Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Strategic Mapping of University Research Portfolios Using NLP and Machine Learning: An Integrated Analysis of Research Projects

**Üzeyir Pala**

uzeyir.pala@izu.edu.tr

İstanbul Sabahattin Zaim Üniversitesi

**Bahar Birelli**

İstanbul Sabahattin Zaim Üniversitesi

**Ayşenur Yalçın**

İstanbul Sabahattin Zaim Üniversitesi

**Research Article**

**Keywords:** Research Projects Portfolio Management, Natural Language Processing (NLP), Machine Learning (ML), Unsupervised Clustering, Text Mining

**Posted Date:** August 27th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10615568/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Strategic Mapping of University Research Portfolios Using NLP and Machine Learning: An Integrated Analysis of Research Projects

Üzeyir Pala[1*], Bahar Birelli[2], Ayşenur Yalçın[3]

[1*] İstanbul Sabahattin Zaim Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Bölümü, İstanbul, Türkiye, (ORCID: 0000-0001-6231-0846), uzeyir.pala@izu.edu.tr

[2] İstanbul Sabahattin Zaim Üniversitesi, Fen Bilimler Enstitüsü, Büyük Veri ve İş Analitiği Anabilim Dalı, İstanbul, Türkiye, (ORCID: 0009-0004-9293-5440), bahar.birelli@izu.edu.tr

[3] İstanbul Sabahattin Zaim Üniversitesi, Fen Bilimler Enstitüsü, Büyük Veri ve İş Analitiği Anabilim Dalı, İstanbul, Türkiye, (ORCID: 0009-0005-1715-019X), aysenur.yalcin@izu.edu.tr

**Abstract**

This study aims to develop an integrated analysis model for research management by examining the thematic structure, risk profile, and strategic features of a university's Scientific Research Projects (SRPs) portfolio using data-driven methods. Within the scope of the research, a dataset comprising 631 projects from the 2015–2026 period was evaluated, and following data cleaning and preprocessing procedures, analyses were conducted on 593 projects. Notably, the fact that the Faculty of Engineering and Natural Sciences utilizes approximately 62% of the total project budget points to high research productivity in technical fields. Furthermore, it was determined that the current portfolio structure holds significant potential regarding the transition to strategic and high-budget projects. Methodologically, project descriptive texts were transformed into multilingual Sentence-BERT-based embedding vectors, followed by dimensionality reduction using PCA and UMAP methods. During the clustering phase, K-Means and HDBSCAN algorithms were compared, and the optimal number of clusters was determined as C=10 using the Silhouette, Calinski-Harabasz, and Gap Statistic metrics. While the thematic structure was examined through LDA topic modeling, risk analysis was performed via the combined application of Isolation Forest and Local Outlier Factor methods. The results indicate that 33,4% of the portfolio falls into the high-risk category and that the clustering model exhibits high stability with a mean ARI of 0,981. The resulting clusters were reduced to 10 core project profiles; with its approach integrating natural language processing, dimensionality reduction, clustering, and risk modeling, this study offers an original contribution to the research portfolio management literature.

**Keywords:** Research Projects Portfolio Management, Natural Language Processing (NLP), Machine Learning (ML), Unsupervised Clustering, Text Mining.

## 1. Introduction

Higher education institutions (HEIs) worldwide allocate substantial resources to research activities to advance scientific knowledge, address societal challenges, and strengthen their scientific and innovation capacity. A significant proportion of these activities is conducted through Scientific Research Projects (SRPs), commonly referred to as BAP projects in the Turkish higher education

system, which are designed to enhance the research capacity of universities. Technology Transfer Offices (TTOs) play a central role in managing, monitoring, and evaluating project portfolios, as well as facilitating the transformation of research outputs into economic and societal value. However, the increasing number of projects, diversification of research fields, resource constraints, and performance uncertainties have made the management of university research portfolios progressively more complex.

Conventional project monitoring and evaluation approaches primarily focus on individual projects and may be insufficient for identifying thematic patterns, interdisciplinary relationships, clusters of structural risks, and strategic gaps across an entire portfolio. In research portfolios comprising hundreds or thousands of projects, the manual review of project abstracts and content is often both time-consuming and prone to subjective interpretation.

Consequently, growing interest has emerged in the application of Artificial Intelligence (AI)-driven approaches for the systematic analysis of research portfolios. These approaches include Natural Language Processing (NLP) methods (e.g., tokenization, stemming, lemmatization, named entity recognition, and word embeddings), Machine Learning (ML) algorithms (e.g., decision trees, random forests, support vector machines, neural networks, and XGBoost), Text Mining techniques (e.g., TF-IDF, keyword extraction, topic modeling, and document similarity analysis), and Unsupervised Clustering methods (e.g., K-means, hierarchical clustering, and DBSCAN) to identify thematic structures, research trends, interdisciplinary linkages, and portfolio-level risk patterns.

Studies on research performance evaluation have traditionally focused on bibliometric indicators, citation analyses, and productivity-based performance measurement approaches (Garfield, 1979; Abramo et al., 2020). In particular, the performance of TTOs has commonly been assessed using tangible output indicators such as patents, licensing activities, spin-off companies, and research funding (Chapple et al., 2005; Perkmann et al., 2013). However, these approaches primarily focus on realized performance outcomes and provide limited insight into the thematic structures, strategic orientations, and potential risks of ongoing research projects. Previous studies have emphasized the importance of considering the content characteristics and risk profiles of projects for the effective management of research portfolios (Meskendahl, 2010; Teller & Kock, 2013).

Recent advances in AI, ML, and NLP have enabled the evaluation of research activities not only through output-based indicators but also through the content analysis of project and publication texts. Topic modeling, text embedding techniques, and transformer-based language models have been widely employed for the automated classification of research domains, identification of scientific trends, mapping of institutional research profiles, and prediction of emerging research directions (Blei et al., 2003; Devlin et al., 2019; Cohan et al., 2020; Grootendorst, 2022). Furthermore, NLP applications developed for scientific literature can uncover latent knowledge structures within existing publications and contribute to the identification of future research opportunities (Tshitoyan et al., 2019; Zhao et al., 2023).

The development of transformer-based language models has significantly enhanced the ability to represent the semantic structure of research documents with high accuracy. Sentence-BERT-based embedding approaches have demonstrated strong performance in identifying semantic similarities among project abstracts and scientific documents, while BERTopic and related methods enable a more detailed exploration of the thematic structures of research portfolios. In parallel, studies integrating bibliometric data, text mining, and ML techniques have laid the foundation for the development of more dynamic and predictive decision-support systems in research management (Alam et al., 2023; Aria & Cuccurullo, 2020; Zhao et al., 2023). Nevertheless, the existing literature provides limited evidence on studies that examine university-level SRP portfolios through an integrated framework combining ML, NLP, clustering techniques, and risk modeling approaches.

This study aims to address this gap by examining a research portfolio comprising 593 SRPs conducted at a university between 2015 and 2026 through a comprehensive analytical framework encompassing thematic structure, performance characteristics, and risk profiles. Project texts were transformed into semantic embeddings using Sentence-BERT, and topic modeling was performed through Latent Dirichlet Allocation (LDA). Dimensionality reduction was subsequently applied using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), and the resulting representations were clustered using the K-Means and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithms. In addition, Isolation Forest and Local Outlier Factor (LOF) methods were employed to identify anomalous patterns within the project portfolio.

The primary contribution of this study is the development of a unified research portfolio analytics framework that integrates text representation, topic modeling, dimensionality reduction, unsupervised clustering, and anomaly detection techniques. Through the joint analysis of thematic structures, portfolio segmentation patterns, and risk indicators, the framework enables a comprehensive and data-driven evaluation of research portfolios, supporting Technology Transfer Office (TTO) managers in project prioritization, resource allocation, risk monitoring, and strategic decision-making. By uncovering latent portfolio structures and facilitating scalable portfolio-level intelligence, the study advances the methodological foundations of research management and technology transfer literature.

## 2. Literature Review

Research portfolio analysis has undergone a significant transformation from traditional bibliometric approaches toward data-intensive and AI-driven analytical frameworks. Early studies primarily evaluated institutional research performance through publications, citations, and scientific collaboration networks to reveal the structure and evolution of research activities (Garfield, 1979; Narin et al., 1997). However, these approaches rely largely on realized research outputs and provide limited insight into ongoing research activities or projects that have not yet resulted in publications, patents, or other scientific outputs. Consequently, there is a growing need for analytical approaches that enable research portfolios to be monitored and strategically evaluated at earlier stages through project-level data.

Within the topic modeling literature, the Latent Dirichlet Allocation (LDA) model developed by Blei et al. (2003) has become one of the most widely used methods for identifying latent thematic structures in large document collections. Its probabilistic framework enables systematic analysis of topic distributions within academic publications and scientific documents. Subsequent studies extended this approach to project databases maintained by funding agencies, facilitating the identification of thematic clusters, cross-disciplinary interactions, and potential research gaps within research portfolios (Chuang et al., 2012; Talley et al., 2011). However, the limited ability of traditional topic modeling approaches to capture contextual meaning in short and heterogeneous documents has stimulated the development of more advanced methods. In this regard, BERTopic, proposed by Grootendorst (2022), integrates contextual language models into the topic modeling process, enabling more effective identification of semantic relationships and achieving improved interpretability and performance, particularly in short-text datasets.

A major advancement in semantic text representation was introduced through Sentence-BERT, developed by Reimers and Gurevych (2019), which enables highly accurate and computationally efficient measurement of semantic similarity between texts. Compared with conventional language models, Sentence-BERT generates more informative document and sentence representations and has therefore been widely adopted in the semantic analysis of large-scale text collections. When combined with clustering algorithms, semantic embeddings facilitate not only the grouping of similar documents but also the identification of conceptual relationships and distinctions among research domains. Supporting this perspective, Cohan et al. (2020) demonstrated that semantic vector spaces derived from scientific documents provide a robust analytical foundation for examining relationships among research topics and knowledge domains.

In the dimensionality reduction literature, Principal Component Analysis (PCA) has long been established as a fundamental method for representing high-dimensional data in lower-dimensional spaces. Nevertheless, its linear structure may be insufficient for capturing nonlinear patterns and local neighborhood relationships embedded within complex semantic text data. To address this limitation, McInnes et al. (2018) introduced Uniform Manifold Approximation and Projection (UMAP), which preserves both global topological structures and local neighborhood information, thereby producing more meaningful and interpretable representations of high-dimensional data.

Various clustering approaches have been developed to analyze different forms of data structures. Among them, K-Means, introduced by MacQueen (1967), remains one of the most widely used clustering algorithms because of its ability to group similar observations while minimizing within-cluster variance. In contrast, the HDBSCAN algorithm developed by McInnes et al. (2017) can identify clusters with varying densities and irregular shapes without assuming spherical cluster structures. Consequently, both methods are frequently employed as complementary analytical tools in the analysis of high-dimensional text representations.

Within the anomaly detection literature, Isolation Forest (Liu et al., 2008) and Local Outlier Factor (LOF) (Breunig et al., 2000) have been widely adopted for identifying outlier observations in high-dimensional datasets. In portfolio-level analyses, these methods facilitate the early detection of

projects that deviate from typical patterns, unusual resource utilization profiles, and potential risk areas requiring managerial attention.

Research on the performance evaluation of TTOs has traditionally relied on efficiency and output-oriented measurement frameworks. Chapple et al. (2005), for example, combined Data Envelopment Analysis (DEA) with parametric methods to investigate the relative efficiency of TTOs and contributed significantly to methodological developments in the field. Their work helped establish patents, licensing agreements, spin-off companies, and research revenues as key indicators of TTO performance, shaping the output-oriented evaluation paradigm that dominated the technology transfer literature for many years.

A broader perspective on university-industry collaboration and knowledge transfer processes was subsequently provided by Perkmann et al. (2013), who emphasized that TTOs should not be viewed solely as operational units managing research outputs but also as strategic intermediary structures coordinating knowledge flows among universities, industry partners, and other stakeholders. This perspective suggests that research portfolio evaluation should incorporate not only performance indicators but also network structures, collaboration dynamics, and knowledge transfer capacities. Accordingly, effective portfolio analysis requires a more comprehensive framework that extends beyond output-based evaluation and encompasses connectivity, interaction, and impact dimensions.

The integration of AI, ML, and NLP into research analytics has further expanded the possibilities of research evaluation. Zhao et al. (2023) demonstrated that NLP and ML approaches can effectively analyze the thematic structures of scientific publications and support the identification of emerging research trends and developmental trajectories. Similarly, advances in scientific knowledge discovery have shown that semantic representations derived from large-scale scientific corpora can reveal latent relationships and previously unrecognized research opportunities (Tshitoyan et al., 2019). Furthermore, bibliometric and science-mapping tools have enabled the systematic visualization of thematic networks, research trends, and relationships among scientific domains, thereby supporting strategic decision-making in research management (Aria & Cuccurullo, 2020).

Although previous studies have applied topic modeling, semantic embeddings, bibliometric analysis, and ML techniques to various forms of scientific data, few studies have integrated these methods within a unified framework for the analysis of university research project portfolios. Moreover, limited attention has been given to the simultaneous examination of thematic structures, portfolio segmentation patterns, and risk indicators using project-level textual data. This gap is particularly evident in the context of university-managed SRP portfolios, where comprehensive portfolio analytics remain underexplored.

Despite these advances, studies integrating NLP, semantic representation, topic modeling, clustering, and anomaly detection within a unified framework for research portfolio management remain limited. This study seeks to address this gap by proposing an integrated analytical framework for the examination of university-level SRP portfolios. By combining thematic,

structural, and risk-oriented analyses, the proposed approach contributes to the literature on research management and technology transfer while providing practical decision-support mechanisms for portfolio monitoring and strategic planning.

### 2.1. Methodological Transformation in Research Portfolio Management

For TTOs and SRP units, project portfolio management plays a critical role in the effective achievement of institutional strategic objectives. Meskendahl (2010) demonstrated that alignment between project portfolios and organizational strategy has a direct impact on organizational performance and success, while Teller and Kock (2013) emphasized that a risk management perspective must be integrated into portfolio management processes to ensure the sustainability of this alignment. Traditional portfolio evaluation approaches have largely relied on the static analysis of bibliometric indicators and research outputs (Narin et al., 1997), which may be insufficient for capturing the dynamic nature of research activities and emerging trends. In recent years, however, the research management literature has undergone a significant transformation toward the Research Analytics paradigm, characterized by the increasing integration of ML, NLP, and AI-based methods into research portfolio analysis (Zhao et al., 2023). This transformation has enabled research portfolios to be evaluated not only on the basis of past performance but also through the data-driven identification of thematic trends, potential risk areas, and emerging research opportunities.

### 2.2. NLP- and Deep Learning-Based Approaches

Methods for document representation and clustering have undergone a substantial transformation over the past decade, evolving from traditional statistical approaches to deep learning-based language models. During this transition, distance-based clustering methods such as K-Means (MacQueen, 1967) and probabilistic topic modeling approaches such as LDA (Blei et al., 2003) have served as fundamental tools for identifying the structural and thematic characteristics of text corpora. However, the limited capacity of these methods to capture contextual relationships among words has increased the demand for representation techniques capable of more effectively modeling the semantic content of texts.

Recent advances in NLP have led to the widespread adoption of transformer-based models for document representation and semantic analysis. The Bidirectional Encoder Representations from Transformers (BERT) model proposed by Devlin et al. (2019) introduced a major paradigm shift in text analytics through its ability to learn context-dependent word representations. Extending this approach to scientific literature, Cohan et al. (2020) developed the SPECTER model, which generates enriched document representations by jointly leveraging textual content and citation relationships. Similarly, Sentence-BERT (SBERT), introduced by Reimers and Gurevych (2019), enables highly accurate measurement of semantic similarities at both sentence and document levels and has become a widely adopted framework for the semantic analysis of large-scale text collections.

As a natural extension of these developments, Grootendorst (2022) proposed BERTopic, a topic modeling framework that combines semantic embeddings derived from contextual language models with a class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) approach.

Beyond assigning documents to thematic groups, BERTopic facilitates the generation of more interpretable, semantically coherent, and hierarchically structured topic representations, making it one of the most prominent approaches in contemporary topic modeling research. Its strong interpretability and semantic sensitivity are particularly advantageous for the thematic analysis of research portfolios composed of short, heterogeneous, and interdisciplinary texts.

### 2.3. Dimensionality Reduction and Outlier Analysis

The visualization and interpretation of research portfolios require the transformation of high-dimensional data into more accessible and interpretable representations. Chuang et al. (2012) emphasized that interpretability is a fundamental criterion in science mapping studies for revealing relationships among research domains. In this context, UMAP, developed by McInnes et al. (2018), enables the representation of high-dimensional data by preserving both global data structures and local neighborhood relationships, thereby facilitating more meaningful and visually interpretable analyses of research portfolios.

Effective research portfolio management requires not only the identification of dominant thematic areas but also the detection of projects that deviate from prevailing patterns. In this regard, the density-based LOF algorithm (Breunig et al., 2000) and the tree-based Isolation Forest method (Liu et al., 2008) are widely employed for identifying anomalous, unconventional, or potentially high-impact projects within large portfolios. These approaches enable the systematic detection of projects that may be overlooked by conventional evaluation processes and provide valuable decision support for risk monitoring, resource allocation, and strategic prioritization.

Furthermore, interdisciplinary research projects often face challenges in gaining visibility within traditional academic structures and may not be adequately represented by conventional classification systems (Rafols et al., 2012). From this perspective, outlier detection methods serve not only to identify potentially risky projects but also to reveal innovative research initiatives with high transformative potential. Consequently, dimensionality reduction and anomaly detection techniques can be regarded as complementary components of contemporary research portfolio analytics.

### 2.4. Technology Transfer Offices (TTOs) and Research Portfolio Analytics

The performance of TTOs is commonly evaluated through multidimensional indicators encompassing research productivity, organizational efficiency, and technology transfer capacity (Abramo et al., 2020; Chapple et al., 2005). In recent years, these evaluation processes have evolved beyond traditional performance measurement approaches toward the adoption of ML-based predictive models and large-scale research analytics. For example, ML applications have been employed to predict research performance and support decision-making processes (Shahiri & Husain, 2015), while large research funding databases, such as those maintained by the U.S. National Institutes of Health (NIH), have been analyzed using graph-based clustering techniques to examine the structure and dynamics of research portfolios (Talley et al., 2011). These developments have enabled more dynamic and data-driven approaches to research portfolio evaluation.

In parallel, the integration of bibliometric mapping techniques with advanced ML methods has become increasingly important for the effective management of academic research portfolios (Aria & Cuccurullo, 2020). In particular, the application of density-based clustering algorithms such as HDBSCAN to bibliometric and textual data can facilitate a more precise identification of thematic structures, emerging research trends, and strategic focus areas within research portfolios. From this perspective, the integration of advanced analytical methods into research management processes should be viewed not merely as a methodological choice but as a strategic necessity for aligning local research ecosystems with international standards of research management and innovation governance.

## 3. Methodology

### 3.1. Dataset and Variables

Due to the general confidentiality of university Technology Transfer Office (TTO) portfolios, the dataset was restricted to the institution where the authors have access privileges. Consequently, the proposed analytical framework was applied specifically to this institutional data.

The dataset encompasses SRPs (categories BAP-100 through BAP-500) conducted between 2015 and 2026, monitored by the university's TTO. Data were acquired directly from the institutional project management unit. This dataset was selected because it provides detailed, previously unexamined records that directly align with the research objectives. Furthermore, systematic institutional data management practices ensured high data quality and reliability.

The final dataset consists of 631 project records and 19 variables (Table 1).

**Table 1.** Dataset variables used in the study

| # | Variable | # | Variable |
|---|---|---|---|
| 1 | Project Number | 11 | Total Approved Budget |
| 2 | Project Status | 12 | Scholarship Expenses |
| 3 | Project Title | 13 | Travel Expenses |
| 4 | Faculty Name | 14 | Machinery and Equipment Purchase Expenses |
| 5 | Department Name | 15 | Consumable Material Expenses |
| 6 | Supporting Institution | 16 | Service Procurement Expenses |
| 7 | Program Type | 17 | Publication Expenses |
| 8 | Start Date | 18 | Software Expenses |
| 9 | End Date | 19 | Other Expenses |
| 10 | Project Duration *(Months)* | | |

Each row in the dataset represents an independent project. This comprehensive structure enables the simultaneous evaluation of the projects' thematic, financial, and managerial dimensions.

An analysis of the institutional distribution reveals significant variations in project generation capacity across faculties. The Faculty of Engineering and Natural Sciences accounts for the largest

share of the research portfolio, hosting 390 of the total 631 projects. This is followed by the Faculty of Business and Management Sciences (102 projects) and the Faculty of Education (67 projects). The remaining 72 projects are distributed among other faculties (Figure 1). This distribution demonstrates that research activities are concentrated in specific academic units, reflecting varying levels of institutional research capacity development across faculties.

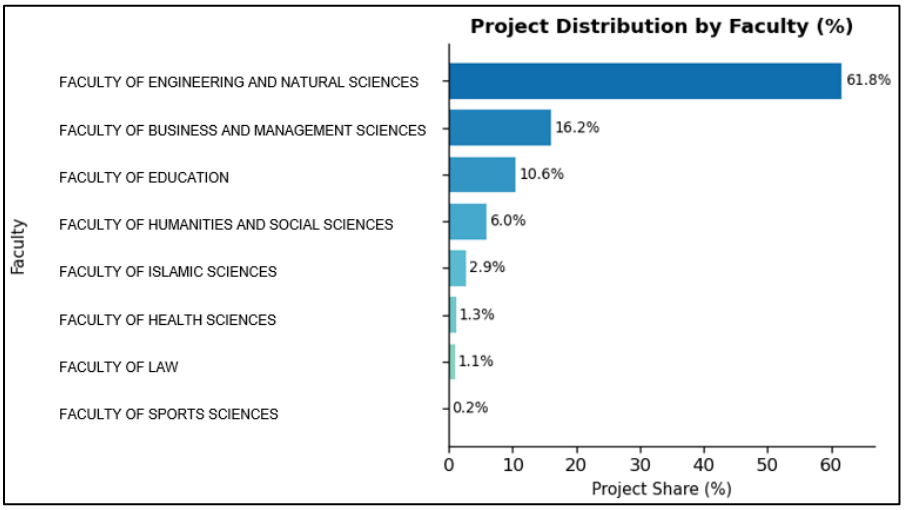



**Figure 1.** Percentage distribution of research projects by faculty

Prior to the analysis, a comprehensive preprocessing procedure was applied to the dataset. Initially, variable names were standardized, and temporal and numerical data were converted into appropriate analytical formats. For missing data management, median imputation was preferred because it produces more robust results against skewed distributions and outliers, as noted by Alam et al. (2023). Accordingly, missing or zero values identified in the budget and duration variables were imputed using their respective medians.

As part of the data cleaning and quality control process, 38 projects with descriptions shorter than three words -which were deemed unsuitable for generating meaningful textual content- were excluded from the scope of the analysis. Consequently, subsequent analyses were conducted on a final sample of 593 projects. Furthermore, to facilitate a more robust comparison of resource utilization intensity across projects, a “budget/month” indicator was created by dividing the total project budget by the project duration. This indicator allowed for the evaluation of the financial intensity of projects with varying durations using a standardized metric, thereby contributing to a more detailed analysis of the budget utilization dynamics within the project portfolio.

To minimize the impact of deviations caused by outliers on the analysis results, Winsorization was applied to the budget variable at the 5th and 95th percentiles. Through this procedure, values in the bottom 5% were set to the 5th percentile value, and those in the top 5% were set to the 95th percentile value, effectively limiting the influence of extreme observations.

Additionally, Robust Scaling was employed to create a data structure more resilient to outliers while preserving the distributional characteristics of the variables. Using this approach, which is

based on the median and interquartile range (IQR), the budget, project duration, and budget/month variables were standardized to generate the *budget_sc*, *duration_sc*, and *budget-duration_sc* variables, respectively, making them suitable for subsequent analyses.

Finally, the projects were examined based on their thematic characteristics and classified into qualitative and quantitative research categories. This classification was incorporated into the analysis process to facilitate the evaluation of the methodological diversity of the project portfolio.

### 3.2. Text Processing

During text preprocessing, project titles underwent a series of standardization and cleaning procedures to prepare them for NLP. All text was converted to lowercase, while punctuation, special characters, and non-informative expressions were removed. To minimize semantic noise, Turkish and English stop-word lists were applied to exclude words with low informational value. This process yielded a streamlined, consistent, and semantically focused data structure, optimizing it for subsequent topic modeling and semantic analyses.

To effectively process linguistic features, the dominant language of each text was identified using a rule-based approach relying on Turkish and English stop-word matching. Following language detection, language-specific stemming was applied to reduce words to their root forms by removing affixes. This step minimized the morphological variance of similar concepts and enhanced the overall semantic representational power of the corpus.

Following these preprocessing steps, texts lacking sufficient content to form a meaningful thematic structure were excluded from the dataset. Specifically, project titles that became empty post-cleaning or contained fewer than three words were omitted. This threshold was established because preliminary reviews indicated that titles shorter than three words generally failed to adequately represent the research topic and provided limited semantic value. Consequently, this filtering ensured that subsequent tasks, including topic modeling, embedding vector generation, and clustering, produced more reliable and robust results.

### 3.3. Natural Language Processing (NLP) - Text Representation and Topic Modeling Methodology

To numerically represent textual data within the scope of this study, text embedding techniques were utilized. Accordingly, project titles were converted into vector representations using the multilingual SBERT (Sentence-BERT) model. Specifically, the *paraphrase-multilingual-MiniLM-L12-v2* model was selected for its efficacy in capturing semantic similarities across texts, thereby generating dense vector representations.

In addition, the TF-IDF approach was implemented to construct a sparse feature matrix as an alternative textual representation. However, SBERT embeddings served as the primary representation method throughout the analysis. By converting sentences into fixed-dimensional vectors, SBERT enables the rapid and efficient computation of semantic similarities. Consequently, SBERT-based vector representations were preferred for all subsequent analyses.

Following the extraction of text representations, topic modeling analyses were conducted to systematically and comprehensively reveal the thematic distribution of project titles. To this end, Latent Dirichlet Allocation (LDA), a widely used topic modeling technique, was initially applied. Prior to LDA modeling, bigrams were generated to identify frequently co-occurring word patterns within the texts. Subsequently, a dictionary and a corpus representing the text collections were prepared, leading into the modeling phase.

To determine the optimal number of topics, C_v coherence scores were calculated for various topic numbers, and the resulting values were compared. Furthermore, the BERTopic methodology was applied to facilitate a deeper examination of the thematic distribution. In this stage, SBERT embeddings were utilized for the semantic representation of texts. Subsequently, the UMAP algorithm was employed to reduce high-dimensional vectors into a lower-dimensional space, and clustering operations were executed using the HDBSCAN algorithm.

As a result of these procedures, project texts were analyzed within topic clusters, ensuring that thematic structures were elucidated in a clear, interpretable, and meaningful manner. This approach allowed for a more effective evaluation of content similarities and thematic relationships among the project titles.

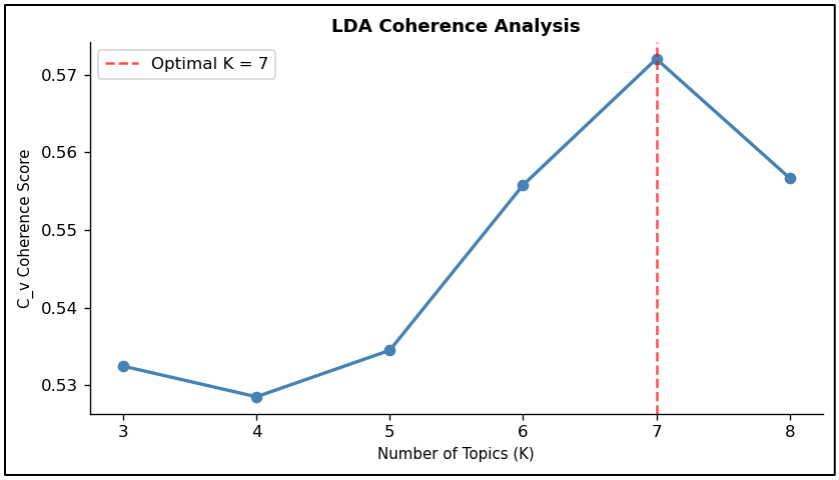



**Figure 2.** LDA C_v coherence score (screening from C=3 to C=8)

Within the LDA framework, C_v coherence scores were calculated for different numbers of topics, ranging from C=3 to C=8. The analyses revealed that the highest C_v value of 0,572 was achieved by the C=7 model (Figure 2). This result indicates that the seven-topic model represents the thematic structure within the project texts more coherently and meaningfully compared to the alternative models. In contrast, the lowest coherence score was observed in the C=4 model. When comparing the C=4 and C=7 models, it is evident that the C=7 model demonstrates superior performance regarding thematic integrity and topic differentiation.

Furthermore, increasing the number of topics to eight resulted in a marginal decline in the C_v score. This suggests that expanding the number of topics does not enhance model quality; rather, it leads to a slight reduction in thematic coherence. Consequently, these findings establish that C=7

is the optimal number of topics for the dataset, allowing the model to most accurately reflect the underlying thematic structure.

### 3.4. Dimensionality Reduction: PCA and UMAP

In the dimensionality reduction phase, an integrated data representation structure was constructed by combining SBERT embeddings with numerical variables. To render this high-dimensional structure more comprehensible and interpretable, dimensionality reduction methods were applied prior to the clustering analyses. Within this scope, Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) techniques were utilized.

The PCA method was employed to reduce dimensionality while preserving the underlying variance structure of the dataset to the greatest extent possible. Accordingly, cumulative explained variance ratios were examined, and the number of components retaining 90% of the total variance was identified as the optimum value. The obtained results were presented visually (Figure 3). Furthermore, a two-dimensional PCA projection was generated to facilitate a more intuitive evaluation of data distribution. This visualization allowed for the examination of data point positions in a lower-dimensional space and the clearer observation of potential patterns.

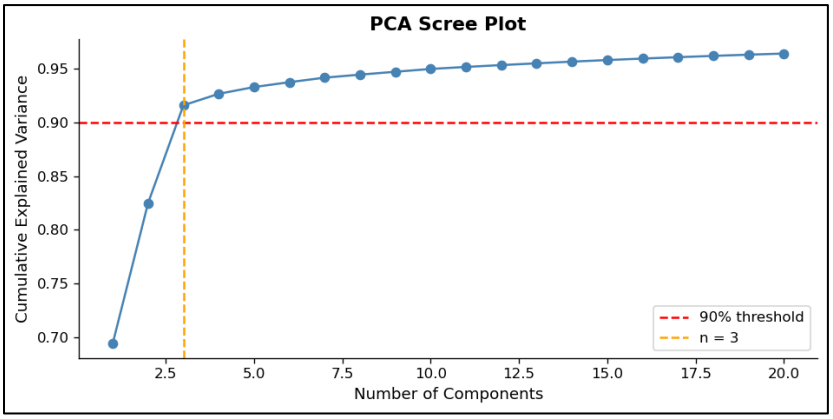



**Figure 3.** PCA scree plot and cumulative explained variance

An examination of Figure 3 reveals that the first principal component explains approximately 70% of the total variance, while the second component provides an additional contribution of approximately 13%. Evaluating the cumulative explained variance curve, it was determined that the total explained variance ratio exceeds the 90% threshold with the addition of the third component. This result indicates that a vast majority of the information within the dataset can be represented by the first three principal components, demonstrating the efficacy of the dimensionality reduction process.

In addition to the PCA analysis, the UMAP method was applied to better uncover the non-linear structures and local neighborhood relationships of the data. UMAP projects the data into a lower-dimensional space while preserving similarity relationships in the high-dimensional space, thereby making potential cluster structures more visible. Accordingly, the two-dimensional projections obtained using the PCA and UMAP methods were compared, and the positions of the data points in the lower-dimensional space were visualized prior to the clustering process (Figure 4). This

comparison contributed to evaluating the extent to which different dimensionality reduction approaches reflect the data structure and to determining the appropriate representation format for the clustering analyses.

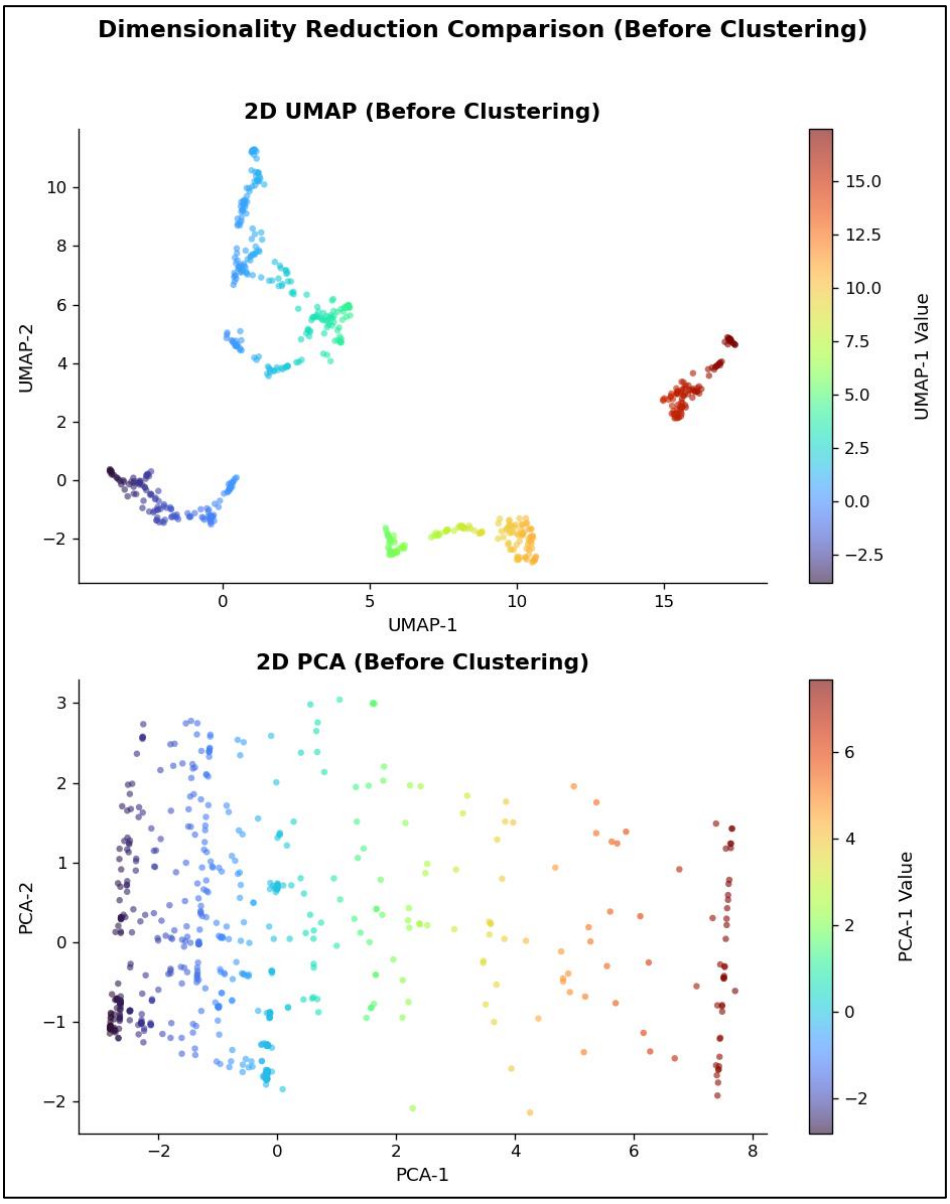



**Figure 4.** Comparison of dimensionality reduction methods: UMAP and PCA 2D projections (before clustering)

The color distributions presented in Figure 4 illustrate the spatial patterns of the project data in UMAP- and PCA-based two-dimensional spaces prior to the clustering process. The colors represent the relative positions of the data points along the axes generated by the respective dimensionality reduction methods.

The color transitions observed in the UMAP plane reflect structural differentiation, particularly along the UMAP-1 axis, whereas the color variations in the PCA plane predominantly indicate variations along the PCA-1 axis. These distributions visualize the similarities and differences among the projects within a low-dimensional space, providing crucial insights into the continuity of the data structure, density regions, and potential divergence tendencies. Consequently, evaluating the spatial organization of the dataset prior to the clustering analysis facilitated a preliminary assessment of potential cluster structures and thematic differentiations.

### 3.5. Clustering: K-Means and HDBSCAN

For visualization purposes, two-dimensional PCA and UMAP projections were calculated separately, and the obtained results were evaluated comparatively (Figure 4). By preserving the topological structure of high-dimensional data, the UMAP method enables the representation of both local and global relationships in a low-dimensional space. Through this feature, similarities and divergences among data points can be demonstrated more distinctly.

An examination of the PCA and UMAP projections presented in Figure 4 reveals that potential cluster structures are more clearly differentiated in the UMAP representation. In contrast, the PCA projection offers a more generalized view by reflecting the fundamental linear variance structure of the data. The primary reason for this discrepancy is that UMAP is a non-linear dimensionality reduction technique that strives to preserve neighborhood relationships among data points as much as possible. Conversely, PCA adopts a linear approach, identifying the linear components that best explain the total variance in the dataset. Therefore, it is expected that the projections generated by these two methods exhibit distinct visual patterns. The resulting projections were utilized to visually inspect the data structure and potential cluster formations prior to the K-Means and HDBSCAN clustering analyses.

The K-Means algorithm was selected as the primary method for the clustering analyses. To determine the optimal number of clusters, the Silhouette score, Calinski-Harabasz index, and Gap statistic were evaluated collectively. To render these performance metrics -calculated on varying scales- comparable, they were normalized to a 0-1 range. Subsequently, in calculating the final evaluation score, weights of 40%, 30%, and 30% were applied to the Silhouette score, Calinski-Harabasz index, and Gap statistic, respectively. To construct the composite clustering performance score, a slightly higher weight (40%) was assigned to the Silhouette index because it simultaneously captures within-cluster cohesion and between-cluster separation, while the Calinski-Harabasz index and Gap statistic were included as complementary validation measures with equal weights (30% each). This approach ensured that clustering performance was analyzed within a multi-criteria evaluation framework rather than relying on a single metric.

In the context of portfolio analysis, it was considered that excessively low cluster counts would inadequately reflect the differences among project groups. Therefore, the analyses were conducted for cluster numbers ranging from C=4 to C=10, and each alternative clustering solution was compared in terms of the relevant performance metrics. This process aimed to determine the optimal number of clusters that best represents the dataset's structure and most meaningfully reveals the thematic differentiation among the projects.

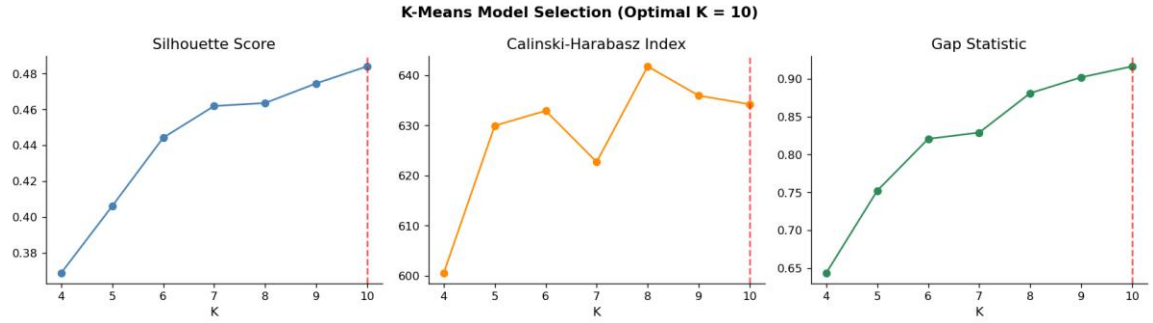



**Figure 5.** K-means metric evaluation: Silhouette score, Calinski–Harabasz index, and gap statistic (C=4-10)

The highest integrated evaluation score, derived from combining the clustering performance metrics, was observed at C=10. Although the LDA analyses identified seven as the optimal number of topics for the dataset, C=10 was selected for the clustering analyses because its integrated score was over 10% higher than that of the C=7 solution, thereby yielding the strongest numerical performance. This approach relies on the premise that topic modeling and clustering analyses serve distinct purposes; thus, their optimal parameters do not necessarily coincide.

To achieve more stable and robust results, the final K-Means model was trained with the parameters *n_init*=30 and *max_iter*=500. Performance evaluation of the model yielded a Silhouette score of 0,484 and a Calinski-Harabasz (CH) index of 634,2. These metrics indicate a meaningful separation between the clusters and demonstrate that project titles within the same cluster exhibit relatively homogeneous characteristics. Consequently, the findings suggest that the generated clusters adequately reflect the structural and thematic variances within the dataset, indicating that the K-Means model delivers an acceptable performance for classifying the project portfolio.

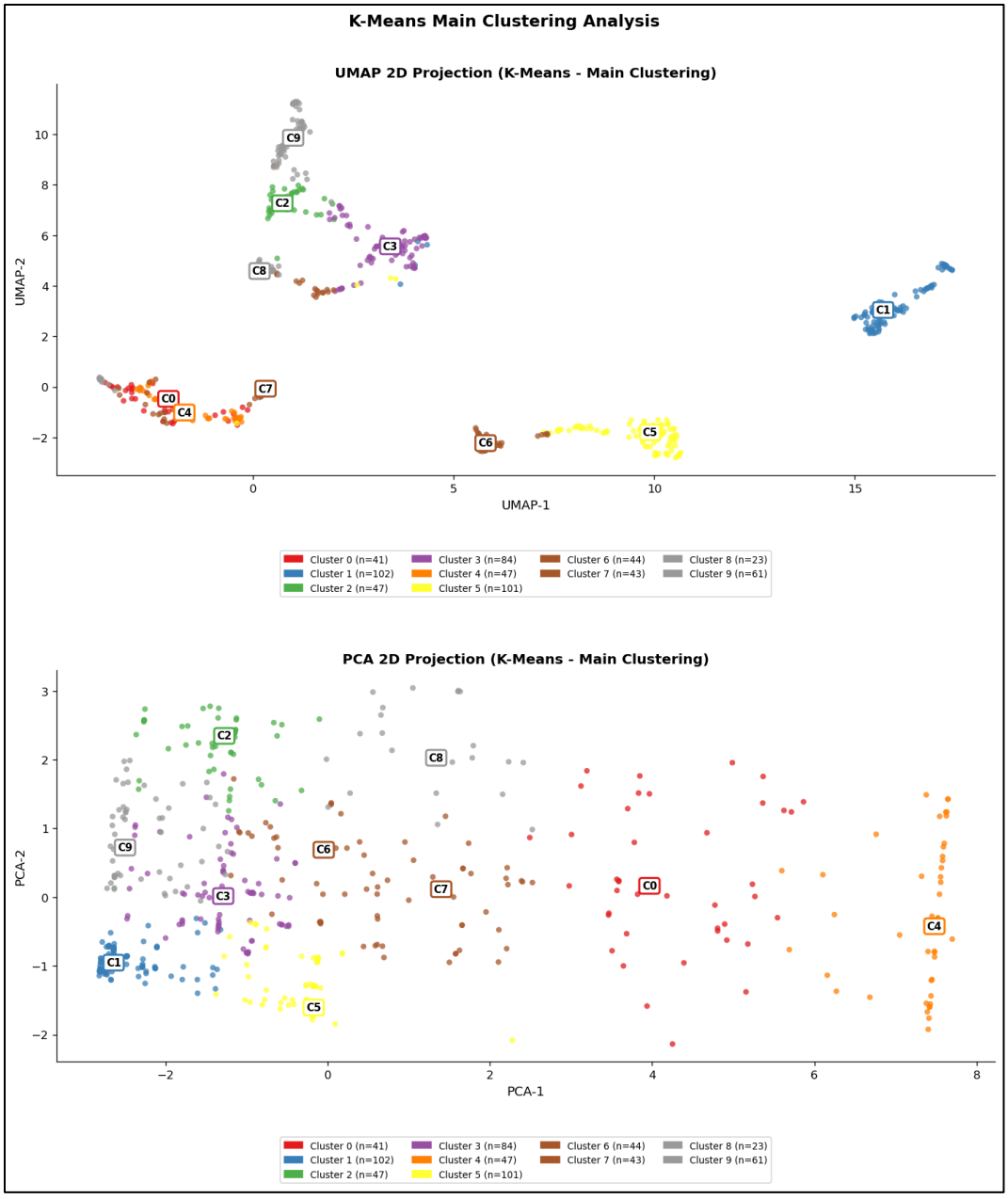



**Figure 6.** K-means cluster structure: PCA and UMAP 2D projections (# of C=10)

The K-Means clustering analysis was executed on the PCA-reduced dataset. Upon determining the optimal number of clusters, the final K-Means model was retrained, and the resulting cluster labels were appended to each project in the dataset. The formed cluster structure was then visualized via two-dimensional PCA and UMAP projections to evaluate the spatial distributions of the clusters. These visualizations enabled a clearer examination of inter-cluster separation and intra-cluster data point positions.

To assess the reliability and robustness of the K-Means cluster structure, a supplementary validation analysis was conducted using the HDBSCAN algorithm, a density-based clustering method. The primary advantage of HDBSCAN is that it bypasses the need for a predetermined number of clusters, instead automatically identifying them based on the dataset's inherent density structure. Consequently, this method is highly effective in detecting heterogeneous data structures and clusters with varying density levels.

Within this study, HDBSCAN served to validate the K-Means findings through an alternative approach. During the analysis, the minimum cluster size parameter (min_cluster_size) was established based on dataset characteristics, allowing the algorithm to automatically generate the final cluster structure. Comparing the K-Means and HDBSCAN results thus facilitated an evaluation of structural consistency and the extent to which the derived clusters accurately reflected the underlying data architecture.

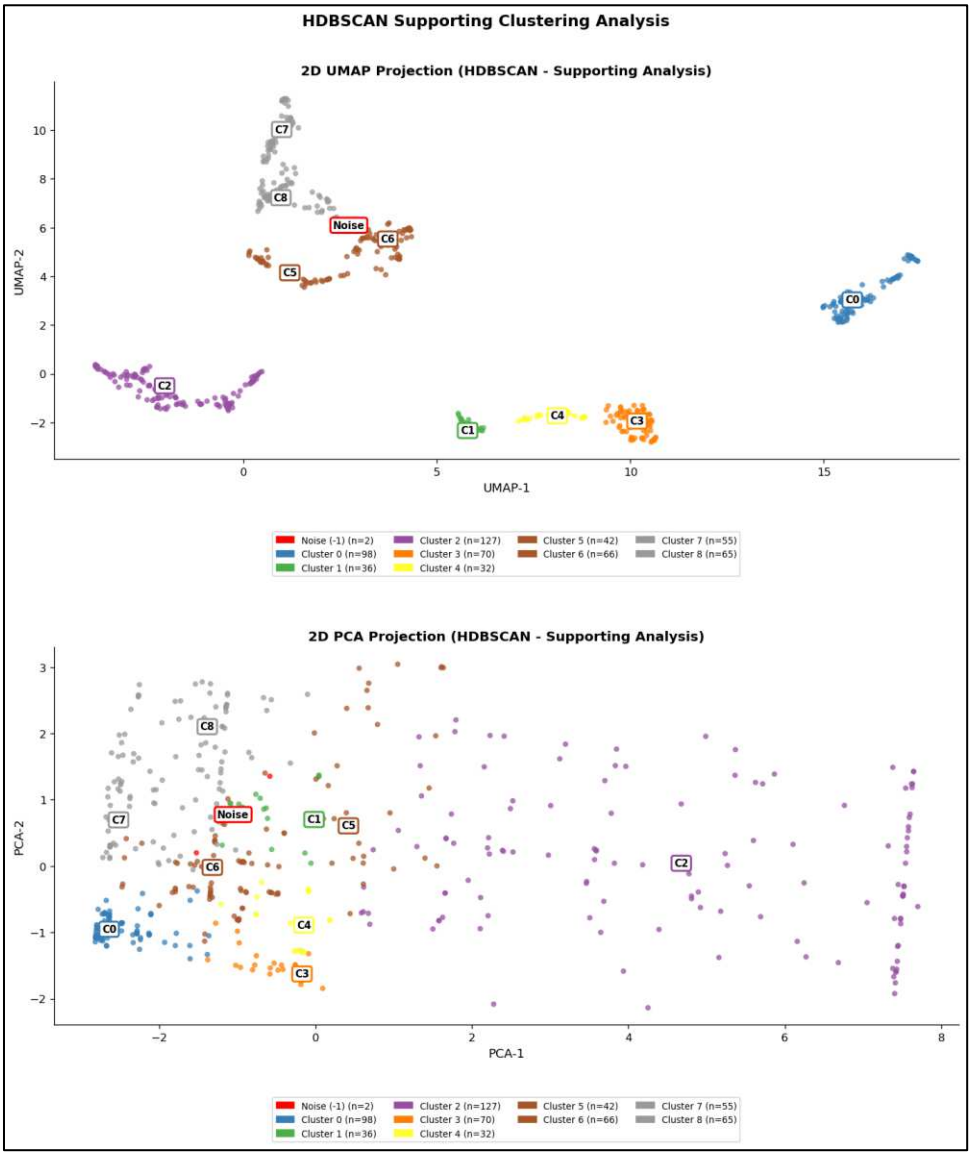



**Figure 7.** HDBSCAN cluster structure: UMAP and PCA 2D projections

Following the HDBSCAN analysis, a total of nine clusters were identified within the two-dimensional UMAP space, with only two data points (0,3%) classified as noise. When these noise observations were excluded from the analysis, the Adjusted Rand Index (ARI), calculated to evaluate the agreement between the K-Means and HDBSCAN methods, yielded a value of 0,6512.

The ARI is a performance metric utilized to measure the similarity between two distinct clustering outcomes, with values ranging from -1 to +1. An ARI value of 1 indicates complete overlap between the clustering results, while a value of 0 signifies that the observed similarity is purely random. In this context, although the calculated ARI of 0,6512 does not indicate perfect agreement, it demonstrates a significant and substantial similarity between the two clustering methods.

These findings indicate that the cluster structure identified by the K-Means algorithm is largely supported by HDBSCAN, an alternative density-based method. Specifically, the concentration of certain clusters in similar regions within the UMAP projection and the identification of analogous data groups by both methods provide additional evidence for the consistency of the clustering results. This demonstrates that the K-Means-derived cluster structure adequately reflects the natural organization of the dataset and that the results are reliable in terms of cross-method validation.

### 3.6. Profiling and Anomaly Detection

In this phase of the study, Isolation Forest and Local Outlier Factor (LOF) methods were applied to determine the risk levels of the projects and identify potential outliers. Variables reflecting the fundamental characteristics of the projects -budget, project duration, and budget/month ratio- were utilized in the risk profiling process. To prevent scale discrepancies among these variables from skewing the results, a standardization procedure was executed initially.

Within the scope of the analysis, the Isolation Forest method was employed to detect projects that deviated significantly from the overall dataset. Additionally, the LOF method was applied to assess the positions of the projects within their local density structures, thereby identifying observations exhibiting anomalous behavior relative to their nearest neighbors. Consequently, outlier analysis was performed concerning both the global dataset structure and local neighborhood relationships.

Following the application of these methods, the resulting anomaly scores were normalized to a 0-1 range to ensure comparability. Subsequently, the normalized scores from both methods were averaged to compute an integrated risk score for each project. By combining the strengths of distinct outlier detection techniques, this approach yielded a more balanced and robust risk assessment.

In the final stage, projects were classified into three risk profiles: low, medium, and high risk, based on their composite risk scores. Rather than relying on predefined fixed thresholds, risk categories were determined from the empirical distribution of the composite anomaly scores. Specifically, the scores were partitioned into three equally sized groups using quantile analysis, and projects were assigned to low-, medium-, or high-risk categories according to their position within the distribution. This data-driven classification enabled a systematic assessment of risk dispersion

across the project portfolio and facilitated the identification of projects requiring priority monitoring and managerial attention.

### 3.7. Statistical Tests and Robustness

At this stage of the study, it was examined whether there were statistically significant differences among the clusters in terms of budget, project duration, and budget/month variables. To this end, the distribution characteristics of the variables were first evaluated; One-Way ANOVA was applied when parametric test assumptions were met, and the non-parametric Kruskal-Wallis test was used when these assumptions were violated. Consequently, it was determined whether the observed differences among the clusters were statistically significant.

For variables where significant differences were identified, Dunn's multiple comparison test with Bonferroni correction was performed to determine between which specific clusters these differences occurred. Through this approach, it was evaluated to what extent the structures derived from the clustering algorithms diverged not only computationally but also statistically. Thus, the levels of significant differentiation among the clusters were validated using objective statistical methods.

Furthermore, a robustness analysis was conducted to evaluate the reliability and reproducibility of the findings. This analysis aims to examine the extent to which the clustering structure obtained from the applied methods remains stable under varying conditions. Within the scope of the robustness assessment, the K-Means model served as the baseline, and the algorithm was executed multiple times using different random seeds. The resulting outcomes were then compared to analyze the stability of the cluster structure.

The Silhouette score and the Adjusted Rand Index (ARI) were utilized to assess consistency among the clusters (Figure 8). Through these metrics, the degree of similarity between clustering results obtained from different runs was determined. In the final stage, to evaluate the impact of different text representation methods on the results, clusters derived using SBERT- and TF-IDF-based representations were compared. Consequently, the effects of the text representation approach on clustering performance and cluster structure were investigated; the sensitivity of the obtained results to the applied methods was assessed, thereby reinforcing the methodological robustness of the analyses.

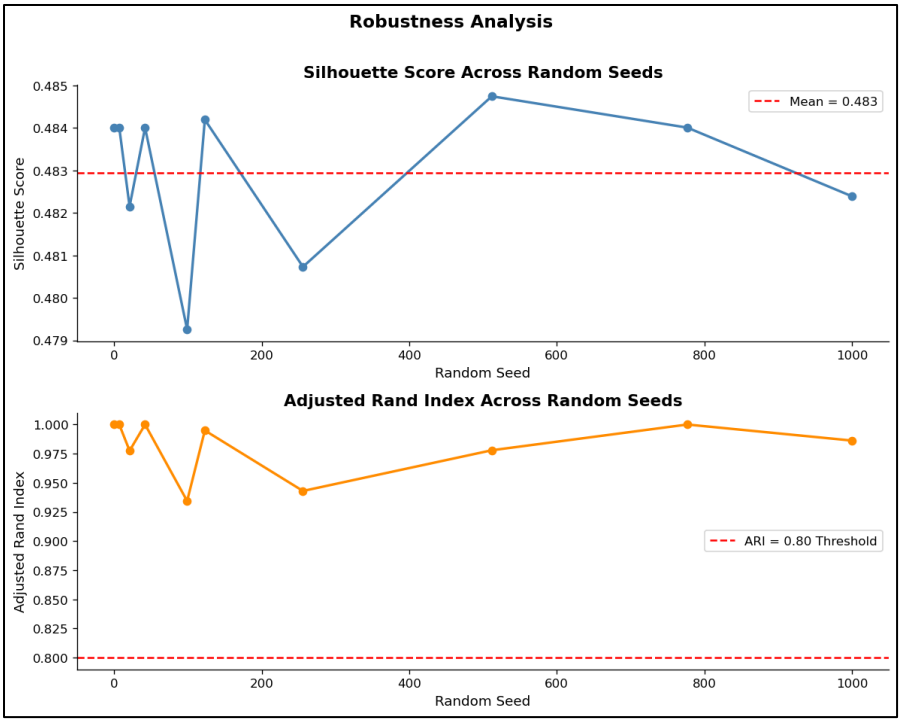



**Figure 8.** Robustness analysis: Silhouette score and ARI stability across different random seed values

The robustness analysis results revealed that the K-Means model exhibits a high degree of stability under different initialization conditions. It was determined that the Silhouette score ranged between 0,479 and 0,485 across all random seed values, with a mean of 0,483 and a standard deviation of 0,0017. Similarly, the ARI values, calculated to assess the consistency of cluster assignments, ranged from 0,934 to 1,000, with a mean ARI of 0,981 and a standard deviation of 0,023. The exceptionally low variability in Silhouette scores, combined with high ARI values, demonstrates that cluster quality and membership largely remain preserved despite varying initial conditions. These findings indicate that the K-Means model is highly resilient to random initialization values and yields reproducible results.

To evaluate the impact of different text representation methods on clustering outcomes, Sentence-BERT (SBERT) and Term Frequency-Inverse Document Frequency (TF-IDF) approaches were compared. The obtained results indicate that the SBERT-based representation performs better in terms of clustering performance. While the Silhouette score for the clustering solution achieved via the SBERT method was calculated as 0,484, this value remained at approximately 0,421 in the TF-IDF-represented model. This finding demonstrates that SBERT more effectively captures semantic relationships among texts and provides a more pronounced separation among clusters.

On the other hand, the ARI, computed in Python to quantify the agreement between the clustering solutions obtained from the two representation methods, was 0,012. Such a low ARI value suggests a near absence of correspondence between the SBERT- and TF-IDF-based cluster assignments, indicating that the two approaches capture markedly different structural patterns within the dataset.

This finding highlights the critical role of text representation in unsupervised learning applications, as the choice of representation method can substantially influence clustering outcomes and the resulting interpretation of thematic structures. Overall, the SBERT-based representation was found to be more appropriate for the purposes of this study, given its superior clustering performance and stronger ability to preserve semantic relationships among texts.

### 3.8. Strategy Map and Visualization

At this stage, a strategy map was generated (Figure 9) to present the analysis results in a more holistic and easily interpretable manner. This map was visualized on two-dimensional layout planes obtained via UMAP and PCA. Each point on the map represents a project, where point colors indicate the clusters to which the projects belong, point sizes reflect project budgets, and point opacities designate budget-per-month intensities. Additionally, projects highlighted with red borders represent those classified in the high-risk category. Cluster centers were also marked on the map, allowing project cluster memberships, budget characteristics, and risk levels to be evaluated simultaneously on the same visual plane. This approach enabled a more comprehensive examination of the structural features of the project portfolio and the differences among clusters.

To support the analyses, distinct word clouds were also prepared for each cluster (Figure 10). The word clouds served to visually summarize prominent concepts and thematic contents within the clusters, contributing to a clearer understanding of the subject areas represented by each cluster. Thus, contextual differences among clusters were rendered more visible, the interpretability of the findings was enhanced, and thematic segregations were clearly demonstrated.

Following the clustering analyses, various comparative visualizations were prepared to reveal the fundamental differences among the clusters. To this end, violin plots and box plots were utilized to compare the distributions of the clusters in terms of budget, duration, and budget/month variables. These charts facilitated a detailed evaluation of distribution characteristics, central tendencies, and variability levels across the clusters.

Furthermore, risk heat maps were generated on the UMAP and PCA planes to examine the distribution of risk scores within the data space. Through these visualizations, the locations of projects with high-risk scores within the data were identified, and the regions where risk concentrations emerged were analyzed. Consequently, the relationship between risk profiles and cluster structures could be visually assessed.

## 4. Results

### 4.1. Strategy Map and Cluster Characterization

An examination of the strategy map presented in Figure 9 reveals that projects are positioned in low-dimensional spaces obtained using UMAP and PCA methods. This visualization enables a holistic evaluation of the structural features, cluster distributions, and risk profiles of the project portfolio.

Each point on the map represents a project, with point colors indicating the clusters to which the projects belong. While point sizes reflect project budgets, point opacities designate resource utilization intensity calculated via the budget/month variable. Consequently, not only the thematic positions of the projects but also their resource utilization and budget magnitudes can be evaluated within the same visualization.

Additionally, points marked with red borders represent projects identified as high-risk by the Isolation Forest and Local Outlier Factor (LOF) methods. This representation allows for a visual examination of the distributions of high-risk projects within clusters, their positions relative to other projects, and potential concentration regions. Ultimately, the strategy map serves as a comprehensive decision-support tool that enables the simultaneous analysis of relationships among cluster memberships, budget magnitudes, resource intensity, and risk profiles.

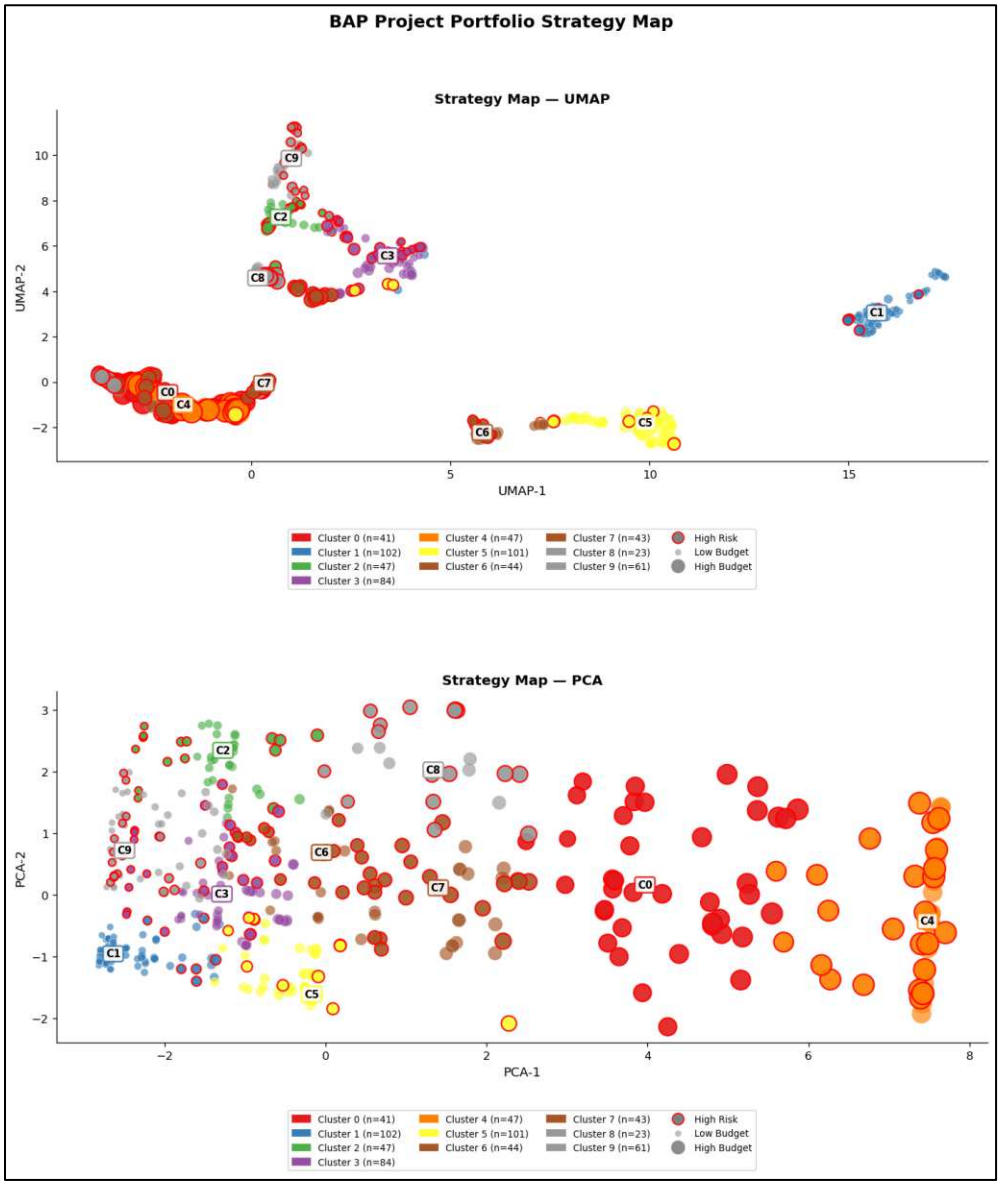



**Figure 9.** Strategy map: UMAP and PCA projections

An examination of the UMAP layout reveals that cluster C1 (the blue-colored cluster in the middle-right region) is spatially distinct and clearly segregated from the other clusters. This indicates that this cluster possesses a more unique and homogeneous thematic structure compared to other project groups in the dataset. In particular, the shared content characteristics of the projects within the cluster have contributed to their close proximity on the UMAP plane and their clearer separation from other clusters.

When evaluated in conjunction with the findings from the word cloud analysis, it is understood that cluster C1 predominantly comprises projects in the fields of materials science, advanced materials, and nanotechnology. Therefore, the prominent separation observed in the UMAP space reflects not merely a spatial difference, but also the unique thematic character inherent to the cluster. The obtained results demonstrate that cluster C1 represents a more specialized research domain within the project portfolio, differing significantly from the other clusters.

### 4.2. Thematic Distribution and Portfolio Structure

An examination of the cluster-based word clouds (Figure 10) clearly reveals that each cluster is formed around specific thematic areas. The size of the words in the word clouds reflects the relative frequency of the corresponding term within the cluster, thereby enabling the visual identification of the dominant subject areas of the cluster.

The obtained results demonstrate that the groups formed through clustering analysis exhibit meaningful differences not only in terms of numerical characteristics but also regarding content and theme. The prominent words in each cluster provide important clues regarding the research or project fields represented by that cluster. This supports the notion that the clustering results are thematically consistent and that project titles are grouped under similar contents.

Furthermore, the word clouds facilitated a visual comparison of contextual differences among the clusters and contributed to a more straightforward interpretation of the characteristic features of each cluster. Thus, they served as a complementary analytical tool for validating and interpreting the thematic structures derived from the clustering analyses.

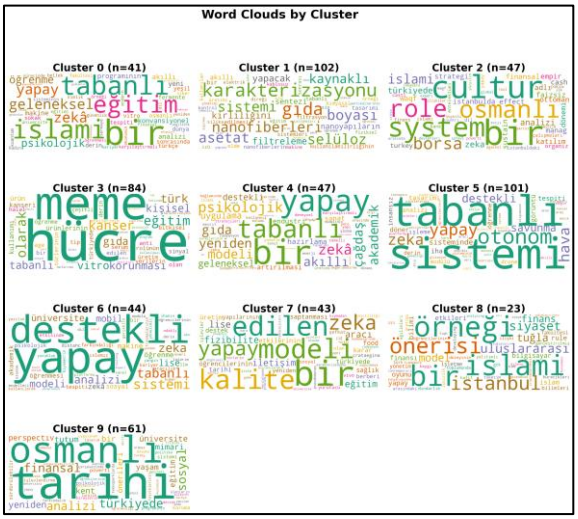



**Figure 10.** Cluster-specific word clouds

Word cloud analysis was utilized as an essential visualization tool supporting the quantitative findings of the clustering results in terms of thematic content. An examination of the obtained word clouds reveals that each cluster concentrates around specific research fields and subject headings. This demonstrates that the groups generated by the clustering algorithm exhibit meaningful segregation not only statistically but also regarding content and theme.

For instance, the prominence of terms such as "nanofibers", "characterization", and "cellulose" in cluster C1 indicates that the projects within this group predominantly focus on materials science, advanced materials, and nanotechnology. Similarly, the high frequency of expressions such as "artificial", "intelligence", "based", and "supported" in cluster C4 shows that this group represents AI and data-driven engineering applications. In cluster C5, the concurrent prominence of concepts such as "autonomous", "defense", "system", and "analysis" reveals that this cluster reflects studies in the fields of autonomous systems, decision support mechanisms, and defense technologies.

These findings demonstrate that the thematic patterns derived from the word clouds are consistent with the clustering results and that each cluster possesses a distinct research focus. Consequently, word clouds make an important contribution as a complementary tool in explaining the contextual characteristics of the clusters and interpreting thematic differentiations within the project portfolio.

### 4.3. Cluster-Based Budget, Duration, and Budget/Month Distributions

In Figure 11, the distributions of budget, project duration, and budget/month variables across clusters are visualized using violin plots with embedded box plots, displaying outliers as well. These visualizations enable the simultaneous evaluation of not only central tendency measures but also data densities, distribution shapes, and potential outlier observations.

An examination of the graphs reveals that Cluster 0 and Cluster 4 diverge significantly from the other clusters, particularly regarding the budget and budget/month variables. The relatively high median values in both clusters indicate that the projects within these groups possess larger budget magnitudes and higher monthly resource utilization intensity. In addition, the shape and width of the violin plots suggest that the distribution structures of these clusters exhibit distinct characteristics compared to the others.

When evaluating the density structures of the distributions, it is observed that data concentrate within certain ranges in some clusters, while exhibiting a wider spread in others. Furthermore, the outliers indicated on the box plots demonstrate that specific projects can deviate significantly from the overall trend within their respective cluster. These findings reveal that the groups derived from the clustering analysis diverge from one another not only thematically but also in terms of budget and resource utilization characteristics. Thus, structural differences among the clusters are supported both visually and statistically.

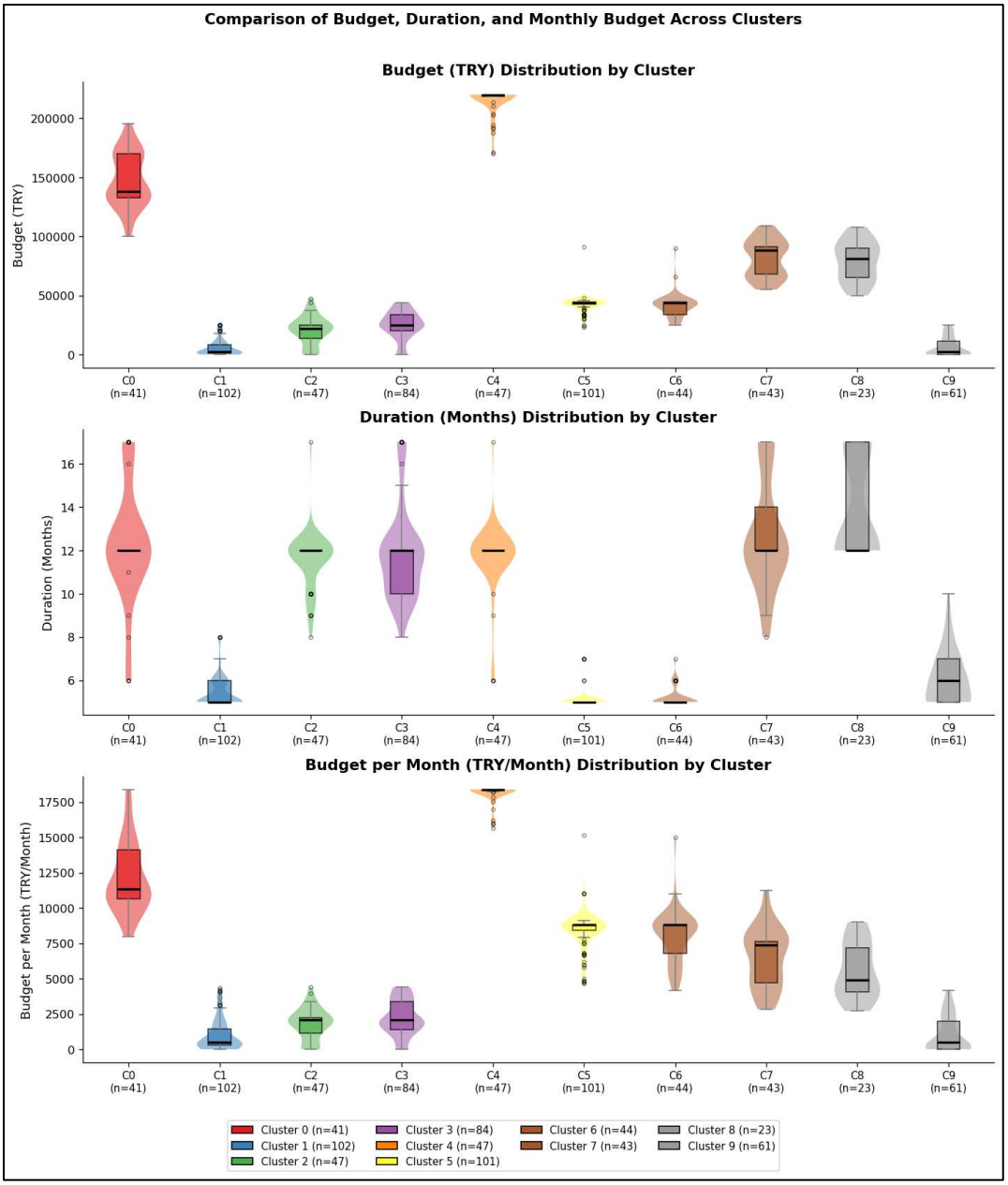



**Figure 11.** Cluster-based distributions of budget, project duration, and budget per month (violin and box plots)

### 4.4. Risk Clusters and Anomaly Heatmap

Figure 12 presents the risk density map generated by combining the results obtained from the Isolation Forest and LOF methods. This visualization enables the evaluation of project risk profiles on a low-dimensional space and uncovers patterns in risk distribution.

An analysis of the risk distribution reveals that high-risk projects are more densely concentrated, particularly in clusters C0, C8, C7 and C4. Cluster C0 stands out as the riskiest group within the

portfolio, with a high-risk proportion of 100%, as all of its projects fall into the high-risk category. Similarly, clusters C8 (70%), C7 (56%), and C4 (53%) are among the groups possessing high risk concentrations. In contrast, the low-risk proportion of only 6% in cluster C1 indicates that this cluster has one of the lowest risk profiles within the portfolio.

It is observed that risk density exhibits a tendency to cluster in specific thematic regions on the UMAP plane. This indicates that certain groups with similar content and project characteristics may share common risk traits. On the PCA plane, however, the risk distribution appears to concentrate around closely positioned and partially overlapping clusters. This discrepancy stems from the distinct ways in which the UMAP and PCA methods represent the data structure.

Overall, it is understood that risk does not exhibit a homogeneous distribution across the project portfolio. Particularly in clusters where high-budget and long-duration projects concentrate, risk levels become more pronounced. This finding suggests that as project scale and resource utilization intensity increase, the risk level also tends to rise. Consequently, the risk density map provides an essential decision-support tool for identifying project groups within the portfolio that necessitate priority monitoring and evaluation.

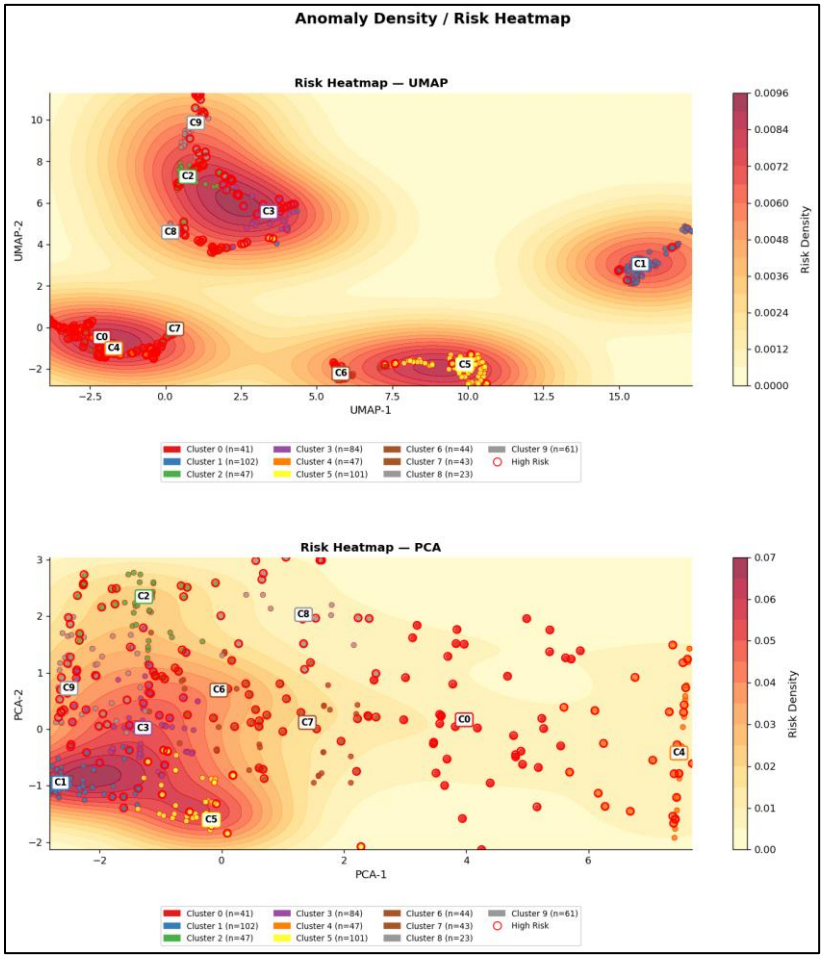



**Figure 12.** Anomaly density and risk heat map: UMAP and PCA projections (LOF and Isolation Forest)

An examination of the risk heat map reveals that risk density tends to cluster in specific regions, particularly within the UMAP space. It is observed that risk scores reach higher values especially in the areas where clusters C2 and C3 are located. This indicates that projects within these clusters exhibit similar risk profiles and share common risk factors in terms of certain structural features. On the PCA plane, however, the risk distribution is predominantly concentrated in the lower-left region, which overlaps with areas where multiple clusters partially intersect.

When analyzing the projects with the highest risk scores, it is noteworthy that risk is not randomly distributed across the dataset; rather, it tends to concentrate around specific clusters and regions. The prominent presence of high-risk projects in certain clusters suggests that these groups may exhibit more vulnerable characteristics not only in terms of thematic content but also regarding budget management, resource utilization, and project execution processes. This finding indicates that project risks may be associated with thematic structure and operational features.

Overall, it is understood that the risk distribution across the project portfolio does not exhibit a homogeneous structure. Risk density becomes more pronounced particularly in clusters CO, C8, C7 and C4, where high-budget and long-duration projects predominate. This result demonstrates that as project scale and resource requirements increase, the risk level also tends to rise. Consequently, closely monitoring the projects within these clusters and prioritizing risk management processes can be evaluated as a significant imperative for portfolio management.

## 5. Discussion

The findings obtained from this study clearly demonstrate the importance of evaluating research portfolios in higher education institutions not merely through individual projects, but also from a clustering and network/topological structure perspective. The findings indicate that comprehensively examining the thematic structures, risk patterns, and resource distributions of project portfolios can make significant contributions to strategic decision-making processes. In particular, the concentration of high-risk projects within specific clusters suggests that TTOs should be positioned not only as supportive units but also as structures equipped with proactive monitoring mechanisms to detect and manage risks at an early stage. These results are also consistent with the findings of previous studies conducted on risk concentration and strategic portfolio management in R&D project portfolios (Teller and Kock, 2013; Meskendahl, 2010).

An examination of the inter-faculty project distribution reveals that the Faculty of Engineering and Natural Sciences holds a distinct superiority in terms of budget share. This indicates that technical disciplines gain a significant advantage regarding project generation and access to external funding. Nevertheless, it is understood that there is a need for strategic initiatives aimed at enhancing project development capabilities in fields such as Health Sciences, Social Sciences, and others. Equipping the university's research portfolio with a more balanced and diversified structure can be considered a key priority for long-term scientific productivity and institutional sustainability.

One of the striking results of the study is the uncovering of thematic structures derived through NLP-based clustering analyses. Thanks to this approach -which transcends traditional faculty- or department-based classifications- implicit thematic relationships among projects were identified.

For instance, despite appearing to belong to different disciplines, certain projects in health, education, and engineering fields were observed to reside within the same cluster under the common theme of AI-supported systems. This finding points to the existence of an interdisciplinary collaboration potential within the university that has not yet been adequately utilized. Similarly, international research assessment studies emphasize that thematic mapping methods are effective in uncovering interdisciplinary relationships (Rafols et al., 2012).

Evaluated from a methodological standpoint, UMAP -a non-linear dimensionality reduction technique- was observed to yield more distinct cluster separations compared to PCA, a linear method. This result demonstrates that, particularly in text-based and high-dimensional datasets, non-linear dimensionality reduction methods can more effectively reflect complex relationships within the data structure. The clearer separation of clusters in UMAP projections indicates that local neighborhood relationships and thematic similarities within the data can be preserved more successfully.

The K-Means model producing a very high mean ARI (ARI=0,981) under different random initialization values demonstrates that the model exhibits a high degree of stability against initialization conditions and provides reproducible results. Furthermore, the ARI value of 0,6512 calculated between the K-Means and HDBSCAN algorithms indicates that the two methods do not generate identical cluster structures, but there is a meaningful level of structural agreement between the clusters. This finding supports the notion that the obtained clustering results are preserved to a certain extent independently of the method and that a natural cluster structure exists within the dataset.

As with any research, this study also has certain limitations. First, the content heterogeneity of project abstracts and the brevity of descriptions in some projects may limit the scope and depth of text-based analyses. Additionally, the fact that project outputs (publications, patents, commercialization, utility models, etc.) are not yet reported within a fully standardized indicator framework may introduce a degree of uncertainty in output-oriented evaluations. Moreover, the unsupervised learning methods utilized in the study inherently require researcher interpretation and domain expert evaluations.

It is recommended that future studies be conducted on larger datasets, supported by comparative analyses encompassing different universities and research institutions, and that project outputs be evaluated within a framework of standardized performance indicators. Such studies will make significant contributions to testing the generalizability of the proposed analytical framework, strengthening its external validity, and further developing risk-based project portfolio management models. In addition, the integration of AI-driven decision support systems into project portfolio management is considered an important research avenue for future studies regarding the more efficient utilization of research resources and the strengthening of strategic planning processes.

## 6. Conclusion and Recommendations

This study has comprehensively and data-driven analyzed a research portfolio consisting of 631 SRPs conducted between the 2015–2026 period, utilizing NLP and ML techniques. Thanks to the

developed methodological framework, the thematic structure, resource utilization patterns, risk profiles, and clustering dynamics of the project portfolio were evaluated from a holistic perspective. When examined in the context of strategic management of research portfolios, evaluation of institutional research performance, and data-driven decision-making processes, the findings reveal the following fundamental conclusions.

### 6.1. Methodological Validation and Strategic Mapping

The semantic analysis and clustering methods applied in the study revealed the presence of distinct thematic and structural patterns within the research portfolio. In this context, the K-Means algorithm (C=10) was utilized as the primary clustering method, while the HDBSCAN density-based clustering approach and the LDA topic modeling method were evaluated as supplementary analytical tools. As a result of the combined application of Isolation Forest and LOF methods in the risk profiling process, a total of 98 projects (16,5%) were determined to exhibit anomalous characteristics.

The findings regarding clustering performance indicate that the model exhibits a high degree of stability and reliability. The obtained Silhouette score of 0,484 and mean Adjusted Rand Index (ARI) of 0,481 demonstrate that similar cluster structures are preserved under varying initialization conditions and that model results are highly reproducible. Furthermore, the superior performance of the Sentence-BERT-based semantic representation approach compared to the TF-IDF-based representation shows that language model-based methods can produce meaningful and reliable results in classifying academic project portfolios.

The findings indicate that the strategic alignment of research portfolios can be examined not only based on expert evaluations but also through data-driven and objective methods. Ten distinct clusters defined in this direction were meaningfully separated and interpreted in terms of research themes, budget structures, resource utilization characteristics, and risk levels. Thus, by evaluating the structural features of the project portfolio within a holistic framework, significant insights concerning institutional research strategies were derived.

### 6.2. Risk Management and Proactive TTO Interventions

The risk analysis results indicate that 198 out of 593 projects (33,4%) in the project portfolio were classified as high-risk projects. This finding highlights the importance of risk-oriented approaches in project portfolio management and indicates that high-budget projects concentrated particularly in clusters C0 and C4 necessitate closer monitoring. The high-budget concentration and risk accumulation observed in clusters C0 and C4 underscore the significance of early warning systems for TTOs and research management units. The findings show that tracking project performance solely through periodic reports may be insufficient and that data-driven decision-making support systems where risk indicators are continuously monitored are required. Through such an approach, potential performance deviations, time delays, and issues related to resource utilization can be identified at early stages, enabling timely preventive measures.

The differences in budget and resource utilization observed among clusters indicate that certain disciplines hold a more advantageous or diligent position regarding access to financial resources within the research ecosystem. The relatively low resource utilization intensity observed

particularly in fields outside of engineering points to the necessity of strengthening support mechanisms aimed at enhancing research capacity. Therefore, in evaluating research portfolios, it is crucial to consider not only performance indicators but also elements such as interdisciplinary diversity, resource distribution equity, and the balance among research fields.

### 6.3. Policy Recommendations and Future Work

In line with the research findings, various strategic recommendations can be formulated for TTOs and research management units.

**Strengthening a Balanced Research Ecosystem**: To achieve a more sustainable and balanced structure for the research portfolio, project development capabilities of faculties outside of engineering fields must be enhanced, their access to financial resources supported, and research incentive mechanisms strengthened. This approach will contribute to the formation of a more inclusive and balanced research ecosystem by reducing the concentration of research activities in specific disciplines.

**Development of Data-Driven and Intelligent Monitoring Systems**: SRP and research project processes should be transitioned from a purely administrative tracking-oriented structure and empowered with research analytics and ML-based decision support mechanisms. Consequently, identifying high-risk projects at an early stage, optimizing resource allocation processes, and continuously monitoring project performance will become feasible. Furthermore, the developed data-driven systems will assist managers in making faster and more reliable evaluations during strategic decision-making processes.

**Recommendations for Future Research**: The model developed in this study can be expanded with more comprehensive data sources whenever available. In future investigations, incorporating performance indicators such as:

- Publication counts,
- Citation performance,
- Patent and utility model outputs,
- Commercialization indicators,
- External funding and international project acquisitions

into the analysis is recommended. Consequently, not only the content and risk dimensions of research portfolios, but also their scientific and economic impacts can be evaluated.

Additionally, utilizing contemporary neural topic modeling approaches such as BERTopic can contribute to uncovering thematic relationships within project descriptions in greater detail. Comparative studies conducted with data obtained from different universities will allow for testing the generalizability of the proposed approach and strengthening its external validity.

In conclusion, this study demonstrates that NLP and ML methods are effective and viable tools in the analysis of research portfolios. The study provides TTOs and research management units with a holistic analytical framework through which research inventories can be evaluated as a strategic

data asset. In this respect, the study delivers significant methodological and practical contributions to the literature on project portfolio management, research analytics, and technology transfer.

## 7. Acknowledgements and Ethical Statement

The data utilized in this study were provided by the Istanbul Sabahattin Zaim University Technology Transfer Office (TTO). The institution is gratefully acknowledged for its support during the data acquisition process.

No financial support was received from any institution, organization, or funding agency for this study. All data used in the research were de-identified and anonymized in accordance with privacy and data security principles prior to analysis.

The authors declare that there are no conflicts of interest regarding this study. Furthermore, all authors have made significant contributions to conceptualization, methodological design, data analysis, interpretation of findings, and reporting processes of the research, and have approved the final version of the manuscript.

No personal data were utilized in the research, and no experimental procedures or data collection processes involving human participants were conducted. Therefore, the study does not require ethics committee approval within the framework of current ethical research principles.

## 8. References

Abramo, G., D'Angelo, C. A., & Felici, G. (2020). The relationship between research productivity and efficiency: A data-driven approach. *Journal of Informetrics, 14*(4), 101017. https://doi.org/10.1016/j.joi.2017.09.007

Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, *9*, 100341. https://doi.org/10.1016/j.dajour.2023.100341

Aria, M., & Cuccurullo, C. (2020). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Scientometrics, 125*, 1–26. https://doi.org/10.1016/j.joi.2017.08.007

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 93–104. https://doi.org/10.1145/342009.335388

Chapple, W., Lockett, A., Siegel, D. S., & Wright, M. (2005). Assessing the relative performance of U.K. university technology transfer offices: Parametric and non-parametric evidence. *Research Policy, 34*(3), 369–384. https://doi.org/10.1016/j.respol.2005.01.007

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74–77. https://doi.org/10.1145/2254556.2254572

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. *Proceedings of ACL*, 2270–2282. 10.18653/v1/2020.acl-main.207

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. 10.18653/v1/N19-1423

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York: Wiley.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, 413–422. https://doi.org/10.1109/ICDM.2008.17

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*(14), 281–297. https://doi.org/10.1109/ICDM.2008.17

McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density-based clustering. *Journal of Open-Source Software, 2*(11), 205. https://doi.org/10.21105/joss.00205

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

Meskendahl, S. (2010). The influence of business strategy on project portfolio management and its success - A conceptual framework. *International Journal of Project Management, 28*(8), 807–817. 10.1016/j.ijproman.2010.06.007

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy, 26*(3), 317–330. https://doi.org/10.1016/S0048-7333(97)00013-9

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., ... & Sobrero, M. (2013). Academic engagement and commercialization: A review of the literature on university–industry relations. *Research Policy, 42*(2), 423–442. https://doi.org/10.1016/j.respol.2012.09.007

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy, 41*(7), 1262–1282. https://doi.org/10.1016/j.respol.2012.03.015

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., ... & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods, 8*(6), 443–444. https://doi.org/10.1038/nmeth.1619

Teller, J., & Kock, A. (2013). An empirical investigation on how portfolio risk management influences project portfolio success. *International Journal of Project Management, 31*(6), 817–829. 10.1016/j.ijproman.2012.11.012

Tshitoyan, V., Dagdelen, J., Weston, L., et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Zhao, X., Li, Y., & Chen, H. (2023). Artificial intelligence for research analytics: A bibliometric and machine learning approach. *Scientometrics.* https://doi.org/10.3390/math10234398