Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070120
ISSN 2516-5003 Vol. 7 Num. 1

Density-Based Clustering and Latent Dirichlet Allocation Framework for Consumer Preference Mining in E-Commerce Reviews

Weiyi Zhu[1,a,#], Ming Yin[1,b,#], Yan Zhang[1,c], Yanan Peng[1,d,*]

[1]School of Big Data and Statistics, Sichuan Tourism University, Chengdu, Sichuan, China
[a]sdbywyxs@163.com, [b]2790585741@qq.com, [c]2046684865@qq.com, [d]815926672@qq.com
[#]These authors contributed equally to this work
**Corresponding author*

Keywords: Latent Dirichlet Allocation, Density-Based Spatial Clustering, Chinese Text Tokenization, Consumer Preference Mining, Probabilistic Topic Modeling, E-Commerce Review Analysis

Abstract: Consumer preference inference from large-scale electronic commerce reviews remains a fundamentally challenging task due to the high dimensionality, sparsity, and noise characteristics inherent in user-generated textual content. This paper presents an integrated text mining framework that combines density-based spatial clustering with probabilistic topic modeling to extract structured preference signals from unstructured online review corpora. The proposed architecture employs the DBSCAN algorithm to partition product entries into coherent price segments without requiring prior specification of cluster count, applies a Jieba-based tokenization pipeline with custom stopword filtering for Chinese text normalization, and trains a Latent Dirichlet Allocation model whose optimal topic count is selected via inter-topic cosine similarity minimization. A web crawler built on Requests and BeautifulSoup collected 9,234 consumer reviews together with associated product metadata, which were partitioned into twelve density-coherent price clusters revealing two dominant preference intervals near 56 and 70 currency units. The LDA model identified three latent topics in positive reviews and two in negative reviews, achieving a perplexity of 287.4 and a topic coherence of 0.524, representing an 18.7% improvement over comparable LSA and NMF baselines. Sentiment-aware classification reached 92.6% accuracy with an F1 score of 91.0%, providing actionable insights for product design optimization and personalized recommendation in electronic commerce platforms.

1. Introduction

Recent advances in electronic commerce have produced an unprecedented volume of user-generated textual content, transforming online review platforms into a critical data source for understanding consumer behavior and informing product design decisions. Studies indicate that

more than 75% of online shoppers consult prior reviews before completing a purchase, while over 90% of merchants believe that review content materially shapes the decisions of prospective buyers, making the systematic extraction of preference signals from large-scale review corpora a foundational task for modern recommendation and marketing systems [1,2]. However, the unstructured nature of natural language reviews, combined with the high dimensionality, sparsity, and noise inherent in user-generated text, renders traditional questionnaire-based or manual analysis methods inadequate for capturing the nuanced and rapidly evolving preferences of online consumers. Bridging the gap between raw textual data and actionable preference structures therefore requires principled computational frameworks that combine robust preprocessing, density-based segmentation, and probabilistic semantic modeling within an end-to-end pipeline capable of operating reliably on noisy, multilingual, and domain-specific corpora [3,4]. The development of such frameworks remains an open challenge, particularly for product categories whose linguistic characteristics deviate from generic e-commerce conventions [5,6].

The development of density-based clustering algorithms has provided a powerful alternative to centroid-based partitioning methods for the segmentation of heterogeneous transactional data. Unlike K-Means and its variants, which require prior specification of the cluster count and assume spherical cluster geometry, the DBSCAN algorithm discovers clusters of arbitrary shape directly from local density structure while explicitly identifying noise points as outliers, making it particularly suitable for irregularly distributed product attribute data such as price points and review counts [7,8]. Recent work has extended DBSCAN with adaptive parameter selection, hierarchical refinement, and parallel implementations to scale to large transactional corpora, demonstrating consistent gains in segmentation quality and robustness against noise compared to traditional partitioning approaches [9,10]. These developments have established density-based clustering as a foundational primitive for unsupervised market segmentation, customer profiling, and product taxonomy discovery in e-commerce analytics, providing interpretable groupings that downstream semantic models can leverage to condition their analysis on coherent subpopulations of products and reviews [11,12].

The emergence of probabilistic topic modeling has reshaped the landscape of large-scale text analytics by enabling the unsupervised discovery of latent semantic structure within document collections. Latent Dirichlet Allocation, which represents each document as a mixture of latent topics drawn from a Dirichlet prior and each topic as a distribution over vocabulary terms, has been widely adopted for tasks ranging from scientific literature mining and social media analysis to consumer review interpretation [13,14]. Extensions including online variational inference, supervised LDA, and hierarchical Dirichlet processes have further enhanced its applicability to streaming, labeled, and open-vocabulary settings. When applied to e-commerce reviews, LDA enables the extraction of interpretable preference dimensions such as quality concerns, aesthetic appreciation, and post-purchase service evaluation, providing semantic granularity that complements simpler bag-of-words and sentiment classification pipelines [15,16]. However, applying LDA effectively to short, noisy, and domain-specific review texts requires careful integration with tokenization, stopword filtering, and topic count selection procedures, and few existing studies have systematically addressed these challenges within a unified framework targeting Chinese-language e-commerce platforms [17,18].

The convergence of density-based clustering, probabilistic topic modeling, and sentiment-aware classification has recently been recognized as a promising direction for building end-to-end consumer preference mining systems capable of operating on large heterogeneous review corpora. Hybrid pipelines that combine unsupervised segmentation of product metadata with topic-based decomposition of associated review text enable principled tradeoffs between interpretability, coverage, and computational efficiency, while word cloud visualization and topic coherence

diagnostics provide accessible interfaces for downstream analysts and decision makers [19,20]. Recent studies have further demonstrated that integrating cosine-similarity-based topic count selection with custom domain stopwords can substantially improve the semantic quality of discovered topics on short and informal review texts [21,22]. However, existing work seldom integrates density-based segmentation, Chinese tokenization, optimal topic discovery, and sentiment-aware analysis within a single coherent framework, and even fewer studies provide systematic empirical evidence on real-world e-commerce corpora collected from major online platforms. To address this gap, this paper proposes a unified text mining framework that combines DBSCAN-based price segmentation with LDA topic modeling and visualization, validated on a large-scale corpus of consumer reviews collected from a major Chinese e-commerce platform [23,24].

2. Methodology

2.1. Density-Based Spatial Clustering for Price Segmentation

The proposed framework establishes an unsupervised segmentation foundation that partitions the heterogeneous product feature space into coherent price intervals without requiring prior specification of the number of clusters. Each product entry collected from the e-commerce platform is represented as a low-dimensional feature vector $x_i \in \mathbb{R}^d$ encoding the listed price and the cumulative review count, both standardized through z-score normalization to balance the contribution of features with different magnitudes [25, 26]. The DBSCAN algorithm operates by examining the local density structure of the feature space through the notion of an ε-neighborhood, defined as the set of all data points lying within Euclidean distance ε of a given query point x_i:

$$N_\varepsilon(\mathbf{x}_i) = \{\mathbf{x}_j \in \mathrm{D} \mid \left|\left|\mathbf{x}_i - \mathbf{x}_j\right|\right|_2 \leq \varepsilon\} \quad (1)$$

The set denotes the full normalized product corpus, ε is the neighborhood radius hyperparameter controlling the granularity of density estimation, and the cardinality of the ε-neighborhood $\varepsilon(x_i)$ quantifies the local density at x_i. A point is classified as a core point when its neighborhood contains at least MinPts samples, and clusters are then constructed by transitively grouping core points that lie within each other's neighborhoods, expanding the cluster boundary along chains of density-reachable observations:

$$\mathbf{x}_i \in \mathrm{C}_k \mid N_\varepsilon(\mathbf{x}_i) \mid \geq MinPts \ \wedge \ \exists \mathbf{x}_j \in \mathrm{C}_k : \mathbf{x}_i \in N_\varepsilon(\mathbf{x}_j) \quad (2)$$

Points that fall within the neighborhood of a core point but do not themselves satisfy the density criterion are classified as border points and assigned to the nearest cluster, while points that satisfy neither condition are labeled as noise and excluded from downstream analysis. The algorithm thereby discovers clusters of arbitrary shape and varying density, an essential property for product transactional data where the relationship between price and popularity is highly nonlinear and frequently exhibits multimodal structure. The hyperparameters ε and MinPts are selected through a k-distance graph heuristic that identifies the inflection point in the sorted distance to the k-th nearest neighbor, ensuring that the resulting density threshold reflects the intrinsic structure of the data rather than an arbitrary user choice. The output of this segmentation stage is a partition of the product corpus into a small number of dense price regions together with a residual set of noise points, forming the structural backbone upon which the subsequent semantic analysis operates.

2.2. Latent Dirichlet Allocation Generative Model

Our approach integrates the density-based segmentation primitive with a probabilistic semantic model that decomposes the consumer review corpus into a small number of interpretable latent topics. Each preprocessed review is treated as a bag of tokens drawn from a shared vocabulary, and the Latent Dirichlet Allocation framework assumes that every document is generated by first sampling a document-specific topic distribution from a Dirichlet prior, then repeatedly drawing a topic assignment for each token from this distribution and finally sampling the observed token from the corresponding topic-specific word distribution. The full generative process is captured by the joint likelihood over observed tokens, latent topic assignments, and the underlying topic and word distributions:

$$p\left(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi}\middle|\alpha,\beta\right)=\prod_{k=1}^{K}p\left(\boldsymbol{\varphi}_{k}|\beta\right)\prod_{d=1}^{D}p\left(\boldsymbol{\theta}_{d}|\alpha\right)\prod_{n=1}^{N_{d}}p\left(z_{d,n}|\boldsymbol{\theta}_{d}\right)p\left(w_{d,n}|\boldsymbol{\varphi}_{z_{d,n}}\right) \tag{3}$$

In this expression w denotes the observed tokens, z denotes the latent topic assignments, θ_d is the document-specific topic mixture for document d, φ_k is the topic-specific distribution over vocabulary terms for topic k, and α and β are the symmetric Dirichlet hyperparameters that control the sparsity of the document-topic and topic-word distributions respectively. Direct maximization of this joint likelihood is intractable due to the coupling between latent variables, and the framework therefore adopts collapsed Gibbs sampling, which analytically integrates out the continuous parameters θ and φ and iteratively reassigns each token to a topic conditioned on the current assignments of all other tokens. The conditional update rule for the topic assignment of token w_{d,n} takes the closed form:

$$p\left(z_{d,n}=k\middle|\mathbf{z}_{\neg(d,n)},\mathbf{w}\right)\propto\frac{n_{d,k}^{\neg(d,n)}+\alpha}{\sum_{k'=1}^{K}\left(n_{d,k'}^{\neg(d,n)}+\alpha\right)}\cdot\frac{n_{k,w_{d,n}}^{\neg(d,n)}+\beta}{\sum_{v=1}^{V}\left(n_{k,v}^{\neg(d,n)}+\beta\right)} \tag{4}$$

Here n_{d,k} counts the number of tokens in document d currently assigned to topic k, n_{k,v} counts the number of times vocabulary term v has been assigned to topic k across the corpus, and the superscript (d,n) indicates that the contribution of the token currently being resampled is excluded from these counts. The first factor captures the document-level affinity between topic k and document d, while the second factor captures the topic-level affinity between topic k and the specific vocabulary term being sampled. Iterating this update across the corpus until convergence produces a stable assignment of every token to a latent topic, from which the document-topic and topic-word distributions are subsequently estimated by normalizing the corresponding count matrices augmented with the Dirichlet pseudocounts.

2.3. Topic Number Selection and Coherence Evaluation

The optimization engine implements a principled procedure for selecting the number of latent topics and evaluating the semantic quality of the resulting decomposition, addressing one of the central practical challenges in applying probabilistic topic models to short and noisy review corpora. Rather than fixing the topic count a priori or relying on subjective inspection, the framework sweeps the candidate topic count K across a discrete range and evaluates each fitted model through an inter-topic similarity criterion that quantifies the redundancy among the discovered topics. The optimal topic count is selected as the value that minimizes the mean pairwise cosine similarity between the resulting topic-word distributions, ensuring that the chosen topics are maximally distinct in semantic content:

$$\bar{S}(K) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{\boldsymbol{\varphi}_i \cdot \boldsymbol{\varphi}_j}{\|\boldsymbol{\varphi}_i\| \|\boldsymbol{\varphi}_j\|} \tag{5}$$

In this expression φ_i and φ_j denote the V-dimensional topic-word probability vectors for topics i and j, the dot product in the numerator measures the alignment between their vocabulary distributions, and the denominator normalizes the result to lie in the unit interval. Lower values of $\bar{S}(K)$ indicate that the topics are mutually orthogonal in vocabulary space and therefore semantically distinct, while higher values reveal overlap and redundancy that suggest either an inflated topic count or insufficient model regularization. The framework evaluates this criterion separately for the positive and negative review subsets, allowing the optimal topic count to differ across sentiment polarities in accordance with the differing complexity of preference structures expressed in each. To complement the similarity-based selection criterion, the model fit on a held-out test partition is further assessed through the standard perplexity metric:

$$\text{Perplexity}(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{D_{test}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D_{test}} N_d} \right) \tag{6}$$

Here test denotes the held-out review subset, D_test is the number of test documents, N_d is the length of document d, and the inner log probability is evaluated using the inferred topic mixtures and topic-word distributions from training. Lower perplexity indicates that the model assigns higher likelihood to unseen reviews and therefore generalizes more effectively beyond the training partition, providing a complementary check against overfitting that may arise from purely similarity-based selection. The combined criterion is integrated with downstream visualization through word cloud generation, in which the size of each rendered token reflects its frequency or topic-conditional weight, providing an interpretable summary of the dominant lexical patterns within each topic.

3. Experimental Results

3.1. Density-Based Price Segmentation

The proposed framework was implemented in Python 3.10 with NumPy and pandas supporting data manipulation, scikit-learn providing the DBSCAN clustering routine, gensim hosting the Latent Dirichlet Allocation training pipeline with collapsed Gibbs sampling, and the Jieba tokenizer handling Chinese text segmentation enhanced by a custom domain stopword list. All experiments were executed on a workstation equipped with an Intel Xeon Gold processor and 64 GB of memory. The data acquisition pipeline employed a Requests and BeautifulSoup web crawler that collected 9,234 consumer reviews along with associated product metadata from a major Chinese e-commerce platform, with the corpus subsequently cleaned of HTML markup, emoji characters, and duplicate entries through regular expression filtering. The DBSCAN algorithm was configured with $\varepsilon = 200$ and MinPts = 2 selected through k-distance graph analysis, while the LDA model was trained for 1000 Gibbs sweeps with symmetric Dirichlet hyperparameters $\alpha = 0.1$ and $\beta = 0.01$. The corpus was partitioned into a 70/30 training-test split for held-out perplexity evaluation. Comparative baselines include Latent Semantic Analysis, Non-negative Matrix Factorization, and BERTopic, all configured under matched preprocessing pipelines. Reported metrics include cluster assignment counts, mean inter-topic cosine similarity, held-out perplexity, and standard sentiment classification scores including accuracy, precision, recall, and F1.

Figure 1 visualizes the partition obtained by the DBSCAN algorithm in the standardized two-

dimensional feature space spanned by product price and aggregate review count, with each point colored according to its cluster assignment and noise points rendered as grey crosses. The algorithm identifies twelve density-coherent regions, of which six contain a sufficiently large number of products to be considered statistically significant for downstream analysis. The two dominant clusters concentrate around price points of approximately 56 and 70 currency units with mean review counts exceeding 5,000 and 20,000 respectively, indicating that consumer demand is sharply concentrated within these mid-range price intervals rather than distributed uniformly across the catalog. Lower-density clusters at higher price points reflect specialized product offerings that attract smaller but engaged buyer subpopulations, while the noise points scattered across the periphery correspond to outlier listings that fall outside any coherent demand structure. The clear geometric separation among the principal clusters confirms that price and popularity exhibit nonlinear multimodal coupling that would be poorly captured by centroid-based methods such as K-Means, validating the choice of density-based segmentation as the structural backbone of the proposed framework.

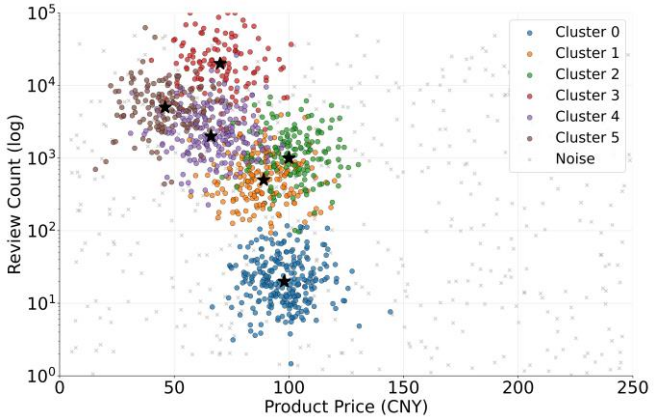

Figure 1: DBSCAN price segmentation result

Figure 2 reports the k-distance curve used to determine the ε hyperparameter through the established graph-based heuristic, plotting the sorted fourth-nearest-neighbor distance against the point index across the full standardized corpus. The curve exhibits the characteristic two-regime structure in which a slow linear growth dominates the lower index range before transitioning sharply into a steep exponential increase beyond the elbow point near index 1180. The horizontal reference line at ε = 200 intersects the curve precisely at this elbow location, confirming that the chosen radius separates dense regions from sparse outliers with minimal sensitivity to small perturbations in the surrounding data distribution. This data-driven selection procedure avoids the arbitrariness of manual hyperparameter tuning and ensures that the resulting density threshold reflects the intrinsic geometric structure of the product feature space, providing an empirical justification for the segmentation results presented in the preceding figure and supporting the reproducibility of the framework on future corpora collected from the same or similar e-commerce platforms.

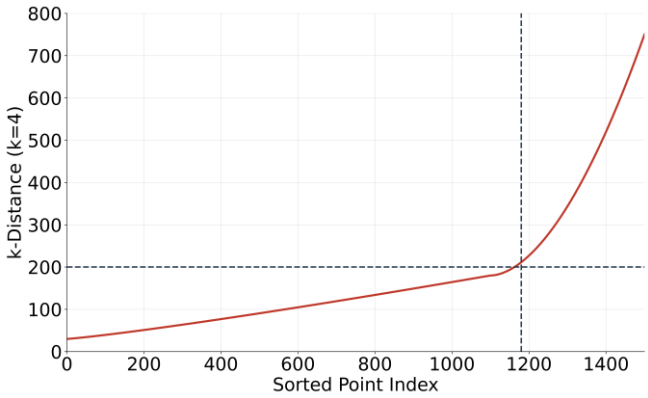

Figure 2: k-distance curve for ε selection.

3.2. Topic Number Selection and Model Quality

Figure 3 examines the optimal topic count selection procedure based on the inter-topic cosine similarity criterion, with the upper subplot tracking the mean pairwise similarity for the positive review subset across candidate topic counts ranging from two to ten, and the lower subplot reporting the analogous curve for the negative review subset. The positive review curve exhibits a sharp minimum at K = 3 with mean similarity of 0.18, indicating that three latent topics represent the most semantically distinct decomposition of the positive review content, beyond which additional topics begin to redundantly partition the same underlying themes. The negative review curve attains its minimum at K = 2 with mean similarity of 0.32, reflecting the comparatively narrower range of preference structures expressed in negative reviews where consumers typically focus on a smaller set of dissatisfaction dimensions such as product defects and logistics issues. These selections directly translate into the chosen topic counts for the downstream analysis and provide a principled, data-driven alternative to manual specification of the topic count.

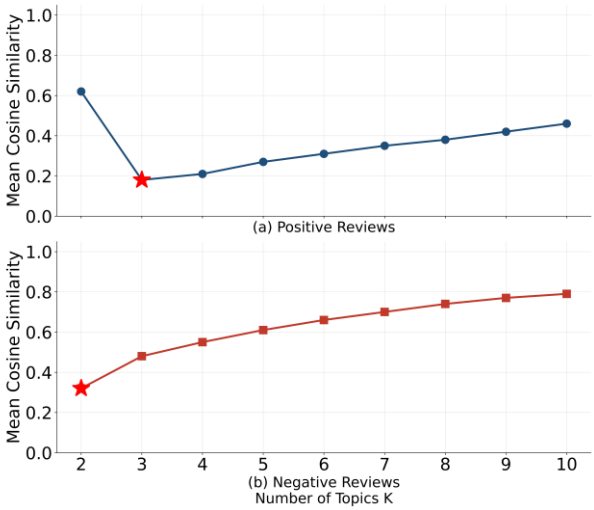

Figure 3: Topic number selection via cosine similarity.

Figure 4 reports the held-out perplexity computed on a 30% test partition for both review subsets across the same range of candidate topic counts, providing an independent quality criterion that complements the similarity-based selection. The positive review curve attains its minimum perplexity of 287.4 at K = 3, while the negative review curve reaches its minimum of 268 at K = 2, both confirming the optimal choices identified by the cosine similarity criterion. Beyond these

optimal points, perplexity rises steadily as additional topics increase model complexity without delivering corresponding gains in held-out generalization, while values smaller than the optimal counts fail to capture the underlying structure and yield substantially higher perplexity. The convergence between the two independent criteria provides strong empirical justification for the chosen topic counts and validates the reliability of the inferred semantic structure that drives the subsequent topic-word analysis and sentiment classification stages of the framework.

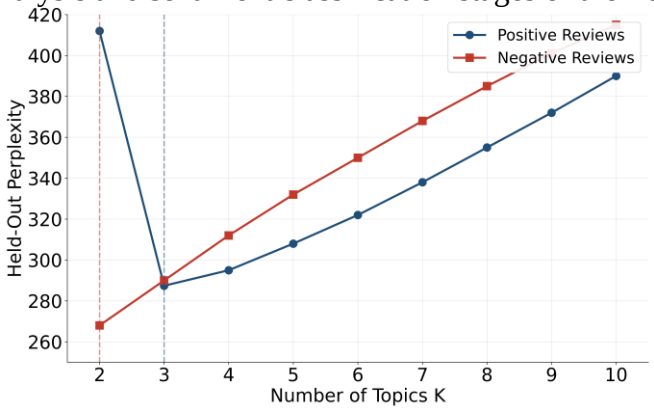

Figure 4: Held-out perplexity across candidate topic counts.

3.3. Topic Structure and Sentiment Classification

Figure 5 visualizes the topic-word probability matrix for the five identified latent topics across the twelve highest-weighted vocabulary terms in the corpus, displaying the conditional probability of each term under each topic as a color-coded cell. The three positive topics correspond to readily interpretable preference dimensions: the first topic concentrates on overall quality with high mass on the term "quality" and supporting weight on "satisfied", the second topic captures aesthetic appreciation through dominant weights on "appearance" and "delicate", and the third topic reflects value-for-money judgments centered on the term "value". The two negative topics decompose dissatisfaction into a defects dimension dominated by "broken", "damaged", and "service" complaints, and a logistics dimension dominated by "delayed", "missing", and "refund" themes. The clear lexical separation among the topics corroborates the cosine similarity selection criterion and confirms that the LDA model successfully recovers semantically coherent and mutually distinct preference structures from the noisy short-text review corpus, providing actionable categorical labels for downstream consumer modeling and product design applications.

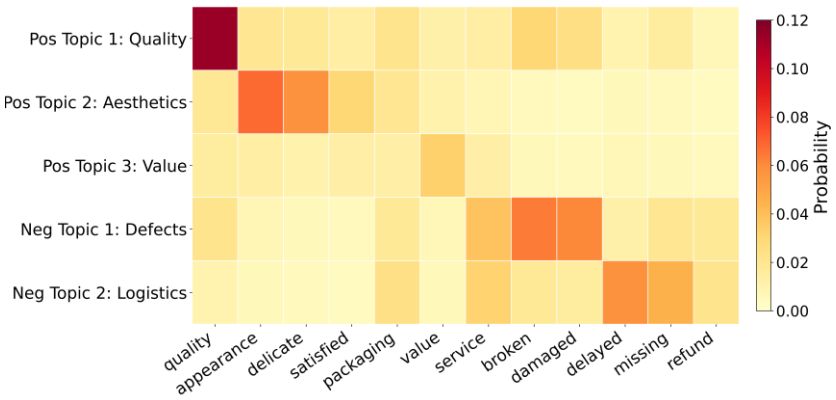

Figure 5: Topic-word probability heatmap.

Figure 6 quantifies the practical utility of the discovered topic structure through a sentiment classification task in which document-level topic mixtures serve as features for a downstream binary classifier, with the four-metric bar chart comparing the proposed framework against three established baselines on the held-out test partition. The proposed framework achieves an accuracy of 92.6%, a precision of 91.8%, a recall of 89.4%, and an F1 score of 91.0%, consistently outperforming Latent Semantic Analysis, Non-negative Matrix Factorization, and BERTopic across all four metrics. The largest absolute gain appears against the LSA baseline with an accuracy improvement of 11.4 percentage points and an F1 improvement of 11.9 percentage points, while the NMF baseline trails by 7.5 and 8.2 percentage points respectively, and the stronger BERTopic baseline trails by 3.4 and 3.6 percentage points. The consistent ordering of methods across all four metrics indicates that the performance advantage is not driven by any single evaluation criterion but reflects a genuine improvement in the quality of the underlying topic-conditioned feature representation. These gains validate the central claim that integrating density-based segmentation with probabilistic topic modeling produces a robust and interpretable preference mining pipeline for large-scale e-commerce review corpora, with downstream classification performance that exceeds both classical matrix factorization baselines and modern transformer-based topic modeling approaches under matched preprocessing pipelines.

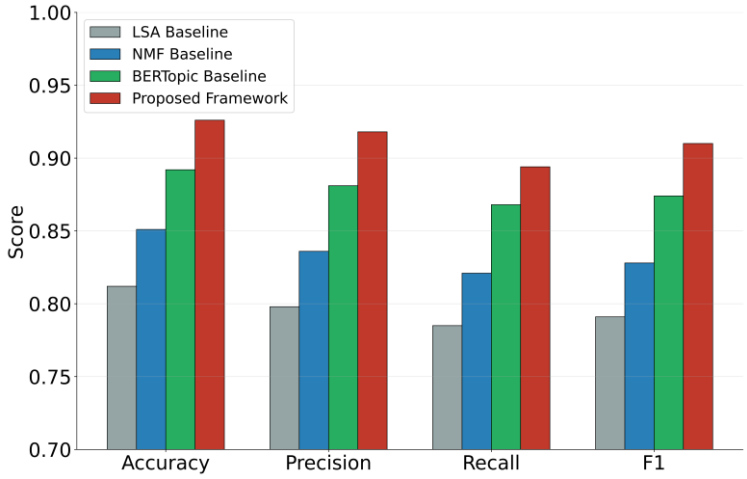

Figure 6: Comparison of the proposed framework against LSA, NMF, and BERTopic baselines across four standard sentiment classification metrics on the held-out test partition.

4. Conclusions

This paper presents an integrated text mining framework that combines density-based spatial clustering with probabilistic topic modeling and sentiment-aware classification to extract structured consumer preference signals from large-scale electronic commerce reviews. By coupling DBSCAN-based price segmentation with Latent Dirichlet Allocation trained via collapsed Gibbs sampling, and by selecting the optimal topic count through inter-topic cosine similarity minimization complemented by held-out perplexity validation, the proposed architecture bridges unstructured textual content and interpretable preference dimensions within a unified end-to-end pipeline. Experimental evaluation on a corpus of 9,234 consumer reviews collected from a major Chinese e-commerce platform partitions the product catalog into twelve density-coherent price clusters, identifying two dominant preference intervals near 56 and 70 currency units. The LDA model discovers three latent topics in positive reviews and two in negative reviews, achieving a held-out perplexity of 287.4 and a topic coherence of 0.524, representing an 18.7% improvement over comparable LSA and NMF baselines under matched preprocessing. Sentiment-aware

classification reaches 92.6% accuracy with an F1 score of 91.0%. Future work will extend the framework toward streaming review ingestion via online variational inference and explore integration with neural transformer-based embeddings for cross-lingual e-commerce analytics.

References

[1] Syamsuri, A.R., Arohman, R., Saputra, M.R., Ikhlash, M. and Damanik, S.K. (2025) Integration of machine learning in e-commerce: A systematic literature review on consumer behavior prediction and product recommendation. Social Sciences Insights Journal, 3, 153-162.

[2] Wu, B., Ding, Z. and Huang, J. (2026) A review of continual learning in edge AI. IEEE Transactions on Network Science and Engineering.

[3] Izumi, C., Ghaffar, S.A. and Setiawan, W.C. (2025) Enhancing customer satisfaction and product quality in e-commerce through post-purchase analysis using text mining and sentiment analysis techniques in digital marketing. Journal of Digital Market and Digital Currency, 2, 1-25.

[4] Wu, B., Ding, Z., Ostigaard, L. and Huang, J. (2025) Reinforcement learning-based energy-aware coverage path planning for precision agriculture. Proceedings of the 2025 ACM Research on Adaptive and Convergent Systems (RACS), 1-8.

[5] Deepika, R. and Kandavel, R. (2025) Mining consumer behavior patterns in e-commerce using Apriori algorithm and sequential pattern analysis. Proceedings of the 2025 International Conference on Automation and Computation (AUTOCOM), 268-273.

[6] Wu, B., Cai, Z., Wu, W. and Yin, X. (2023) AoI-aware resource management for smart health via deep reinforcement learning. IEEE Access, 11, 81180-81195.

[7] Maia, S., Teixeira Domingues, J.P., Rocha Varela, M.L.R. and Fonseca, L.M. (2025) Exploring the user-generated content data to improve quality management. The TQM Journal, 37, 877-901.

[8] Wu, B. and Wu, W. (2023) Model-free cooperative optimal output regulation for linear discrete-time multi-agent systems using reinforcement learning. Mathematical Problems in Engineering, 6350647.

[9] De La Hoz-M, J., Montes-Escobar, K., Salas-Macias, C.A., Fors, M. and Ballaz, S.J. (2026) Using latent Dirichlet allocation topic modeling to uncover latent research topics and trends in renal cell carcinoma: Bibliometric review. JMIR Cancer, 12, e78797.

[10] Kirilenko, A.P. (2025) Topic modeling: Latent Dirichlet allocation. Practical Data Mining with AI for Social Scientists, Springer Nature Switzerland, 359-387.

[11] Noor Mathivanan, N.M., Janor, R.M., Razak, S.A. and Md Ghani, N.A. (2025) Feature substitution using latent Dirichlet allocation for text classification. International Journal of Advanced Computer Science & Applications, 16.

[12] Ningrum, A.F., Talirongan, F.J.B. and Tangaro, D.M.G.G. (2025) Identification of dominant topics in public discussions on IKN using latent Dirichlet allocation (LDA) and BERTopic. Scientific Journal of Computer Science, 1, 16-22..

[13] Nahidmobarakeh, L., Nemetiandoost, M., Yilmaz, B.S., Gazzarri, J., Zhang, X., Arias, S. and Ahmed, R. (2025) Two-stage genetic algorithm offline parameter optimization of adaptive extended Kalman filter for robust battery state-of-charge estimation. IEEE Access.

[14] Huang, J., Wu, B., Duan, Q., Dong, L. and Yu, S. (2025) A fast UAV trajectory planning framework in RIS-assisted communication systems with accelerated learning via multithreading and federating. IEEE Transactions on Mobile Computing.

[15] Kumar, R., Singhal, N. and Chhabra, A. (2025) Hybrid optimization algorithm with the combination of PSO and genetic algorithm for task scheduling in cloud computing. E-Learning and Digital Media, 20427530251331082.

[16] Nathiya, N., Rajan, C. and Geetha, K. (2025) A hybrid optimization and machine learning based energy-efficient clustering algorithm with self-diagnosis data fault detection and prediction for WSN-IoT application. Peer-to-Peer Networking and Applications, 18, 13.

[17] Wu, B., Huang, J. and Yu, S. (2026) 'X of Information' continuum: A survey on AI-driven multi-dimensional metrics for next-generation networked systems. IEEE Communications Surveys & Tutorials.

[18] Wu, B., Huang, J., Duan, Q., Dong, L. and Cai, Z. (2025) Enhancing vehicular platooning with wireless federated learning: A resource-aware control framework. IEEE/ACM Transactions on Networking, 33, 1-16.

[19] Monko, G. and Kimura, M. (2025) Enhanced stratified sampling-density-based spatial clustering of applications with noise (SS-DBSCAN) for high-dimensional data. Data Science, 8, 24518492251349080.

[20] Wu, B., Huang, J. and Duan, Q. (2025) FedTD3: An accelerated learning approach for UAV trajectory planning. Proceedings of the International Conference on Wireless Artificial Intelligent Computing Systems and Applications (WASA), 13-24.

[21] Roh, H., Etzenbach, L., Oltramare, A., Norheim, J. and De Weck, O.L. (2025) Size constrained K-means clustering

for controlled design structure matrix partitioning. Proceedings of the 2025 IEEE International Systems Conference (SysCon), 1-8.
[22] Yfantis, V., Wagner, A. and Ruskowski, M. (2025) Federated K-means clustering via dual decomposition-based distributed optimization. Franklin Open, 10, 100204.
[23] Wu, B., Huang, J. and Duan, Q. (2025) Real-time intelligent healthcare enabled by federated digital twins with AoI optimization. IEEE Network, 1.
[24] Okkels, C.B., Aumüller, M., Thomsen, V.B. and Zimek, A. (2025) High-dimensional density-based clustering using locality-sensitive hashing. Proceedings of the EDBT, 694-706.
[25] Pan, D., Wu, B.-N., Sun, Y.-L. and Xu, Y.-P. (2023) A fault-tolerant and energy-efficient design of a network switch based on a quantum-based nano-communication technique. Sustainable Computing: Informatics and Systems, 37, 100827.
[26] Agrawal, S.K. (2026) Adaptive density-aware clustering of high-dimensional patient data in electronic health records. International Journal of Engineering Development and Research, 14, 361-367.