Hugo, Moreira
hugoalmeidamoreira@gmail.com

Sérgio, Vilela
jose.sergio.vilela@tecnico.ulisboa.pt

ABSTRACT

This report presents a systematic exploration of unsupervised learning architectures applied to two distinct analytical domains: Portuguese media discourse regarding Artificial Intelligence (Dataset A) and Customer Personality segmentation (Dataset B).

For Dataset (A), we implemented an advanced semantic pipeline utilizing transformer-based embeddings (`Qwen2.5`) and `UMAP` dimensionality reduction to construct a "semantic topography." This work introduces a robust three-step unsupervised outlier identification pipeline—addressing global, local, and structural anomalies—to isolate high-density narrative cores. Pattern discovery was further enhanced through zero-shot `LogProb` centrality scoring, allowing for a multi-dimensional "blueprint" analysis of thematic intensity. Parallelly, for Dataset (B), we conducted a comparative analysis of partitioning and hierarchical clustering, demonstrating that targeted feature selection significantly enhances cluster separation, increasing the Silhouette score from 0.16 to 0.40. Outlier detection benchmarking in Dataset (B) using Mahalanobis Distance, LOF, and Isolation Trees further identified specific niche behaviors while preserving dataset integrity.

Comparative results show that while Dataset (A) benefits from high-dimensional topology for complex discourse discovery, Dataset (B) relies on rigorous feature engineering for distinct segmentation. Both domains validate that hybrid regimes, combining manifold learning with domain-specific pruning, yield more stable and interpretable categorical insights than baseline approaches.

Keywords: Unsupervised Learning, LLMOps, Semantic Topography, Feature Selection, Clustering, AI News Analysis.

1. DATA PROFILING

1.1 Descriptive statistics

Dataset (A) - AI News The dataset(A) comprises 11,922 portuguese news articles published between 2022 and 2024. No null values are present and every record contains the expression "Inteligência artificial" and/or "AIAct" in either the title or the description. For this work, we focus on the `title` and `description` fields, which correspond to the news headline and its content, respectively.

Table 1: Descriptive statistics of the dataset(A).

Variable	Mean	Med.	Std	O.1-3σ	O.$> 3\sigma$
`char_count`	6518.1	3937.5	9389.0	6.0%	1.6%
`ai_mentions`	2.05	1.0	2.83	7.4%	1.6%

Given the LLM-intensive nature of this project, monitoring text length and keyword frequency is useful for managing computational costs and context window limits. Table 1 summarizes metrics for the character count and AI-related mentions within the news descriptions.

The character count shows a large amplitude and extreme skewness, which is explained by the presence of long-form articles. However, since this work is not strictly stochastic in nature, we found no need to exclude these outliers based on length. Although that decision can be made if on a computational budget.

Dataset (B) - Customer Analysis The dataset in study is the publicly available Customer Personality Analysis present in the following Kaggle URL.

A brief analysis is presented below:

- Dataset has 2240 rows and 28 columns including the class "Response".
- 24 numeric variables and 3 symbolic and a binary class variable.
- Variable **Income** contains 24 missing values.
- Class distribution is 331 true and 1881 false observations

Prior to any clustering or outlier analysis the following data preparation steps were followed:

1. Feature selection analysis through low variance. Variables identified for later removal during clustering and outlier analysis
2. Rows with missing values were dropped
3. Date column was dropped due to ot being relevant in this process

4. InterQuartile Range (IQR) and standard Deviation (stdev) outlier analysis identified that all the dataset's variables contain multiple observations that could be identified as outliers

5. Data was scaled according to the MinMax approach

2. DATA REPRESENTATION

Dataset (A) - AI News Analysis

2.1 Embedding Methodology

Textual data was processed for embedding using the `Qwen2.5-8B-Instruct` model, generating high-dimensional vectors ($D = 4096$) that capture both structural and semantic nuances.

Vector generation was orchestrated via the `vLLM` inference engine, optimized for Blackwell architecture with FP16 batching. The resulting embeddings are persisted in a `PostgreSQL 16.x` instance equipped with the `pgvector` extension ($v0.7.0$).

2.2 Dimensionality Reduction

To mitigate the curse of dimensionality inherent to 4096-dimensional vectors and improve clustering density estimation, we employ **UMAP** (Uniform Manifold Approximation and Projection) [8]. The reduction is twofold:

1. **Topological Projection ($D = 5$)**: Vectors are reduced to 5 components to serve as the input space for density-based clustering (HDBSCAN). This dimension aligns with the intrinsic dimensionality of the dataset, estimated at $d \approx 4.11$ using the TwoNN algorithm [5]. This step counteract sparsity in high-dimensional space which degrades density estimation [8], preserving local manifold structure while discarding noise.

2. **Visual Projection ($D = 2$)**: A further reduction to 2 components is generated strictly for visualization purposes, creating a "semantic topography" of the news landscape (Fig. 1).

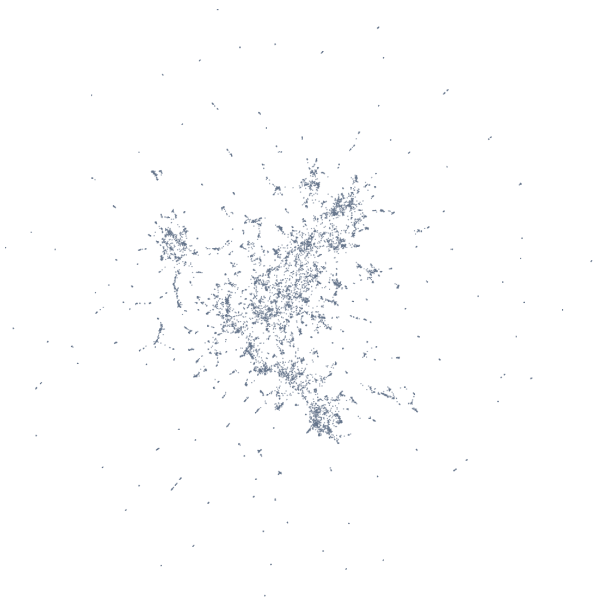


Figure 1: Semantic topography of the dataset(A) using UMAP (D=2).

The UMAP manifold is constructed using the *Cosine* metric with $n\_neighbors = 15$ and $min\_dist = 0.1$, ensuring that the angular relationships captured by the Qwen2.5 model are preserved in the lower-dimensional space.

2.3 Data Transformations: LogProb Centrality

To enrich the representation beyond raw embeddings, we apply a domain-guided transformation to assess the thematic centrality of each news item.

The centrality scoring employs a zero-shot generative prompt-weighting technique using the `Qwen2.5-7B-Instruct` model. Each news item is evaluated against the target theme to produce a continuous metric $S \in [0, 1]$ leveraging the **logprobs** extraction methodology as described in *"The Mean-Difference"* [10].

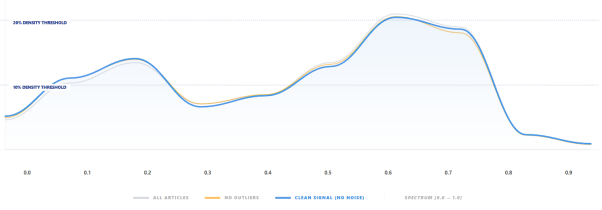


Figure 2: Semantic topography overlaid with LogProb Centrality scores.

As illustrated in Fig. 2, the dataset exhibits a strong semantic centrality towards the theme "Artificial Intelligence". The majority of the news items cluster in the high-confidence region ($S > 0.7$), validating the effectiveness of the initial keyword-based filtering

while providing a nuanced gradient of relevance for items on the conceptual periphery.

This methodology was extended, but with `mesolitica/Qwen2.5-72B-Instruct-FP8` (because the scaling problem is more soft and needs a model with more semantic capacity), to map seven additional semantic dimensions of the news identity: *Opportunity vs. Risk*, *Regulatory Pressure*, *Economic Momentum*, *Ethics vs. Utility*, *Technical Depth*, *Geopolitical Scope*, and *Urgency*. The integrated summary and complete distribution of these dimensions are available in Appendix and .

Dataset (B) - Customer Personality Analysis

...

3. CLUSTERING

Dataset (A) - AI News Topography

3.1 Density-Based Approach (HDBSCAN)

We applied **HDBSCAN** ($min\_cluster\_size = 30$, $min\_samples = 10$, EOM) to identify high-density semantic cores. However, applied to the 11,922 articles, this approach classified 5450 items as noise.

This high rejection rate highlights the extreme volatility and rotation of the AI media narrative, where "AI" is frequently referenced in transient or disperse contexts that defy dense aggregation. Consequently, HDBSCAN was discarded as the primary segmentation tool in favor of K-Means to ensure a comprehensive structural mapping of the domain. The density-based results were retained solely for the outlier analysis discussed in Section 4.

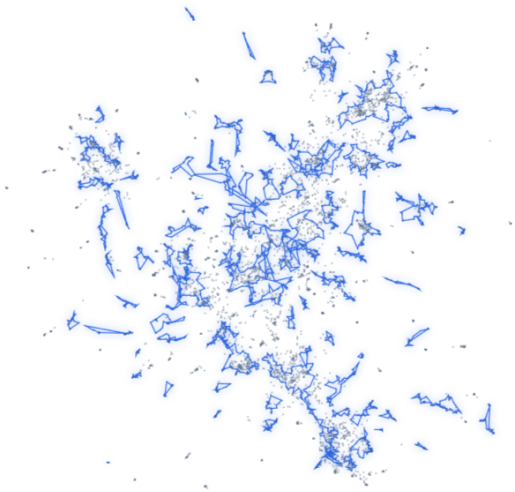


Figure 3: Density-based clustering (HDBSCAN) revealing only the strongest semantic cores, excluding 45% of the corpus as noise.

The visual comparison between the raw clustering with noise (Appendix A2) and the pruned structure (Fig. 3) highlights the effectiveness of this pipeline in distilling the core narrative.

3.2 Structural Segmentation (K-Means)

After the initial density-based approach, we employed **K-Means** clustering with $K = 15$.

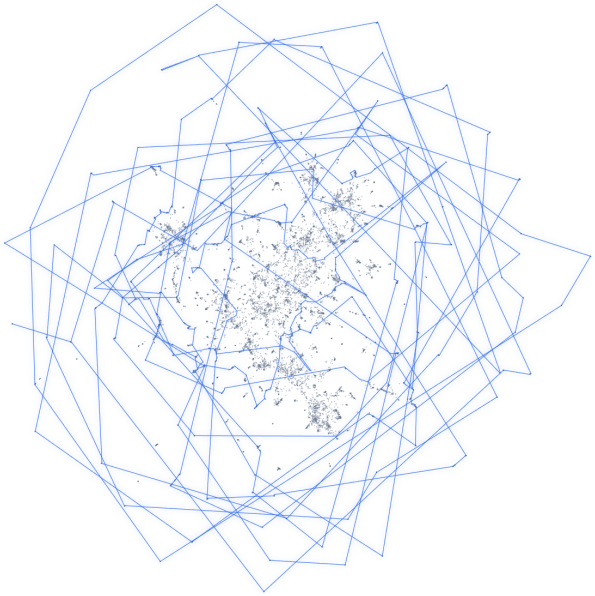


Figure 4: K-Means partitioning of the semantic topography. The final solution refined the initial 15-cluster sweep into 13 stable thematic regions.

Detailed visualizations of the initial regions identified are provided in Appendix .

3.3 Major Findings (Knowledge Acquisition)

Analyzing the topology of these clusters proves to be an effective solution for mapping the semantic space of the Portuguese media ecosystem regarding Artificial Intelligence. This approach reveals a highly fragmented landscape, where core narratives (e.g., Institutional Regulation, Technical Enterprise Solutions) are surrounded by a vast periphery of transient or niche reports.

The hybrid strategy of using density-based methods for noise identification followed by K-Means for structural mapping proved essential. It allowed for the preservation of narrative clusters that are semantically distinct but not sufficiently dense for traditional

density-based algorithms. Ultimately, the 15 identified regions provide a stable and interpretable framework for monitoring AI discourse and its diverse manifestations in Portuguese news.

Dataset (B) - Customer Personality Analysis

3.4 Reference clustering solutions

Clustering in dataset 2 was tested with the following approaches:

- K-means clustering
- Hierarchical clustering

As mentioned in Section 1, the dataset went through a feature selection process and both the scenarios were tested in the above mentioned approaches.

3.5 Visualization and description

Dataset 2 clustering related visualizations are presented in Figures 5 and 6. In these figures, four scenarios are shown:

1. Elbow method - Process used to identify k number of clusters through cohesion analysis.
2. Silhouette score - Process used to identify k number of clusters through silhouette analysis.
3. K-means cluster (k=3) visualization including medians and medoids.
4. Hierarchical cluster (k=3) visualization including medians and medoids.

3.6 Distances and methods

3.6.1 D2 | Pre-feature selection dataset

Elbow Method Analysis: As seen in Figure 5a), Cohesion (or inertia) decreases sharply from k=1 to k=3. After k=3, the rate of decrease becomes less significant, forming an 'elbow' around k=3 and k=4.

Silhouette Score Analysis: As seen in Figure 5b), the silhouette score is highest at k=2, and k=3, and then generally decreases or fluctuates at higher k values.

Considering both methods, k=3 should be the choice for the optimal number of clusters.

3.6.2 D2 | Post-feature selection dataset

Elbow Method Analysis: As seen in Figure 6a), the cohesion decreases sharply from k=1 to k=2. After k=2, the rate of decrease becomes less significant, forming an 'elbow' around k=2 or k=3.

Silhouette Score Analysis: As seen in Figure 6b), the silhouette score is highest at k=2 (approximately 0.39) and then it decreases for higher k values.

Considering both methods, k=2 was selected as the optimal number of clusters for the feature-selected data given that the elbow plot shows a clear bend at k=2, and the silhouette score is highest at k=2.

3.7 Number of clusters

As discussed in the previous section, the number of clusters selected for dataset 2 was k=3 for the pre-feature selection data and k=2 for post-feature selection.

3.8 Preprocessing impact

Results for dataset 2 approaches analysis can be found in the table below. It can be seen that the silhouette value has increased which suggests strong clustering and distinct separation between clusters.

Moreover, in the case of the K-means its cohesion has reduced which suggests that points are close to their respective centroids, suggesting compact clusters.

In conclusion, there was an overall improvement after preprocessing for both the K-means and Hierarchical approaches.

Approach	Metric	Original	After
K-means	Silhouette	0.160	0.397
K-means	Cohesion	1999.8	902.09
Hierarchical	Silhouette	0.139	0.397

3.9 Detailed assessment

In dataset 2 scenario the following overall conclusions can be taken of the clustering analysis exercise:

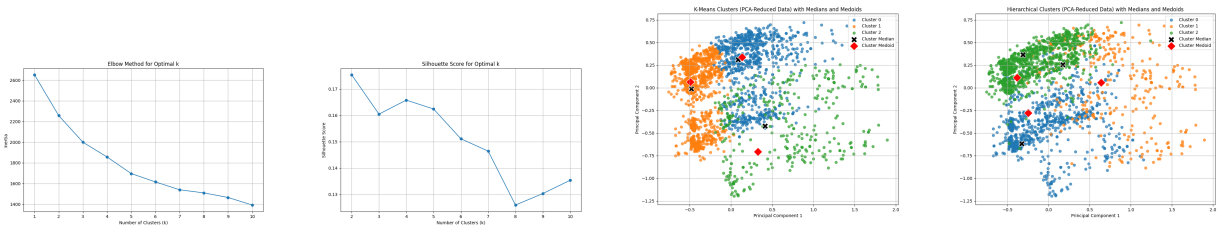
For the K-Means clustering approach: With the Original data the (k=3) clusters defined appeared somewhat overlapping and less distinct with Medians and Medoids' plot, as seen in the 5c).

Figure 6c) which was developed from pre-processed data shows a reduction in clusters (k=2) with much clearer separation and distinct boundaries, indicating more compact and well-defined groups.

For the Hierarchical clustering approach: With the Original Data the (k=3) clusters in the Figure 5d) were not perfectly separated.

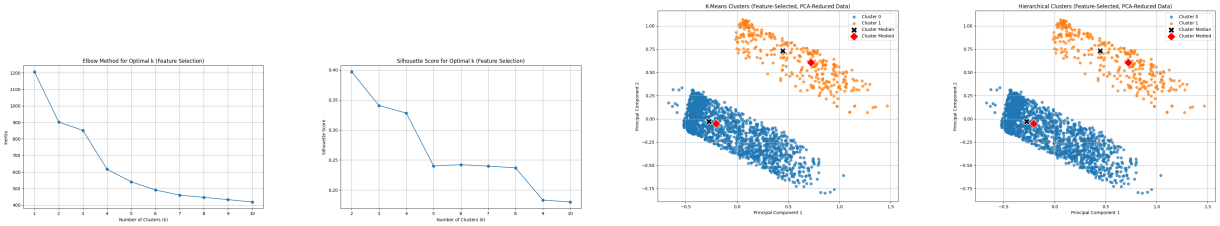
Figure 6d) which was developed from pre-processed data shows a reduction in clusters (k=2) with better visual separation, aligning with the improved Silhouette Score.

Figure 5: D2 | K-Means and Hierarchical cluster analysis



(a) Elbow method (b) Silhouette score (c) K-Means cluster (d) Hierarchical cluster

Figure 6: D2 | K-Means and Hierarchical cluster analysis - Post feature selection



(a) Elbow method (b) Silhouette score (c) K-Means cluster (d) Hierarchical cluster

4. OUTLIER/PATTERN ANALYSIS

Dataset (A) - AI News Topography To ensure the integrity of the semantic clusters and understand the peripheral boundaries of the news landscape, we implemented a three-step unsupervised multivariate outlier analysis.

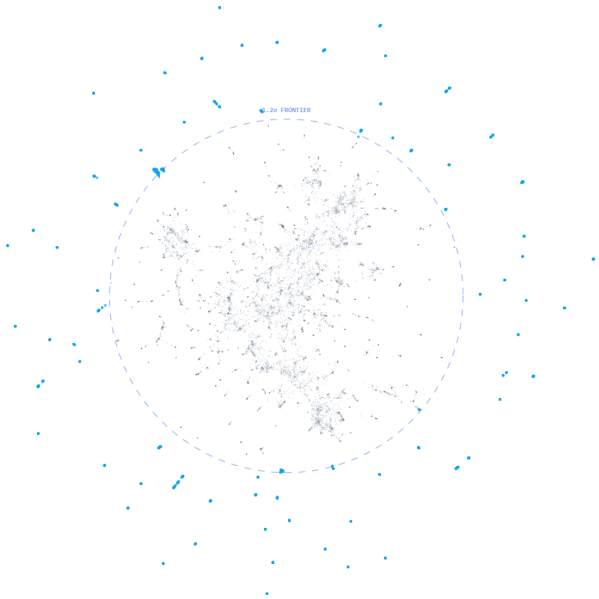


Figure 7: Global outlier detection showing items beyond the 1.2σ continental radius.

4.1 Step 1: Global Outlier Detection

The initial stage identifies "global strays" — news items positioned significantly far from the primary semantic body of the dataset. This is achieved by calculating the Euclidean distance of each article from the global centroid within the 2D UMAP projection [9].

Observations exceeding a threshold of **1.2σ (Continental Radius)** are flagged as global outliers. This criterion successfully isolates **1282** objects, representing highly niche or semantically isolated news reports.

4.2 Step 2: Local Maverick Extraction

Following the global sweep, we performed a cluster-conditional analysis within each K-Means region. This step identifies "Mavericks" — articles that belong to a specific thematic cluster but are positioned as outliers relative to that cluster's local centroid.

The methodology mirrors the global approach, calculating Euclidean distances within the local reference frame of each cluster but with a threshold of **1.8σ**. This localized pruning identified **944** mavericks that

represent conceptual deviations within otherwise cohesive narratives.

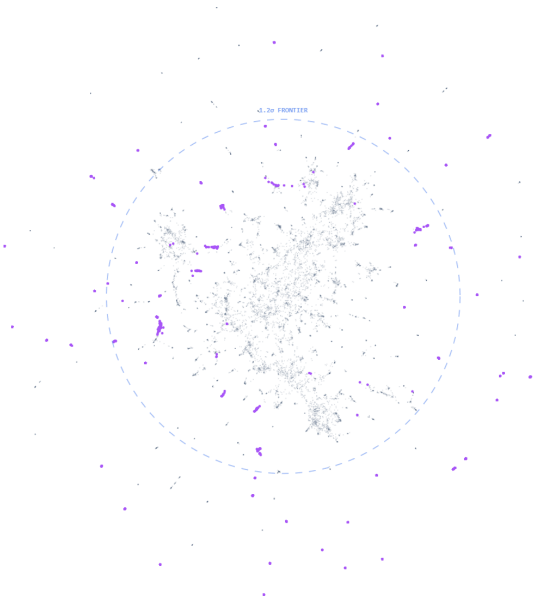


Figure 8: Local outlier detection (Mavericks) using a 1.8σ threshold within each cluster.

4.3 Step 3: Graph-Based Structural Pruning

The final validation phase utilizes a graph-based connectivity analysis to isolate **structural outliers**. This technique, inspired by the SCAN algorithm [11], prunes news items that, despite passing local and global filters, lack sufficient proximity to the high-density cores of the K-Means clusters.

By evaluating the reachability of each object within the semantic manifold, this step identifies artifacts that effectively function as "fragile bridges" or bridge-noise between narratives. This ensures that the final segmentation captures the most representative and stable clusters of the Portuguese media sphere.

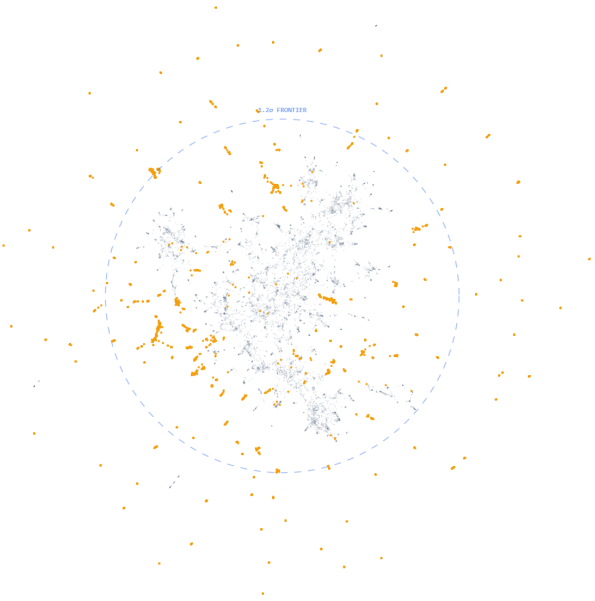


Figure 9: Structural outlier identification using graph-based reachability analysis.

4.4 Density-Based Semantic Noise (HDBSCAN)

We identified **Semantic Noise** using the non-aggregated objects from the HDBSCAN pipeline. This layer comprises approximately **5,450** items that defy dense structural grouping. This noise was not pruned to preserve the broader K-Means topology, but can be useful in some applications, for example reducing semantic redundancy.

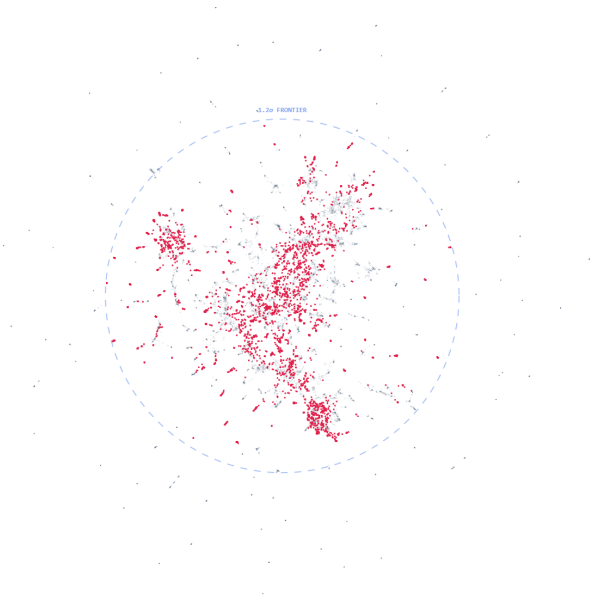


Figure 10: Distribution of Semantic Noise (non-aggregated objects) across the UMAP topography.

4.5 Final Assessment & Impact

The three-step outlier pipeline described above successfully isolated **2565** unique objects from the corpus. While the individual steps (global, local, and structural) operate on different semantic scales, their juxtaposition ensures a multi-layered validation of the news domain's boundaries.

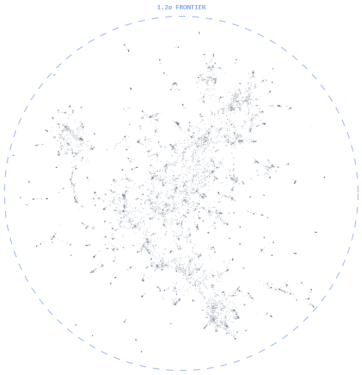


Figure 11: Final semantic topography after the integrated 3-step pruning and structural validation.

The final visual result of this integrated pruning is shown in Fig. 11. This process resulted in a significant improvement in the structural mapping (Fig. 12), where the semantic clusters exhibit higher thematic cohesion and more distinct boundaries.

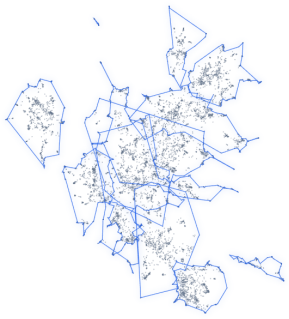


Figure 12: Impact of outlier removal on K-Means ($K = 15$) stability and cluster definition.

A detailed visualization of the refined semantic core regions is provided in Appendix , representing the definitive structural mapping of the Portuguese AI news landscape.

Dataset (B) - Customer Personality Analysis

4.6 Reference pattern/outlier solutions

...

Outlier analysis in dataset 2 was tested with the following approaches:

- Mahalanobis Distance - Detects global outliers. It measures how far a point is from the mean of the entire distribution (or a cluster's mean, if applied per cluster), taking into account the covariance structure of the data. Outliers are points that deviate significantly from the central tendency of the overall dataset. It assumes a multivariate Gaussian distribution for effective detection.

- Local Outlier Factor (LOF) - This method detects local outliers. It quantifies the degree of isolation of a data point with respect to its neighbors. A point is considered an outlier if its local density is significantly lower than that of its neighbors.

- Isolation Tree - It works on the principle that anomalies are few and different and therefore easier to isolate than normal data points. The algorithm builds a number of isolation trees. Given their difference, outliers typically require fewer random partitions to be isolated.

As mentioned in Section 1, the dataset went through a feature selection process and both the scenarios were tested in the above mentioned approaches.

4.7 Preprocessing impact

...

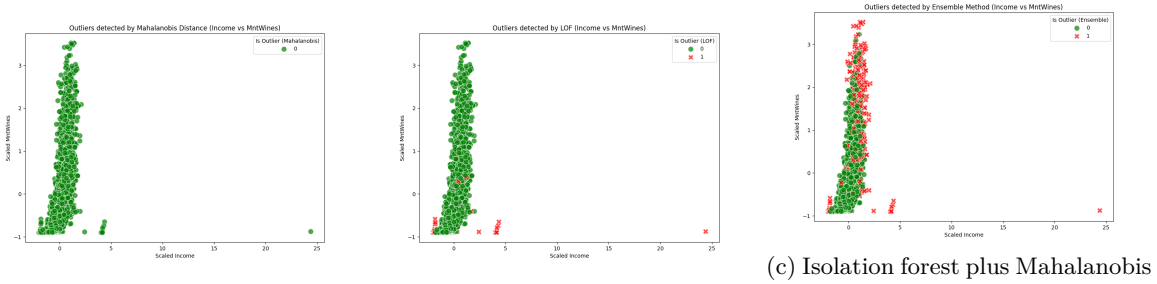
Same preprocessing was applied as in the previous section.

4.8 Class-conditional outliers/patterns (optional)

...

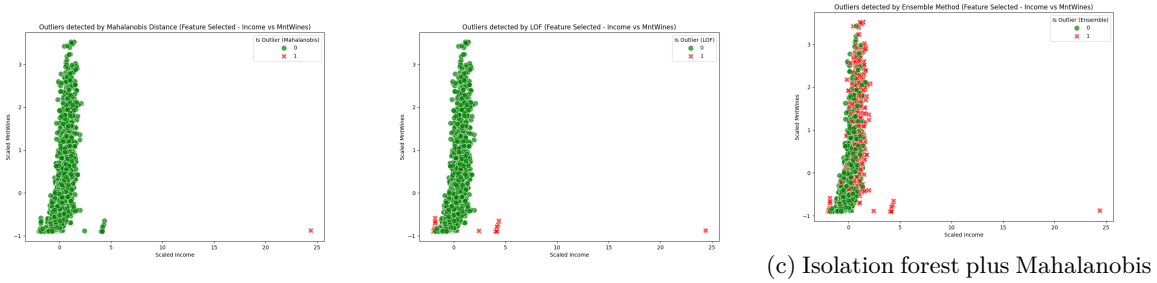
See Figures 13 and 14. Insights obtained from these results can be seen in the next sections.

Figure 13: D2 | Outliers Methods comparison - Income/Mntwines



(a) Mahalanobis distance (b) Local Outlier Factor (c) Isolation forest plus Mahalanobis and LOF

Figure 14: D2 | Outliers Methods comparison after preprocessing - Income/Mntwines



(a) Mahalanobis distance (b) Local Outlier Factor (c) Isolation forest plus Mahalanobis and LOF

4.9 Detailed assessment

...

As seen in Figures 13 and 14 the tree methods in study have severe differences in terms of results.

In Figure 13a) and 14a) when using the distance method only the most distant scenario is identified as an outlier (and only in the preprocessed data). This is aligned with the definition of the method which states that this approach is good for global outliers.

In Figure 13b) and 14b) the LOF method improves upon the distance method by finding outliers closer to the majority of observations. Of particular interest are the few outliers detected inside the green cluster of observations which were reduced in the preprocessed data.

In Figure 13c) and 14c) the Isolation forest was used and its results were added to the previous two methods. This approach has identified a large number of outliers when comparing to the previous two methods. While this is an advantage it is unclear if all of the outliers detected are valid results.

4.10 Major findings (knowledge acquisition)

In dataset 2 the ensembled approach of joining the three methods together can be a way of identifying further amounts of outliers, however, the fact that the LOF and distance methods are more conservative in this scenario can also be an advantage.

Given that for this dataset the looks mostly clean the goal will be mostly to remove exceptional data points and understand niche segments meaning that the few Mahalanobis and LOF outliers should be the ones to be used. The isolation forest approach is better for a high-level cleanup of data which is not the present scenario. ...

5. PATTERN DISCOVERY

Dataset (A) - AI News Topography In this phase, we transition from anomaly detection to the identification of latent semantic patterns within the media ecosystem. Our approach leverages zero-shot **LogProb** scores obtained from the transformer-based representation layer to "paint" the semantic manifold, revealing the thematic intensity and categorical distribution of narratives.

5.1 Thematic Centrality Patterns

The **Semantic Centrality** score serves as a proxy for thematic density. By projecting these LogProb weights onto the topography (Fig. 15), we identify the "gravitational centers" of the Portuguese AI discourse.

Areas of high centrality coincide with stabilized nar-

ratives where "Artificial Intelligence" is the primary subject, while the periphery reflects a more incidental or transient usage of the term.

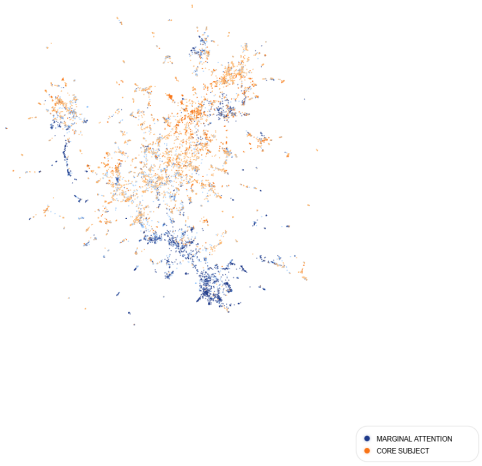


Figure 15: Topographic distribution of Semantic Centrality, highlighting the core thematic intensity.

5.2 Semantic Identity (Blueprint) Analysis

To further express the narrative structure, we used the semantic "Blueprint" previously presented in data representations section. This blueprint of the corpus across seven critical dimensions: *Economic Momentum*, *Ethics vs. Utility*, *Regulatory Pressure*, *Opportunity vs. Risk*, *Geopolitical Scope*, *Technical Depth*, and *Urgency*.

Our analysis focuses on the **semantic extremes** — objects scoring below **0.25** or above **0.75** for each dimension. This filtering reveals the most polarized and conceptually pure instances of each trait, mapping the ideological boundaries of the conversation. A detailed gallery of all dimensions and the integrated blueprint summary is provided in Appendix .

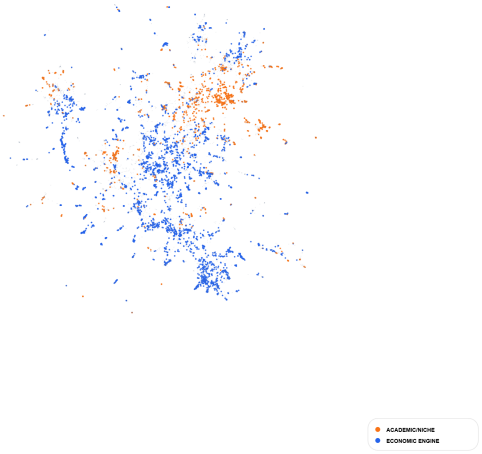


Figure 16: Example of a polarized Blueprint dimension: Economic Momentum distribution (< 0.25 and > 0.75).

6. MAJOR FINDINGS & CONCLUSIONS

The integration of advanced clustering techniques with LLMOps methodologies has shown significant potential for the systematic preparation and deep analysis of large-scale textual corpora.

Our findings indicate that a **three-step unsupervised outlier identification pipeline** — targeting semantic anomalies at global, local, and structural levels — is a highly effective and robust mechanism for pruning noise and isolating stable narrative cores. Furthermore, the use of **LogProb-based pattern identification** revealed to be useful for topological painting of the semantic manifold. This approach allowed a clear "semantic blueprint" that defines the conceptual character and ideological boundaries of each identified region within the media ecosystem that can be used in diverse applications.

REFERENCES

References

[1] Charu C. Aggarwal. *Outlier Analysis*. Cham: Springer, 2017. 488 pp. ISBN: 978-3-319-47577-6.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. *Representation Learning: A Review and New Perspectives*. Apr. 23, 2014. DOI: 10.48550/arXiv.1206.5538. arXiv: 1206.5538

[cs]. URL: http://arxiv.org/abs/1206.5538 (visited on 01/10/2026). Pre-published.

[3] Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Cham: Springer International Publishing, 2024. ISBN: 978-3-031-45467-7 978-3-031-45468-4. DOI: 10.1007/978-3-031-45468-4. URL: https://link.springer.com/10.1007/978-3-031-45468-4 (visited on 01/17/2026).

[4] Nadia Burkart and Marco F. Huber. "A Survey on the Explainability of Supervised Machine Learning". In: *Journal of Artificial Intelligence Research* 70 (Jan. 19, 2021), pp. 245–317. ISSN: 1076-9757. DOI: 10.1613/jair.1.12228. arXiv: 2011.07876 [cs]. URL: http://arxiv.org/abs/2011.07876 (visited on 01/10/2026).

[5] Elena Facco et al. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *Scientific Reports* 7.1 (2017). ISSN: 2045-2322. DOI: 10.1038/s41598-017-11873-y. URL: http://dx.doi.org/10.1038/s41598-017-11873-y.

[6] Rui Henriques and Sara C. Madeira. "Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey". In: *ACM Comput. Surv.* 51.5 (Sept. 18, 2018), 95:1–95:43. ISSN: 0360-0300. DOI: 10.1145/3195833. URL: https://dl.acm.org/doi/10.1145/3195833 (visited on 01/10/2026).

[7] *Learning Deep Representations of Data Distributions*. URL: https://ma-lab-berkeley.github.io/deep-representation-learning-book/ (visited on 01/10/2026).

[8] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018). URL: https://arxiv.org/abs/1802.03426.

[9] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

[10] Eric Wallace et al. "The Mean-Difference: A Simple and Effective Method for Zero-Shot Classification". In: *arXiv preprint arXiv:2403.14859* (2024). URL: https://arxiv.org/abs/2403.14859.

[11] Xiaowei Xu et al. "Scan: a structural clustering algorithm for networks". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 824–833.

[12] Mohammed J. Zaki and Wagner Meira Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press, 2020. 776 pp. ISBN: 978-1-108-47398-9.

APPENDIX

A. Blueprint Summary Mapping

B. HDBSCAN Noise Distribution

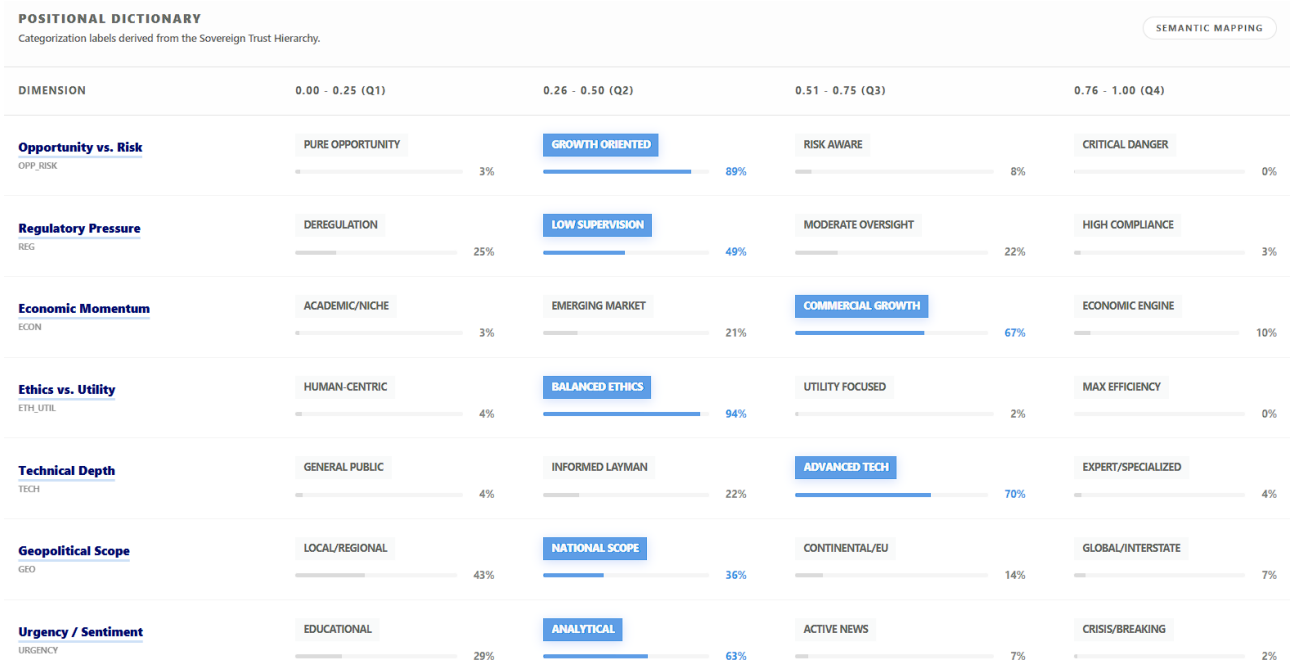


Figure A1: Integrated Blueprint summary showing the distribution of seven polarized semantic traits across the topography.

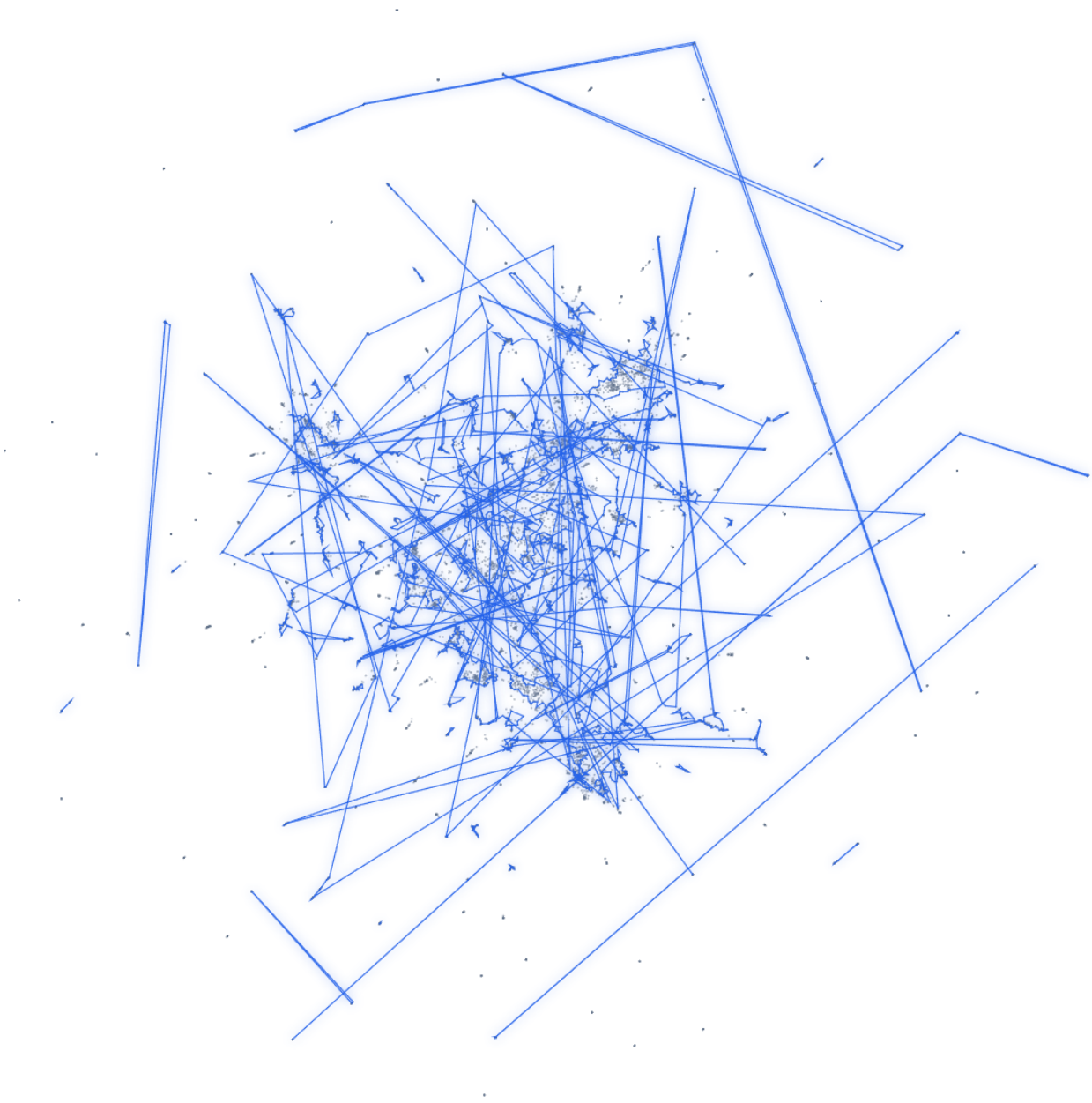
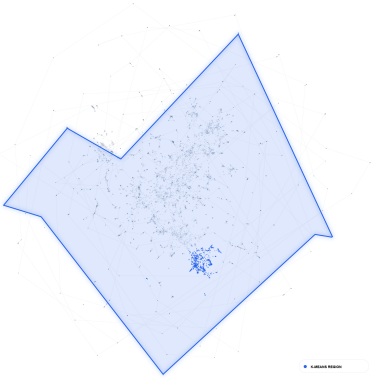
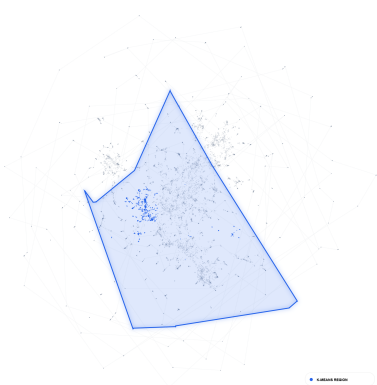


Figure A2: Full topography including peripheral noise and outliers flagged by the HDBSCAN pipeline.

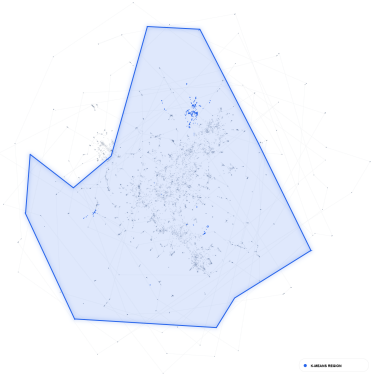
C. Structural Breakdown (Initial K-Means)



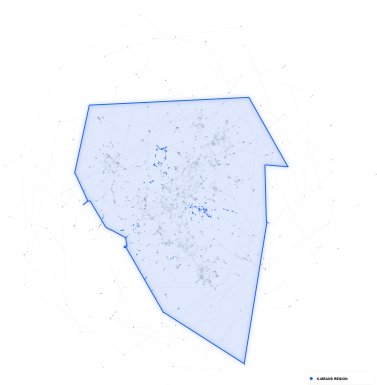
Region 0: Global Regulatory & Policy Frameworks



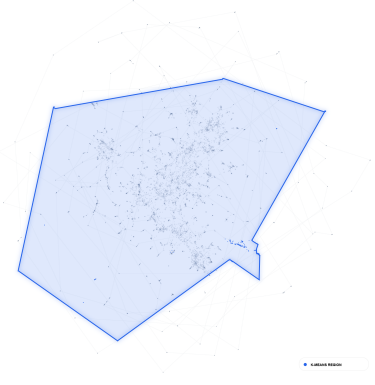
Region 1: Educational Narratives



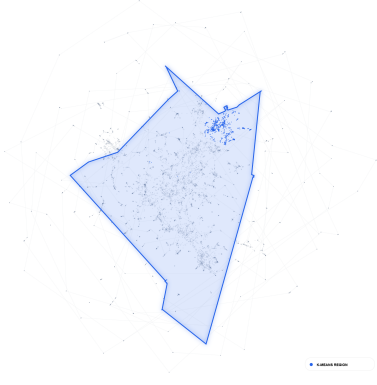
Region 2: Institutional Alignment & Compliance



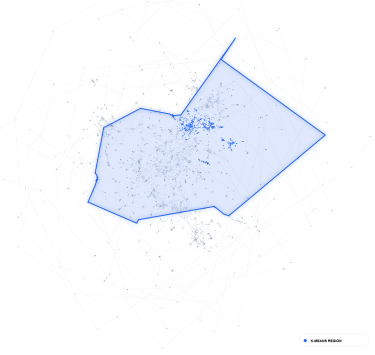
Region 3: Market Expansion



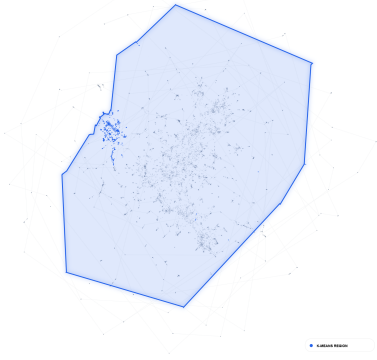
Region 4: Localized Innovation



Region 5: Continental Strategy

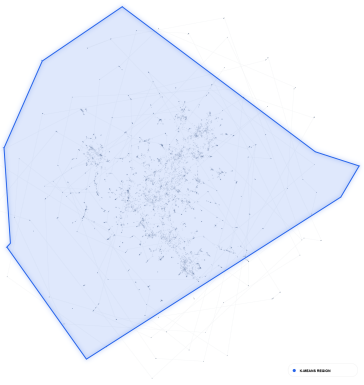


Region 6: Human-Centric Literacy

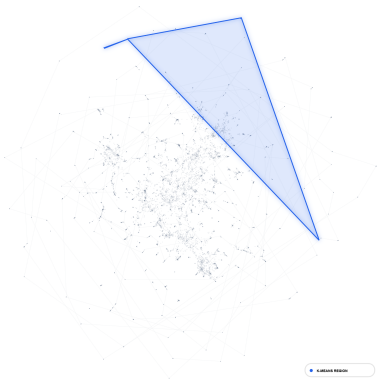


Region 7: Enterprise Solutions

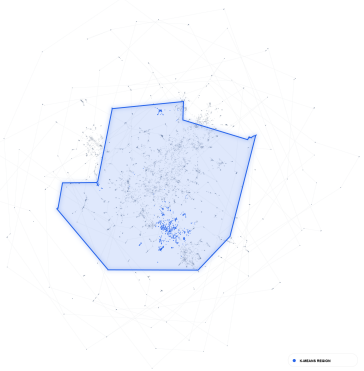
Figure A3: Initial structural breakdown of semantic regions (0-7).



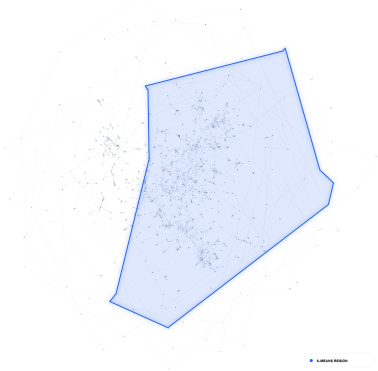
Region 8: Utility & Service AI



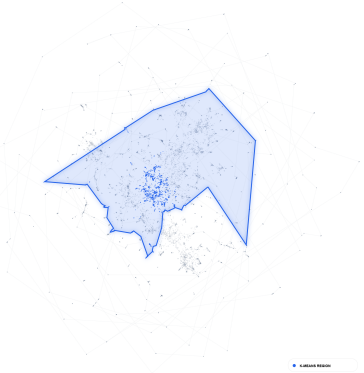
Region 9: Algorithmic Optimization



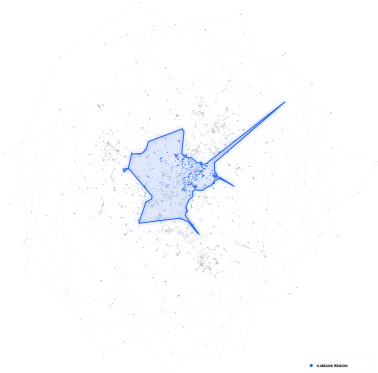
Region 10: R&D Infrastructure



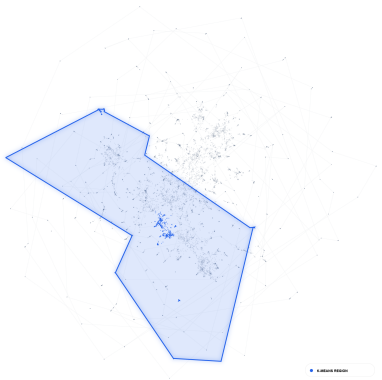
Region 11: Global Tech Frontier



Region 12: Synthetic Intelligence



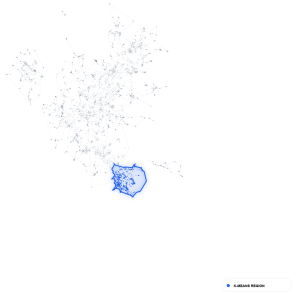
Region 13: Economic Transformation



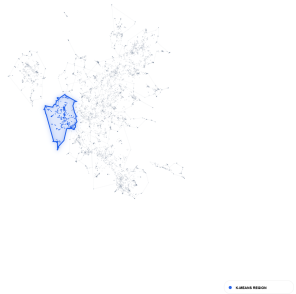
Region 14: Cross-Border Integration

Figure A4: Initial structural breakdown (8-14).

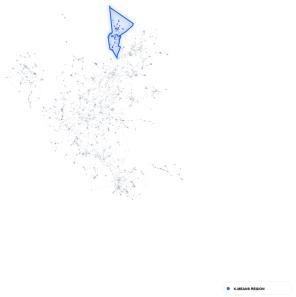
D. Refined Semantic Cores (Final Solution)



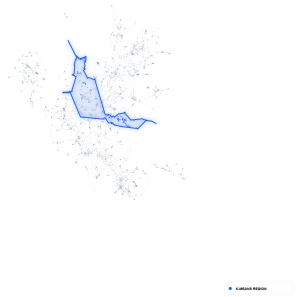
Region 0: Regulatory & Policy Frameworks



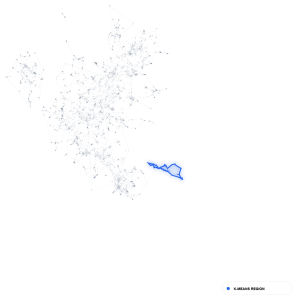
Region 1: Educational Narratives



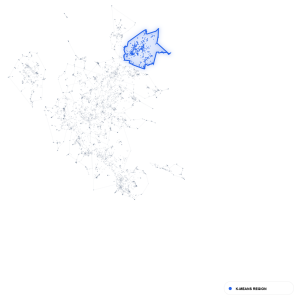
Region 2: Institutional Alignment & Compliance



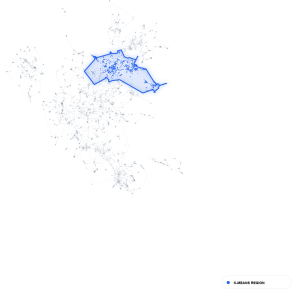
Region 3: Market Expansion



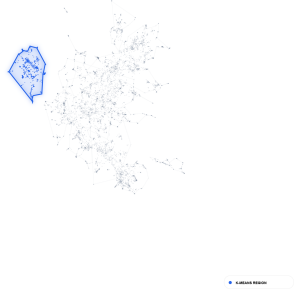
Region 4: Localized Innovation



Region 5: Continental Strategy

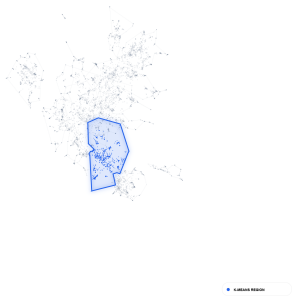


Region 6: Human-Centric Literacy

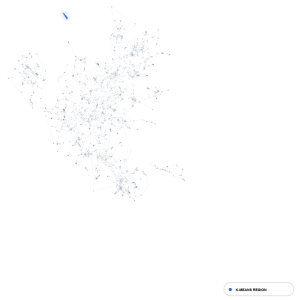


Region 7: Enterprise Solutions

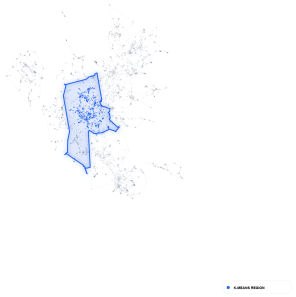
Figure A5: Structural breakdown of optimized core regions (0-7).



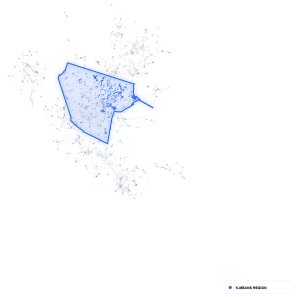
Region 10: R&D Infrastructure



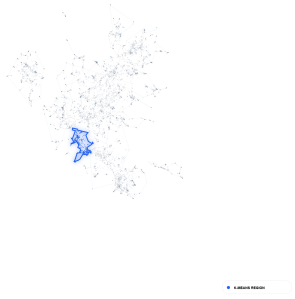
Region 11: Global Tech Frontier



Region 12: Synthetic Intelligence



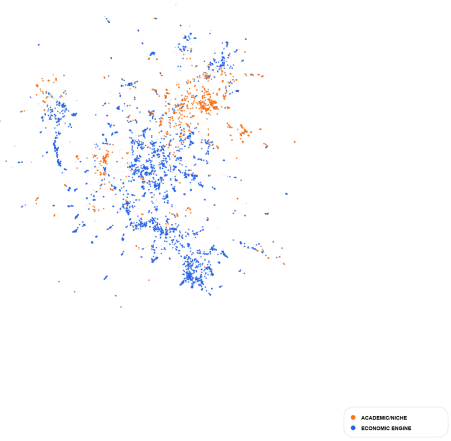
Region 13: Economic Transformation



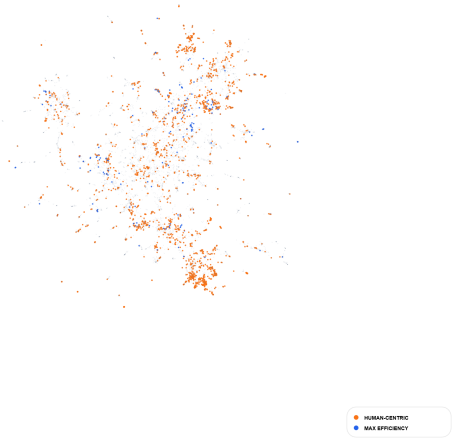
Region 14: Cross-Border Integration

Figure A6: Structural breakdown of optimized core regions (10-14).

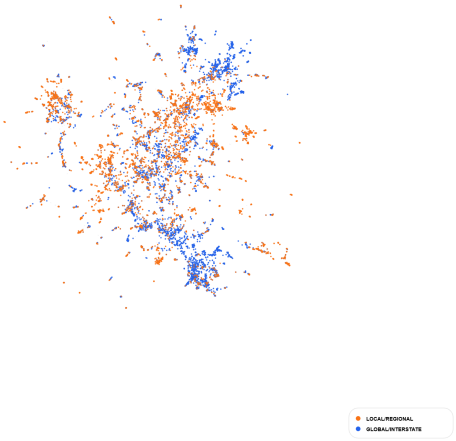
E. Semantic Blueprint Dimensions Gallery



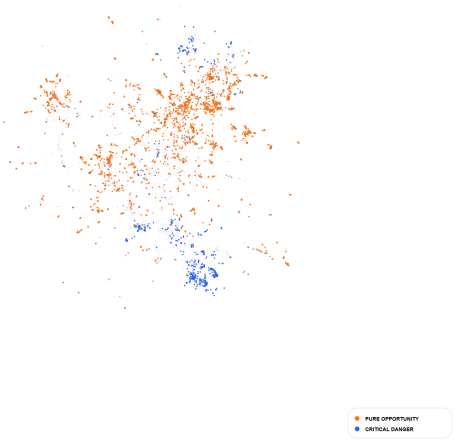
Dimension: Economic Momentum



Dimension: Ethics vs. Utility

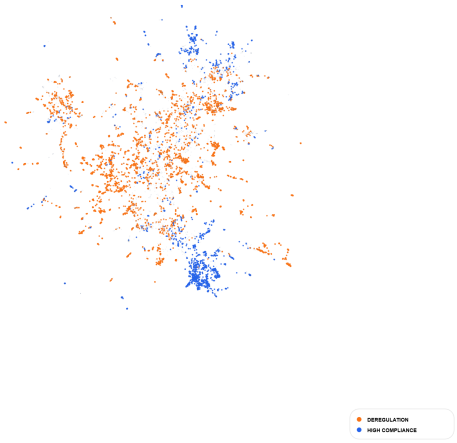


Dimension: Geopolitical Scope

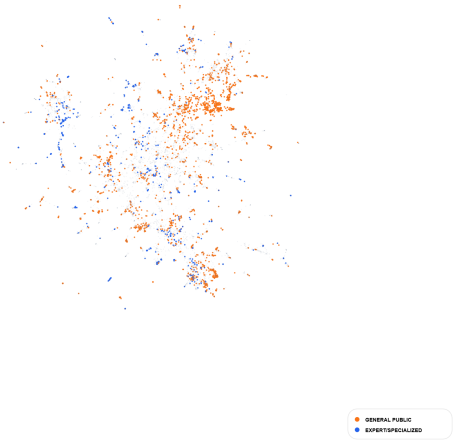


Dimension: Opportunity vs. Risk

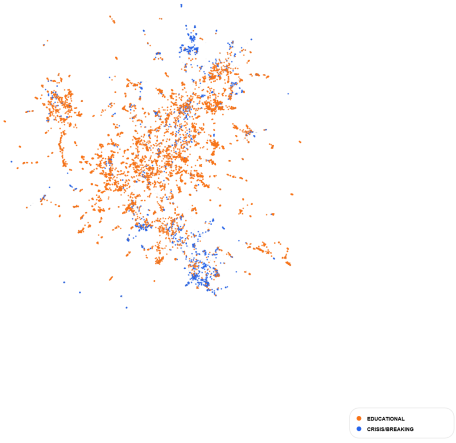
Figure A7: Breakdown of Semantic Blueprint dimensions (1-4).



Dimension: Regulatory Pressure



Dimension: Technical Depth



Dimension: Urgency / Sentiment

Figure A8: Detailed breakdown of the remaining Semantic Blueprint dimensions (< 0.25 and > 0.75).