# Web-Based Data Enrichment Using Multi-Agent AI for Enhanced Company Clustering

Yagiz Erdem*†, Sahin Uregil*†, Sultan Turhan‡, Günce Keziban Orman‡, Doğa Yağmur Yılmaz†
†*Technology&Innovation*, *Kariyer.net*, Istanbul, Turkey
Email: yagiz.erdem@kariyer.net, sahin.uregil@kariyer.net, doga.yilmaz@kariyer.net
‡*Department of Computer Engineering*, *Galatasaray University*, Istanbul, Turkey
Email: sturhan@gsu.edu.tr, korman@gsu.edu.tr
*Equal contribution

***Abstract*—Traditional company clustering approaches relying on manually curated taxonomies often fail to capture true business operations, particularly for innovative companies in rapidly evolving markets. This study presents an automated, semantically-aware approach for clustering companies based on their actual business operations. The analysis is based on a dataset primarily composed of Turkish and multinational companies, curated from the corporate listings of Kariyer.net. Our methodology employs a multi-agent artificial intelligence (AI) system to enrich company datasets with product descriptions through web-based information retrieval, then generates semantic embeddings using transformer models, followed by dimensionality reduction and clustering analysis. Through systematic evaluation of 14 configurations across different clustering algorithms, distance metrics, and dimensionality reduction techniques, we demonstrate that Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with Euclidean distance on t-distributed Stochastic Neighbor Embedding (t-SNE)-reduced 2D embeddings achieves optimal performance. The resulting clusters exhibit strong intra-sector cohesion and meaningful inter-sector relationships, validated through multiple clustering metrics. Our approach produces semantically coherent industry groupings with practical effectiveness for business intelligence applications including salary benchmarking and market segmentation.**

***Index Terms*—Company clustering, semantic embeddings, multi-agent systems, DBSCAN, t-SNE, dimensionality reduction, sector classification**

## I. Introduction

Effective clustering of companies by their operational sectors is fundamental for meaningful comparative analyses in business intelligence applications, including salary benchmarking, market segmentation, and competitive analysis [1], [2]. Traditional approaches to company clustering often rely on manually curated taxonomies, which frequently fail to the true nature of business operations because existing taxonomies are designed for mature markets, which may not be appropriate for small companies with innovative business models [3]. Additionally, emerging markets are fast-developing, so static business taxonomies cannot promptly reflect new features [3]. This limitation becomes particularly pronounced in modern corporate landscapes where companies may adopt names that bear little resemblance to their actual products or services, or where large holding companies encompass diverse subsidiaries spanning multiple industries [4].

The primary objective of this study is to develop an automated, scalable, and semantically-aware approach for clustering companies based on their actual business operations. Our analysis is grounded in a dataset derived from Kariyer.net's[1] corporate listings, which includes a wide range of companies operating in Turkey, along with a subset of multinational firms. Specifically, we aim to create a robust pipeline that can accurately group companies into meaningful sector-based clusters by leveraging contextual information about their products and services, thereby enabling more precise comparative analyses (such as salary benchmarking and market segmentation) for recruitment platforms, particularly in emerging business environments like Turkey. Our goal is to overcome the limitations of taxanomy based clustering while addressing the computational and methodological challenges inherent in processing high-dimensional semantic representations of business descriptions.

The challenge of automated company clustering is compounded by the need to extract meaningful semantic representations from limited textual information. While recent advances in natural language processing (NLP), particularly transformer-based models [5], [6], have demonstrated remarkable capabilities in capturing semantic relationships within text, their application to company classification requires careful consideration of contextual information and domain-specific challenges. Furthermore, the high-dimensional nature of modern text embeddings necessitates sophisticated dimensionality reduction and clustering techniques that can preserve semantic relationships while enabling interpretable visualizations and stable cluster assignments [7].

This study addresses these challenges by developing a comprehensive pipeline that combines multi-agent AI for contextual enrichment, state-of-the-art text embedding models for semantic representation, and advanced clustering algorithms for sector-based grouping. Our work makes several key contributions to the field of automated company classification and semantic clustering. First, we introduce a novel multi-agent AI system for systematic contextual enrichment of company datasets, demonstrating how web-based information retrieval can overcome the limitations of sparse company metadata. Second, we provide empirical evidence that dimensionality

[1]https://www.kariyer.net

reduction through t-SNE [8], significantly improves cluster quality in semantic embedding spaces compared to traditional linear methods. Third, we conduct a comprehensive comparative analysis of clustering algorithms and distance metrics in the context of company sector classification. Finally, we demonstrate the practical effectiveness of our approach through rigorous evaluation using multiple clustering validation metrics, showing that our pipeline produces semantically coherent industry groupings that maintain both intra-sector cohesion and meaningful inter-sector relationships.

## II. Related Works

The automation of company classification has gained increasing attention as traditional taxonomies like the Global Industry Classification Standard (GICS)[2] and North American Industry Classification System (NAICS)[3] struggle to keep pace with evolving industries and new business models [9]. Recent research leverages machine learning and natural language processing to build more scalable and flexible classification systems. Below, we discuss two closely related studies that inform and contrast with our approach.

A closely related study by Husmann et al. [10] clusters companies based on stock return data, aiming to enhance portfolio optimization. They use t-SNE for dimensionality reduction followed by spectral clustering to form groups that reflect economic similarity. While their approach is innovative in connecting machine learning with financial strategy, it is constrained by its reliance on historical return data and does not account for the semantic nature of company operations. In contrast, our method addresses this gap by leveraging transformer-based semantic embeddings of product descriptions, enriched via a multi-agent AI pipeline. This enables us to capture nuanced, real-world business contexts, making our clusters more meaningful for modern applications like market segmentation and benchmarking.

Another closely related study is by Rizinski et al. [9], which evaluates various NLP-based approaches to automate company classification, traditionally performed using expert-defined standards like GICS. They assess zero-shot, multi-class, and One-vs-Rest classifiers, including transformer models like Robustly Optimized BERT Approach (RoBERTa) [11], on a large dataset of company descriptions, achieving F1 scores above 0.80 in their best configurations. While the study highlights the effectiveness of deep learning models for structured classification tasks, it relies heavily on labeled datasets and static taxonomies. To address dataset scarcity, the authors generate synthetic company descriptions using ChatGPT[4]; however, this method is inherently limited. Large language models (LLMs), when used without access to real-time external data, are prone to hallucinations, fabricating plausible but incorrect content [12]. In contrast, our system employs a multi-agent AI architecture that includes a web-enabled information retrieval pipeline, allowing it to ground its outputs in actual publicly available data. This prevents hallucinations and results in more accurate, up-to-date, and semantically rich product and service descriptions. As a result, our approach enables unsupervised clustering that reflects the real operational landscape of companies, making it more suitable for dynamic or emerging market analysis.

In summary, prior work demonstrates the value of both financial signal-based clustering and NLP-driven classification, but each comes with significant limitations. Our system advances the field by integrating web-enabled, multi-agent AI with semantic modeling to offer a scalable and up-to-date solution for industry classification in increasingly complex and fluid market environments.

## III. Method

### A. Dataset Overview

The initial dataset consisted of a list of company names sourced from Kariyer.net's own resources, primarily including companies from Turkey as well as some global firms. The dataset contains 400 companies, of which 45 were conglomerates with an average of 25.44 subsidiaries per conglomerate (6.45 subsidiaries per company overall). Based on feedback and guidance from Kariyer.net's recruitment domain experts, it was determined that company names themselves were insufficient for high-quality clustering, as they often lacked sufficient contextual information. Therefore, a field called `product_description` was needed for each company to enrich the dataset with more semantically meaningful content. Moreover, experts also noted that conglomerate companies, due to their multi-sector nature, could introduce significant noise into the clustering process by linking otherwise unrelated companies. To address this issue, another field called `subsidiaries` was needed to enable the use of subsidiaries as more focused units of analysis, enabling disaggregation from their parent conglomerates.

To that end, we implemented a procedure whose pseudocode is given in Algorithm 1 which utilizes a multi-agent AI system, for the contextual enrichment of each company in the dataset. This system systematically aggregated and analyzed publicly accessible web-based information to generate structured and semantically rich metadata.

Each company was enriched with two key metadata fields: `product_description` and `subsidiaries`. The `product_description` field is a string that captures a semantically rich summary of the company's main products and services, constructed based on publicly available information about the company's areas of activity. The `subsidiaries` field is a list of strings that includes the names of subsidiary companies or sub-brands that operate under the parent company, derived from information about the company's organizational structure. Additionally, the company name was updated according to the agentic system's research and stored in the `company_name`.

The AI system responsible for context enrichment was designed as a modular pipeline of agents, each specializing in a different task: query generation, web search, cleaning of

[2]https://www.msci.com/indexes/index-resources/gics
[3]https://www.census.gov/naics/
[4]https://chatgpt.com

raw content, summarization and information mapping. As is shown in Algorithm 1, there are five major steps to the process which can be listed as:

- *generateQueries*: This agent is responsible for creating internet search queries based on a given company name and a specified output schema, which serves as a guiding framework.
- *search*: This procedure utilizes the queries from the preceding step to conduct web searches via DuckDuckGo[5], employing Selenium WebDriver[6] for browser automation. The raw content from the retrieved webpages is subsequently downloaded and transferred to the next stage.
- *clean*: In this procedure, webpage content is initially parsed from HTML format using BeautifulSoup[7]. Subsequently, each query, along with its corresponding webpages, is processed by an LLM to extract information pertinent to the query.
- *summarize*: This procedure involves an LLM extracting relevant information from the cleaned content, guided by the provided schema.
- *map*: This final procedure uses an LLM to map the extracted notes into a JSON document, adhering to the specified schema.

Contextual enrichment using `product_description` and structural refinement using `subsidiaries` allowed us to construct a clean, semantically rich dataset. This dataset served as a solid foundation for our subsequent embedding and clustering pipeline, enabling more meaningful and accurate groupings reflective of the true industrial landscape.

**Algorithm 1** Company Context Enrichment

**Input:** List of companies, output schema
**Output:** List of companies enriched with structured information conforming to the specified output schema
**Begin**
$enriched \leftarrow \emptyset$
**for** each $company$ in $companies$ **do**
  $contents \leftarrow \emptyset$
  $queries \leftarrow$ generateQueries$(company, schema)$
  **for** each $query$ in $queries$ **do**
    $raw \leftarrow$ search$(query)$
    $processed \leftarrow$ clean$(raw)$
    $contents \leftarrow contents \cup \{processed\}$
  **end for**
  $extractedNotes \leftarrow$ summarize$(contents, schema)$
  $structured \leftarrow$ map$(extractedNotes, schema)$
  $enriched \leftarrow enriched \cup \{structured\}$
**end for**
**return** $enriched$
**End**

[5]https://duckduckgo.com/
[6]https://www.selenium.dev/
[7]https://www.crummy.com/software/BeautifulSoup/

To explain this procedure more clearly, we present the following sample output generated by the agentic system in response to the input 'T.C. Garanti Bankası', which is a company operating in the banking industry:

- `company_name`: T.C. Garanti Bankası (Garanti BBVA)
- `product_description`: Garanti BBVA offers a wide range of financial products and services, including: checking and savings accounts, investment products (gold, mutual funds, stocks, derivatives), loans (personal, auto, mortgage), credit and debit cards, online and mobile banking, insurance and retirement products, payment systems, private banking services, leasing, fleet, and factoring services, investment banking and financing services, ATM and cash management services
- `subsidiaries`: Garanti Asset Management, Garanti Bank International N.V., Garanti Factoring, Garanti Leasing, Garanti Mortgage, Garanti Technology, etc.

As a result, we had a final dataset of 1,500 companies (355 from the original dataset plus 1,145 subsidiaries extracted from the holding structures), with each having a relevant `product_description` field.

### *B. Text Embedding Generation*

After enriching the dataset with contextual information via the `product_description` field, we then decomposed conglomerates into their subsidaries to prevent them from causing mergers between different sectors. The next step was to transform unstructured textual `product_description` of companies in the new set into high-dimensional numerical embedding vectors. To achieve this, we explored several multilingual sentence similarity models based on pretrained transformer architectures [5], including Bidirectional Encoder Representations from Transformers (BERT) [6] and RoBERTa variants [13]–[16], to identify the model that best captured the semantic relationships among companies based on their descriptions.

Transformer models revolutionized NLP by employing self-attention mechanisms that efficiently capture contextual relationships in text [5]. BERT leverages masked language modeling to develop bidirectional contextual representations [6], while RoBERTa refines this approach through optimized training methodology [11]. These models are used for generating semantically rich embeddings that effectively capture linguistic relationships.

To prevent name-based biases from skewing the clustering, we deliberately excluded the original `company_name` field from the embedding phase. Instead, we relied entirely on the `product_description` field, which offered a more accurate and functionally grounded view of each company.

Each model was used to encode the `product_description` field into a 768-dimensional embedding using the `SentenceTransformers` library[8]. The embedding pipeline concatenated text from the `product_description` feature and passed it through

[8]https://www.sbert.net/

the selected model, with the output embeddings normalized to unit length. This normalization step was critical for the clustering phase, where distance-based similarity metrics were used [17].

### C. Dimensionality Reduction

The high dimensionality of the embeddings (768 dimensions) posed challenges for visualization and clustering [7]. To reduce the dimensions, we applied t-SNE to project the vectors into a two-dimensional space. t-SNE preserves local neighborhood structures by converting high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities, then optimizing these relationships in the lower-dimensional space [8].

Although t-SNE is not traditionally used as a preprocessing step for clustering, due to its emphasis on preserving local rather than global structure [18], it proved to be highly effective in our context. The structure of the semantic text embedding space is inherently non-linear, and thus, better preserved through a non-linear projection like t-SNE. While acknowledging the possible distortions that t-SNE may introduce, the transformed representations may provide opportunities for enhanced interpretability and could facilitate improved clustering outcomes. Nevertheless, linear dimensionality reduction techniques such as Principal Component Analysis (PCA) [19] were also explored.

### D. Clustering Methodology

Following dimensionality reduction, we applied clustering algorithms to group companies based on the semantic similarity of their descriptions. Initially, we experimented with several clustering algorithms, including k-means [20], Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [21], and DBSCAN [22]. Traditional methods like k-means require the number of clusters to be specified a priori, which is a significant limitation in our context where the number and structure of sectors are not known in advance. In contrast, density-based algorithms such as HDBSCAN and DBSCAN could be more applicable to situations in which there is a need for discovering clusters of varying shapes and densities without requiring such assumptions.

DBSCAN allowed us to explicitly control the granularity of clustering through the $\epsilon$ parameter, which defines the maximum distance between two samples for them to be considered part of the same neighborhood. This tunability provided better alignment with our objective of grouping companies by sector.

DBSCAN labels low-density points as outliers, assigning them a cluster label of `-1`. To handle outliers, we adopted a centroid-based postprocessing step: each outlier was assigned to the nearest cluster centroid based on Euclidean distance. Cluster centroids were iteratively updated as new members were added, ensuring that the growing clusters remained representative of their members.

### E. Evaluation of Clustering Performance

To evaluate the quality of the clusters produced by our pipeline, we employed a set of widely used internal clustering validation metrics: the *Silhouette Score* [23], the *Davies–Bouldin Index (DBI)* [24], and the *Calinski–Harabasz Index (CHI)* [25]. These metrics provide different perspectives on cluster validity by quantifying intra-cluster cohesion and inter-cluster separation, without requiring access to external ground-truth labels.

*a) Silhouette Score.:* The *Silhouette Score* measures how similar a sample is to its own cluster compared to other clusters. It is defined for each sample $i$ as we presented in Equation 1:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{1}$$

Here $a(i)$ is the mean intra-cluster distance (i.e., the average distance between $i$ and all other points in the same cluster), and $b(i)$ is the mean nearest-cluster distance (i.e., the average distance from $i$ to the points in the nearest different cluster). The overall score is the mean of $s(i)$ across all points. Values range from $-1$ (poor clustering) to $+1$ (ideal clustering), with values near 0 indicating overlapping clusters.

*b) Davies–Bouldin Index.:* The Davies–Bouldin Index evaluates cluster compactness and separation, and is defined as we presented in Equation 2:

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d_{ij}} \right), \tag{2}$$

Here, $k$ is the number of clusters, $\sigma_i$ is the average distance between points in cluster $i$ and the cluster centroid, and $d_{ij}$ is the distance between the centroids of clusters $i$ and $j$. Lower values indicate better clustering, with minimal intra-cluster distances and large inter-cluster separations.

*c) Calinski–Harabasz Index.:* The Calinski–Harabasz Index, also known as the Variance Ratio Criterion, is calculated as we presented in Equation 3:

$$\text{CHI} = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{n - k}{k - 1}, \tag{3}$$

Here, $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix, $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix, $n$ is the number of data points, and $k$ is the number of clusters. Higher values of *CHI* indicate better-defined clusters.

The goal is to identify configurations that maximize the *Silhouette Score* (ranging from -1 to 1, with higher values indicating better clustering) and *CHI* (unbounded and positive, with higher values denoting better separation) while minimizing *DBI* (non-negative, with values approaching zero indicating optimal clustering). This approach favors well-separated and compact clusters. However, it's important to interpret these metrics in the context of our high-dimensional semantic embeddings since achieving optimal scores becomes significantly more difficult in high-dimensional or sparsely distributed data due to the curse of dimensionality [26].

## IV. Results

### A. Experimental Setup

Our experimental evaluation comprised a systematic comparison of various clustering configurations to determine the optimal approach for grouping companies based on their semantic embeddings. We run 14 distinct configurations, exploring the effects of different algorithms (k-means vs. HDBSCAN vs. DBSCAN), distance metrics (Euclidean vs. cosine), dimensionality reduction techniques (t-SNE vs. PCA), and reduced dimensionality (2D vs. 4D vs. original 768D).

For configurations using DBSCAN, we tuned the $\epsilon$ parameter individually for each setting to account for variations in the data distribution after different dimensionality reduction and distance metric combinations. For HDBSCAN, we fixed the `min_cluster_size` parameter to 2 to ensure that each identified cluster (i.e., potential industry group) contained at least two companies. This choice reflects our preference for avoiding singleton clusters while still allowing fine-grained industry distinctions to emerge.

To assess clustering performance, we employed three complementary internal validation metrics that were previously introduced in Section III-E. However, due to the way *DBI* and *CHI* are constructed, they operate solely in euclidiean space which is why for setups in which we used cosine distance as our distance metric in clustering, we necessarily excluded these indices from our evaluation framework.

Because our final clustering was performed in a lower-dimensional t-SNE space (rather than the original 768-dimensional embedding space), we also computed evaluation metrics on the same t-SNE-reduced data. For this purpose, we re-applied t-SNE (with a fixed random seed for reproducibility) to project the full set of embeddings, and used the resulting coordinates for metric computation.

### B. Quantitative Evaluation of Clustering Methods

The results of our clustering experiments, as summarized in Table I, reveal several important trends and trade-offs across different algorithmic configurations. Our analysis focused on identifying the most effective combination of clustering algorithm, distance metric, dimensionality reduction technique, and hyperparameter setting.

**Choice of Embedding Model.** In this study, we ultimately selected the `Alibaba-NLP/gte-multilingual-base` [14] model for embedding generation. Although several models were evaluated, including `BAAI/bge-m3` [16], the final choice was based primarily on empirical observations rather than formal quantitative superiority. Specifically, during the clustering and visualization phases, the embeddings generated by the `Alibaba-NLP/gte-multilingual-base` model led to more semantically coherent and thematically consistent groupings. While `BAAI/bge-m3` showed competitive performance, it was significantly more resource-intensive. It is worth noting that our assessment on embedding model selection did not solely rely on objective clustering metrics such as the *Silhouette Score* or *DBI*; instead, it was driven by qualitative interpretability and human evaluation of cluster cohesion. This practical, insight-driven approach guided our decision to favor the `Alibaba-NLP/gte-multilingual-base` model for all subsequent stages in the pipeline.

**Choice of Clustering Algorithm.** Both HDBSCAN and DBSCAN consistently outperformed k-means across various configurations. However, DBSCAN emerged as our preferred choice due to its granularity control through the `eps` parameter. While HDBSCAN correctly identified major industry sectors such as banking, it often produced excessive fragmentation within these sectors, dividing semantically cohesive industries into multiple subclusters without clear interpretability. In contrast, DBSCAN allowed us to directly control cluster granularity, achieving industry-level clustering that aligned with our objective of grouping companies by their primary business domain.

**Distance Metric Comparison.** Contrary to conventional wisdom for embedding spaces, Euclidean distance consistently outperformed cosine similarity in our experiments. Although cosine similarity is the typically preferred metric when working with high-dimensional text embeddings [27], it exhibited high sensitivity to hyperparameter changes in our context. Minor adjustments to the $\epsilon$ parameter with cosine distance resulted in dramatic shifts in cluster composition, either creating a single dominant cluster that absorbed semantically unrelated companies, or fragmenting industries into numerous incoherent subclusters. Euclidean distance, particularly in the t-SNE reduced space, provided notably more stable and visually interpretable clusters, with clearer boundaries between distinct industry sectors.

**Dimensionality Reduction Impact.** The impact of dimensionality reduction proved substantial, with t-SNE significantly outperforming PCA across all metrics. As evident in Table I, configurations using PCA achieved the lowest *Silhouette Scores* (-0.7132 and -0.3746) and poor *DBI* and *CHI*, indicating poorly separated and internally inconsistent clusters. This aligns with our observations in the Method section regarding the non-linear structure of the embedding space, which t-SNE successfully preserves while PCA fails to capture.

As illustrated in Figure 1, the t-SNE projection reveals distinct clusters that are not apparent in the PCA projection, highlighting the advantages of non-linear dimensionality reduction for our embedding space. When inspected through an interactive visualization, the t-SNE projection reveals semantically meaningful groupings, for example, the rightmost cluster predominantly contains companies from the banking sector, while a nearby cluster slightly to the left and below consists mainly of insurance firms. Such industry-specific structures are not evident in the PCA projection, further underscoring the suitability of t-SNE for uncovering non-linear relationships in the embedding space.

**Metric Trade-offs.** Our analysis revealed an interesting trade-off between different evaluation metrics. The configuration with the highest *Silhouette Score* (0.3329, achieved by HDBSCAN with Euclidean distance in 2D t-SNE space) simultaneously produced relatively poor *DBI* and *CHI* (3.8162

| Method | Silhouette Score | Davies-Bouldin Index | Calinski-Harabasz Index |
|---|---|---|---|
| HDBSCAN (cosine) | 0.1300 | - | - |
| HDBSCAN (euclidean) | 0.0826 | 1.8799 | 4.7345 |
| HDBSCAN (cosine, dim. reduced to 2, t-SNE) | -0.4678 | - | - |
| HDBSCAN (cosine, dim. reduced to 4, t-SNE) | 0.0346 | - | - |
| HDBSCAN (euclidean, dim. reduced to 2, t-SNE) | 0.3329 | 3.8162 | 111.0373 |
| HDBSCAN (euclidean, dim. reduced to 4, t-SNE) | 0.2067 | 1.6007 | 45.6189 |
| DBSCAN ($\epsilon = 0.203125$, cosine) | 0.0008 | - | - |
| DBSCAN ($\epsilon = 0.65$, euclidean) | -0.0401 | 1.0019 | 1.6885 |
| DBSCAN ($\epsilon = 0.000075$, cosine, dim. reduced to 2, t-SNE) | -0.2210 | - | - |
| DBSCAN ($\epsilon = 0.01$, cosine, dim. reduced to 4, t-SNE) | 0.0124 | - | - |
| DBSCAN ($\epsilon = 2.0$, euclidean, dim. reduced to 2, t-SNE) | 0.2450 | 0.5509 | 853.6933 |
| DBSCAN ($\epsilon = 4.5$, euclidean, dim. reduced to 4, t-SNE) | 0.0569 | 0.9305 | 56.2889 |
| DBSCAN ($\epsilon = 0.01$, euclidean, dim. reduced to 2, PCA) | -0.7132 | 13.4321 | 2.8777 |
| DBSCAN ($\epsilon = 0.04$, euclidean, dim. reduced to 4, PCA) | -0.3746 | 2.6240 | 2.6686 |

TABLE I: Clustering evaluation metrics for various approaches

and 111.0373, respectively). Conversely, the configuration with the best *DBI* and *CHI* scores (0.5509 and 853.6933, respectively) recorded a somewhat lower *Silhouette Score* (0.2450). This illustrates the multifaceted nature of cluster quality and the importance of considering multiple metrics when evaluating clustering performance.

**Optimal Configuration.** Based on both quantitative metrics and qualitative assessments of cluster coherence, we identified DBSCAN ($\epsilon = 2.0$) with Euclidean distance on t-SNE-reduced 2D embeddings as the optimal configuration. This approach achieved the second-best *Silhouette Score* (0.2450) while simultaneously recording the best *DBI* (0.5509) and *CHI* (853.6933) among all tested configurations. Moreover, visual inspection of the resulting clusters revealed excellent alignment with industry boundaries, with companies from the same sector forming cohesive and well-separated groups in the embedding space. The method also demonstrated consistent behavior across multiple runs, indicating robustness and reliability under varied initialization conditions. This balance of strong quantitative performance, qualitative interpretability, and operational stability made this configuration our definitive choice for the final clustering solution.

### C. Interpreting Company Clusters

Figure 1b offers a bird's-eye view of our optimized clustering solution, DBSCAN on the 2D t-SNE projection, where spatially contiguous points denote semantically coherent groups.

First, we observe that companies within the same sector form tight, well-separated clusters in the t-SNE map. For instance, a distinct indigo region in the center-right corresponds to banking institutions, while a neighboring magenta region captures insurance firms. Their adjacency quantitatively confirms the semantic closeness of finance subdomains. Similarly, clusters of energy producers and energy distributors occupy adjacent zones in the lower half of the plot, reflecting their functional linkage yet distinct operational roles.

Beyond purely intra-sector cohesion, the layout uncovers meaningful inter-sector proximities: related industries, such as mining and heavy manufacturing, appear in contiguous clusters, as do various branches of the media sector (e.g., television broadcasters versus radio broadcasters).

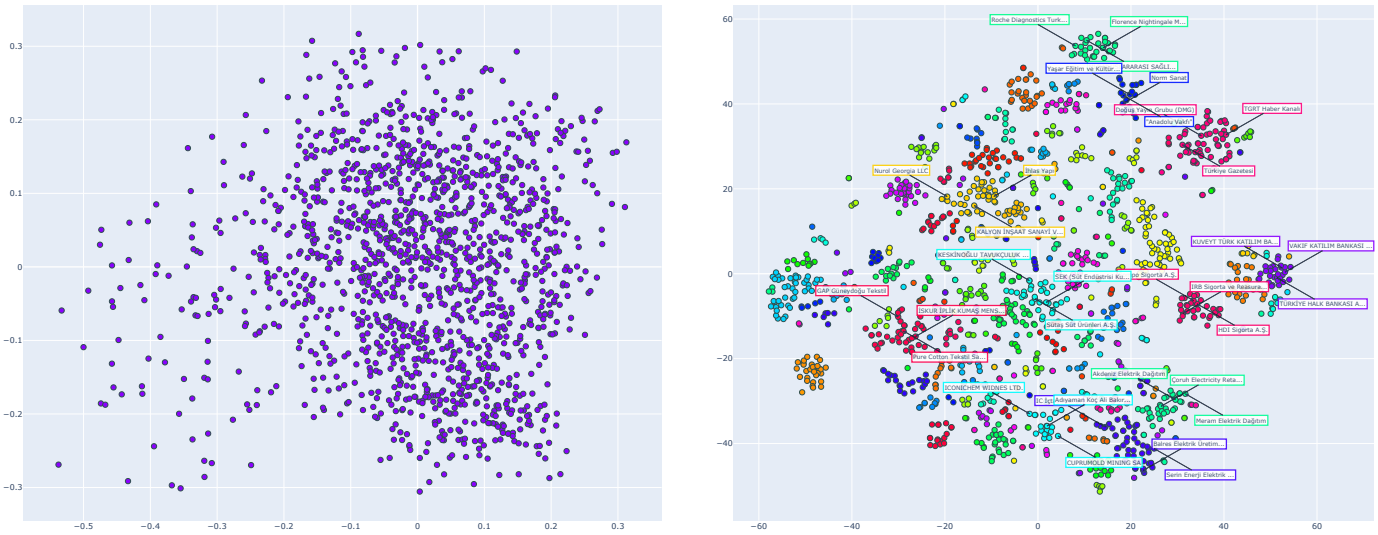
These patterns demonstrate that our embedding–dimensionality reduction pipeline not only recovers clear sectoral groupings but also preserves non-linear semantic relationships between industries. The ability to discern both tight intra-cluster cohesion and sensible inter-cluster topology validates our choice of t-SNE for visualization and DBSCAN (with Euclidean distance) for clustering. Overall, the plot confirms that our methodology yields an interpretable and semantically faithful representation of the corporate landscape.

Overall, our clustering methodology emphasized interpretability, semantic coherence, and adaptability. All of which are crucial for practical downstream applications such as salary benchmarking or market segmentation.

## V. Conclusion

This study aimed to develop an automated, scalable approach for clustering companies based on their actual business operations rather than relying on potentially misleading company names or traditional taxonomies. We implemented a comprehensive pipeline that combined multi-agent AI for contextual enrichment, advanced text embedding models for semantic representation, and sophisticated clustering algorithms for sector-based grouping. Through systematic experimentation with 14 distinct configurations, we explored the effects of different clustering algorithms (k-means, HDBSCAN, DBSCAN), distance metrics (Euclidean vs. cosine), and dimensionality reduction techniques (t-SNE vs. PCA). Our results demonstrated that DBSCAN with Euclidean distance on t-SNE-reduced 2D embeddings achieved optimal performance, producing semantically coherent industry groupings with strong intra-sector cohesion and meaningful inter-sector relationships. Notably, our findings indicate that t-SNE outperforms PCA for dimensionality reduction, and that the Euclidean distance demonstrates superior performance compared to cosine similarity within the reduced embedding space.

Future work could explore the integration of external validation through expert-labeled ground truth datasets to complement our internal clustering metrics and examine the temporal




(a) Visualization of data points without clustering in 2D PCA space.

(b) Visualization of data points and their clusters with randomly chosen examples in 2D t-SNE space.

Fig. 1: Side-by-side comparison of PCA and t-SNE projections of embedding vectors obtained through `Alibaba-NLP/gte-multilingual-base`. PCA fails to reveal clear structure, whereas t-SNE uncovers distinct clusters, indicating the presence of non-linear relationships in the semantic embedding space. Although similar colors may appear due to the presence of 233 clusters in the second plot, spatial proximity indicates related groups. Conversely, groups that are far apart represent distinct clusters, even if their colors look similar.

stability of clusters as companies evolve their business models over time. Additionally, incorporating multilingual capabilities more extensively and developing domain-specific embedding models trained on business and industry-specific corpora could further enhance clustering accuracy and semantic coherence.

## REFERENCES


[1] A. Bose, A. Munir, and N. Shabani, "A comparative quantitative analysis of contemporary big data clustering algorithms for market segmentation in hospitality industry," 2017. [Online]. Available: https://arxiv.org/abs/1709.06202

[2] T. Reutterer and D. Dan, *Cluster Analysis in Marketing Research*. Cham: Springer International Publishing, 2019, pp. 1–29. [Online]. Available: https://doi.org/10.1007/978-3-319-05542-8_11-1

[3] H. Bai, F. Z. Xing, E. Cambria, and W.-B. Huang, "Business taxonomy construction using concept-level hierarchical clustering," 2019. [Online]. Available: https://arxiv.org/abs/1906.09694

[4] B. L. Bendell and E. K. Kristal, "Five naming strategies to help tell your organization's story," *Business Horizons*, vol. 66, no. 3, pp. 387–404, 2023, sPECIAL ISSUE: STRATEGIC STORYTELLING. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681323000277

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[7] M. Steinbach, L. Ertöz, and V. Kumar, *The Challenges of Clustering High Dimensional Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 273–309. [Online]. Available: https://doi.org/10.1007/978-3-662-08968-2_16

[8] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[9] M. Rizinski, A. Jankov, V. Sankaradas, E. Pinsky, I. Mishkovski, and D. Trajanov, "Comparative analysis of nlp-based models for company classification," *Information*, vol. 15, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2078-2489/15/2/77

[10] S. Husmann, A. Shivarova, and R. Steinert, "Company classification using machine learning," 2020. [Online]. Available: https://arxiv.org/abs/2004.01496

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[12] H. Kang, J. Ni, and H. Yao, "Ever: Mitigating hallucination in large language models through real-time verification and rectification," 2024. [Online]. Available: https://arxiv.org/abs/2311.09114

[13] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," 2022. [Online]. Available: https://arxiv.org/abs/2007.01852

[14] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, and M. Zhang, "mgte: Generalized long-context text representation and reranking models for multilingual text retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2407.19669

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020. [Online]. Available: https://arxiv.org/abs/2002.10957

[16] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[17] A. Han and H. Du, "How does normalization impact clustering?" in *Recent Advances in Next-Generation Data Science*, H. Han, Ed. Cham: Springer Nature Switzerland, 2024, pp. 34–47.

[18] D. Kobak and P. Berens, "The art of using t-sne for single-cell transcriptomics," *Nature Communications*, vol. 10, no. 1, p. 5416, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-13056-x

[19] K. P. and, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720

[20] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, L. M. Le Cam and J. Neyman, Eds. Berkeley, CA: University of California Press, 1967, pp. 281–297.

[21] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Trans. Knowl. Discov. Data*, vol. 10, no. 1, Jul. 2015. [Online]. Available: https://doi.org/10.1145/2733381

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

[23] P. Rousseeuw, "Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 11 1987.

[24] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[25] T. Caliński and J. H. and, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

[26] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory — ICDT'99*, C. Beeri and P. Buneman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 217–235.

[27] K. You, "Semantics at an angle: When cosine similarity works until it doesn't," 2025. [Online]. Available: https://arxiv.org/abs/2504.16318