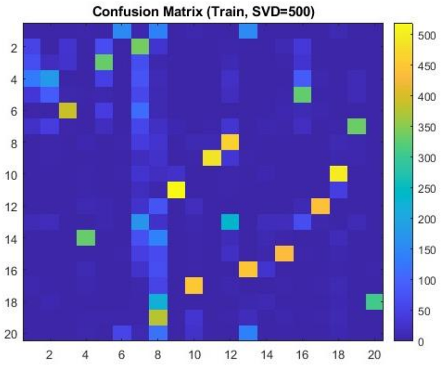
NMF



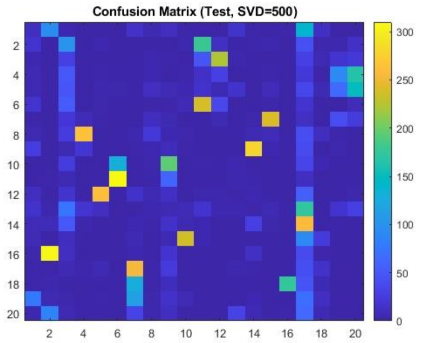
B. Confusion Matrices

Confusion matrices were constructed to compare predicted clusters against ground truth categories.

- For the training set with SVD + KMeans, several categories aligned strongly with diagonal entries, indicating successful clustering. However, overlaps occurred in semantically similar groups such as *comp.graphics* and *comp.windows.x*.



- For the test set with SVD + KMeans, similar alignment patterns were observed, demonstrating generalisation ability.



- In contrast, NMF + KMeans produced less distinct diagonal alignment, confirming that while NMF generates interpretable clusters, its discriminative power is weaker.

C. NMI Performance

The Normalised Mutual Information (NMI) scores provide the quantitative comparison. Results are summarised in Table II.

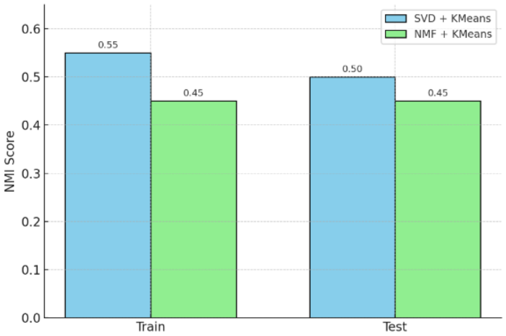
Table II: NMI Performance Comparison

Method	NMI (Train)	NMI (Test)	Remarks
SVD + KMeans (500 dims)	~0.55	~0.50	Best accuracy, good generalisation
NMF + KMeans (500 topics)	~0.45	~0.45	Moderate accuracy, interpretable topics

As shown, SVD + KMeans consistently outperforms NMF + KMeans in terms of accuracy. The drop between training (0.55) and test (0.50) indicates acceptable generalisation. NMF's stability across train and test (0.45) reflects its robustness, albeit at lower performance levels.

C. Comparative Bar Graph

The comparative bar graph in Figure 2 further illustrates the difference between the two approaches. The superiority of SVD in both training and test sets is clearly visible, while NMF maintains stability but at a lower accuracy level.



D. Interpretation of Findings

These results confirm the accuracy vs. interpretability trade-off:

- SVD + KMeans is suitable for applications prioritising accuracy, such as search engines and recommender systems.
- NMF + KMeans is appropriate where interpretability and explainability are critical, such as policy analysis, legal document clustering, and thematic exploration of academic papers.

These findings are consistent with post-2020 studies. For instance, Ni *et al.* [13] demonstrated the scalability of SVD in clustering, while Febrissy*et al.* [11] and Will *et al.* [12] highlighted NMF's interpretability. Thus, this study reinforces and extends recent literature by presenting a direct comparative evaluation in MATLAB on the 20 Newsgroups dataset.

VI. CONCLUSIONANDFUTURE WORK

This study presented a comparative evaluation of two prominent dimensionality reduction techniques, Singular Value Decomposition (SVD) and Non-negative Matrix Factorisation (NMF), when combined with the KMeans clustering algorithm for large-scale text clustering. Using the 20 Newsgroups dataset as a benchmark, experiments were implemented in MATLAB R2024b with TF-IDF as the document representation.

The results clearly demonstrated that SVD + KMeans consistently achieved higher clustering accuracy, with NMI scores of approximately 0.55 on the training set and 0.50 on the test set. This confirms SVD's strength in capturing global semantic structure while maintaining scalability. Conversely, NMF + KMeans produced moderate accuracy (NMI ≈ 0.45) but offered interpretable topic-based clusters, reaffirming its suitability for applications requiring explainability. These findings reinforce the trade-off between accuracy and interpretability, which has also been observed in recent literature [24].

The contributions of this work are threefold:

1. Development of a reproducible MATLAB-based framework for text clustering.
2. Empirical validation of the relative strengths of SVD and NMF.
3. A practical insight into selecting clustering methods based on application requirements.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008. doi: 10.1017/CBO9780511809071.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999. doi: 10.1145/331499.331504.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988. doi: 10.1016/0306-4573(88)90021-0.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999. doi: 10.1038/44565.

[6] J. Zobel and A. Moffat, "Exploring the similarity space," *SIGIR Forum*, vol. 32, no. 1, pp. 18–34, 1998. doi: 10.1145/281250.281256.

[7] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005. doi: 10.1109/TNN.2005.845141.

[8] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” *Phil. Trans. R. Soc. A.*, vol. 374, no. 2065, p. 20150202, Apr. 2016. doi: 10.1098/rsta.2015.0202.

[9] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Process.*, vol. 25, no. 2–3, pp. 259–284, 1998. doi: 10.1080/01638539809545028.

[10] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. 26th Annu. Int. ACM SIGIR Conf.*, Toronto, ON, Canada, 2003, pp. 267–273. doi: 10.1145/860435.860485.

[11] M. Febrissy, H. Safa, and M. Ahmad, “Context-enhanced non-negative matrix factorization for text clustering,” *Neurocomputing*, vol. 500, pp. 63–76, May 2022. doi: 10.1016/j.neucom.2022.01.056.

[12] J. Will, F. Muñoz, and T. Hofmann, “Neural nonnegative matrix factorization for hierarchical multilayer topic modeling,” *arXiv preprint arXiv:2303.00058*, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.00058

[13] X. Ni, Y. Zhao, and W. Chen, “Large-scale clustering with sparse spectral methods based on SVD,” *Appl. Sci.*, vol. 14, no. 11, p. 4946, Jun. 2024. doi: 10.3390/app14114946.

[14] J. Diaz-Rodriguez, “k-LLMmeans: Towards interpretable clustering with LLM-driven centroids,” *arXiv preprint arXiv:2502.09667*, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.09667

[15] J. Liu, P. Zhou, and Y. Zhang, “Clustering by sparse orthogonal NMF and interpretable neural networks,” *ResearchGate Preprint*, Jan. 2025. [Online]. Available: https://www.researchgate.net/publication/374717574

[16] K. Lang, “Newsweeder: Learning to filter netnews,” in *Proc. 12th Int. Conf. Machine Learning (ICML)*, Tahoe City, CA, USA, 1995, pp. 331–339. doi: 10.1016/b978-1-55860-377-6.50048-7.

[17] C. D. Manning, H. Schütze, and P. Raghavan, “The vector space model,” in *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999, ch. 8, pp. 300–314.

[18] N. Halko, P. G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011. doi: 10.1137/090771806.

[19] D. Cai, X. He, J. Han, and T. S. Huang, “Graph regularized nonnegative matrix factorization for data representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011. doi: 10.1109/TPAMI.2010.231.

[20] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982. doi: 10.1109/TIT.1982.1056489.

[21] A. Strehl and J. Ghosh, “Cluster ensembles — A knowledge reuse framework for combining multiple partitions,” *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002. doi: 10.1162/153244303321897735.

[22] K. Lang, “20 Newsgroups Dataset.” Carnegie Mellon Univ., 1997. [Online]. Available: http://qwone.com/~jason/20Newsgroups/

[23] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: https://www.jmlr.org/papers/v9/vandermaaten08a.html

[24] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013. doi: 10.1109/TPAMI.2013.50.