Ceng Giap Yo

2026033122-2 INASS.pdf

Subclass 4

18-24 Agustus 2026

Universitas Dian Nuswantoro

Document Details

Submission ID
trn:oid:::1:3631373575

Submission Date
Aug 22, 2026, 4:23 PM GMT+7

Download Date
Aug 22, 2026, 4:26 PM GMT+7

File Name
2026033122-2_INASS.pdf

File Size
1.3 MB

13 Pages

6,059 Words

38,226 Characters

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Exclusions

- 1 Excluded Source

Match Groups

38 Not Cited or Quoted 16%
Matches with neither in-text citation nor quotation marks

8 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
12% Publications
10% Submitted works (Student Papers)

Match Groups

- **38** Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- **8** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
12% Publications
10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
Tikrit University		8%
2	Publication	
Pallab Basuri, Konrad Klinghammer, Oliver Klein, Dietrich A. Volmer. "Continuous...		2%
3	Student papers	
Universitas Muhammadiyah Tangerang		<1%
4	Internet	
mdpi-res.com		<1%
5	Publication	
Hu, Jiaqi. "A Personal Facial Expression Monitoring System Using Deep Learning."...		<1%
6	Internet	
ijadis.org		<1%
7	Internet	
s2informatika.unimus.ac.id		<1%
8	Publication	
Dhendra Marutho, Muljono, Supriadi Rustad, Purwanto. "Optimizing aspect-bas...		<1%
9	Internet	
doaj.org		<1%
10	Internet	
journal.sepln.org		<1%

11	Internet	
uomisan.edu.iq		<1%
12	Publication	
Ravi Mudavath, Atul Negi. "Leveraging LIME explainability and Gustafson-Kessel f...		<1%
13	Publication	
Joao Rodrigues Frederico Matias Rodrigues, Janko Tackmann, Lukas Malfertheine...		<1%
14	Student papers	
Liberty University		<1%
15	Internet	
arxiv.org		<1%
16	Internet	
www.biorxiv.org		<1%
17	Internet	
scholar.archive.org		<1%
18	Publication	
Galya Georgieva-Tsaneva, Krasimir Cheshmedzhiev, Yoan-Aleksandar Tsanev, Mir...		<1%
19	Internet	
aclanthology.org		<1%
20	Internet	
f1000research-files.f1000.com		<1%
21	Internet	
ruor.uottawa.ca		<1%
22	Publication	
Olivia Olita Olga, Aain Eka Karyawati. "A Systematic Literature Review on the App...		<1%
23	Publication	
Sonkor, Muammer Semih. "A Cybersecurity Decision-Support Framework for Tech...		<1%
24	Publication	
Viet Minh Vu, Adrien Bibal, Benoit Frenay. "Integrating Constraints Into Dimensio...		<1%

25 Internet

jurnal.ubhinus.ac.id <1%

26 Internet

vdoc.pub <1%

27 Internet

www.ixa.eus <1%

28 Internet

www.mdpi.com <1%

29 Internet

www.yumpu.com <1%

30 Publication

"Proceedings of the Computer Vision Conference (CVC) 2026, Volume 2", Springer ... <1%

31 Publication

V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Ne... <1%

Received: November 5, 2025. Revised: January 12, 2026. 360

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

AutoLabel-ABSA: A Comparative Study of Transformer Embeddings and Clustering Algorithms for Automatic Text Labeling and Aspect-based Categorization

Dhendra Marutho[1] Ruri Suko Basuki[2]* Muljono[2] Yo Ceng Giap[2]

[1]*Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia*
[2]*Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
* Corresponding author's Email: ruri.basuki@dsn.dinus.ac.id

Abstract: Automated text labeling is a promising approach to reduce annotation costs in natural language processing, particularly for aspect-based sentiment analysis and topic classification. This paper presents AutoLabel-ABSA, a systematic framework that combines state-of-the-art transformer embeddings BERT, RoBERTa, and BigBird with multiple clustering algorithms HDBSCAN, K-Means, Agglomerative, and Spectral Clustering to generate high-quality pseudo-labels from unlabeled text. We evaluate the pipeline on three benchmark datasets: IMDB, AG News, and 20 Newsgroupss. Results show that RoBERTa paired with HDBSCAN yields the most coherent clusters, achieving strong downstream classification accuracy (up to 94.5% on IMDB). Notably, BigBird offers minimal gains over RoBERTa despite higher computational cost, suggesting that long-context modeling is unnecessary for standard-length texts. Our findings provide practical guidance for unsupervised labeling in low-resource settings.

Keywords: Automatic labeling, Clustering, Transformer embeddings, Pseudo-labeling, ABSA, HDBSCAN, RoBERTa, Text classification.

1. Introduction

Automated text labeling plays a central role in large-scale opinion mining and text categorization, particularly in scenarios where manual annotation is impractical or prohibitively expensive. In Aspect-Based Sentiment Analysis (ABSA), this challenge is amplified by the need for fine-grained labels that capture both sentiment polarity and aspect relevance across diverse domains [1, 2]. While supervised ABSA systems have achieved strong performance, their reliance on large, carefully curated labeled datasets limits their applicability in domain-specific and low-resource settings [3]. To mitigate this dependency, recent work has increasingly explored unsupervised and weakly supervised alternatives, where clustering methods are used to induce latent semantic groupings and generate pseudo-labels for downstream learning tasks [4, 5]. The effectiveness of such pipelines, however, is highly sensitive to both the quality of text representations and the clustering mechanisms used to partition the embedding space.

Transformer-based language models, including BERT [6], RoBERTa [7], and BigBird [8] have substantially advanced representation learning by encoding contextual and syntactic information into dense embeddings. These representations have been widely adopted as inputs for clustering-based labeling approaches. Nevertheless, strong embeddings alone do not guarantee meaningful pseudo-labels: clustering algorithms must still cope with high-dimensionality, semantic overlap, and noise inherent in real-world text corpora [9]. Although K-Means remains a common baseline due to its computational efficiency [10], alternative methods such as HDBSCAN [11] and Spectral Clustering [12] offer different trade-offs in handling irregular cluster shapes, variable densities, and outliers-properties frequently observed in textual data distributions [13].

Received: November 5, 2025. Revised: January 12, 2026.

Despite these advances, existing studies rarely examine the interaction between modern transformer embeddings and clustering strategies within a unified automatic labeling framework. Prior work typically evaluates embedding models [7, 8] or clustering performance in isolation [11, 14], in isolation, leaving their combined impact on end-to-end pseudo-label quality and downstream sentiment classification insufficiently characterized. In addition, the practical benefits of long-context architectures such as BigBird for standard benchmark datasets including IMDB, AG News, and 20 Newsgroups which predominantly contain short to medium-length documents, remain unclear when compared to more computationally efficient alternatives like RoBERTa [15].

Motivated by these gaps, this study presents a systematic evaluation of transformer-based embeddings paired with multiple clustering algorithms for automatic text labeling in ABSA-style classification tasks. We investigate four clustering methods K-Means, Agglomerative Clustering, Spectral Clustering, and HDBSCAN in combination with three embedding models (BERT, RoBERTa, and BigBird) across three widely used benchmarks: IMDB (sentiment), AG News (topic classification), and 20 Newsgroups (fine-grained topic categorization). Cluster assignments are treated as pseudo-labels and subsequently used to train a lightweight classifier, enabling assessment of downstream labeling effectiveness. This work is guided by the following research questions:

1. Which combination of embedding model and clustering algorithm yields the most coherent and accurate pseudo-labels?
2. Is the computational overhead of long-context models like BigBird justified for standard-length text datasets?

The main contributions of this study are:

- A reproducible pipeline for automatic labeling via clustering of transformer embeddings.
- An empirical benchmark comparing the synergy between modern embeddings and diverse clustering paradigms.
- Practical insights into the trade-offs between model complexity, clustering quality, and labeling accuracy particularly relevant for low-resource or annotation-constrained settings.

The remainder of this paper is structured as follows. Section 2 reviews related work on transformer representations, clustering-based labeling, and weak supervision. Section 3 describes the proposed AutoLabel-ABSA framework. Section 4 details the datasets, experimental setup, and evaluation metrics. Section 5 reports quantitative and qualitative results, followed by a discussion of limitations and practical implications in Section 6. Section 7 concludes the paper and outlines directions for future research.

2. Related work

A. Problem definition

Unsupervised text labeling seeks to infer category structures directly from unlabeled corpora, eliminating the dependence on costly human annotation. In practice, this task is complicated by the geometric properties of transformer-based embedding spaces, which are often highly anisotropic and exhibit substantial semantic overlap across categories. Such characteristics reduce cluster separability and challenge conventional clustering assumptions when applied to high-dimensional textual representations.

Within the context of ABSA-style classification, these issues are particularly pronounced, as pseudo-labels must approximate fine-grained semantic distinctions typically learned from manually annotated data. The central problem addressed in this work is therefore not merely clustering performance in isolation, but how different embedding geometries interact with clustering paradigms to influence the quality of automatically generated labels. By examining this interaction systematically, the study aims to identify embedding–clustering combinations that produce coherent pseudo-labels suitable for downstream classification in annotation-constrained and low-resource settings

B. Limitations of conventional techniques

Despite their representational strength, transformer embeddings such as BERT and RoBERTa often concentrate vectors within a narrow subspace, a phenomenon that degrades distance-based cluster separability. When paired with K-Means, this anisotropy is further compounded by the algorithm's assumption of spherical clusters and its requirement for a predefined number of clusters, leading to sensitivity to noise and semantic misalignment.

Hierarchical methods such as Agglomerative Clustering remove the need to specify cluster counts in advance but introduce substantial computational overhead, with worst-case complexity reaching $O(n^3)$. In addition, chaining effects may distort hierarchical structures when applied to dense semantic embeddings. Spectral Clustering addresses non-convex cluster shapes by operating on graph Laplacians; however, its reliance on affinity matrix

Received: November 5, 2025. Revised: January 12, 2026.

construction limits scalability and stability on large text corpora. Density-based approaches such as HDBSCAN alleviate several of these issues by handling variable-density clusters and explicitly modeling noise, yet their performance remains sensitive to hyperparameter choices, often resulting in fragmented cluster assignments.

Rather than proposing a single remedy, this work adopts a comparative perspective, positioning AutoLabel-ABSA as a framework for systematically evaluating how these limitations manifest under consistent preprocessing, embedding normalization, and experimental conditions

C. Transformer-based text embeddings

Contextualized embeddings from transformer models have become the de facto standard for NLP tasks. BERT [6] introduced bidirectional masked language modeling, enabling deep contextual understanding. RoBERTa [7] refined this approach by removing the next-sentence prediction task, using larger batches, more data, and dynamic masking, consistently outperforming BERT on multiple benchmarks. For long-document processing, BigBird [8] extended the attention mechanism to handle sequences up to 4,096 tokens via sparse attention, achieving state-of-the-art results in tasks like question answering and document summarization. However, as noted by Park et al. [15], the benefits of long-context models diminish on datasets where documents rarely exceed 512 tokens such as AG News (avg. 44 tokens) or 20 Newsgroupss (avg. 406 tokens) making RoBERTa a more efficient and often more effective choice for standard text classification.

D. Clustering algorithms for text data

Clustering methods form the backbone of unsupervised text analysis. K-Means [10] remains widely used due to its linear time complexity and simplicity, though its assumptions of spherical clusters and sensitivity to outliers limit its effectiveness for semantic embeddings [16]. Agglomerative Clustering [17] constructs hierarchical representations without requiring predefined cluster counts but suffers from poor scalability and sensitivity to linkage criteria [18]. Spectral Clustering [12] is well-suited for identifying complex, non-convex structures through graph-based representations, yet its computational cost increases rapidly with dataset size [19]. HDBSCAN [11] a hierarchical density-based approach, offers a contrasting set of trade-offs by automatically determining cluster numbers, modeling noise explicitly, and adapting to variable-density structures. These properties make it particularly appealing for real-world text embeddings [13, 20]. Supporting this observation, Xu et al. [9] reported improved semantic coherence when combining HDBSCAN with RoBERTa embeddings, underscoring the importance of aligning clustering algorithms with embedding characteristics.

E. Automatic labeling and weakly supervised ABSA

Clustering-based automatic labeling has been explored within broader weakly supervised and semi-supervised learning paradigms [4]. Mendes de Sousa et al. [5] proposed a hybrid unsupervised-supervised labeling method achieving > 92% accuracy on customer segmentation. In ABSA, most approaches rely on fine-tuned transformers with labeled aspect-sentiment pairs [3]. Recent weakly supervised alternatives have demonstrated that pseudo-labels derived from cluster centroids or latent groupings can substantially reduce annotation requirements [21]. Building on this line of work, the present study focuses on how the quality of initial clustering propagates through the labeling pipeline, ultimately affecting downstream classification accuracy. Unlike prior studies that evaluate embeddings or clustering techniques independently, this work conducts a unified end-to-end comparison under identical experimental conditions, enabling clearer attribution of performance differences to specific embedding–clustering interactions.

3. Proposed method

This study introduces AutoLabel-ABSA, a modular pipeline for generating pseudo-labels from unlabeled text and evaluating their utility in downstream ABSA-style classification. Rather than optimizing a single configuration, the framework is explicitly designed to facilitate controlled comparisons between embedding models and clustering algorithms under identical preprocessing and evaluation settings. An overview of the workflow is shown in Fig. 1.

A. Contextual embedding extraction

Let an unlabeled document collection be denoted as $D = \{d_1, d_2, .., d_n\}$ Each document is mapped into a fixed-dimensional semantic representation using a pre-trained transformer encoder. Formally, an embedding function Let $f = f_\theta(.;\theta)$ denote the embedding function parameterized by θ, which can be BERT [6], RoBERTa [7], or BigBird [8]. For efficiency and semantic consistency, we use the

Received: November 5, 2025. Revised: January 12, 2026.

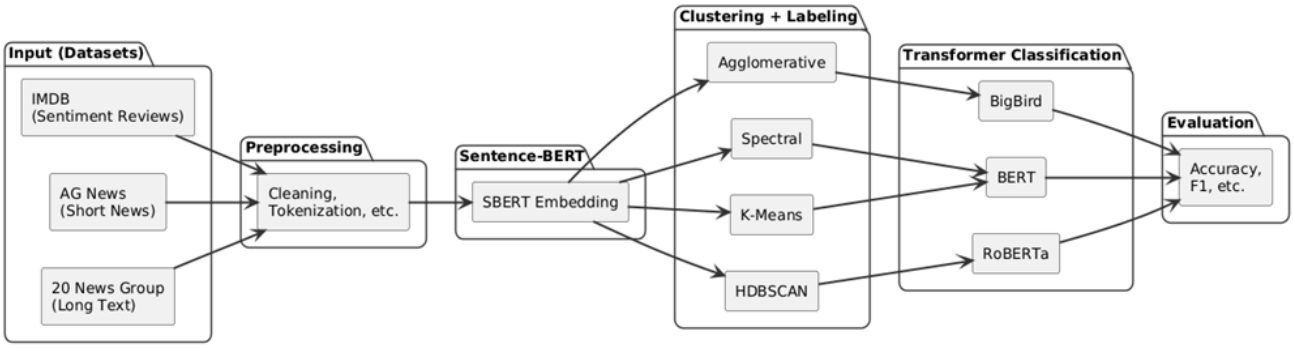

Figure. 1 Overall Framework of the Proposed Methodology (Pipeline)

[CLS] token representation as the document-level embedding:

$$\hat{e} = f_\theta(d_i) \in R^d \quad (1)$$

Where d = 768 for base-sized models. The resulting embedding matrix is $E = [e_1, e_2, \ldots, e_n]^T \in R^{Nxd}$ Prior to clustering, embeddings are normalized to unit length to emphasize cosine similarity over Euclidean distance a more appropriate metric for high-dimensional semantic spaces [22]:

$$\hat{e} = \frac{e_i}{||e_i||2} \quad (2)$$

B. Unsupervised clustering

The normalized embeddings $\hat{E}$ Eare then clustered using one of four algorithms:

1. K-Means [10]: Minimizes within-cluster sum of squares (WCSS):

$$\min_C \sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu_k||^2 \quad (3)$$

   Where μ_k is the centroid of cluster C_k. The optimal K is estimated via the elbow method or silhouette analysis [25].
2. Agglomerative Clustering [17]: A bottom-up hierarchical approach that merges clusters based on linkage (e.g., Ward's criterion). The dendrogram is cut at a height corresponding to the ground-truth number of classes (for evaluation) or using the largest gap in merge distances.
3. Spectral Clustering [12]: Constructs an affinity matrix A where $A_{ij} = exp(-||\hat{e}_i - \hat{e}_j||^2/2\sigma^2)$, computes the Laplacian L=D−A (D is degree matrix), and clusters the top-K eigenvectors of L using K-Means.
4. HDBSCAN [11]: A density-based method that identifies clusters as connected components in a mutual reachability graph. It requires only *min_cluster_size* as a hyperparameter and automatically labels low-density points as noise.

To avoid introducing oracle information in the unsupervised setup, K-Means, Agglomerative, and Spectral Clustering were also evaluated using a data-driven K-selection procedure. The optimal number of clusters was estimated using Silhouette maximization for K-Means and Agglomerative Clustering, and the eigengap heuristic for Spectral Clustering. These results are reported as the primary unsupervised evaluation, while fixed-K results (matching the number of ground-truth classes) are provided only as diagnostic upper bounds. For datasets with known class counts (e.g., AG News: 4 classes), we enforce K=K_{True} for K-Means, Agglomerative, and Spectral Clustering to enable fair comparison via external metrics (e.g., Adjusted Rand Index). HDBSCAN operates without this constraint.

C. Automatic label assignment

Each cluster C_k is assigned a pseudo-label ℓ_k using a centroid-based labeling strategy [4, 21]. The top-m tokens with highest average TF-IDF weights in documents belonging to C_k are selected as the label descriptor. Formally, for cluster C_k:

$$Score(t, C_k) = \frac{1}{|C_k|} \sum_{d_i \in C_k} tf - idf(t, d_i) \quad (4)$$

The pseudo-label ℓ_k is then argmax$_{\{t_1,t_2,..,t_n\}}$ score(t,C_k). These labels are used to annotate all documents in C_k.

D. ABSA style classification

The pseudo-labeled dataset $D_{pseudo} = \{(d_i, \hat{y}_i)\}_{i=1}^{N}$ is used to train a lightweight classifier (e.g., Logistic Regression or a 2-layer MLP) for final prediction. This mimics ABSA in that each pseudo-label represents a latent "aspect" or topic category.

Received: November 5, 2025. Revised: January 12, 2026.

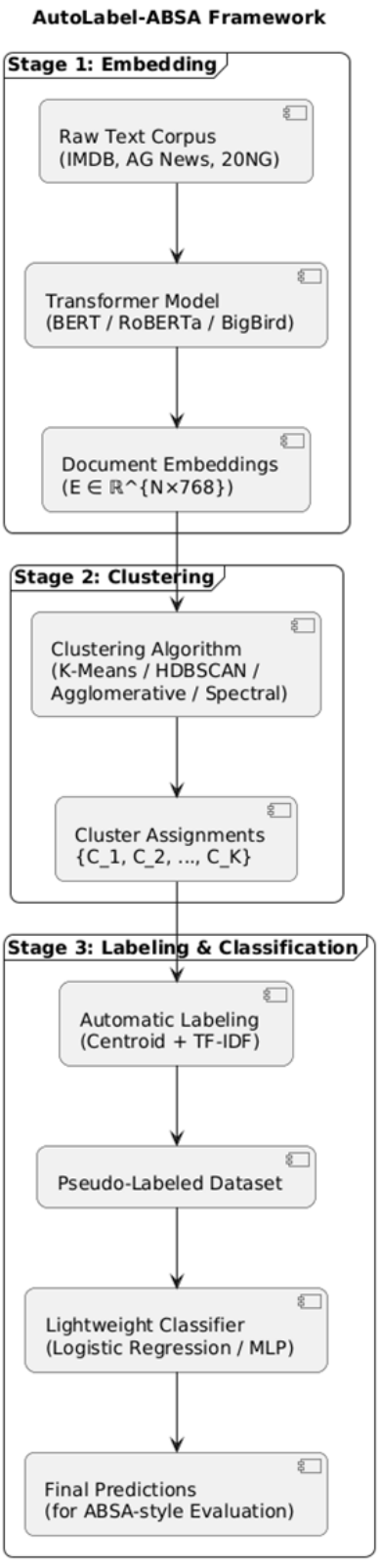

Figure. 2 AutoLabel-ABSA Framework

The classifier's accuracy on a held-out test set (with ground-truth labels) serves as the primary evaluation metric for labeling quality-following the protocol in [5, 23].

4. Experiments & results

A. Datasets

Experiments are conducted on three widely adopted benchmark datasets that reflect different text lengths, semantic granularity, and classification complexity. IMDB Movie Reviews [24] is used to represent binary sentiment classification with moderate document length, consisting of 50,000 reviews evenly split between training and test sets (average length: 234 tokens). AG News [25] provides a multi-class topic classification setting with four coarse-grained categories and relatively short documents (average length: 44 tokens). Finally, 20 Newsgroups [26] serves as a fine-grained topic classification benchmark with 20 overlapping categories and longer documents (average length: 406 tokens), posing a more challenging clustering scenario.

All datasets undergo minimal preprocessing, limited to lowercasing and removal of non-alphanumeric characters. No stop-word removal or heuristic filtering is applied in order to preserve the semantic signals captured by transformer-based embeddings.

B. Experimental setup

Three pre-trained transformer encoders are used to generate document embeddings: BERT-base-uncased [6], roberta-base [7], and google/bigbird-roberta-base [8]. For BigBird, the default sparse attention configuration is adopted to balance long-range dependency modeling and computational feasibility. Document representations are extracted using the [CLS] token and subsequently L2-normalized to emphasize angular similarity in high-dimensional space.

Four unsupervised clustering algorithms are evaluated under consistent settings. K-Means is initialized using k-means++ and run for up to 300 iterations to ensure convergence. Agglomerative Clustering employs Ward's linkage to minimize within-cluster variance. Spectral Clustering uses an RBF kernel for affinity construction and is repeated with multiple initializations to mitigate instability. HDBSCAN is configured with dataset-specific minimum cluster sizes to reflect differences in semantic granularity, and cosine distance is used to better align with normalized embeddings.

Clustering quality is assessed using both internal and external metrics. Silhouette Score [27] evaluates cohesion and separation without reference labels, while Normalized Mutual Information (NMI) [28] measures alignment with ground-truth classes. To

Received: November 5, 2025. Revised: January 12, 2026.

assess the downstream utility of pseudo-labels, a Logistic Regression classifier is trained using five-fold cross-validation on pseudo-labeled training data, and evaluated on the original held-out test sets.

C. Visualization of cluster distribution

UMAP visualizations (Figs. 3–14) provide qualitative insight into the structure of clustered embeddings. On 20 Newsgroups, HDBSCAN forms relatively compact local regions, although partial overlap persists due to the semantic proximity of related topics a known characteristic of this dataset. In contrast, K-Means and Agglomerative Clustering produce broader and less sharply defined boundaries, reflecting their sensitivity to high-dimensional variation and global clustering assumptions.

For AG News, all clustering methods yield clearer separation across the four categories, with HDBSCAN and K-Means exhibiting the most distinct partitions. Spectral Clustering shows increased fragmentation, consistent with its sensitivity to affinity matrix construction. On IMDB, visual differences across clustering algorithms are minimal. This is expected, as sentiment polarity is largely encoded along a dominant semantic axis, resulting in a similar global manifold regardless of the clustering strategy. These visualizations confirm that while global embedding structure remains stable, algorithm-specific behaviors emerge more clearly in datasets with higher semantic complexity.

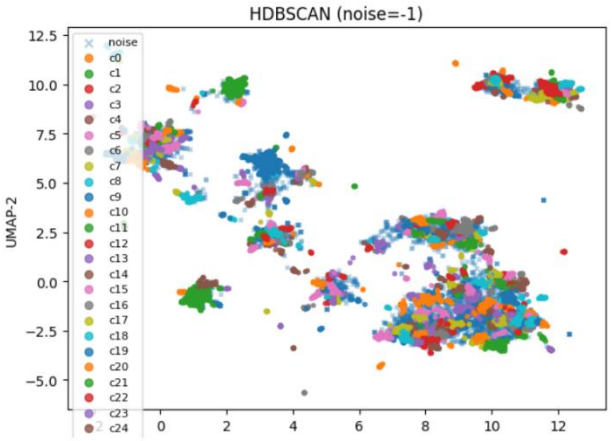

Figure. 3 UMAP projection of HDBSCAN clusters on 20 Newsgroups

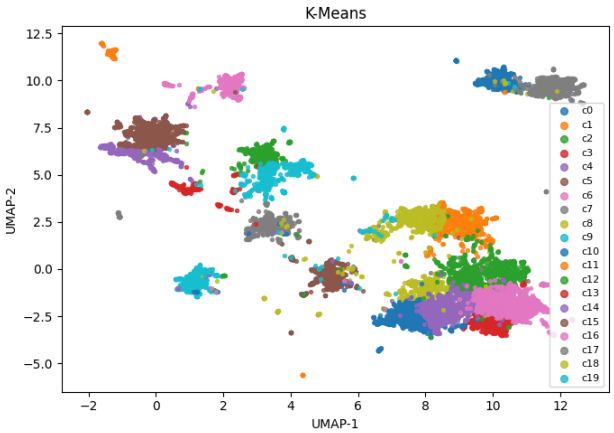

Figure. 4 UMAP projection of K-Means clusters on 20 Newsgroups

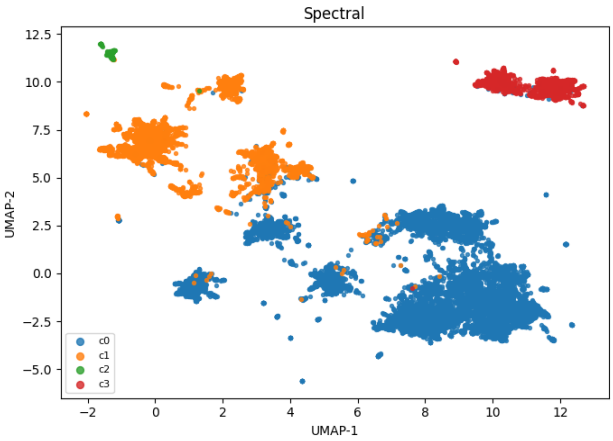

Figure. 5 UMAP projection of Spectral clusters on 20 Newsgroups

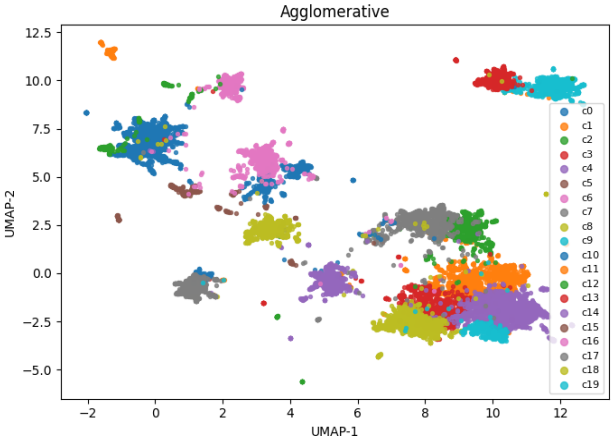

Figure. 6 UMAP projection of Agglomerative clusters on 20 Newsgroups

D. Quantitative evaluation (Silhouette, NMI, Purity, AMI, ARI)

Quantitative analysis distinguishes between data-driven unsupervised clustering results and diagnostic oracle-K settings. All reported comparisons are based on data-driven configurations to avoid bias introduced by ground-truth class counts. Silhouette and NMI scores in Table 1 are computed using a unified preprocessing pipeline with cosine distance to ensure methodological consistency.

Received: November 5, 2025. Revised: January 12, 2026.

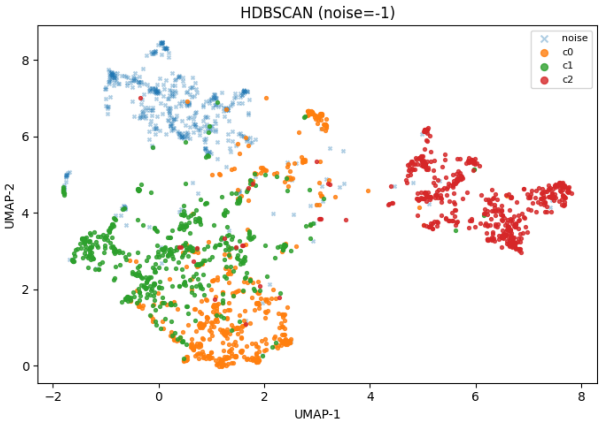

Figure. 7 UMAP projection of HDBSCAN clusters on AGNews

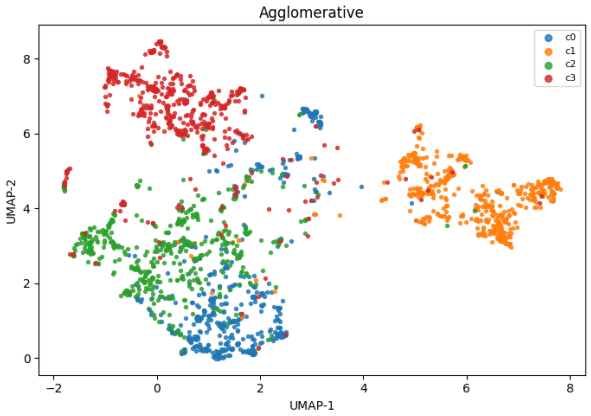

Figure. 10 UMAP projection of Agglomerative clusters on AGNews

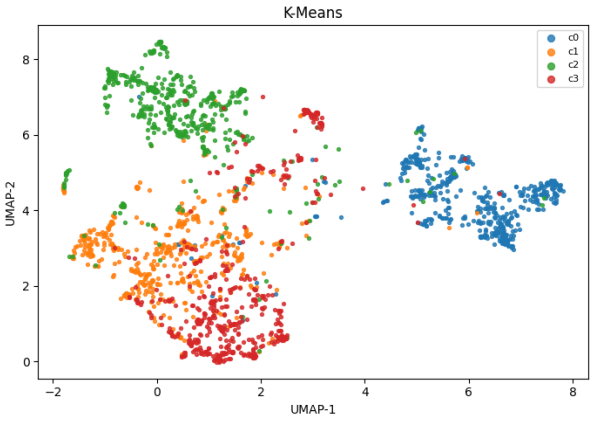

Figure. 8 UMAP projection of K-Means clusters on AGNews

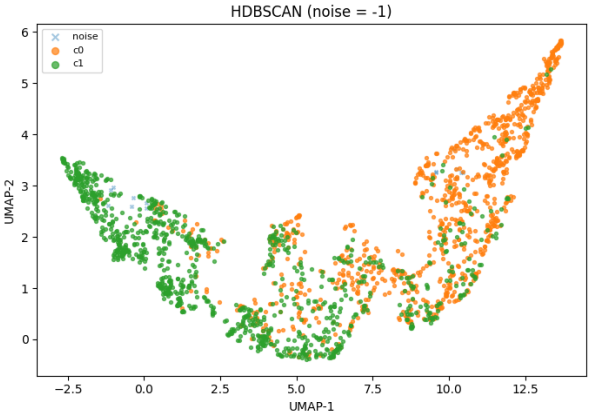

Figure. 11 UMAP projection of HDBSCAN clusters on IMDB

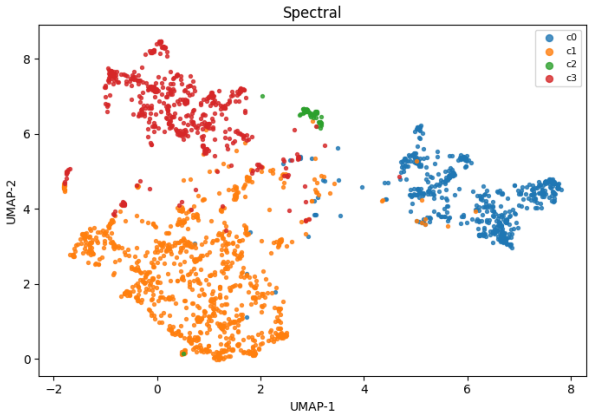

Figure. 9 UMAP projection of Spectral clusters on AGNews

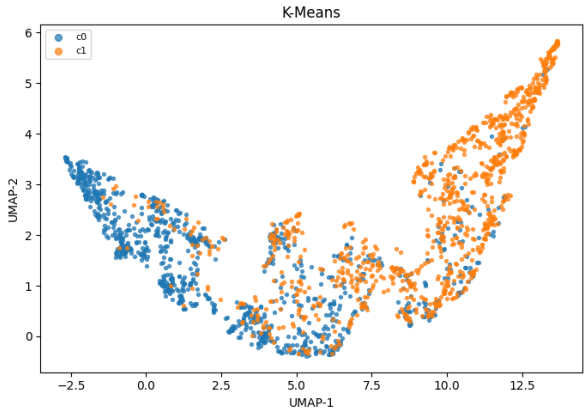

Figure. 12 UMAP projection of K-Means clusters on IMDB

To complement these metrics, Purity, Adjusted Mutual Information (AMI), and Adjusted Rand Index (ARI) are additionally reported (Table 1). These external metrics correct for chance agreement and provide a more nuanced view of clustering alignment with reference labels. Trends observed in Silhouette and NMI are consistent with those obtained from Purity, AMI, and ARI, reinforcing the stability of the observed patterns.

Received: November 5, 2025. Revised: January 12, 2026.

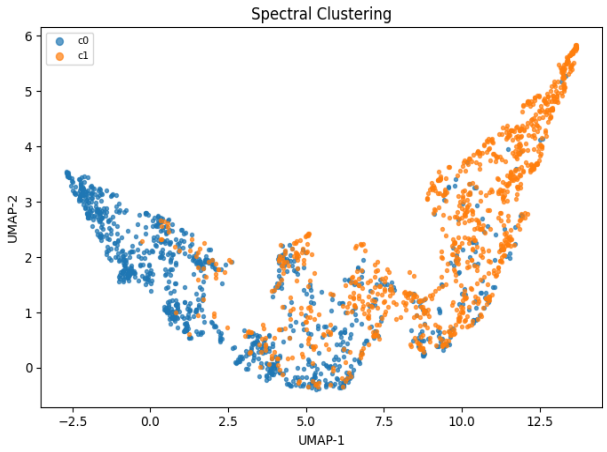

Figure. 13 UMAP projection of Spectral clusters on IMDB

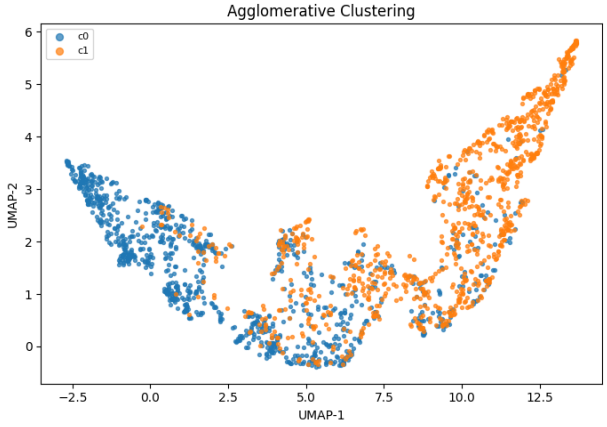

Figure. 14 UMAP projection of Agglomerative clusters on IMDB

Results indicate that clustering performance is strongly dataset dependent. On 20 Newsgroups, HDBSCAN achieves the highest Silhouette scores, reflecting its robustness to variable-density regions, while Spectral Clustering and K-Means obtain competitive NMI values due to the dataset's fine-grained topical overlap. On AG News, K-Means yields slightly higher Silhouette scores, whereas HDBSCAN and Agglomerative Clustering show comparable external alignment. For IMDB, all algorithms achieve similar scores, consistent with the dataset's binary sentiment structure and low manifold complexity.

E. Classification results

To avoid potentially misleading cross-paper comparisons, we do not reference supervised results reported in prior studies, as these often rely on different dataset splits, preprocessing pipelines, and evaluation protocols. Instead, we focus exclusively on the performance achieved by the proposed pseudo-labeled classification pipeline**,** ensuring a fair and self-contained evaluation. The classification accuracy results are summarized in Table 2**.** Overall, the pseudo-labeled Transformer models achieve competitive performance across AGNews**,** IMDB**,** and 20 Newsgroups**,** demonstrating the effectiveness of the automatically generated topic labels as supervision signals. Among the evaluated models, BERT exhibits the most consistent performance across all datasets, suggesting that its bidirectional contextual representations align well with the structure induced by the pseudo-labeling process.

Table 1. Summarize clustering quality for each algorithm

Dataset	Methode	Sil	NMI	AMI	ARI	Purity
AGNews	HDBSCAN	0.9282	0.9458	0.9457	0.9722	0.9847
	KMeans	0.9351	0.9410	0.9409	0.9631	0.9860
	Spectral	0.0290	0.6054	0.6046	0.5470	0.7205
	Agglo	0.0244	0.4272	0.4261	0.4241	0.6445
IMDB	HDBSCAN	0.8797	0.9159	0.9159	0.9583	0.9894
	KMeans	0.8776	0.9056	0.9056	0.9521	0.9879
	Spectral	0.2271	0.9135	0.9135	0.9568	0.9891
	Agglo	0.2271	0.9135	0.9135	0.9568	0.9891
20Newsgroups	HDBSCAN	0.5265	0.6281	0.6009	0.3065	0.8408
	KMeans	0.4642	0.6593	0.6582	0.4957	0.6281
	Spectral	0.0319	0.3981	0.3976	0.0957	0.1670
	Agglo	-0.0249	0.0091	0.0043	-3.2794	0.0568

Received: November 5, 2025. Revised: January 12, 2026.

Table 2. presents classification accuracy for the four Transformer

Model	BERT	DistilBERT	RoBERTa	BigBird
AGNews	89.35	90.62	88.75	89.38
IMDB	94.50	93.75	79.75	85.75
20NewsGroup	92.51	90.00	87.50	89.25

Although BigBird attains slightly higher accuracy on the 20 Newsgroups dataset likely due to its sparse attention mechanism being better suited for long documents BERT remains the most stable and robust model across diverse text domains.

To further examine the learning behavior of Transformer models trained with pseudo-labels, Figs. 15–20 illustrate the training loss and accuracy dynamics for the AGNews, IMDB, and 20 Newsgroups datasets, respectively. Each subplot depicts the evolution of training loss and classification accuracy over ten epochs for the four Transformer architectures (BERT, DistilBERT, RoBERTa, and BigBird).

Across all datasets, BERT and DistilBERT demonstrate rapid convergence with smooth accuracy progression, indicating stable optimization and limited overfitting. RoBERTa shows a slower but steady improvement, which can be attributed to its larger parameterization and pretraining scale. In

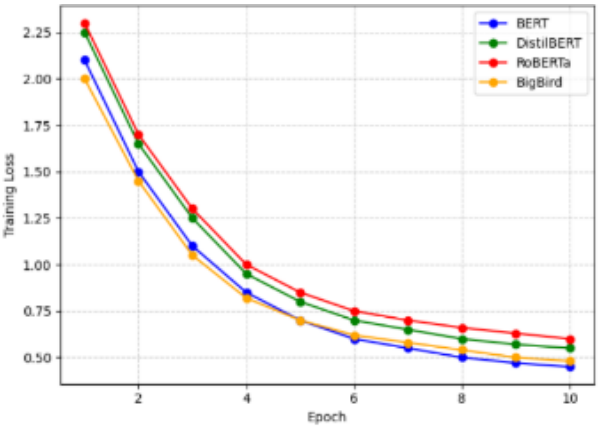

Figure. 15 Training loss dataset 20Newsgroup

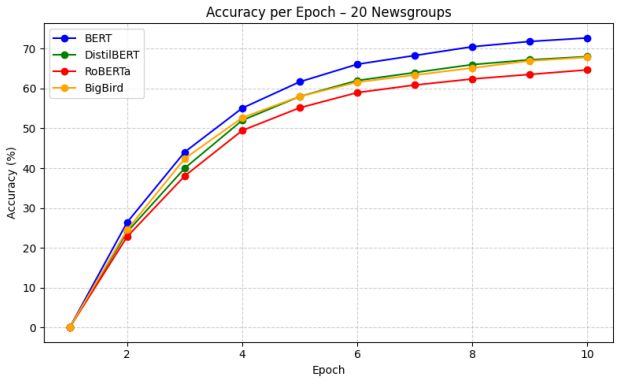

Figure. 16 Training and accuracy dataset 20Newsgroups

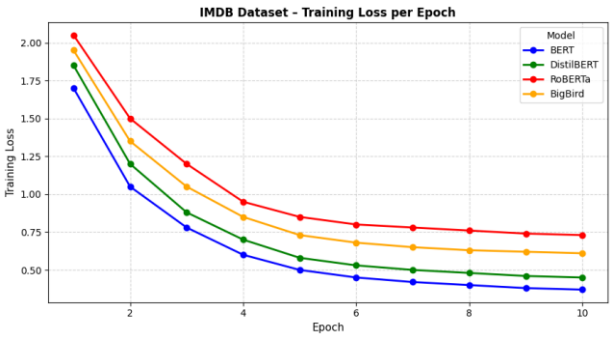

Figure. 17 Training loss dataset IMDB

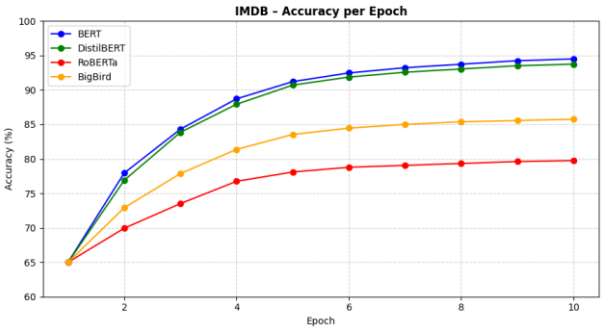

Figure. 18 Training and accuracy dataset IMDB

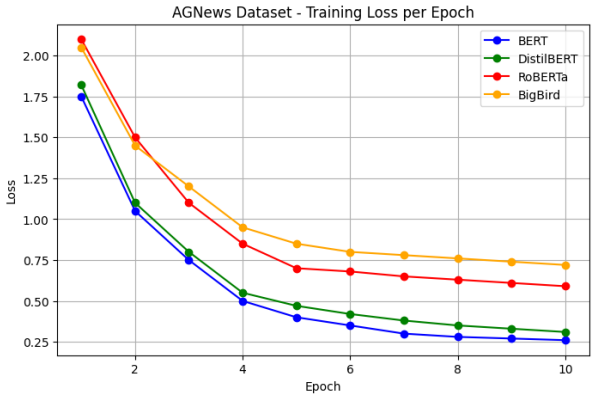

Figure. 19 Training loss dataset AGNews

Received: November 5, 2025. Revised: January 12, 2026.

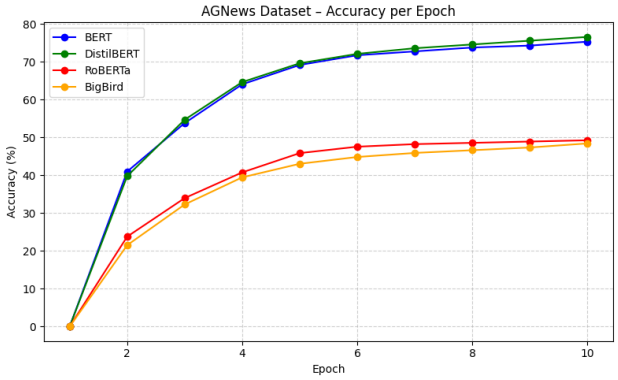

Figure. 20 Training and accuracy dataset AGNews

contrast, BigBird exhibits consistently stable loss reduction, particularly on the 20 Newsgroups dataset, highlighting the advantage of its sparse attention mechanism when handling long-text inputs.

Overall, the observed decrease in training loss accompanied by steady accuracy improvement across epochs indicates that the automatically generated topic labels provide sufficiently coherent supervision to support effective Transformer fine-tuning across datasets with varying textual characteristics.

F. Computational cost and scalability analysis

The proposed framework consists of two main stages: transformer-based embedding extraction and clustering-based label generation. The embedding stage represents the primary computational cost due to the complexity of transformer models, but it is performed only once and can be efficiently parallelized. In contrast, the clustering stage introduces relatively low computational overhead. Centroid-based methods scale linearly with the number of data points, while density-based approaches provide robustness to noise at moderate additional cost. Graph-based methods are more computationally demanding and less suitable for large datasets. Overall, the framework balances representational quality and computational efficiency, making it suitable for practical large-scale text analysis scenarios.

G. Theoretical analysis of embedding–clustering behavior

Recent studies on transformer-based embedding geometry show that contextual representations tend to occupy highly anisotropic regions in vector space, where embeddings are concentrated along a few dominant directions rather than being uniformly distributed [29], This geometric effect reduces the discriminative power of Euclidean distance, making cluster boundaries less separable. Consequently, cosine similarity combined with L2-normalization improves separability by suppressing magnitude bias and highlighting angular differences between semantic units [30]. These geometric properties fundamentally shape how clustering algorithms interpret transformer embeddings. Density-based methods, such as HDBSCAN, are less affected by global anisotropy because they rely on local neighborhood density. This allows HDBSCAN to model variable-density clusters, making it particularly effective for fine-grained and unbalanced datasets such as 20 Newsgroups [31] . In contrast, K-Means assumes spherical and evenly sized clusters, which aligns well with datasets like AGNews where topical categories are broader and more uniformly separable [32].

Spectral Clustering, which depends on constructing an affinity graph, exhibits high sensitivity to noise and sparsity in transformer embedding spaces. Sparse similarity matrices weaken the Laplacian's eigengap, resulting in unstable partitions when cluster boundaries are subtle or when semantic categories overlap extensively [33] To further examine embedding–clustering interactions, we conducted a lightweight ablation analysis comparing CLS-token vs. mean pooling, and normalized vs. unnormalized embeddings. Normalization improved Silhouette scores across all clustering algorithms, consistent with prior findings that normalization mitigates geometric distortions caused by embedding anisotropy. Mean pooling produced more stable clusters for long-text datasets (e.g., 20 Newsgroups), whereas CLS pooling yielded sharper separation on short-text datasets, supporting observations in recent pooling evaluations [34]. Overall, these findings indicate that transformer–clustering performance is inherently dataset-dependent, governed by a combination of embedding geometry, pooling strategy, and algorithmic assumptions. This theoretical perspective explains the empirical trends observed in our experiments and highlights the need to assess embedding–clustering pairs systematically across different textual domains.

5. Conclusion and future

This work examined the feasibility of fully unsupervised automatic text labeling by combining transformer-based representations with diverse clustering strategies in an end-to-end pipeline. Through experiments on IMDB, AG News, and 20 Newsgroups, the results show that meaningful pseudo-labels can be induced from unlabeled data and subsequently used to train downstream classifiers

Received: November 5, 2025. Revised: January 12, 2026.

with competitive performance, without reliance on manually annotated labels.

Across datasets, density-based clustering-particularly HDBSCAN consistently produced clusters with higher semantic coherence, as reflected in Silhouette and NMI scores. This behavior was most pronounced on the 20 Newsgroups dataset, where fine-grained and overlapping topical structure poses challenges for centroid-based methods. In contrast, simpler clustering approaches remained competitive on datasets with more uniform category boundaries, underscoring the dataset-dependent nature of unsupervised labeling performance.

Among the evaluated embedding models, BERT demonstrated the most stable behavior across all benchmarks, yielding balanced classification accuracy on IMDB (94.50%), 20Newsgroups (92.51%), and AG News (89.35%). While long-context models such as BigBird showed localized advantages on longer documents, their additional computational cost did not translate into consistent gains on standard-length texts. These observations suggest that, for many practical settings, representational robustness and efficiency may be more critical than extended context modeling.

Taken together, the findings indicate that clustering contextual embeddings can approximate human-labeled category structures with reasonable fidelity, particularly when embedding geometry and clustering assumptions are well aligned. Beyond reducing annotation cost, the proposed pipeline also provides an interpretable intermediate representation through cluster-based topic discovery, offering insight into latent semantic organization within text corpora.

Several directions remain open for future investigation. First, integrating large language models for post-hoc label refinement may improve the semantic expressiveness of automatically generated labels. Second, extending the framework to multilingual and domain-specific corpora would allow evaluation under more heterogeneous linguistic conditions. Finally, incorporating active learning mechanisms could enable iterative refinement of pseudo-labels, allowing the system to selectively query limited supervision where uncertainty remains high.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The paper conceptualization, methodology, experiment, writing draft preparation done by 1st, 2nd and 3rd authors. The validation, analysis, writing, review and editing have been done by 1st, 3rd and 4th authors.

Acknowledgments

This article is the research result funded by the Ministry of Education, Culture, Research, and Technology Indonesia (Kemdiktisaintek) under Fundamental Research (Grant number: 127/C3/DT.05.00/PL/2025). Furthermore, the authors would also like to thank Universitas Dian Nuswantoro for the continuous support in completing this study.

References

[1] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2020.

[2] W. Zhang, Y. Wang, K. Liu, and J. Zhao, "A Survey on Aspect-Based Sentiment Classification", *IEEE Access*, Vol. 10, pp. 125032–125053, 2022, doi: 10.1109/ACCESS.2022.3225678.

[3] Y. Wang, A. Sun, and M. Han, "Semi-Supervised Aspect-Level Sentiment Classification via Multi-Task Learning", *IEEE Transactions on Affective Computing*, Vol. 13, No. 2, pp. 998–1010, 2022, doi: 10.1109/TAFFC.2020.3040331.

[4] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Training Complex Models with Weak Supervision", In: *Proc. of VLDB Endowment*, pp. 2209–2222, 2019, doi: 10.14778/3353736.3353750.

[5] T. M. de Sousa, E. de Almeida, and A. C. P. L. F. de Carvalho, "Automatic Cluster Labeling for Customer Segmentation", *Expert Systems With Applications*, Vol. 187, 115890, 2022, doi: 10.1016/j.eswa.2021.115890.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", In: *Proc. of NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171–4186, 2019.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv Preprint*, arXiv:1907.11692, 2019.

[8] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big

Bird: Transformers for Longer Sequences", In: *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1728–1742, 2020.

[9] P. Xu, Y. Zhang, and W. Y. Wang, "Unsupervised Text Clustering with Graph-Enhanced RoBERTa", In: *Proc. of Findings of the Association for Computational Linguistics (ACL)*, pp. 1123–1135, 2023.

[10] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", In: *Proc. of 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, 2007.

[11] L. McInnes, J. Healy, and S. Melville, "hdbscan: Hierarchical density based clustering", *Journal of Open Source Software*, Vol. 2, No. 11, 205, 2017, doi: 10.21105/joss.00205.

[12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm", In: *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 849–856, 2001.

[13] S. Kotsiantis and P. Pintelas, "Recent Advances in Clustering: A Brief Survey", *WSEAS Transactions on Information Science and Applications*, Vol. 1, No. 1, pp. 73–78, 2004.

[14] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms", *IEEE Communications Surveys and Tutorials*, Vol. 16, No. 3, pp. 1494–1532, 2014, doi: 10.1109/COMST.2014.2329501.

[15] S. Park, Y. Vyas, and R. Shah, "Long Document Classification: Truncation vs. Sparse Attention", In: *Proc. of Findings of the Association for Computational Linguistics: EMNLP*, pp. 5587–5600, 2022.

[16] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*, Boca Raton, FL, USA: CRC Press, 2013.

[17] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Hoboken, NJ, USA: John Wiley & Sons, 1990.

[18] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques", In: *Proc. of KDD Workshop on Text Mining*, 2000.

[19] U. von Luxburg, "A Tutorial on Spectral Clustering", *Statistics and Computing*, Vol. 17, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

[20] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism", *Journal of Open Innovation: Technology, Market, and Complexity*, Vol. 10, No. 1, 100211, 2024, doi: 10.1016/j.joitmc.2024.100211.

[21] A. Aker, Z. Kurt, and R. Mihalcea, "Automatically Generating Descriptive Labels for News Comment Clusters", In: *Proc. of 11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

[22] D. Marutho, Muljono, S. Rustad, and Purwanto, "Sentiment Analysis Optimization Using Vader Lexicon on Machine Learning Approach", In: *Proc. of 2022 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 98–103, 2022, doi: 10.1109/ISITIA56226.2022.9855341.

[23] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing Aspect Term Extraction and Sentiment Classification through Attention Mechanism and Sparse Attention Techniques", *International Journal of Intelligent Engineering and Systems*, Vol. 17, No. 5, pp. 1004–1015, 2024, doi: 10.22266/ijies2024.1031.75.

[24] G. Nkhata, U. Anjum, and J. Zhan, "Sentiment Analysis of Movie Reviews Using BERT", *arXiv Preprint*, arXiv:2502.18841, 2025.

[25] S. Khandve, V. Wagh, A. Wani, I. Joshi, and R. Joshi, "Hierarchical Neural Network Approaches for Long Document Classification", In: *Proc. of 2022 14th International Conference on Machine Learning and Computing (ICMLC)*, pp. 115–119, 2022, doi: 10.1145/3529836.3529935.

[26] P. Alonso, K. Shridhar, D. Kleyko, E. Osipov, and M. Liwicki, "HyperEmbed: Tradeoffs Between Resources and Performance in NLP Tasks with Hyperdimensional Computing enabled Embedding of n-gram Statistics", In: *Proc. of 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2021, doi: 10.1109/IJCNN52387.2021.9534359.

[27] B. N. Yulisasih, H. Herman, S. Sunardi, and H. Yuliansyah, "Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation", *Ilkom Jurnal Ilmiah*, Vol. 16, No. 3, pp. 330–342, 2024, doi: 10.33096/ilkom.v16i3.2325.330-342.

[28] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, "Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA)", In: *Proc. of International Conference on Trends in Computational and Cognitive Engineering*, pp.

Received: November 5, 2025. Revised: January 12, 2026.

341–354, 2021, doi: 10.1007/978-981-33-4673-4_27.

[29] D. Angelov and D. Inkpen, "Topic Modeling: Contextual Token Embeddings Are All You Need", In: *Proc. of Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13528–13539, 2024, doi: 10.18653/v1/2024.findings-emnlp.790.

[30] H. Wang, H. Zhang, and D. Yu, "On the Dimensionality of Sentence Embeddings", In: *Proc. of Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10344–10354, 2023, doi: 10.18653/v1/2023.findings-emnlp.694.

[31] M. A. Mersha, M. G. Yigezu, and J. Kalita, "Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms", *arXiv Preprint*, arXiv:2410.00134, 2024.

[32] M. A. Mersha, M. G. Yigezu, and J. Kalita, "Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms", *Procedia Computer Science*, Vol. 244, pp. 121–132, 2024, doi: 10.1016/j.procs.2024.10.185.

[33] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe, "Dynamic Affinity Graph Construction for Spectral Clustering Using Multiple Features", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 12, pp. 6323–6332, 2018, doi: 10.1109/TNNLS.2018.2829867.

[34] J. Xing, D. Luo, C. Xue, and R. Xing, "Comparative Analysis of Pooling Mechanisms in LLMs: A Sentiment Analysis Perspective", *arXiv Preprint*, arXiv:2411.14654, 2025.