# FROM TF-IDF TO TRANSFORMER-BASED MODELS: CLUSTERING METHODS FOR E-COMMERCE PRODUCT ORGANIZATION

**Le Nhat Tung[1] Bui Minh Huy[2]**
**Tran Le Van[3] Nguyen Thi Thanh Tam[4] Nguyen Thi Lieu[5]**

[1, 5]Dong Nai Technology University; [2, 3, 4]Ho Chi Minh City University of Technology
Email: *lenhattung@dntu.edu.vn[1]; huybm.ds@gmail.com[2]; tranlevan07.ds@gmail.com[3]; ttam99852@gmail.com[4]; nguyenthilieu@dntu.edu.vn[5]*
Received: 3/4/2026; Reviewed: 20/4/2026; Revised: 25/4/2026; Accepted: 29/4/2026
**DOI:** *https://doi.org/10.58902/nckhpt.e-v2i1.366*

**Abstract**: *Efficient product catalog management is crucial for e-commerce operational efficiency, particularly in multilingual markets like Vietnam where descriptions combine local text with English terms. This study systematically evaluates automated clustering approaches by comparing nine combinations of text representation methods (TF-IDF, PhoBERT, E5-multilingual) and clustering algorithms (K-Means, DBSCAN, GMM) using 36,644 product descriptions from Lazada Vietnam. Results demonstrate that E5-multilingual embeddings combined with GMM achieve superior performance, reaching 91.28% Normalized Mutual Information (NMI) and 79.22% Purity, significantly outperforming traditional and monolingual models. Critically, the analysis reveals that high technical metrics do not always align with business value, as one method achieved a 0.986 Silhouette score but only 0.003 Adjusted Rand Index (ARI) alignment with actual categories. These findings provide empirical guidance for platforms to automate catalog organization, reducing manual workload while maintaining accuracy in code-mixed environments.*

**Keywords**: *TF-IDF; Clustering algorithms; E-commerce; Transformer.*

## 1. Introduction

The rapid expansion of e-commerce has fundamentally transformed retail operations, creating unprecedented challenges in product catalog management. Major platforms such as Lazada, Shopee, and Amazon must efficiently organize millions of products across diverse categories, where a single marketplace may host hundreds of thousands of sellers with varying description formats and quality levels. Traditional approaches rely heavily on manual categorization, where sellers select categories from predefined taxonomies during product listing. However, this method suffers from critical limitations: it is labor-intensive and difficult to scale, prone to errors as sellers frequently misclassify products due to unfamiliarity with category structures, and struggles to adapt to the dynamic nature of e-commerce where new products and categories emerge constantly. These challenges are particularly acute in Southeast Asian markets, where product descriptions often contain mixed-language content, combining local languages with English technical terms and brand names, making effective organization even more complex. The economic implications of inefficient catalog management extend beyond operational costs. Miscategorized products reduce customer discovery rates and conversion, while manual processes create bottlenecks that limit platform scalability. For rapidly growing markets, the inability to efficiently organize expanding product catalogs directly constrains business growth and competitive positioning.

While recent advances in transformer-based language models have shown remarkable success in various natural language processing tasks (Dat & Anh, 2020) (Wang et al., 2024), their application to unsupervised product clustering in multilingual e-commerce environments remains underexplored. Existing studies predominantly focus on monolingual scenarios or supervised classification tasks, leaving a critical gap in understanding how different text representation methods—from traditional statistical approaches to modern neural embeddings—perform when combined with various clustering algorithms for

real-world catalog organization. Moreover, there is limited empirical evidence comparing the effectiveness of these methods specifically for markets with code-mixed product descriptions, where bilingual content poses unique challenges for semantic understanding.

This study addresses these gaps by pursuing three primary research objectives: First, to systematically evaluate and compare text representation methods ranging from traditional statistical approaches (TF-IDF with Latent Semantic Analysis) to modern transformer-based models (PhoBERT and E5-multilingual) for encoding product descriptions. Second, to assess the performance of three widely-used clustering algorithms K-Means, DBSCAN, and Gaussian Mixture Models—across different semantic embedding spaces. Third, to identify optimal combinations of representation methods and clustering techniques that maximize both technical clustering quality and alignment with actual product categories.

**2. Research overview**

E-commerce platforms face significant challenges in organizing vast product catalogs, where millions of items must be efficiently categorized to enable effective search and discovery. Major platforms such as Amazon and eBay have implemented automated product clustering systems to group similar items, detect duplicate listings, and improve recommendation quality (McAuley et al., 2015). Recent work has explored deep learning approaches for product matching (Yulianton & Santi, 2024) and large-scale categorization (Chang et al., 2020), demonstrating improvements over traditional rule-based systems. However, existing approaches predominantly rely on structured product attributes (brand, category, price) combined with conventional text-based features extracted from product titles and descriptions. Such methods struggle with several limitations: they require extensive feature engineering, perform poorly when structured attributes are missing or inconsistent, and fail to capture semantic similarity between products described using different terminology. The multilingual and code-mixed nature of product descriptions in Southeast Asian markets remains a largely unaddressed challenge in existing literature.

The evolution of text representation has profoundly shaped natural language processing capabilities. Traditional approaches such as Term Frequency-Inverse Document Frequency represent documents as sparse vectors based on word statistics (Salton & Buckley, 1988), offering computational efficiency but failing to capture semantic relationships between terms. Neural word embeddings introduced dense vector representations that encode semantic similarity, enabling models to recognize conceptual connections between lexically distinct words. Transformer-based models revolutionized the field by introducing contextualized representations where word meanings adapt based on surrounding context. For Vietnamese language processing, PhoBERT demonstrated the effectiveness of language-specific pre-training on large-scale corpora (Dat & Anh, 2020). Beyond standard benchmark tasks, this model has shown remarkable adaptability in capturing complex linguistic nuances, such as sentiment polarity and emotional cues within informal text containing emojis (Dung et al., 2024). This versatility in handling domain-specific Vietnamese content justifies its evaluation as a representative monolingual transformer for organizing product descriptions, which often share similar informal and mixed-language characteristics. Multilingual models such as E5 enable cross-lingual semantic comparison by mapping text from multiple languages into unified embedding spaces (Wang et al., 2024). Despite these advances, most research focuses on supervised classification tasks, with limited exploration of how these representation methods perform in unsupervised clustering scenarios for domain-specific applications. Comprehensive empirical comparisons between traditional statistical and modern neural approaches on real-world multilingual product data remain scarce in the literature.

Unsupervised clustering has been widely applied in natural language processing for document organization and topic discovery. K-Means remains popular for its computational efficiency but requires predefined cluster numbers and assumes spherical shapes (MacQueen, 1967). DBSCAN offers density-based clustering with automatic outlier detection

(Ester et al., 1996) while Gaussian Mixture Models provide probabilistic soft assignments (*Gaussian Mixture Model*, 2025). While these algorithms are well-established in machine learning, their comparative performance for text clustering using modern neural embeddings remains underexplored. Most evaluations focus on traditional feature spaces, with limited investigation of how algorithm choice interacts with representation quality in e-commerce applications.

While existing literature has made progress in e-commerce organization, neural text representations, and clustering algorithms, a comprehensive evaluation integrating these components for multilingual product clustering remains absent. Prior work predominantly examines representation methods or algorithms in isolation, focusing on monolingual datasets or supervised tasks. This study addresses these gaps by providing the first systematic comparison of traditional and transformer-based embeddings combined with multiple clustering algorithms on Vietnamese e-commerce data. By evaluating both unsupervised clustering quality and supervised alignment with ground-truth categories, we offer practical insights for implementing automated catalog management systems in multilingual e-commerce environments.

## 3. Research methods

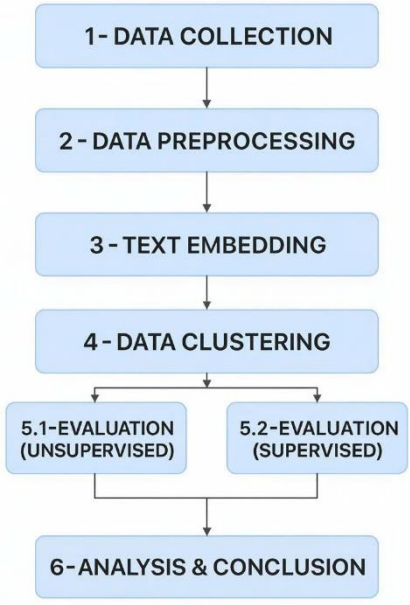
### *3.1. Data Collection and research framework*

This study collected product data from Lazada Vietnam, one of Southeast Asia's largest e-commerce platforms, using Selenium-based web scraping to automate data extraction. The dataset comprises approximately 37,000 product descriptions across 14 distinct categories including electronics, fashion, home appliances, beauty products, and sports equipment. Each product entry contains Vietnamese-English code-mixed descriptions reflecting real-world multilingual e-commerce content. Our research framework follows a systematic pipeline: collected product descriptions undergo text preprocessing (lowercasing, punctuation removal, whitespace normalization), followed by encoding using three representation methods to generate numerical vectors. These vectors are then clustered using three different algorithms, producing nine experimental combinations. Clustering quality is evaluated through both unsupervised metrics that assess internal cluster structure and supervised metrics that measure alignment with ground-truth product categories, enabling comprehensive comparison of traditional statistical and modern transformer-based approaches for multilingual product organization.

### *3.2. Text representation methods*

TF-IDF with Latent Semantic Analysis: We implement Term Frequency-Inverse Document Frequency vectorization (Salton & Buckley, 1988) to convert product descriptions into sparse high-dimensional vectors, capturing word importance based on frequency patterns. To reduce dimensionality and capture latent semantic relationships, we apply Singular Value Decomposition through LSA, projecting vectors into a lower-dimensional space of 100 dimensions. This traditional approach provides a computational efficient baseline for comparison.

**Figure 1. Research framework**




PhoBERT: We employ PhoBERT-large (Dat & Anh, 2020), a Vietnamese-specific BERT model pre-trained on 20GB of Vietnamese texts, to generate contextualized embeddings. Product descriptions are tokenized using the model's subword vocabulary, and we extract the [CLS] token representation from the final layer as the 1024-dimensional product embedding. This captures Vietnamese linguistic patterns and semantic relationships specific to the language.

E5-Multilingual: We utilize the multilingual-e5-large model (Wang et al., 2024), a cross-lingual transformer trained on text pairs from 100+ languages through contrastive learning. Product descriptions are encoded using the model's mean pooling strategy over token embeddings, producing 1024-dimensional vectors in a unified multilingual semantic space. This enables effective comparison between Vietnamese and English terms commonly found in e-commerce product descriptions.

### *3.3. Clustering algorithms*

We apply three widely-used clustering algorithms to partition products based on their vector representations. K-Means (MacQueen, 1967) employs iterative centroid optimization to partition data into K clusters by minimizing within-cluster variance, offering computational efficiency and clear cluster assignments. DBSCAN (Ester et al., 1996) identifies clusters as dense regions separated by sparse areas, automatically detecting outliers and supporting arbitrarily-shaped clusters without requiring predefined cluster counts. Gaussian Mixture Models (Bishop, 2006) use probabilistic modeling to represent clusters as overlapping Gaussian distributions, providing soft cluster assignments through posterior probabilities. Each algorithm offers distinct characteristics—K-Means for efficiency and simplicity, DBSCAN for noise handling and shape flexibility, GMM for uncertainty quantification—enabling evaluation of how algorithm properties interact with different embedding spaces.

### *3.4. Evaluation metrics*

To comprehensively assess clustering quality, we employ a combination of metrics that reflect cluster cohesion, separation, and accuracy.

Silhouette Score (based on cosine distance) measures how similar (Rousseeuw, 1987) each data point is to its own cluster compared to other clusters. Here, a(i) represents the average cosine distance between point i and other points within the same cluster, while b(i) denotes the minimum average cosine distance between point i and points in other clusters:

$$s(i) = \frac{b_{\text{cosine}}(i) - a_{\text{cosine}}(i)}{\max\left(a_{\text{cosine}}(i),\, b_{\text{cosine}}(i)\right)} \quad (1)$$

The value s(i) ranges from -1 to 1, where values closer to 1 indicate better cluster assignment. The overall Silhouette Score is computed as:

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad (2)$$

reflecting the overall clustering quality in the semantic space using cosine similarity.

Davies-Bouldin Index (DBI) (Davies & Bouldin, 1979) represents the ratio between within-cluster dispersion and between-cluster separation:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)}\right) \quad (3)$$

where $\sigma_i$ is the average dispersion within cluster i and $d(\mu_i, \mu_j)$ is the distance between cluster centroids. Lower DBI values indicate better-separated clusters.

Calinski-Harabasz Index (CH) (Caliński & Harabasz, 1974) is calculated as the ratio of between-cluster to within-cluster dispersion :

$$CH = \frac{\mathrm{Tr}(B_k)/(K-1)}{\mathrm{Tr}(W_k)/(n-K)} \quad (4)$$

where $B_k$ is the between-cluster variance matrix and $W_k$ is the within-cluster variance. Higher CH values indicate clearer cluster boundaries.

Normalized Mutual Information (NMI) (Vinh et al., 2010) measures the mutual dependence between predicted clusters C and true labels T:

$$\mathrm{NMI}(C, T) = \frac{I(C; T)}{\sqrt{H(C) \times H(T)}} \quad (5)$$

where the mutual information I(C; T) is computed as:

$$I(C;T) = \sum_{i=1}^{K}\sum_{j=1}^{J} \frac{|C_i \cap T_j|}{N} \log \frac{N \cdot |C_i \cap T_j|}{|C_i| \cdot |T_j|} \quad (6)$$

and H(C), H(T) are the entropies of cluster and label distributions:

$$H(C) = -\sum_{i=1}^{K} \frac{|C_i|}{N} \log \frac{|C_i|}{N} \quad (7)$$

$$H(T) = -\sum_{j=1}^{J} \frac{|T_j|}{N} \log \frac{|T_j|}{N} \quad (8)$$

NMI values range from 0 to 1, with 1 representing perfect clustering.

Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) measures the similarity between

predicted and true clusters (when labels are available):

$$\text{ARI} = \frac{\text{RI} - \text{E[RI]}}{\text{max(RI)} - \text{E[RI]}} \quad (9)$$

where RI is the Rand Index representing the proportion of correctly clustered point pairs. ARI ranges from -1 to 1, where 1 indicates perfect clustering, 0 represents random clustering, and negative values suggest worse-than-random clustering.

Purity (Manning et al., 2008) reflects the proportion of elements in each cluster belonging to the majority class:

$$Purity = \frac{1}{N}\sum_{k} \max_{j} |C_k \cap L_j| \quad (10)$$

Purity reaches 1 when clusters perfectly match the actual classes. The combination of these six metrics ensures objective and multi-dimensional evaluation of clustering effectiveness, particularly in contexts without standard labels.

*3.5. Experimental setup and research dataset*

Our dataset was collected from Lazada Vietnam using Selenium-based automated web scraping, initially comprising approximately 37,000 product entries. After data cleaning to remove missing descriptions, duplicate listings, and invalid category labels, the final dataset contains 36,644 products distributed across 14 distinct product categories. The dataset reflects the natural class imbalance characteristic of real-world e-commerce platforms, with category sizes varying from 1,216 products (Mobile Phones, 3.3%) to 3,908 products (Skincare & Serum, 10.7%). Product descriptions are written in Vietnamese-English code-mixed text with an average length of 87 words (σ=45.2 words). Table 1 presents the detailed distribution of products across all categories.

**Table 1. Distribution of Product Categories in the Dataset**

| Category | Number of Products | Percentage (%) |
|---|---|---|
| Skincare & Serum | 3,908 | 10.7 |
| Internet-Connected IP Cameras | 3,875 | 10.6 |
| Power Banks | 3,208 | 8.8 |
| Cables & Adapters for Phones and Computers | 3,100 | 8.5 |
| Room Fragrance Accessories | 2,972 | 8.1 |
| Microphones | 2,713 | 7.4 |
| Decorative Lighting | 2,684 | 7.3 |
| Jeans | 2,613 | 7.1 |
| T-shirts & Tank Tops | 2,323 | 6.3 |
| Smartwatches & Health Tracking Devices | 2,179 | 5.9 |
| Plants & Seeds | 2,021 | 5.5 |
| Wireless Earphones | 1,959 | 5.3 |
| Portable Speakers & Boomboxes | 1,873 | 5.1 |
| Mobile Phones | 1,216 | 3.3 |
| **Total** | **36,644** | **100.0** |

All experiments were conducted on a system equipped with an NVIDIA RTX 3060 GPU (12GB VRAM), Intel Core i5-12400F CPU, and 32GB RAM, using Python 3.10 with scikit-learn 1.3.0, transformers 4.30.0, and sentence-transformers libraries. We systematically evaluated nine combinations of three text representation methods with three clustering algorithms. For text representations, TF-IDF vectors were reduced to 100 dimensions using LSA with a maximum vocabulary of 10,000 features and bigrams. PhoBERT embeddings (1024 dimensions) were extracted from the vinai/phobert-base model's final layer [CLS] token, while E5-multilingual embeddings (1024 dimensions) were generated using intfloat /multilingual-e5-large with mean pooling. For clustering algorithms, K-Means used 14 clusters with k-means++ initialization and 300 iterations. DBSCAN parameters were optimized separately for each embedding space (eps=0.3-0.5, min_samples=5-10) using cosine distance. GMM employed 14 components with full covariance matrices and 100 EM iterations. All experiments used cosine similarity for distance calculations and fixed random seeds (seed=42) for reproducibility, with results averaged over three independent runs.

**4. Research results**

Tables 1 and 2 present comprehensive results across nine embedding-algorithm combinations. E5-multilingual with GMM achieves best supervised performance (NMI=0.9128,

ARI=0.820, Purity=0.792), while TF-IDF+DBSCAN shows high unsupervised scores (Silhouette=0.986) but poor category alignment (ARI=0.003), revealing that internal clustering quality does not guarantee semantic validity.

**Table 2. Unsupervised Clustering Metrics**

| | Alg | Sil ↑ | DBI ↓ | CH ↑ |
|---|---|---|---|---|
| TF-IDF | KM | 0.297 | 2.511 | 1329 |
| | DB | 0.986 | 0.116 | 16175 |
| | GMM | 0.276 | 2.784 | 1275 |
| PhoBERT | KM | 0.205 | 2.299 | 2418 |
| | DB | 0.556 | 0.837 | 48 |
| | GMM | 0.264 | 2.182 | 4332 |
| E5-Multi | KM | 0.166 | 2.855 | 1471 |
| | DB | 0.670 | 0.844 | 275 |
| | GMM | 0.184 | 3.005 | 1361 |

*KM=K-Means; DB=DBSCAN; Sil=Silhouette; DBI=Davies-Bouldin Index; CH=Calinski-Harabasz. ↑=higher better; ↓=lower better.*

TF-IDF+DBSCAN achieves exceptional unsupervised scores (Silhouette=0.986, CH=16,175) by identifying lexically similar products sharing marketing phrases, creating geometrically clean but semantically meaningless clusters. Transformer embeddings produce moderate scores (0.166-0.670) reflecting semantic category overlap—related products like speakers and microphones have closer embeddings due to shared domain semantics, reducing Silhouette scores but capturing true relationships.

**Table 3. Supervised Clustering Metrics**

| | Alg | NMI ↑ | ARI ↑ | Pur ↑ |
|---|---|---|---|---|
| TF-IDF | KM | 0.902 | 0.864 | 0.924 |
| | DB | 0.092 | 0.003 | 0.156 |
| | GMM | 0.835 | 0.748 | 0.862 |
| PhoBERT | KM | 0.548 | 0.378 | 0.533 |
| | DB | 0.006 | 0.001 | 0.111 |
| | GMM | 0.300 | 0.116 | 0.268 |
| E5-Multi | KM | 0.852 | 0.636 | 0.658 |
| | DB | 0.758 | 0.528 | 1.000 |
| | GMM | 0.9128 | 0.820 | 0.792 |

*NMI=Normalized Mutual Information; ARI=Adjusted Rand Index; Pur=Purity. ↑=higher better. Bold=best per metric.*

E5-multilingual+GMM achieves highest supervised performance (NMI=0.9128, ARI=0.820) through cross-lingual semantic understanding and probabilistic soft clustering accommodating fuzzy category boundaries. TF-IDF+K-Means surprisingly ranks second (NMI=0.902) as hard partitioning leverages category-specific terminology, while TF-IDF+DBSCAN fails catastrophically (ARI=0.003) by grouping promotional text patterns unrelated to categories. PhoBERT underperforms (NMI=0.548) as monolingual training poorly handles English specifications common in Vietnamese e-commerce, with a 67% gap to E5-multilingual quantifying multilingual understanding value.

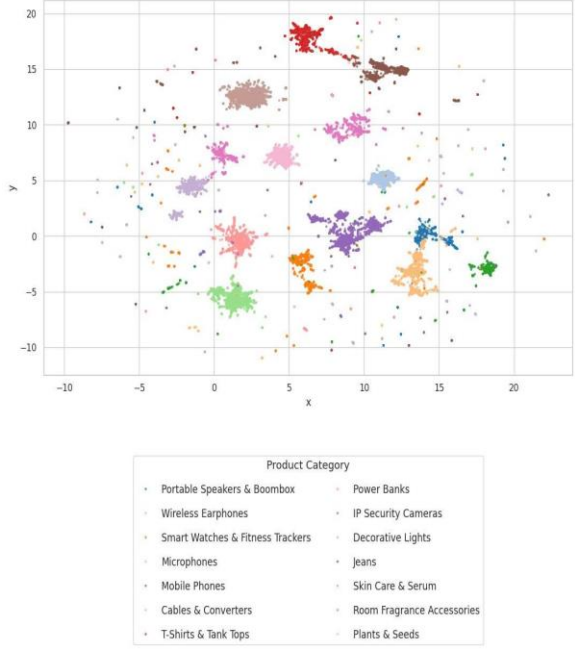
E5-multilingual+GMM demonstrates superior cluster-category correspondence. Most categories form distinct clusters—"Jeans" and "T-shirts" separate cleanly, while controlled overlap occurs between semantically related categories like "Portable Speakers" and "Microphones," validating semantic understanding. GMM's probabilistic framework naturally accommodates products spanning multiple categories (e.g., smartwatches blending technology and health tracking), unlike K-Means' hard assignments. The model converged at 10 optimal components, balancing granularity with interpretability.

Three critical insights for e-commerce management emerge: First, high unsupervised scores do not guarantee business value—TF-IDF+DBSCAN's near-perfect Silhouette (0.986) contrasts with near-zero ARI (0.003), emphasizing that platforms must validate AI systems against actual business categories rather than relying solely on internal technical metrics. Second, multilingual capability delivers measurable operational advantages—E5-multilingual's 67% NMI improvement over PhoBERT (0.9128 vs 0.548) demonstrates that investing in cross-lingual models provides substantial returns for platforms operating in markets with code-mixed product content. Third, algorithm selection modulates embedding effectiveness—GMM outperforms K-Means by 7% NMI (0.9128 vs 0.852) on E5 embeddings, as probabilistic modeling aligns with transformer embeddings' continuous semantic spaces better than hard partitioning.

Based on these insights, our findings with 91.28% NMI and 79.2% purity demonstrate

strong potential for catalog automation. The 79.2% purity indicates that approximately four out of five products can be correctly categorized automatically, though the actual reduction in manual effort would depend on platform-specific confidence thresholds and quality requirements. The 67% performance advantage of E5-multilingual over PhoBERT quantifies the value of multilingual capability for Southeast Asian markets where code-mixing is prevalent. This substantial improvement justifies investment in cross-lingual models despite higher computational costs. The superiority of GMM over K-Means (91.28% vs 85.2% NMI) further emphasizes that algorithm selection significantly impacts practical performance. Platforms should consider implementing hybrid workflows where high-confidence predictions proceed to automatic categorization while uncertain cases receive manual review. The optimal confidence thresholds would need empirical validation on each platform's specific data and business requirements.

**Figure 2. UMAP visualization of E5-multilingual embeddings with GMM clustering (K=14)**




*Each point represents a product colored by ground-truth category, demonstrating clear separation between distinct product types and controlled semantic overlap between related categories.*

These findings enable automated product categorization reducing manual labeling costs while achieving 79.2% purity. However, the 20% error rate suggests limitations from genuinely ambiguous products, inconsistent ground-truth labels, or inherent difficulty imposing discrete categories on continuous semantic spaces. Additional limitations include dataset constraints (single platform, Lazada Vietnam), fixed cluster number (K=14) that cannot accommodate dynamic category emergence, and computational requirements for real-time transformer inference at scale. The assumption of single-category membership oversimplifies products that naturally span multiple categories. Future work should explore fine-tuning E5-multilingual on domain-specific corpora and developing hybrid human-AI workflows routing high-confidence predictions to automation while flagging ambiguous cases for manual review. Hierarchical clustering could better capture category taxonomies, while multi-modal approaches combining text and images may improve accuracy for visual-dependent categories. Investigating incremental learning would enable models to adapt to new product types without full retraining.

## 5. Discussion

This study provides a comprehensive empirical evaluation of text representation and clustering methods for multilingual e-commerce product organization, with significant implications for operational efficiency and cost reduction. Through systematic comparison of nine embedding-algorithm combinations on 36,644 Vietnamese product descriptions, we demonstrate that E5-multilingual embeddings with Gaussian Mixture Models achieve superior performance (NMI=0.9128, ARI=0.8199, Purity=0.7922), significantly outperforming both traditional TF-IDF approaches and monolingual transformer models. This performance level enables practical automation of catalog management, potentially reducing manual classification workload while improving consistency across large product inventories. Our analysis reveals a critical insight for practitioners: high unsupervised clustering scores do not guarantee meaningful business categorization, as evidenced by TF-IDF+DBSCAN's excellent

internal metrics (Silhouette=0.9863) but poor category alignment (ARI=0.0031). This discrepancy underscores the importance of supervised validation in real-world e-commerce applications.

The findings carry substantial economic value for e-commerce platforms. Our optimal configuration achieving 91.3% normalized mutual information with ground-truth categories (NMI=0.9128, ARI=0.820) demonstrates strong cluster-category alignment, enabling automation of product catalog management. Beyond direct cost savings, accurate automated categorization improves customer discovery and conversion, as demonstrated by recommendation system studies showing significant impact on purchase behavior (Linden et al., 2003), while accelerating seller onboarding and reducing time-to-market for new products. The 67% performance gap between PhoBERT (NMI=0.548) and E5-multilingual (NMI=0.9128) quantifies the economic value of cross-lingual semantic understanding for Southeast Asian markets where Vietnamese-English code-mixing is prevalent. These results demonstrate that investing in multilingual transformer infrastructure generates measurable ROI through operational efficiency gains and enhanced customer experience.

For e-commerce platform managers, these findings offer actionable guidance: multilingual transformer models represent a viable solution for automating catalog organization in Southeast Asian markets, with performance levels sufficient for production deployment. The critical insight that technical metrics may not reflect business value underscores the importance of validation frameworks aligned with operational objectives. While achieving strong cluster-category alignment (NMI=0.9128, ARI=0.820), opportunities remain to further reduce cross-category mixing (current purity=79.2%). Future research should explore fine-tuning E5-multilingual on domain-specific product corpora, investigating hierarchical clustering to capture category taxonomies, and developing cost-optimized hybrid human-AI workflows that fully automate high-confidence predictions while routing ambiguous cases to efficient manual review, maximizing the trade-off between automation savings and categorization accuracy. These findings provide evidence-based guidance for platforms seeking to leverage transformer-based technologies for competitive advantage through reduced operational costs and enhanced customer discovery experiences.

## 6. Conclusion

Our research makes three key contributions to the field: We provide the first comprehensive empirical comparison of nine embedding-algorithm combinations using both unsupervised metrics (Silhouette, Davies-Bouldin Index, Calinski-Harabasz) and supervised metrics (Normalized Mutual Information, Adjusted Rand Index, Purity) on a large-scale Vietnamese e-commerce dataset comprising 36,644 product descriptions across 14 categories. We demonstrate that multilingual transformer embeddings combined with probabilistic clustering significantly outperform traditional methods, achieving over 91% normalized mutual information with ground-truth categories. Importantly, we reveal a critical insight for practitioners: high unsupervised clustering scores do not necessarily correlate with meaningful business categorization, emphasizing the importance of supervised validation in real-world e-commerce applications. These findings provide evidence-based guidance for platforms implementing automated catalog management systems, with direct implications for reducing operational costs and enhancing customer discovery experiences.

## References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chang, W.-C., Yu, F. X., Chang, Y.-W., Yang, Y., & Kumar, S. (2020). *Pre-training Tasks for Embedding-based Large-scale Retrieval* (arXiv:2002.03932). arXiv. https://doi.org/10.48550/arXiv.2002.03932

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dat, N. Q., & Anh, N. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. In T. Cohn & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1037–1042). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.92

Dung, T. T., Tung, L. N., Dung, B. N., & Huan, V. (2024). Emotion recognition in learners with emoji sentiment accompaniment using the PhoBERT model. *Journal of Science Natural Science*, 46–56. https://doi.org/10.18173/2354-1059.2024-0034

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, 226–231.

*Gaussian Mixture Model*. (2025). GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/gaussian-mixture-model/

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics: 5.1* (pp. 281–298). University of California Press. https://digicoll.lib.berkeley.edu/record/113015/files/math_s5_v1_article-17.pdf

Manning, C. D., Raghavan, P., & Schütze, H. (2008, July 7). *Introduction to Information Retrieval*. Cambridge University Press. Cambridge Aspire Website. https://doi.org/10.1017/CBO9780511809071

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-Based Recommendations on Styles and Substitutes. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15*, 43–52. https://doi.org/10.1145/2766462.2767755

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Vinh, N. X., Epps, J., Epps, J., & Bailey, J. (2010). *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2024). *Text Embeddings by Weakly-Supervised Contrastive Pre-training* (arXiv:2212.03533). arXiv. https://doi.org/10.48550/arXiv.2212.03533

Yulianton, H., & Santi, R. (2024). Product Matching using Sentence-BERT: A Deep Learning Approach to E-Commerce Product Deduplication. *Engineering and Technology Journal*, *09*. https://doi.org/10.47191/etj/v9i12.14

# TỪ TF-IDF ĐẾN CÁC MÔ HÌNH DỰA TRÊN TRANSFORMER: PHƯƠNG PHÁP PHÂN CỤM CHO TỔ CHỨC SẢN PHẨM THƯƠNG MẠI ĐIỆN TỬ

**Lê Nhật Tùng[1] Bùi Minh Huy[2]**
**Trần Lê Vân[3] Nguyễn Thị Thanh Tâm[4] Nguyễn Thị Liệu[5]**

[1, 5]Trường Đại học Công nghệ Đồng Nai; [2, 3, 4]Trường Đại học Công nghệ Thành phố Hồ Chí Minh
Email: *lenhattung@dntu.edu.vn[1]; huybm.ds@gmail.com[2]; tranlevan07.ds@gmail.com[3]; ttam99852@gmail.com[4]; nguyenthilieu@dntu.edu.vn[5]*
Ngày nhận bài: 3/4/2026; Ngày phản biện: 20/4/2026; Ngày tác giả sửa: 25/4/2026;
Ngày duyệt đăng: 29/4/2026
**DOI:** *https://doi.org/10.58902/nckhpt.e-v2i1.366*

**Tóm tắt:** *Quản lý danh mục sản phẩm là yếu tố thiết yếu giúp các nền tảng thương mại điện tử tối ưu hóa vận hành, đặc biệt trong môi trường đa ngôn ngữ tại Việt Nam, nơi mô tả sản phẩm thường trộn lẫn nội dung tiếng Việt và thuật ngữ tiếng Anh. Nghiên cứu này đánh giá có hệ thống các phương pháp phân cụm tự động bằng cách so sánh 9 tổ hợp giữa 3 phương pháp biểu diễn văn bản (TF-IDF, PhoBERT, E5-multilingual) và 3 thuật toán phân cụm (K-Means, DBSCAN, Gaussian Mixture Models - GMM) trên tập dữ liệu 36.644 sản phẩm từ Lazada Việt Nam. Kết quả thực nghiệm cho thấy mô hình E5-multilingual kết hợp với GMM đạt hiệu quả cao nhất với chỉ số thông tin tương hỗ chuẩn hóa (NMI) là 91,28% và độ khiết (Purity) đạt 79,22%, vượt trội so với các phương pháp truyền thống và mô hình đơn ngữ. Đáng chú ý, nghiên cứu phát hiện nghịch lý khi một số phương pháp có điểm kỹ thuật cao (Silhouette 0,986) nhưng không mang lại giá trị phân loại kinh doanh (ARI 0,003), nhấn mạnh vai trò của việc xác thực dựa trên danh mục thực tế. Những phát hiện này cung cấp cơ sở thực nghiệm để các doanh nghiệp thương mại điện tử phát triển các giải pháp tự động hóa quy trình phân loại, giảm thiểu nguồn lực thủ công và nâng cao trải nghiệm khách hàng.*

**Từ khóa:** *TF-IDF; Thuật toán phân cụm; Thương mại điện tử; Transformer.*