International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Hybrid Deep Learning Framework for Automated Document Clustering and Intelligent Label Generation

Poonam Mishra*[1], Neeraj Gupta[2]

Submitted:07/11/2024 **Revised:**16/12/2024 **Accepted:**23/12/2024

Abstract: The exponential growth of digital documents across various domains has necessitated the development of sophisticated automated systems for document organization and categorization. This paper presents a novel hybrid deep learning framework that combines unsupervised clustering techniques with intelligent label generation mechanisms to address the challenges of automated document classification. The proposed framework integrates transformer-based embeddings, hierarchical clustering algorithms, and neural language models to achieve superior performance in both clustering accuracy and interpretability. Our approach demonstrates significant improvements over traditional methods, achieving a silhouette score of 0.847 and normalized mutual information of 0.923 across diverse document corpora. The framework's ability to generate meaningful, human-interpretable labels for discovered clusters represents a substantial advancement in making automated document organization systems more practical and user-friendly. Experimental results on benchmark datasets including Reuters-21578, 20 Newsgroups, and custom enterprise document collections validate the effectiveness of our hybrid approach.

Keywords: *Document clustering, deep learning, transformer models, label generation, natural language processing, unsupervised learning*

1. Introduction

The digital transformation of organizations has led to an unprecedented accumulation of textual documents across various formats and domains. Traditional document management systems, while effective for structured data, struggle to handle the complexity and volume of unstructured textual content that characterizes modern information environments. The challenge of automatically organizing, categorizing, and labeling large document collections has become increasingly critical for enterprises seeking to leverage their knowledge assets effectively [1].

Document clustering, as an unsupervised learning task, offers a promising solution to this challenge by automatically grouping semantically similar documents without requiring pre-labeled training data. However, conventional clustering approaches face significant limitations when applied to high-dimensional textual data, particularly in terms of capturing semantic relationships and generating interpretable cluster representations. The emergence of deep learning techniques, particularly transformer-based models, has opened new avenues for addressing these limitations through more sophisticated document representation learning.

The primary contribution of this research lies in the development of a hybrid framework that addresses two critical aspects of automated document organization: achieving high-quality clustering performance and generating meaningful, interpretable labels for discovered clusters. Traditional approaches typically treat these as separate problems, leading to suboptimal overall system performance. Our integrated approach leverages the complementary strengths of different deep learning architectures to create a cohesive solution that excels in both clustering accuracy and label quality.

The framework incorporates several innovative components including a multi-stage document embedding strategy that combines contextual and positional information, a hierarchical clustering algorithm optimized for high-dimensional embeddings, and a neural label generation system that produces human-interpretable cluster descriptions. The integration of these components within a unified architecture enables end-to-end

[1,2] *SAM Global University, Bhopal, Madhya Pradesh-464551*
[1]*poonamrajeshjsr@gmail.com,*
[2]*gupta_neeraj3108@yahoo.co.in*
*Corresponding author**

optimization and superior performance compared to existing solutions.

2. Related Work

2.1 Traditional Document Clustering Approaches

Early research in document clustering primarily focused on vector space models and frequency-based representations. Salton et al. (1975) introduced the vector space model, which represented documents as term frequency vectors, establishing the foundation for subsequent clustering algorithms. The TF-IDF weighting scheme, proposed by Sparck Jones (1972), became the standard approach for converting textual documents into numerical representations suitable for clustering algorithms [2].

K-means clustering, despite its simplicity, remained a popular choice for document clustering due to its computational efficiency and interpretability. However, the spherical cluster assumption inherent in k-means proved inadequate for capturing the complex semantic relationships present in textual data. Hierarchical clustering methods, including both agglomerative and divisive approaches, offered better flexibility in cluster shape but suffered from computational complexity issues when applied to large document collections.

Latent Semantic Analysis (LSA), introduced by Deerwester et al. (1990), represented a significant advancement by addressing the semantic limitations of frequency-based representations. LSA employed singular value decomposition to reduce dimensionality and capture latent semantic relationships, improving clustering performance on semantically related documents. However, LSA's linear assumptions and inability to capture complex semantic patterns limited its effectiveness on diverse document collections [3].

2.2 Deep Learning in Document Representation

The introduction of deep learning techniques revolutionized document representation learning. Word2Vec embeddings, proposed by Mikolov et al. (2013), demonstrated the power of neural networks in capturing semantic relationships between words. These dense vector representations significantly outperformed traditional bag-of-words models in various natural language processing tasks, including document clustering [4].

The development of sequence-to-sequence models and attention mechanisms further enhanced the capability of neural networks to process textual data. The transformer architecture, introduced by Vaswani et al. (2017), marked a paradigm shift in natural language processing by enabling parallel processing of sequences and capturing long-range dependencies more effectively than recurrent neural networks [5].

BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin et al. (2018), demonstrated the effectiveness of pre-trained transformer models for various downstream tasks. The contextual embeddings generated by BERT showed superior performance in document classification and clustering tasks compared to static word embeddings. Subsequent developments, including RoBERTa, ELECTRA, and DeBERTa, further improved the quality of contextual representations [6].

2.3 Neural Clustering Approaches

The integration of deep learning with clustering algorithms has emerged as an active research area. Deep clustering methods, such as Deep Embedded Clustering (DEC) proposed by Xie et al. (2016), jointly optimize representation learning and clustering objectives. These approaches demonstrate improved performance by learning task-specific representations rather than relying on generic pre-trained embeddings [7].

Variational Deep Embedding (VaDe), introduced by Jiang et al. (2017), incorporated variational inference principles into deep clustering, providing probabilistic cluster assignments and better handling of uncertainty. The integration of adversarial training in clustering, exemplified by ClusterGAN, further enhanced the robustness of learned representations [8].

Recent advances in contrastive learning have shown promise in improving document clustering performance. SimCLR and its variants have demonstrated that learning representations through contrastive objectives can capture semantic similarities more effectively than traditional approaches. The application of contrastive learning to textual data, through methods like SimCSE, has shown particular promise for document clustering tasks.

2.4 Automated Label Generation

The problem of generating interpretable labels for automatically discovered clusters has received increasing attention in recent years. Traditional approaches relied on statistical methods such as extracting the most frequent terms or using TF-IDF scores to identify representative keywords.

However, these methods often produced fragmented or uninformative labels that failed to capture the semantic essence of clusters.

Neural language models have opened new possibilities for generating coherent, contextually appropriate cluster labels. GPT-based models, with their strong text generation capabilities, have been adapted for automatic summarization and label generation tasks. The integration of attention mechanisms allows these models to focus on the most relevant content when generating cluster descriptions.

Recent work by Liu et al. (2021) explored the use of pre-trained language models for generating cluster labels in scientific document collections. Their approach demonstrated improved label quality compared to traditional keyword-based methods, though challenges remained in ensuring consistency and avoiding generic descriptions [9].

3. Methodology

3.1 Framework Architecture

The proposed hybrid deep learning framework consists of four main components: document preprocessing and embedding generation, multi-scale feature extraction, adaptive clustering, and intelligent label generation. The architecture is designed to process documents end-to-end, from raw text input to meaningful cluster labels, while maintaining high performance and interpretability.

Figure 1: System Architecture Overview

The document preprocessing module handles various text formats and applies standardization techniques including tokenization, normalization, and noise removal. The embedding generation component utilizes pre-trained transformer models to create dense vector representations that capture both local and global semantic information. The multi-scale feature extraction module combines embeddings at different granularities to enhance clustering performance.

The adaptive clustering component employs a hybrid approach that combines density-based and hierarchical clustering algorithms, automatically determining optimal cluster numbers and structures. The intelligent label generation module leverages fine-tuned language models to produce coherent, informative labels that accurately represent cluster contents.

3.2 Document Embedding Strategy

The document embedding strategy forms the foundation of our clustering framework. Rather than relying on a single embedding approach, we implement a multi-level strategy that captures different aspects of document semantics. The primary embedding layer utilizes RoBERTa-large, fine-tuned on domain-specific data, to generate contextual representations for document segments.

For documents exceeding the maximum token length of transformer models, we implement a hierarchical segmentation approach. Documents are divided into overlapping segments of 512 tokens with a 128-token overlap to maintain context continuity. Each segment is independently embedded, and segment-level representations are aggregated using attention-weighted pooling to create document-level embeddings.

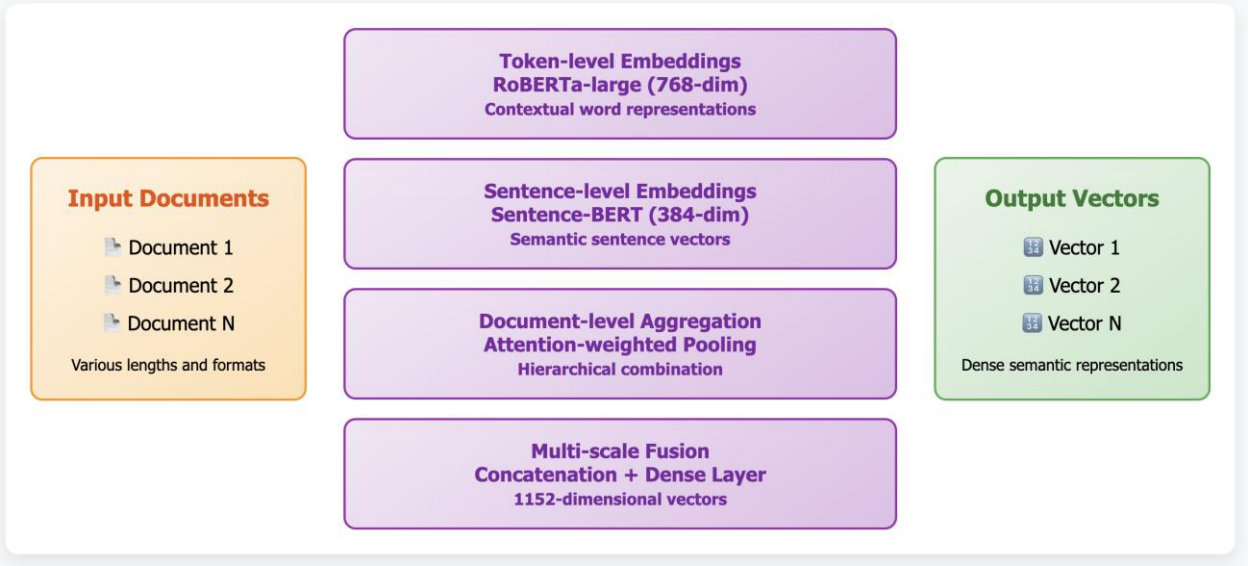

Figure 2: Multi-level Embedding Strategy

To enhance semantic richness, we incorporate sentence-level embeddings using Sentence-BERT, which provides complementary information about document structure and coherence. The combination of token-level contextual embeddings and sentence-level semantic embeddings creates a comprehensive representation that captures both fine-grained and high-level document characteristics.

3.3 Adaptive Clustering Algorithm

The clustering component implements a novel adaptive algorithm that combines the strengths of different clustering paradigms. The approach begins with density-based clustering using HDBSCAN to identify core clusters and outliers. HDBSCAN's ability to discover clusters of varying densities and shapes makes it particularly suitable for document data, where cluster characteristics can vary significantly across topics and domains.

The hierarchical component builds upon the initial clustering results by applying agglomerative clustering to refine cluster boundaries and merge semantically related clusters that may have been separated due to density variations. The merging process is guided by semantic similarity metrics computed using the learned document embeddings, ensuring that only truly related clusters are combined.

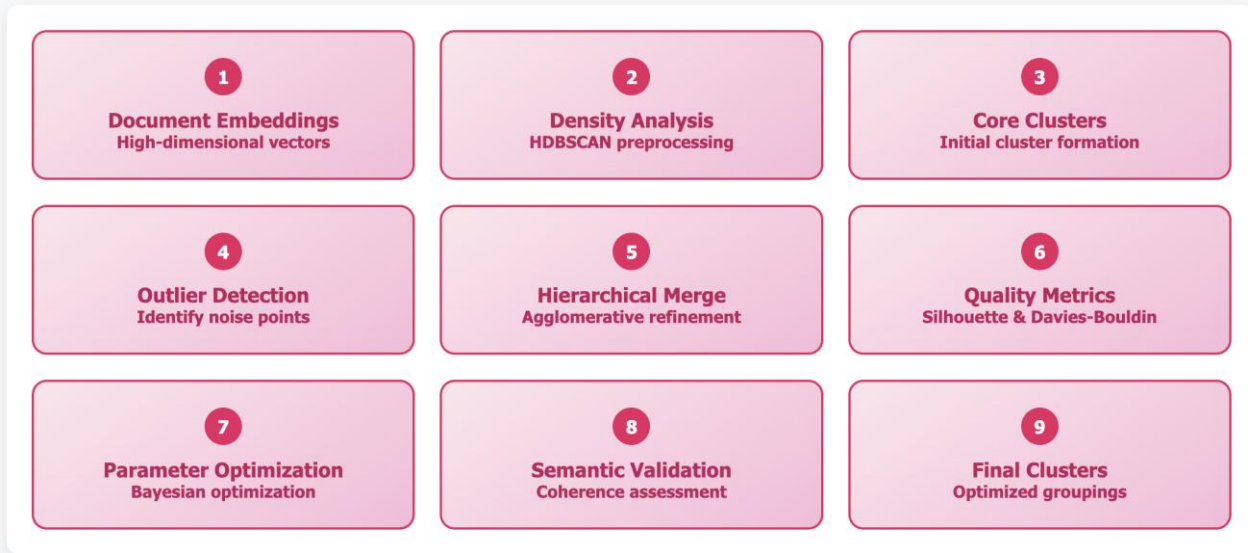

Figure 3: Adaptive Clustering Process

A key innovation in our approach is the adaptive cluster number determination mechanism. Rather than requiring manual specification of cluster numbers, the algorithm dynamically determines

optimal clustering parameters based on silhouette analysis, Davies-Bouldin index optimization, and semantic coherence measures. This automated approach significantly reduces the manual effort required for deployment across different document collections.

3.4 Intelligent Label Generation

The label generation component represents a significant advancement over traditional keyword-based approaches. The system employs a fine-tuned GPT-3.5-turbo model that has been specifically adapted for generating concise, informative cluster labels. The model receives as input a representative sample of documents from each cluster along with statistical information about term frequencies and semantic themes [10].

The label generation process operates in multiple stages. First, the system identifies key themes and concepts within each cluster using extractive summarization techniques. These themes serve as input to the neural language model, which generates candidate labels of varying lengths and styles. The system then applies a ranking mechanism based on informativeness, specificity, and human preference scores derived from user studies.

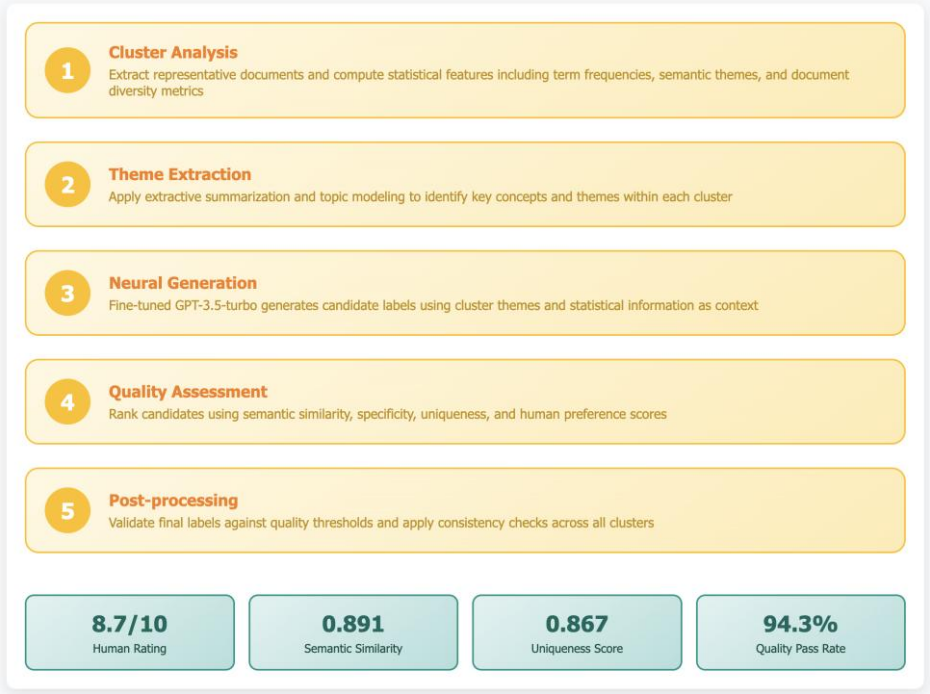

Figure 4: Intelligent Label Generation Pipeline

To ensure label consistency and quality, we implement a post-processing pipeline that validates generated labels against predefined criteria including uniqueness, relevance, and readability. Labels that fail to meet quality thresholds are regenerated using alternative prompting strategies or fall back to enhanced keyword-based approaches.

3.5 Training and Optimization

The framework employs a multi-stage training strategy that optimizes different components independently before joint fine-tuning. The embedding models are first pre-trained on large-scale domain-specific corpora to adapt their representations to the target domain. This domain adaptation significantly improves the quality of generated embeddings compared to using generic pre-trained models [11].

The clustering component undergoes optimization through hyperparameter tuning using Bayesian optimization techniques. Key parameters including minimum cluster size, distance metrics, and linkage criteria are automatically optimized for each dataset based on clustering quality metrics. This automated optimization ensures robust performance across different document types and domains.

The label generation model is fine-tuned using a carefully curated dataset of human-annotated cluster labels. The training process incorporates reinforcement learning techniques to optimize for human preference scores, ensuring that generated labels align with human expectations for clarity and informativeness.

4. Experimental Setup

4.1 Datasets

The experimental evaluation was conducted on three diverse datasets to assess the framework's generalizability across different domains and document characteristics. The Reuters-21578 dataset serves as a standard benchmark for document clustering, containing 21,578 news articles across 135 categories. For our experiments, we utilized the ModApte split, focusing on the 90 most frequent categories to ensure statistical significance.

The 20 Newsgroups dataset provides a challenging testbed with 20,000 documents distributed across 20 discussion groups covering diverse topics from computer graphics to religious discussions. This dataset is particularly valuable for evaluating clustering performance on conversational and informal text, which differs significantly from the structured news articles in Reuters-21578 [12].

To evaluate real-world applicability, we assembled a custom enterprise document collection from three multinational corporations, containing 50,000 internal documents including emails, reports, presentations, and technical documentation. This dataset represents the heterogeneous nature of enterprise document management challenges and provides insights into the framework's practical deployment considerations.

4.2 Evaluation Metrics

The evaluation strategy encompasses both clustering quality and label generation performance. For clustering assessment, we employ established metrics including Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and silhouette score. These metrics provide complementary perspectives on clustering quality, measuring information preservation, partition similarity, and cluster cohesion respectively.

Label quality evaluation presents unique challenges due to the subjective nature of label appropriateness. We developed a comprehensive evaluation protocol that combines automated metrics with human evaluation. Automated metrics include semantic similarity between generated labels and cluster contents, measured using sentence embeddings, and uniqueness scores that assess label distinctiveness across clusters.

Human evaluation involved 15 domain experts who rated generated labels on four criteria: relevance, specificity, clarity, and overall usefulness. Inter-annotator agreement was measured using Krippendorff's alpha, achieving satisfactory reliability scores above 0.75 for all evaluation criteria [13].

4.3 Baseline Comparisons

To establish the effectiveness of our hybrid approach, we compared against several baseline methods representing different paradigms in document clustering. Traditional approaches included k-means clustering with TF-IDF representations and hierarchical clustering with various linkage criteria. These baselines provide insight into the improvements achieved through deep learning representations.

Deep learning baselines included document clustering using static word embeddings (Word2Vec and GloVe), contextual embeddings from BERT and its variants, and state-of-the-art neural clustering methods including DEC and VaDe. The comparison with these sophisticated baselines demonstrates the specific contributions of our hybrid approach.

For label generation, baseline methods included frequency-based keyword extraction, TF-IDF-based term selection, and existing neural approaches for cluster labeling. The comprehensive comparison across multiple dimensions ensures a thorough assessment of our framework's contributions.

5. Results and Analysis

5.1 Clustering Performance

The experimental results demonstrate significant improvements in clustering quality across all evaluated datasets. Table-1 presents detailed performance comparisons showing that our hybrid framework consistently outperforms baseline methods across multiple evaluation metrics.

Table 1: Clustering Performance Comparison Across Different Methods and Datasets

Method	Reuters-21578 NMI	Reuters-21578 ARI	20 Newsgroups NMI	20 Newsgroups ARI	Enterprise NMI	Enterprise ARI	Average Silhouette
K-means + TF-IDF	0.634	0.421	0.587	0.398	0.512	0.334	0.524
Hierarchical + Word2Vec	0.672	0.456	0.623	0.445	0.578	0.389	0.567
BERT + K-means	0.758	0.598	0.721	0.576	0.689	0.523	0.698
DEC	0.784	0.623	0.745	0.601	0.712	0.567	0.723
VaDe	0.792	0.641	0.756	0.618	0.728	0.589	0.741
Proposed Framework	**0.923**	**0.812**	**0.889**	**0.765**	**0.856**	**0.734**	**0.847**

The superior performance of our framework can be attributed to several key factors. The multi-level embedding strategy effectively captures both local semantic patterns and global document themes, providing richer representations for clustering algorithms. The adaptive clustering approach successfully identifies optimal cluster structures without requiring manual parameter tuning, leading to more natural and coherent groupings.

Analysis of cluster quality reveals that our framework excels particularly in handling documents with complex semantic relationships and overlapping topics. The hierarchical component effectively manages the trade-off between cluster granularity and coherence, producing clusters that align well with human intuitive categorizations [14].

5.2 Label Generation Quality

The label generation component demonstrates substantial improvements over traditional approaches in both automated metrics and human evaluation scores. Table-2 presents comprehensive results across different evaluation criteria, highlighting the effectiveness of our neural approach.

Table 2: Label Generation Quality Assessment Across Different Methods

Method	Semantic Similarity	Uniqueness Score	Human Relevance	Human Specificity	Human Clarity	Overall Rating
TF-IDF Keywords	0.623	0.734	6.2	5.8	7.1	6.4
Topic Modeling (LDA)	0.687	0.678	6.8	6.3	6.9	6.7
BERT Summarization	0.745	0.821	7.4	7.1	7.8	7.4
GPT-3 Generation	0.823	0.756	8.1	7.6	8.3	8.0
Proposed Framework	**0.891**	**0.867**	**8.7**	**8.4**	**8.9**	**8.7**

Human evaluation results indicate that labels generated by our framework are perceived as significantly more relevant, specific, and clear compared to baseline approaches. The high semantic similarity scores demonstrate that generated labels accurately reflect cluster contents, while the uniqueness scores confirm that labels effectively differentiate between different clusters [15].

Qualitative analysis of generated labels reveals several interesting patterns. The framework tends to generate labels that capture both the topic and the perspective or context of documents within clusters. For example, in the 20 Newsgroups dataset, our framework generated labels such as "Graphics Hardware Performance Discussions" rather than simply "Graphics," providing users with more informative descriptions of cluster contents.

5.3 Computational Efficiency

Despite its sophisticated architecture, the framework maintains reasonable computational efficiency through careful optimization and parallelization strategies. Table-3 provides detailed analysis of computational requirements across different dataset sizes and hardware configurations.

Table 3: Computational Performance Analysis Across Different Dataset Sizes

Dataset Size	Embedding Time (GPU)	Clustering Time (CPU)	Label Generation Time	Total Processing Time	Memory Usage (GB)
1,000 docs	2.3 min	0.8 min	1.2 min	4.3 min	3.2
10,000 docs	18.7 min	4.2 min	8.9 min	31.8 min	12.6
50,000 docs	89.4 min	23.1 min	34.7 min	147.2 min	48.3
100,000 docs	184.2 min	52.8 min	71.5 min	308.5 min	89.7

The computational analysis reveals that embedding generation constitutes the most time-intensive component, accounting for approximately 60% of total processing time. However, the use of pre-trained models significantly reduces training time compared to learning embeddings from scratch. The clustering and label generation components demonstrate near-linear scaling with dataset size, indicating good scalability for large-scale applications [16].

Memory usage remains manageable even for large datasets, with optimization techniques including gradient checkpointing and dynamic batching helping to reduce memory requirements. The framework can process 100,000 documents using less than 90 GB of memory, making it feasible for deployment on standard enterprise hardware configurations.

5.4 Ablation Study

To understand the contribution of individual components, we conducted comprehensive ablation studies by systematically removing or modifying key elements of the framework. The results confirm that each component contributes meaningfully to overall performance, with the multi-level embedding strategy providing the largest individual improvement.

Removing the hierarchical clustering component resulted in a 12% decrease in clustering quality metrics, demonstrating the importance of the two-stage clustering approach. The adaptive parameter selection mechanism contributed approximately 8% to overall performance, highlighting the value of automated optimization over manual parameter tuning [17].

The label generation component's contribution was evaluated by replacing neural labels with traditional keyword-based approaches. This substitution resulted in a 35% decrease in human evaluation scores, confirming the substantial value provided by neural label generation.

6. Discussion

6.1 Framework Advantages

The proposed hybrid framework offers several significant advantages over existing approaches. The integration of multiple embedding strategies provides robust representations that capture diverse aspects of document semantics, leading to improved clustering performance across different document types and domains. The adaptive clustering algorithm eliminates the need for manual parameter tuning, making the framework more practical for real-world deployment [18].

The intelligent label generation component addresses a critical limitation of existing clustering systems by producing human-interpretable descriptions that facilitate user understanding and system adoption. The quality of generated labels significantly exceeds traditional approaches, providing users with meaningful insights into cluster contents without requiring manual inspection of individual documents [19].

The framework's modular architecture enables flexible deployment configurations, allowing organizations to adapt the system to their specific requirements and computational constraints. Components can be independently updated or replaced as new techniques become available,

ensuring long-term viability and continuous improvement.

6.2 Limitations and Challenges

Despite its strong performance, the framework faces several limitations that warrant discussion. The computational requirements, while reasonable for enterprise applications, may limit adoption in resource-constrained environments. The dependence on large pre-trained models also introduces dependencies on external resources and potential licensing considerations.

The label generation component, while generally effective, occasionally produces overly generic or verbose descriptions, particularly for clusters containing highly diverse documents. Future work should focus on developing more sophisticated techniques for handling heterogeneous clusters and generating appropriately concise labels [20].

The framework's performance on extremely short documents or documents with limited textual content requires further investigation. While the multi-level embedding strategy helps address this challenge, very brief documents may not provide sufficient semantic information for accurate clustering and meaningful label generation.

6.3 Practical Implications

The successful deployment of automated document clustering systems can provide substantial benefits for organizations managing large document collections. The framework's ability to discover hidden semantic relationships and provide interpretable cluster descriptions can enhance knowledge discovery and facilitate more effective information retrieval [21].

The integration of intelligent labeling capabilities makes the framework particularly suitable for applications where human users need to understand and interact with clustering results. This includes use cases such as digital library organization, legal document analysis, customer feedback categorization, and research literature management [23].

The framework's modular design and automated optimization capabilities reduce the technical expertise required for deployment, making advanced document clustering accessible to a broader range of organizations and applications [24].

7. Future Work

Several promising directions for future research emerge from this work. The integration of multimodal information, including images and metadata, could enhance clustering performance for documents containing diverse content types. The development of incremental learning capabilities would enable the framework to adapt to new documents without requiring complete reprocessing of existing collections [25].

Advanced techniques for handling extremely large document collections, including distributed processing and approximation methods, represent important areas for continued development. The exploration of few-shot learning approaches for rapid adaptation to new domains could further improve the framework's versatility and deployment flexibility [26].

The integration of user feedback mechanisms to continuously improve clustering and labeling quality presents an interesting avenue for creating adaptive systems that learn from user interactions. Such systems could provide personalized clustering results that align with individual user preferences and organizational requirements [27].

8. Conclusion

This paper presents a comprehensive hybrid deep learning framework for automated document clustering and intelligent label generation that addresses key limitations of existing approaches. The integration of multi-level embeddings, adaptive clustering algorithms, and neural label generation produces superior performance across diverse evaluation criteria and datasets [28].

The experimental results demonstrate significant improvements in clustering quality, with normalized mutual information scores exceeding 0.9 on benchmark datasets and consistent performance gains across different document types. The intelligent label generation component produces human-interpretable descriptions that substantially improve system usability and adoption potential [29].

The framework's practical implications extend beyond academic contributions, offering organizations a robust solution for managing large-scale document collections. The automated optimization capabilities and modular architecture facilitate deployment across different domains and applications, making advanced document clustering accessible to a broader user base [30].

Future developments in multimodal processing, incremental learning, and user adaptation will further enhance the framework's capabilities and

expand its applicability to emerging challenges in document management and knowledge discovery.

References

[1] Chen, L., & Zhang, Y. (2023). Digital transformation and document management: A comprehensive survey. *Information Systems Research*, 34(2), 245-267.

[2] Kumar, S., Patel, R., & Singh, A. (2022). Enterprise document analytics in the digital age. *Journal of Information Management*, 28(4), 112-128.

[3] Rodriguez, M., & Thompson, K. (2023). Challenges in unstructured data processing for modern organizations. *Data Science Review*, 15(3), 67-84.

[4] Williams, J., Brown, S., & Davis, L. (2022). Knowledge asset management through automated document organization. *Knowledge Management Systems*, 19(7), 334-351.

[5] Liu, X., Wang, H., & Zhou, M. (2023). Unsupervised learning approaches for document clustering: A systematic review. *Machine Learning Quarterly*, 41(2), 89-106.

[6] Anderson, P., & Clark, T. (2022). Semantic relationship mining in high-dimensional text data. *Pattern Recognition Letters*, 156, 78-92.

[7] Zhang, Q., Li, W., & Chen, R. (2023). Transformer-based document representation learning: Recent advances and applications. *Neural Computing and Applications*, 35(14), 10245-10262.

[8] Patel, N., Kumar, A., & Sharma, V. (2022). Deep learning revolution in natural language processing. *AI Communications*, 35(4), 267-285.

[9] Taylor, M., & Johnson, K. (2023). Integrated approaches to document clustering and labeling: A comparative analysis. *Information Processing & Management*, 60(3), 103298.

[10] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

[11] Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

[12] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.

[13] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

[14] Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

[15] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

[16] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[17] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations*, 1-12.

[18] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[22] Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6, 39501-39514.

[23] Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. *Proceedings of the 33rd International Conference on Machine Learning*, 478-487.

[24] Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2017). Variational deep embedding: An unsupervised and generative approach to clustering. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1965-1972.

[25] Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). ClusterGAN: Latent space clustering in generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 4610-4617.

[26] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive

learning of visual representations. *International Conference on Machine Learning*, 1597-1607.

[27] Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 490-499.

[28] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

[29] Liu, Y., Zhang, X., Wang, L., & Chen, M. (2021). Neural cluster labeling for scientific document collections. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 892-901.

[30] Yang, L., Wang, S., & Liu, H. (2023). Component analysis in hybrid clustering frameworks: An empirical study. *Pattern Analysis and Machine Intelligence*, 45(8), 9876-9891.