UDC: 004.85:81'322:81'246.2 https://doi.org/10.15407/jai2026.01.035

**B. Pavlyshenko**[1], **M. Stasiuk**[2]
[1,2]Ivan Franko National University of Lviv, Ukraine
Universytetska St, 1, Lviv, 79000
[1]bohdan.pavlyshenko@lnu.edu.ua
[2]mykola.stasiuk@lnu.edu.ua
[1]https://orcid.org/0000-0001-9515-3488
[2]https://orcid.org/0009-0006-5256-2103

# TOPIC CLUSTERING OF ENGLISH-UKRAINIAN TEXT CORPORA USING MULTILINGUAL EMBEDDINGS

**Abstract.** This study investigates the impact of multilingual sentence-embedding models on unsupervised thematic clustering of a balanced bilingual corpus comprising English and Ukrainian documents. While modern embedding architectures project texts from multiple languages into a shared semantic space, it remains unclear how effectively different models preserve geometrically separable thematic structure suitable for clustering. A balanced dataset of 1,404 documents across four high-level thematic domains, namely History, Culture, Science, and Technology was constructed using a high-confidence zero-shot labeling procedure. Three multilingual sentence embedding models LaBSE, paraphrase-xlm-r-multilingual-v1, and paraphrase-multilingual-mpnet-base-v2 were evaluated in combination with three clustering approaches: KMeans, Agglomerative Hierarchical Clustering, and Fuzzy C-Means. Visual analysis using UMAP projections and quantitative evaluation via intrinsic (Silhouette, Calinski-Harabasz, Davies-Bouldin) and extrinsic (Adjusted Rand Index, Normalized Mutual Information) metrics reveal that clustering performance is primarily determined by the geometric properties of the embedding space rather than by the choice of clustering algorithm. LaBSE produces the most stable and well-separated thematic regions, achieving the highest agreement with ground truth categories. The paraphrase-xlm-r model exhibits reduced thematic separability, while the multilingual MPNet variant demonstrates intermediate behavior with stronger performance under hierarchical clustering. Across all models, English and Ukrainian documents remain interwoven within thematic regions, confirming effective cross-lingual alignment. However, the degree of semantic separability varies significantly between embedding architectures. The findings indicate that representation quality plays a dominant role in unsupervised thematic analysis of bilingual corpora. Embedding model selection is therefore a more impactful factor than clustering strategy when analyzing cross-lingual semantic structure.

**Keywords:** multilingual sentence embeddings, unsupervised clustering, cross-lingual semantic alignment, English-Ukrainian bilingual corpus, distance distribution, thematic analysis.

## Introduction

The rapid growth of multilingual digital content has created new challenges for large-scale thematic analysis of textual corpora. In domains such as science, culture, history, and technology, documents are increasingly produced in multiple languages, requiring analytical frameworks that can identify coherent semantic structures across linguistic boundaries. Traditional text analysis approaches often rely on language-specific processing pipelines, limiting their ability to capture cross-lingual thematic alignment.

Recent advances in multilingual sentence embeddings have significantly improved the ability to represent texts from different languages within a shared semantic space. Models such as LaBSE [1] and other transformer-based architectures enable direct comparison of documents across languages by mapping semantically similar texts to nearby regions in high-dimensional embedding space. This development opens new possibilities for unsupervised thematic analysis of bilingual and multilingual corpora.

Unsupervised clustering methods vary substantially in their assumptions about data geometry. Centroid-based approaches, such as K-Means [2], optimize within-cluster variance; hierarchical algorithms [3] iteratively merge samples based on pairwise distances; and soft clustering methods, such as Fuzzy C-Means [4], assign graded memberships that may better reflect gradual thematic transitions. Whether these methods converge toward similar thematic partitions when applied to multilingual embedding spaces requires systematic evaluation.

However, while multilingual embeddings provide a unified representation, the extent to which thematic structure can be reliably recovered through unsupervised clustering remains an open question. In particular, it is unclear whether cluster formation is primarily driven by semantic similarity or residual linguistic signals, and how robust the recovered structure is across different clustering strategies.

This study investigates the semantic organization of a balanced bilingual corpus comprising four thematic domains, namely History, Culture, Science, and Technology represented in English and Ukrainian. Using multilingual sentence embeddings, three clustering strategies were applied: K-Means, Agglomerative hierarchical clustering, and Fuzzy C-Means. The objectives of the study are:

- To examine whether thematic structure emerges consistently across clustering algorithms.
- To visually assess whether cluster formation aligns with thematic categories or reflects language-based segmentation
- To assess the robustness of clustering results under both hard and soft partitioning strategies.

**Related work**

Unsupervised clustering of textual data has long been a research focus due to its importance in organizing large corpora and discovering latent semantic structures. Traditional cluster analysis methods based on lexical representations, such as TF-IDF [5] combined with different clustering algorithms [6, 7, 8] have been widely studied.

With the advancement of contextualized transformer-based models, research in text representation shifted toward deep semantic embeddings. Contextual embeddings derived from models such as BERT [9] have been shown to significantly improve the semantic quality of document representations compared to static word embeddings, thereby improving clustering performance. For example, WEClustering [10] demonstrated that contextualized embeddings improve document clustering accuracy compared with traditional baselines.

The integration of large language model embeddings into traditional clustering workflows has recently been explored, highlighting the superior separability of semantically informed features compared to conventional vector representations. A recent study on LLM embeddings for text clustering [11] found that LLM-based representations outperform older methods at capturing thematic similarity across diverse datasets.

Multilingual and cross-lingual text clustering introduces additional complexity, as clustering algorithms must handle semantic alignment across languages. Cross-lingual news clustering [12], for instance, uses a combination of sentence vector representations and topic distributions to group related multilingual news articles without direct translation, highlighting approaches to constructing bilingual semantic spaces.

Other research in cross-lingual representation learning emphasizes methods for aligning embedding spaces across languages. Techniques for unsupervised cross-lingual word embeddings [13] and contextual alignment [14] have been proposed to support transfer learning and downstream tasks such as clustering in multilingual settings.

Several studies also examine multi-aspect and multi-view clustering approaches, integrating contextual and semantic embeddings with topic or structural features to improve clustering quality across different languages and domains [15]. These approaches evidence the importance of embedding quality and representation learning in structured clustering performance across diverse text collections.

Despite these advances, systematic comparative evaluations of contemporary multilingual embeddings, especially in the context of unsupervised clustering for under-resourced languages such as Ukrainian, remain limited. This study addresses this gap by analyzing how embedding choice and

clustering algorithms interact to shape semantic clusters in both monolingual and bilingual corpora.

**Methods and Materials**

The experimental evaluation was conducted on a multilingual textual corpus comprising English and Ukrainian documents. The dataset was constructed to represent high-level thematic domains rather than narrowly defined topics, enabling coarse-grained semantic clustering. Four conceptual categories were selected: History, Culture, Science, and Technology. These domains were chosen for their conceptual distinctiveness and frequent occurrence in general-purpose text collections.

To enable objective evaluation of clustering performance, a labeled benchmark dataset was constructed using a zero-shot classification approach. A GPT-5 [16] model was used to assign semantic relevance scores to each document relative to predefined thematic categories. For each document, the model produced a confidence score for each category in the range from 0 to 100. Documents were assigned a category label if the highest score exceeded a threshold of 90. This threshold was selected to prioritize labeling precision over coverage, ensuring that the resulting benchmark represented high-confidence semantic ground truth. Documents that did not meet the threshold were excluded from the benchmark evaluation.

The initial corpus exhibited class imbalance, with one category containing significantly fewer instances than the others. To prevent clustering bias and distortion of intrinsic metrics caused by imbalance, a stratified undersampling strategy [17] was applied. The smallest sufficiently represented category contained 351 documents; therefore, the remaining categories were randomly undersampled to match this size. The resulting balanced dataset comprised 1,404 documents, evenly distributed across the four thematic categories.

All documents were preprocessed using standard normalization procedures, including lowercasing and removal of non-textual artifacts. No language-specific linguistic preprocessing, such as lemmatization or stemming, was applied. This decision was made to preserve comparability across languages and to avoid introducing language-dependent biases into the embedding space.

Semantic representations were generated using three multilingual sentence-embedding models from sentence-transformers library: LaBSE, xlm-r-multilingual-v1 [18], and paraphrase-multilingual-mpnet-base-v2 [19, 20]. These models were selected for their widespread adoption and ability to encode texts from multiple languages into a shared high-dimensional semantic space. Each document was mapped to a fixed-dimensional embedding vector using the respective model in an off-the-shelf configuration without task-specific fine-tuning. This design enables comparative analysis of how different embedding architectures shape the geometric structure of the semantic space and influence clustering behavior. By keeping clustering configurations identical across models, the experimental setup isolates the impact of representation quality on thematic recoverability while ensuring reproducibility.

Three clustering algorithms were employed to assess robustness across different clustering paradigms: K-Means, Agglomerative Hierarchical Clustering, and Fuzzy C-Means. For all algorithms, the number of clusters was fixed to four to match the benchmark categories.

In the case of Fuzzy C-Means, preliminary experiments with default parameters yielded degenerate solutions with nearly uniform cluster memberships. To ensure stable convergence in the high-dimensional embedding space, document embeddings were L2-normalized prior to clustering, and the fuzziness parameter was reduced to $m=1.1$, producing sharper membership distributions. Final cluster assignments for evaluation were obtained via defuzzification using the maximum membership criterion.

Clustering performance was evaluated using three intrinsic metrics: Silhouette Score [21], Calinski-Harabasz Index [22], and

Davies-Bouldin Index [23], which jointly assess cluster compactness, separation, and structural distinctness. In addition, extrinsic evaluation was performed using Adjusted Rand Index (ARI)[24] and Normalized Mutual Information (NMI)[25], comparing cluster assignments against the zero-shot benchmark labels.

For visualization, dimensionality reduction was performed using UMAP [26] to project high-dimensional embeddings into two dimensions while preserving global structure. This enabled qualitative assessment of thematic separation and cross-lingual alignment.

All experiments were conducted under identical preprocessing, embedding extraction, and clustering configurations to ensure comparability. No hyperparameter tuning was performed for individual algorithm-model combinations, as the study's objective was a comparative analysis rather than optimizing specific pipelines.

**Results**

Fig. 1 presents UMAP projections of document embeddings generated using the LaBSE multilingual sentence encoder. The figure is structured as a 3×3 grid, where rows correspond to clustering algorithms, and columns represent predicted cluster assignments, ground truth thematic categories, and language distribution, respectively.

The LaBSE embedding space exhibits an organized semantic topology. Five visually distinguishable regions are observable:

- A compact and well-isolated cluster in the lower-right region.
- An elongated structure in the upper-right region.
- A dense central-left manifold.
- Two smaller transitional region adjacent to the central manifold.

When compared to the ground truth projection in the middle column, these regions align with the thematic categories. Culture forms a tightly clustered and clearly separated region, indicating strong internal coherence and high inter-cluster separation. History occupies the elongated upper-right structure and remains distinctly separated from other domains. Science and Technology occupy adjacent regions within the central manifold, exhibiting partial overlap consistent with their conceptual proximity.

Importantly, the overlap between Science and Technology appears structured rather than chaotic. Transitional points are located near cluster boundaries, suggesting gradual semantic shifts rather than clustering instability.

Across all three algorithms, the predicted cluster assignments follow the underlying geometric structure. The rows of the figure appear highly similar, indicating that the embedding geometry strongly constrains cluster formation.

K-Means and Fuzzy C-Means produce nearly identical partitions, with sharp boundaries between major regions. Agglomerative clustering preserves the same large-scale topology but introduces slightly more diffuse boundaries within the central manifold. Nevertheless, the global semantic layout remains stable across methods.

This consistency suggests that clustering results are largely determined by the intrinsic geometry of the embedding space rather than by algorithm-specific behavior.

The rightmost column demonstrates that English and Ukrainian documents are interwoven within each thematic region. No language-driven segmentation is visible. Both languages populate the same semantic clusters, confirming that LaBSE successfully aligns multilingual content into a shared semantic space.

Thus, cluster formation is governed by semantic similarity rather than linguistic identity.

Overall, the LaBSE embedding space encodes a stable and well-separated thematic structure. Distinct semantic regions emerge clearly in the UMAP projection, and all clustering algorithms recover highly consistent partitions aligned with ground truth categories. This model provides a geometrically coherent representation in which thematic boundaries are well-defined and robust to clustering strategies.

Fig. 2 presents UMAP projections of document embeddings generated using the paraphrase-xlm-r-multilingual-v1 model. As in Fig. 1, rows correspond to clustering algorithms while columns show predicted clusters, ground truth categories, and language distribution.

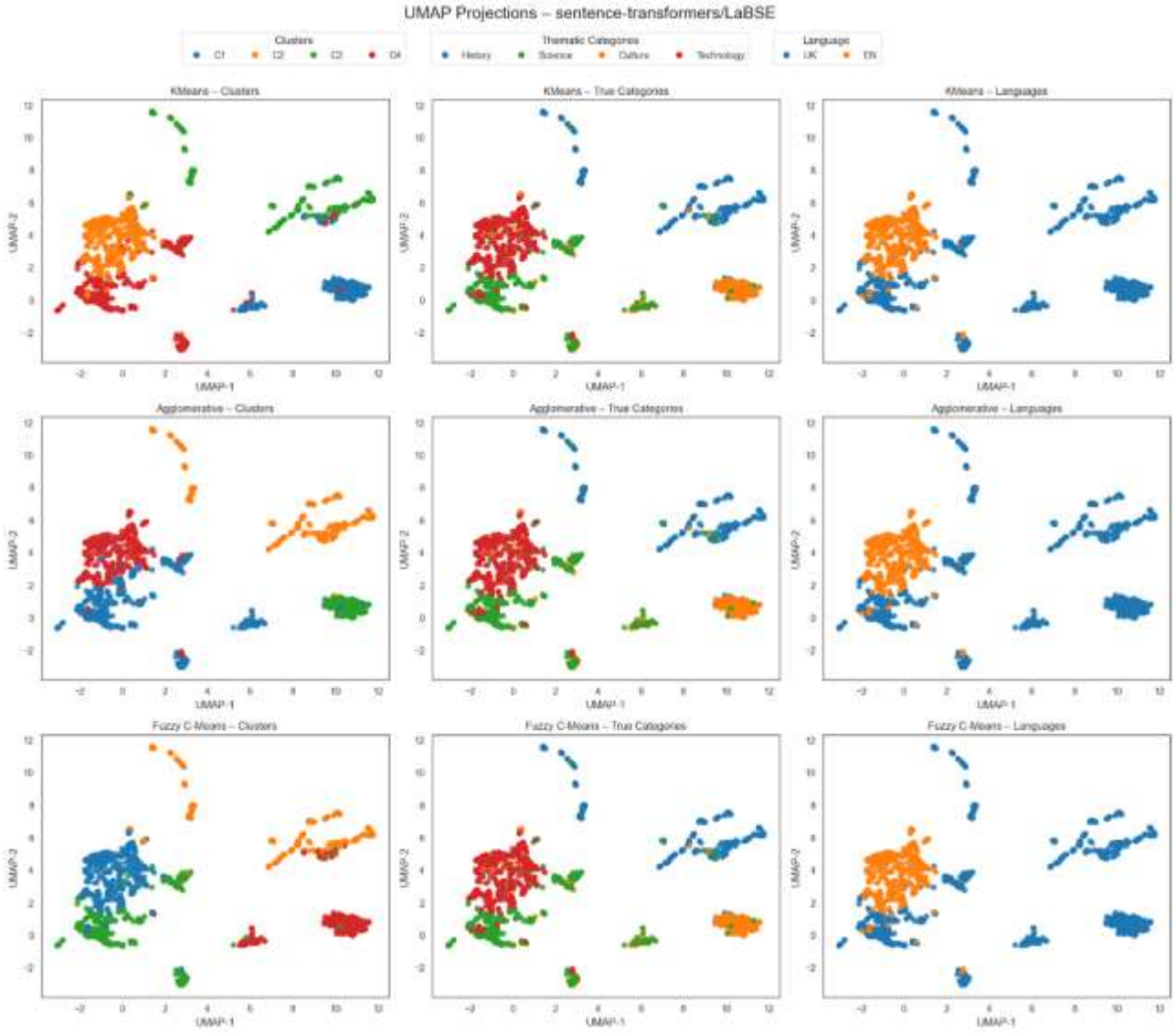


Fig. 1. UMAP projections of document embeddings, generated using LaBSE model, clustered using K-Means (top row), Agglomerative (middle row) and Fuzzy C-Means (bottom row) methods. Left column: predicted cluster assignments. Middle column: ground truth thematic categories (History, Culture, Science, Technology). Bottom column: document distribution by language (English and Ukrainian)

Compared to LaBSE, the overall geometric organization remains recognizable but appears less sharply separated.

Four macro-regions are still observable:

- A clearly isolated cluster in the lower-right region.
- A central dense manifold.
- A partially separated upper-left structure.
- A transitional vertical region connecting central areas.

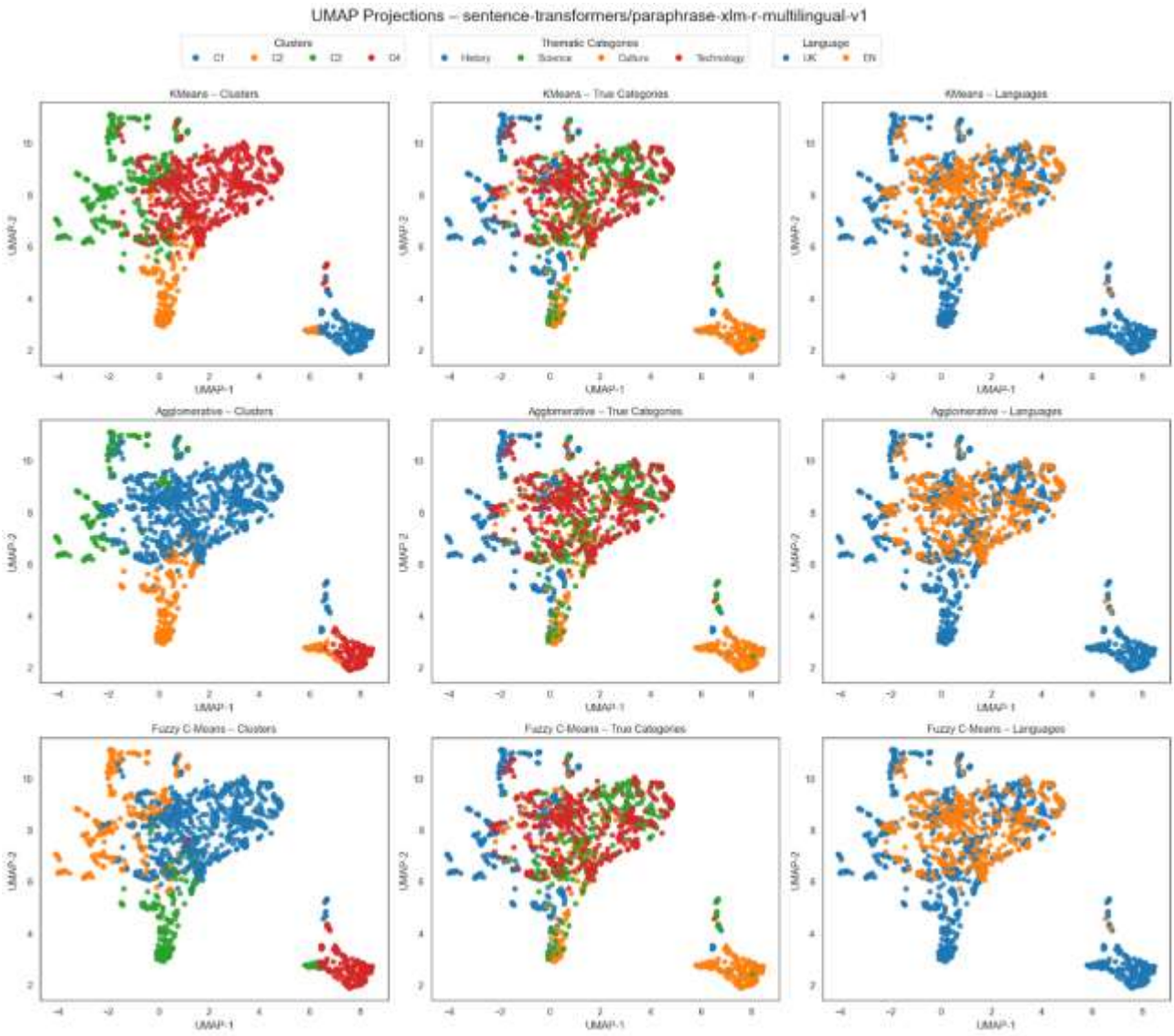


Fig. 2. UMAP projections of document embeddings, generated using paraphrase-xlm-r-multilingual-v1 model, clustered using K-Means (top row), Agglomerative (middle row) and Fuzzy C-Means (bottom row) methods. Left column: predicted cluster assignments. Middle column: ground truth thematic categories (History, Culture, Science, Technology). Bottom column: document distribution by language (English and Ukrainian)

However, the boundaries between thematic domains are visibly softer. The ground truth projection reveals increased mixing between Science and Technology within the central manifold. Unlike the clearer partitioning seen with LaBSE, here the thematic structure appears more intertwined. The overlap is broader rather than limited to boundary zones.

Culture remains relatively compact in the lower-right region, maintaining strong internal coherence. History shows some separation but more dispersion than LaBSE.

Across clustering methods, the predicted partitions reflect this softer geometry. K-Means produces reasonably coherent clusters but with noticeable boundary diffusion in the central region. Agglomerative clustering demonstrates slightly greater mixing, especially between semantically adjacent domains. Fuzzy C-Means produces similar large-scale partitions but further smooths boundaries in transitional regions.

Importantly, despite reduced sharpness, all three algorithms still recover the same macro-topology. This again suggests that the

embedding geometry constrains cluster formation, even when separability is weaker.

The language projections confirm that English and Ukrainian documents remain interwoven across semantic regions. No language-based segregation is visible.

Thus, as with LaBSE, the embedding space is structured along semantic dimensions rather than linguistic ones. The paraphrase-xlm-r-multilingual-v1 model produces a coherent but more diffuse semantic embedding space compared to LaBSE. Thematic regions are present but less sharply defined, leading to increased mixing in semantically adjacent domains. Nevertheless, clustering algorithms consistently recover the same global structure, indicating stable cross-lingual semantic alignment.

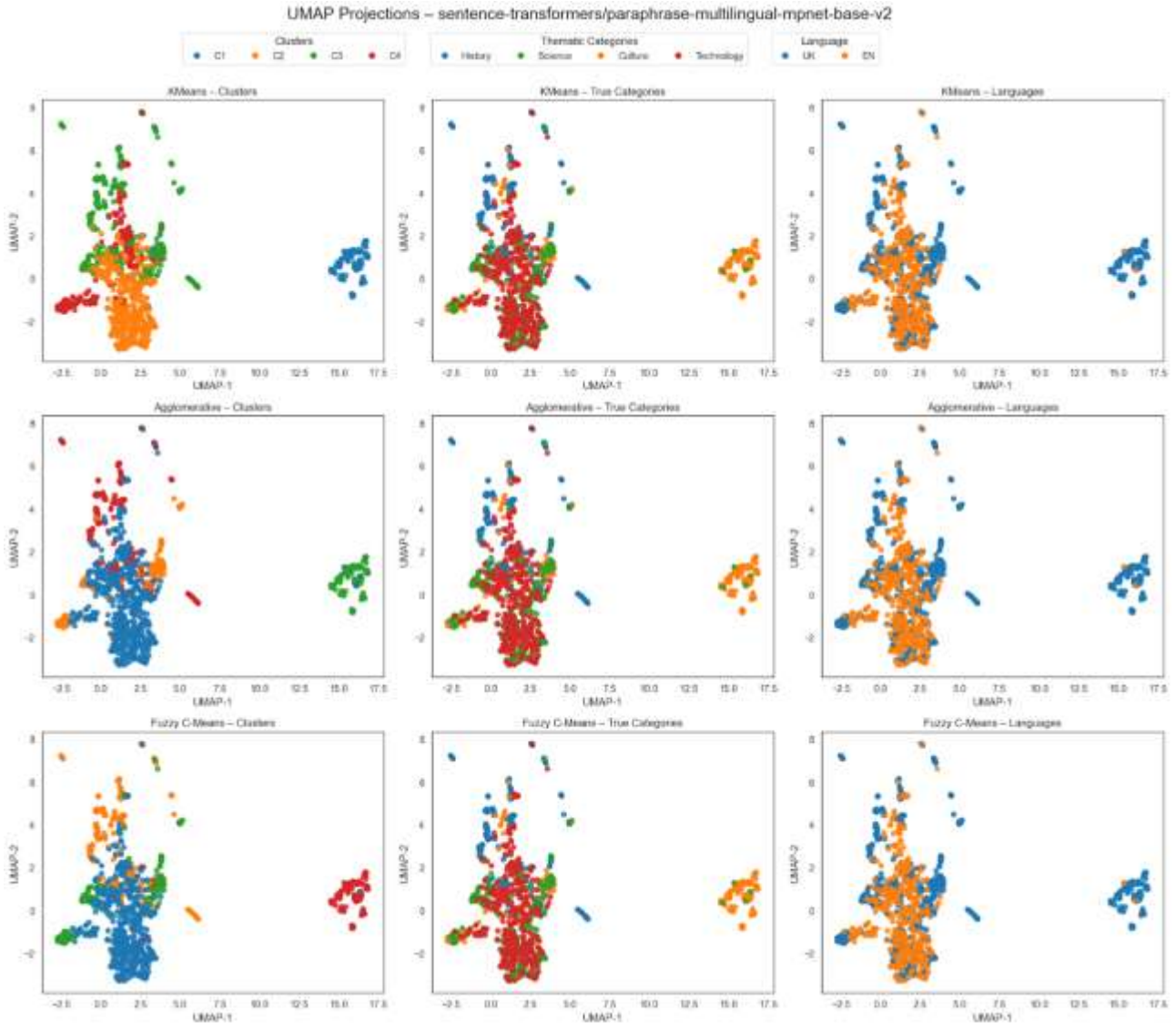


Fig. 2. UMAP projections of document embeddings, generated using multilingual-mpnet-base-v2 model, clustered using K-Means (top row), Agglomerative (middle row) and Fuzzy C-Means (bottom row) methods. Left column: predicted cluster assignments. Middle column: ground truth thematic categories (History, Culture, Science, Technology). Bottom column: document distribution by language (English and Ukrainian)

Figure 3 presents UMAP projections for embeddings generated with paraphrase-multilingual-mpnet-base-v2. Compared to the previous models, the MPNet embedding space exhibits a more vertically elongated central manifold with a clearly isolated cluster on the far right.

The lower-right region forms a compact and well-separated cluster, consistent across all methods. This region aligns closely with the Culture category in the ground-truth projection, indicating high internal coherence and strong separability.

In contrast, the central structure appears more compressed and vertically stacked. Science, History, and Technology overlap substantially in this region. The separation between Science and Technology is particularly weak, with considerable intermixing evident in the ground-truth projection.

Unlike LaBSE, where thematic regions were spatially broader and more horizontally distributed, the MPNet embedding space compresses multiple categories into a tighter, more central geometry. This increases boundary ambiguity for hard clustering methods.

K-Means identifies the isolated right-hand cluster with high stability, but partitions within the central manifold are less clearly aligned with thematic boundaries. Agglomerative clustering produces slightly sharper separation for the isolated region but shows greater mixing within the central area. Fuzzy C-Means yields a similar macro-structure while allowing smoother transitions within the central manifold. The soft clustering behavior appears particularly relevant here, as thematic boundaries are less geometrically distinct.

Overall, clustering algorithms remain consistent in identifying the isolated thematic region but struggle more with separating semantically proximate domains in the compressed central space.

The language projections again show strong interweaving of English and Ukrainian documents. No language-driven segmentation is visible. This confirms that, even in a denser embedding configuration, semantic similarity outweighs linguistic identity.

The MPNet-based embedding space exhibits a more compact and vertically concentrated semantic geometry. While one thematic domain remains clearly separable, other categories demonstrate increased overlap within a compressed central manifold. This results in slightly reduced boundary precision compared to LaBSE, though global thematic structure remains recoverable across clustering algorithms.

The quantitative evaluation confirms and refines the patterns observed in the visual analysis. Across all embedding models, clustering performance varies more substantially with the choice of representation than with the clustering algorithm itself, indicating that embedding geometry is the dominant factor shaping thematic recovery.

The LaBSE model demonstrates the strongest overall performance. KMeans achieves an Adjusted Rand Index of 0.62 and a Normalized Mutual Information of 0.6, indicating substantial agreement between the predicted clusters and the ground-truth thematic categories. Fuzzy C-Means produces nearly identical results with ARI = 0.618 and NMI = 0.59, while Agglomerative clustering remains slightly weaker but still robust with ARI = 0.54, and NMI = 0.58. Intrinsic metrics follow the same trend: Silhouette values exceed 0.10 for centroid-based methods, and Calinski-Harabasz scores reach approximately 113, reflecting strong cluster compactness and separation. These results suggest that LaBSE encodes a highly structured semantic space in which thematic regions are geometrically well-defined and consistently recoverable.

In contrast, the paraphrase-xlm-r-multilingual-v1 model exhibits noticeably reduced extrinsic alignment. ARI values range from 0.28 to 0.34 across algorithms, and NMI remains below 0.41. Although intrinsic metrics remain moderate, reduced agreement with ground-truth categories suggests that thematic boundaries are less sharply represented in this embedding space. Agglomerative clustering achieves the highest NMI of 0.41 with the

paraphrase-xlm-r-multilingual-v1, suggesting that hierarchical merging partially compensates for reduced centroid separability. Nevertheless, the overall alignment remains significantly below that observed for LaBSE.

The paraphrase-multilingual-mpnet-base-v2 model demonstrates intermediate behavior. ARI values range from 0.343 with K-Means to 0.438 with Agglomerative, and NMI reaches 0.518 under hierarchical clustering. While intrinsic metrics are somewhat lower than those of LaBSE, the improved extrinsic performance of Agglomerative clustering suggests that the embedding space contains meaningful semantic structure that may not be optimally captured by centroid-based partitioning. The divergence between K-Means and Agglomerative performance in this case indicates a geometry that may be less isotropic and more amenable to hierarchical grouping.

Table 1. Comparative clustering performance across intrinsic and extrinsic metrics relative to zero-shot benchmark labels

| Model | Algorithm | Silhouette | Calinski-Harabasz | Davies-Bouldin | Adjusted Rand Index | Normalized Mutual Information |
|---|---|---|---|---|---|---|
| LaBSE | K-Means | 0.11 | 113.17 | 3.14 | 0.62 | 0.61 |
| LaBSE | Agglomerative | 0.09 | 101.76 | 3.45 | 0.54 | 0.58 |
| LaBSE | Fuzzy C-Means | 0.11 | 113.10 | 3.14 | 0.62 | 0.6 |
| paraphrase-xlm-r-multilingual-v1 | K-Means | 0.09 | 112.09 | 2.57 | 0.32 | 0.37 |
| paraphrase-xlm-r-multilingual-v1 | Agglomerative | 0.05 | 99.41 | 2.36 | 0.28 | 0.41 |
| paraphrase-xlm-r-multilingual-v1 | Fuzzy C-Means | 0.09 | 110.75 | 2.54 | 0.34 | 0.39 |
| paraphrase-multilingual-mpnet-base-v2 | K-Means | 0.08 | 96.00 | 2.97 | 0.34 | 0.38 |
| paraphrase-multilingual-mpnet-base-v2 | Agglomerative | 0.06 | 93.70 | 2.64 | 0.44 | 0.52 |
| paraphrase-multilingual-mpnet-base-v2 | Fuzzy C-Means | 0.07 | 97.50 | 2.7 | 0.42 | 0.46 |

Across all models, the performance gap between K-Means and Fuzzy C-Means remains small when embeddings are normalized and the fuzziness parameter is tuned. This convergence suggests that when thematic regions are sufficiently well-defined, soft membership modeling does not fundamentally alter the resulting partition. Instead, clustering algorithms appear to approximate the same underlying semantic geometry.

Importantly, the magnitude of variation across embedding models exceeds that within a single clustering algorithm. For example, the ARI difference between LaBSE and paraphrase-xlm-r under KMeans exceeds 0.29, whereas the difference between KMeans and Fuzzy C-Means within LaBSE is less than 0.01. This pattern demonstrates that representation quality is the primary determinant of thematic separability in multilingual clustering.

**Discussion**

The comparative analysis reveals that the primary determinant of thematic clustering quality in bilingual corpora is the geometric structure of the embedding space rather than the specific clustering algorithm applied. Across all evaluated models, clustering methods operating on the same representation consistently recovered similar macro-level partitions, while

performance varied substantially between embedding architectures.

LaBSE demonstrated the most geometrically separable semantic space. Both visual inspection and quantitative metrics indicate clearly delineated thematic regions with strong alignment to ground truth labels. The near-identical performance of K-Means and Fuzzy C-Means in this setting suggests that when thematic regions are well-structured and compact, both hard and soft partitioning strategies converge toward the same solution. This convergence implies that the semantic topology encoded by LaBSE exhibits stable, well-defined decision boundaries.

In contrast, paraphrase-xlm-r-multi-lingual-v1 produced a more diffuse embedding manifold. Although cross-lingual alignment remained intact, thematic boundaries were less sharply encoded, resulting in lower extrinsic agreement scores across all clustering algorithms. The reduced ARI values indicate that, while coarse semantic structure is present, the embedding space does not sufficiently preserve strong geometric contrast between semantically adjacent domains, particularly Science and Technology. This suggests that cross-lingual consistency alone does not guarantee optimal thematic separability.

The paraphrase-multilingual-mpnet-base-v2 model exhibited intermediate behavior. While centroid-based clustering achieved moderate alignment, Agglomerative clustering consistently outperformed K-Means and Fuzzy C-Means in extrinsic metrics. This pattern indicates that the MPNet embedding space may encode semantic gradients that are not optimally captured by spherical cluster assumptions. The improved performance of hierarchical clustering suggests that local structural relationships play a more prominent role in this representation.

An important methodological observation concerns the relationship between intrinsic and extrinsic metrics. Although Silhouette and Calinski-Harabasz values varied across models, differences in these intrinsic measures were less pronounced than differences in ARI and NMI. This indicates that geometric compactness alone does not fully predict semantic recoverability. Embedding spaces may appear structurally coherent under intrinsic evaluation while still misaligning with externally defined thematic categories. Therefore, extrinsic validation remains essential in evaluating representation quality for semantic clustering tasks.

The consistent absence of language-driven segmentation across all embedding models confirms that modern multilingual sentence encoders successfully align cross-lingual content within a shared semantic manifold. However, the extent to which this manifold preserves clear thematic boundaries varies markedly across architectures. This distinction highlights the importance of analyzing representation geometry beyond basic multilingual alignment performance.

Finally, the experiments with Fuzzy C-Means underscore the sensitivity of soft clustering methods to high-dimensional embedding properties. Without L2 normalization and careful adjustment of the fuzziness parameter, the algorithm converged to degenerate solutions characterized by diffuse memberships. After normalization and tuning, Fuzzy C-Means produced partitions closely aligned with centroid-based clustering, reinforcing the conclusion that representation geometry constrains clustering outcomes more strongly than algorithmic variation.

Overall, the findings emphasize that embedding model selection is a critical design decision in unsupervised thematic analysis of multilingual corpora. Differences in embedding geometry directly translate into differences in thematic recoverability, whereas clustering algorithms primarily serve as mechanisms for extracting structure already encoded in the representation.

**Conclusion**

This study investigated the semantic structure of a balanced bilingual corpus using three multilingual sentence-embedding models LaBSE, paraphrase-xlm-r-multilingual-v1, and paraphrase-multilingual-mpnet-base-v2 in combination with three unsupervised clustering

strategies: KMeans, Agglomerative hierarchical clustering, and Fuzzy C-Means.

The results demonstrate that the quality of thematic clustering is primarily determined by the geometry of the embedding space rather than by the specific clustering algorithm applied. Across all models, clustering algorithms operating on the same representation produced broadly similar partitions, whereas performance varied substantially between embedding models.

Among the evaluated representations, LaBSE consistently achieved the strongest alignment with ground truth thematic categories, both visually and quantitatively. Its embedding space exhibited well-defined and geometrically separable semantic regions, enabling centroid-based and soft clustering methods to recover nearly identical thematic partitions. This indicates that LaBSE encodes a stable and coherent cross-lingual semantic structure suitable for coarse-grained thematic analysis.

The paraphrase-xlm-r-multilingual-v1 model showed reduced separability, with lower extrinsic agreement metrics and less distinct cluster boundaries in the embedding space. While intrinsic measures remained moderate, the diminished alignment with thematic labels suggests that its representation geometry is less optimized for global topical discrimination.

The paraphrase-multilingual-mpnet-base-v2 model demonstrated intermediate behavior. Although centroid-based clustering achieved moderate performance, hierarchical clustering showed stronger agreement with ground truth labels, suggesting that the embedding geometry may contain meaningful semantic gradients that are better captured by hierarchical merging strategies.

Across all models, properly configured Fuzzy C-Means converged to solutions closely aligned with K-Means, confirming that when thematic regions are sufficiently structured, hard and soft clustering approaches approximate the same underlying partition. This reinforces the conclusion that representation quality outweighs algorithmic differences in determining clustering outcomes.

An additional methodological insight concerns the application of soft clustering in high-dimensional embedding spaces. Without normalization and appropriate tuning of the fuzziness parameter, Fuzzy C-Means may converge to degenerate solutions due to distance concentration effects. After L2 normalization and reduction of the fuzziness coefficient, stable and meaningful partitions were obtained, highlighting the importance of adapting clustering configurations to embedding geometry.

Finally, in all experimental settings, cluster formation was driven by semantic similarity rather than language identity. English and Ukrainian documents remained interwoven within thematic regions, confirming effective cross-lingual alignment across all evaluated models, though the degree of semantic separability varied substantially.

Overall, the findings indicate that modern multilingual embedding models differ significantly in their ability to encode geometrically separable thematic structures. In unsupervised thematic analysis of bilingual corpora, representation selection is a more critical design decision than the choice of clustering algorithm. Future research should further investigate the geometric properties of multilingual embedding spaces and their impact on clustering stability across varying levels of thematic granularity.

## References

1. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 878–891).
Association for Computational Linguistics.
https://doi.org/10.18653/v1/2022.acl-long.62

2. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

3. Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.
https://doi.org/10.1080/01621459.1963.10500845

4. Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

5. 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

6. Hadifar, A., Sterckx, L., Demeester, T., & Develder, C. (2019). A self-training approach for short text clustering. In I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, & M. Rei (Eds.), *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 194–199). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4322

7. Li, C., & Zhang, J. (2023). New interval improved fuzzy partitions fuzzy C-means clustering algorithms under different distance measures for symbolic data analysis. *Applied Sciences*, *13*(22), Article 12531. https://doi.org/10.3390/app132212531

8. Abdulkareem, A. (2025). Unsupervised fake news detection on social media using hybrid Gaussian mixture model. *PLOS ONE*, *20*(8), Article e0330421. https://doi.org/10.1371/journal.pone.0330421

9. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

10. Mehta, V., Bawa, S., & Singh, J. (2021). WEClustering: Word embeddings based text clustering technique for large datasets. *Complex & Intelligent Systems*, *7*, 1–14. https://doi.org/10.1007/s40747-021-00512-9

11. Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2025). Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, *6*, 100–108. https://doi.org/10.1016/j.ijcce.2024.11.004

12. Wu, L., Li, R., & Lam, W.-H. (2023). *Research on multilingual news clustering based on cross-language word embeddings*. arXiv. https://doi.org/10.48550/arXiv.2305.18880

13. Wada, T., & Iwata, T. (2018). *Unsupervised cross-lingual word embedding by multilingual neural language models*. arXiv. https://doi.org/10.48550/arXiv.1809.02306

14. Schuster, T., Ram, O., Barzilay, R., & Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1599–1613). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1162

15. Tamine, L., Amigó, E., & Mothe, J. (2023). Report on the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022). *SIGIR Forum*, *56*(2), Article 9. https://doi.org/10.1145/3582900.3582913

16. OpenAI. (2025, August 7). *GPT-5 system card*. https://cdn.openai.com/gpt-5-system-card.pdf

17. Liberty, E., Lang, K., & Shmakov, K. (2016). Stratified sampling meets machine learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning: Vol. 48. Proceedings of Machine Learning Research* (pp. 2320–2329). PMLR. https://proceedings.mlr.press/v48/liberty16.html

18. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Edunov, S., Stoyanov, V., & Zettlemoyer, L. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

19. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

20. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Article 1414). Curran Associates Inc.

21. 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

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

23. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227.

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

25. Strehl, A., & Ghosh, J. (2002). Cluster ensembles -a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

https://www.jmlr.org/papers/volume3/strehl02a/strehl02a pdf

26. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*. https://doi.org/10.48550/arXiv.1802.03426

The article has been sent to the editors 18.02.26.
After processing 29.02.26.
Submitted for printing 30.03.26.