Advancing Idea Management Systems with Topic Modeling: Radical Innovation Patterns in Student-Driven Industry Cases

Serena Leka
Department of Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
sela@ece.au.dk

Aman Shah
Department of Computer Science
Tufts University
Boston, United States of America
aman.shah@tufts.edu

***Abstract*—The integration of Topic Modeling (TM) in Idea Management Systems (IMSs) is gaining attention as organizations seek structured approaches to idea classification, trend detection, and knowledge discovery. This study investigates the application of BERTopic and K-Means clustering in an academic IMS, analyzing 5492 innovation ideas generated by multidisciplinary engineering students at Aarhus University as they tackled 47 real-world industry challenges in Denmark. By structuring idea generation through the Creative, Idea, Solution (CIS) framework and leveraging the Rosetta IMS, this research examines how radical innovation patterns might emerge and whether topic modeling can enhance the flow of activities in an innovation process.**

Findings indicate that BERTopic effectively categorizes industry-driven ideas into 19 meaningful themes, aligning with key sectors such as sustainability, automation, waste management, and health-tech. In contrast, K-Means clustering identifies 7 broader innovation clusters, highlighting mainstream vs. underdeveloped idea trajectories throughout the innovation process. TF-IDF analysis was used to track idea evolution across IMS stages, revealing that solution-oriented keywords dominate later phases, while many early-stage concepts are abandoned before full development.

Results suggest that TM-based classification provides structure to innovation landscapes efficiently but requires qualitative validation from human interactions to confirm the emergence of radical ideas. The combination of BERTopic, K-Means, and GPT-based summarization enhances interpretability and trend forecasting, offering an initial AI-enabled framework for IMS implementation. This study contributes to innovation management research and practical IMS applications, demonstrating how AI models can streamline idea management and align student-driven innovations with industry needs.

***Index Terms*—clustering algorithms, innovation management, engineering education, business**

INTRODUCTION

In a time where organizations are eager to innovate rapidly to maintain competitive advantage and keep up with technological advancements, Idea Management Systems (IMSs) have emerged as relevant tools to systematically support innovation processes. IMSs are structured frameworks, primarily digital, used to facilitate innovation steps from idea generation to evaluation and implementation. The effectiveness of an IMS is often attributed to organizational culture, which is characterized by creativity, entrepreneurial mindset, and adaptability to market changes [1]. In other contexts, IMSs promote knowledge sharing among employees to enhance not only incremental but also radical innovation [2]. This highlights the importance of an IMS in shaping the innovation landscape within organizations.

As organizations generate increasingly large volumes of innovation ideas, due to different activities both internally (i.e., organization-wide employee challenges) and externally (i.e., open innovation), navigating and structuring these contributions becomes challenging. This often results in idea overload and lack of direction in innovation pipelines. To address this challenge, Topic Modeling (TM)—a powerful unsupervised machine learning (ML) technique—has been proposed as a method for extracting thematic structures from unstructured text, enabling better understanding, classification, and prioritization of ideas within IMS.

In this study, we analyze results from an IMS adapted to an educational context to support student-driven innovation processes in collaboration with industry partners. While innovation management has been a popular practice to include employees and other stakeholders in the product development process, we further recognize its potential in an educational context. This shift leverages strengths of IMSs such as knowledge absorption, structured ideation, and innovation portfolio management, while also creating an educational environment where students engage in real-world problem-solving for industry. Thus, IMSs become a bridging platform between academia and industry, fostering mutual learning and knowledge transfer while leading to practical innovation outcomes from student solutions.

For students, working with these systems offers a realistic and industry-aligned experience that requires navigating the complexity of multidisciplinary teamwork, structured pro-

979-8-3315-0969-9/25/$31.00 ©2025 IEEE

cesses, and radical innovation thinking. The latter is achieved through the choice of the innovation model, in this case, the Creative-Idea-Solution (CIS) model. This model organizes the work of the student teams to apply theoretical knowledge to real cases, developing the practical innovation skills demanded by modern industries. On the other hand, for industry partners, collaborating through an IMS allows them to outsource early-stage ideation to a pool of multidisciplinary engineering talent, gaining access to novel solutions and radical innovation possibilities. In addition, a structured IMS, preferably AI-enabled, ensures that organizations receive well-documented idea pipelines which can later be evaluated for potential development and commercialization. Given these considerations and potential benefits, it is important to understand how student innovation processes are being conducted and what could be improved. This need motivated us to explore TM techniques as a way to analyze the unstructured results from IMSs.

Background

Topic Modeling

Recent research on topic modeling (TM) has highlighted its diverse applications in fostering innovation across various sectors. TM has emerged as a powerful natural language processing (NLP) and machine learning (ML) technique [3] for analyzing large volumes of textual data, uncovering hidden patterns, and extracting meaningful insights that can drive innovation. IBM defines topic modeling as a text-mining technique that applies unsupervised learning to produce a summary of terms or clusters representing the data collection's commonalities. Topic modeling is one of many natural language processing (NLP) methods, but it is particularly useful as a machine learning (ML) method to annotate large text corpora thematically [3]. Its popularity is due to its structured approach to categorizing large volumes of unstructured text data to identify patterns and trends. The use cases for topic models vary from bioinformatics [4] to hate speech detection on social media [5].

This paper aims to investigate the integration of topic modeling algorithms in IMSs to enhance idea classification, trend detection, and knowledge discovery in innovation processes. By clustering data based on semantic similarities, innovative concepts can be categorized into interpretable themes, which aid in our understanding of the landscape of ideas being generated. Among the variety of AI-powered IMSs, research has shown that TM assists in describing idea classification and idea selection in an innovation process, in addition to synthesizing novel and radical ideas [6]. Furthermore, previous research has been conducted on the role of TM in mapping innovations in geographical areas, like the Flemish business community [7]. In their approach, TM aided in identifying innovative trends and supporting strategic decision-making for local businesses. This study will focus on implementing an IMS in a higher-education setting, where engineering student teams organize their innovation process to propose final radical solutions to real cases in Danish industry.

Radical Innovation

As noted above, innovations can be categorized into radical and incremental. Radical ideas are significant advancements that involve the creation of entirely new markets while challenging the organization to adapt its core activities and capabilities. On the other hand, incremental ideas propose gradual improvements that enhance what is already offered without disrupting the market [8]. Radical innovations are not easy to achieve as they often require substantial investments and a higher risk tolerance. Knowledge discovery identifies hidden patterns, novel insights, or trends organizations seek when implementing open innovation and collaborating with academia through research, education, or technology transfer [9]. In this study, Danish industry seeks new knowledge and radical ideas across the pool of master engineering students by providing their challenges to the Applied Innovation in Engineering course at Aarhus University. Interdisciplinary student teams designed and proposed solutions after going through the Creative, Idea, Solution (CIS) innovation process [10].

Rosetta is the IMS used across the years in the Applied Innovation course at Aarhus University. It is a systematic portfolio tool specifically designed to support radical innovation processes. It operates as the core idea management system aligned with the CIS model, primarily focusing on the Pre-ject phase (fig. 1), where early-stage idea generation and development occur. Unlike typical idea management systems that focus on decision-making and evaluation, Rosetta emphasizes action and knowledge creation, deliberately avoiding early assessments or judgments. Instead, it collects, organizes, and develops ideas, concepts, and designs without deleting or merging them, ensuring a transparent and evolving repository of innovation knowledge. Rosetta facilitates knowledge gathering, assimilation, and integration, supporting both internal and external information processing, reflecting the notion of absorptive capacity.

Given the increasing role of IMS in facilitating radical innovations, the integration of TM offers a structured way to enhance idea classification, detect emerging trends, and support knowledge discovery. This study aims to demonstrate TM's value in academic IMS frameworks and how it can bridge knowledge creation for Danish industry challenges. By analyzing the thematic structures of student-generated ideas, this research will assess whether radical innovation patterns emerge and how TM-based clustering can offer the industry a systematic approach to harnessing student-driven innovations.

Methodology

A computational approach to text analysis is employed to assess the implications of using TM for idea clustering and similarity analysis. The study is based on data collected from Aarhus University's "Applied Innovation in Engineering" Course in 2023, where cross-disciplinary teams of master's students conduct real-world innovation challenges provided by Danish industry partners. The Creative, Idea, Solution (CIS) framework guides the students' innovation process, making it an ideal setting to analyze how radical innovation patterns emerge [11].

This innovation process is structured in the Rosetta idea management system, where students outline their journey from ideation inputs to final solution proposals. The data extracted from Rosetta includes all student ideas across the stages of their journey in the CIS framework (i.e., Input, Idea, Concept, Design) proposed for each industry case (Fig. 1). We then used multiple modeling tools including BERTopic, K-Means Clustering, and Term Frequency-Inverse Document Frequency (TF-IDF) to analyze the data.

BERTopic is an advanced topic modeling technique that leverages contextual embeddings, to capture potential semantic relationships between words. Comparable to other topic models, BERTopic preserves word meaning within context and improves clustering using dimensionality reduction and density-based clustering [12]. On the other hand, K-Means clusters similar topics based on their embeddings by minimizing intra-cluster distance while maximizing inter-cluster separation, thus helping to structure unorganized textual ideas [13].

The choice of BERTopic modeling in this study is due to its ability for more accurate topic representation in order to align with the highly specific areas that the industry needs are representative of [14]. The choice of K-Means is inspired by its possibility to group idea representations in a way that helps identify highly related or outlier topics. This would support the goal of identifying potential radical ideas. In addition, considering the nature of data provided by Rosetta, which is unordered categorical output data in an unsupervised learning environment [15], K-Means is a recommended tool for leveraging NLP in an innovation process.

Lastly, TF-IDF is leveraged as a statistical weighting technique to evaluate the importance of a term to specific projects or stages in a dataset. Based on the Rosetta IMS and the innovation model CIS, students propose ideas in the following order: Input, Idea, Concept, and Final Design (fig. 1). Possibly, this study will help showcase how TF-IDF can help track idea evolution over time. So far, research has used it and TM in general to handle the problem of information overload in the idea generation phases, but less so for idea development phases [6].

Using the tools above, this study explores the role of topic modeling in facilitating not only idea classification, but also idea evolution. The following hypotheses will be addressed:

- H1: Topic modeling effectively classifies and clusters innovation ideas into distinct, interpretable themes.
- H2: Topic modeling can assist in identifying radical ideas and supporting data-driven activities in IMS.

H1 is based on the potential that topic models, particularly those using transformer-based embeddings (i.e., BERTopic) and unsupervised clustering algorithms (i.e., K-Means), can identify latent thematic structures within innovation datasets. Research in NLP suggests that topic models outperform traditional manual classification methods, enabling efficient and scalable organization of ideas [16].

H2 focuses on the predictive power of topic modeling in detecting emerging innovation trends. Prior studies indicate that topic modeling can identify early signals of technological advancements and market shifts by analyzing historical text corpora [17] to identify radical ideas. By leveraging topic models, IMS can provide actionable insights to organizations, helping them align their innovation strategies with future market demands.

The dataset extracted from the Rosetta IMS comprises 5492 ideas scattered across 4 stages of the model CIS. The ideas are related to 47 Danish industry case studies from 27 companies given to students covering a variety of engineering fields, such as mechanical and production engineering, biotechnology and biomedicine, construction and civil engineering, and electrical and computer engineering. The ideas are proposed by 181 students, organized in 48 multidisciplinary teams. Only 12 teams followed all four stages of CIS in Rosetta, indicating that the other 36 chose another tool as their idea management systems (i.e. word document, excel sheet, etc.). All analysis can be found at https://github.com/shahamana/rosetta-topic-model.

Data Preprocessing

Each student entry consisted of a Heading (Danish: Overskrift) and Description (Danish: Beskrivelse), where the Heading is the idea they're trying to record and Description elaborates on that idea. To analyze the data, we first filtered out all student entries without Headings as we interpreted those entries as incomplete. Considering that the Description built on the ideas in the Heading, we decided to combine the two and record each student idea as one string (Heading + Description). The data was then tokenized, lowercase, and cleaned for punctuation using the Natural Language Toolkit (NLTK) [18, 19].

While students were expected to complete the assignment in English, and companies provided cases in English, some students included Danish vocabulary during their use of Rosetta. Given the potential inaccuracies with translation, we made the decision to filter out as many of the Danish words as possible. While NLP resources and datasets for non-English languages are less common, the Alexandra Institute has provided a validated list of possible sources [20]. After careful consideration of the available options, we decided to assemble a list of Danish words using the Europarl7 dataset, a corpus of Proceedings from the European Parliament from 1996-2011 [21]. Given the highly technical nature of some of the cases, traditional corpuses from fiction and books would have likely missed many of the more specific Danish words. Parliamentary proceedings often cover a significant variety of topics that are relevant to the current state of affairs, and as such we could assemble a more comprehensive list. All Danish words found in the dataset were filtered out from the student data.

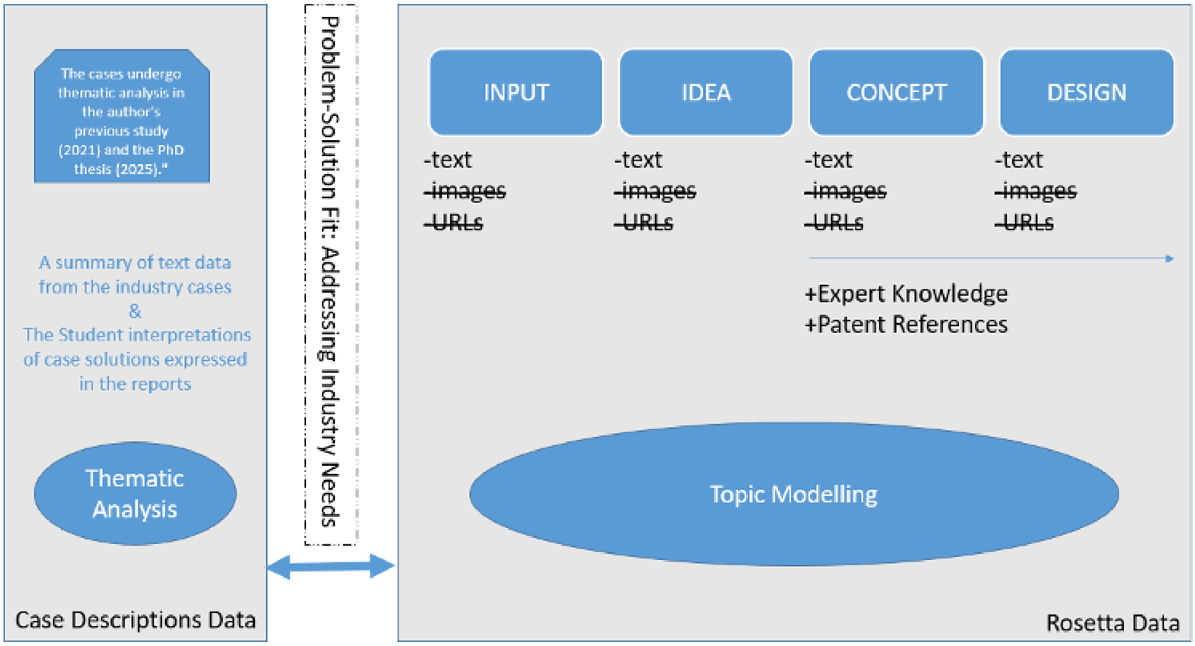
The cases undergo thematic analysis in the author's previous study (2021) and the PhD thesis (2025)."
A summary of text data from the industry cases & The Student interpretations of case solutions expressed in the reports
Thematic Analysis
Case Descriptions Data
Problem-Solution Fit: Addressing Industry Needs
INPUT
IDEA
CONCEPT
DESIGN
-text
~~-images~~
~~-URLs~~
-text
~~-images~~
~~-URLs~~
-text
~~-images~~
~~-URLs~~
-text
~~-images~~
~~-URLs~~
+Expert Knowledge
+Patent References
Topic Modelling
Rosetta Data

Fig. 1: The CIS Model in the Rosetta IMS: Integrating Thematic Analysis and Topic Modelling to Address Industry Needs

BERTopic

After the initial pre-processing steps, we conducted a BERTopic analysis on the data. Since this model relies on establishing semantic and contextual relationships, most stop words, verbs, and other non-topic-related vocabulary were important to preserve the accuracy of the analysis [12]. We did filter out English stop words using BERTopic's native library, which processes stop words after clustering the documents. Given the size of the dataset, we found that clustering into 19 topics provided specificity in topic clusters while still showing trends in significant amounts of the data.

KMeans Clustering

Further pre-processing was needed before conducting the K-Means analysis. Using the NLTK native library, we filtered out all English stop words, all tokens of length 1 or smaller, and all words that were not nouns or adjectives [19]. The data was then lemmatized, and all words fewer than two or longer than 15 characters were discarded [20]. Additionally, we used the gensim Phrase model to identify bigrams and add them to the dataset [22].

Before conducting K-Means analysis, we used Word2Vec to generate 100-dimensional vector representations for every term in the dataset. We then created K-Means clusters using scikit-learn [23]. The Elbow Method was used to determine the number of clusters (See Figure 2) [24]. Even though the method recommended five clusters, we chose to use seven to allow for closer comparison to the eight clusters from a previous study of the author which conducted manual thematic analysis [25]. Using Principal Components Analysis [26], the term vectors were reduced to two dimensions and graphed with colors representing their cluster. Word clouds were also generated for each cluster, with more frequent terms appearing larger in each word cloud.

Term Frequency - Inverse Document Frequency

The same pre-processing steps as K-Means were applied to a subset of the data, particularly the 12 projects where students interacted with Rosetta in all 4 stages of the process. We found that two projects, Project 43 and Project 13 had no terms related to stage 4 as they had all been filtered out through the pre-processing steps. The remaining projects were then separated, and each project's entries were split by stage. TF-IDF was then applied to each stage of each project individually and the results analyzed. TF-IDF was also applied to the data for each stage across all projects and the results were analyzed.

On the same subset of data, we also analyzed the TF-IDF for each stage for each project, but during pre-processing we kept all parts of speech, including verbs. Projects 43 and 13 were still removed from analysis as their stage 4 entries did not have tokens left, signaling that those entries were likely a mix of Danish and stop words.

Results

We present the findings of our topic modeling analysis by applying BERTopic, K-Means clustering, and TF-IDF, to categorize, cluster and analyze ideas within Rosetta as the IMS of choice in an educational setting.

BERTopic was used to classify and organize student-generated ideas into interpretable themes. The model generated 19 topics and 1 outlier category, where each topic was characterized by semantically related keywords. The most frequent topics were related to sustainability, automation, waste management and high-tech monitoring, thus reflecting certain

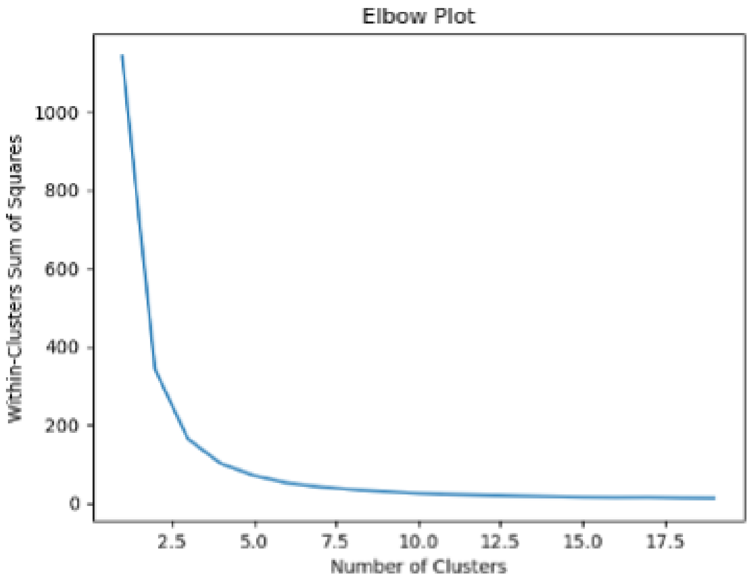

Fig. 2: Within-Clusters Sum of Squares Plot

popular industry needs. Some topics appear underrepresented, indicating potential niche ideas or possible knowledge gaps in ideation. These topics signal unexplored concepts, requiring further qualitative validation from idea creators or other stakeholders, like teachers or industry contacts.

To further analyze the relationship between topics, K-Means clustering was applied to group similar ideas into coherent categories. It provided insight into the distribution of idea clusters and their degree of thematic diversity. The model identified 7 distinct clusters, each representing a broad innovation approach in the proposed ideas. Some clusters were densely packed, indicating overlapping solutions or well-established themes. Yet, a few clusters contained outliers, suggesting unconventional approaches to the industry cases, possibly as novel or radical ideas. Fig. 3 uses PCA to collapse each word to its most important dimensions, allowing for a 2D representation of the cluster characteristics [26]. Right-side clusters (blue, brown, orange, and pink) contain closely related ideas, indicating mainstream innovation trends, across all proposals. Left-side clusters (green, red, and purple) contain unique and underdeveloped ideas, potentially radical ones.

Clustering reflects solution diversity, though only for a limited number of cases. Not all clusters are equally developed, reflecting potential gaps where some cases are underexplored while others are more tangible.

Topic Analysis

After the BERTopic and K-Means analysis, word clouds were generated for the respective 19 topics and 7 clusters. Given the high dimensionality of text embeddings, manual interpretations of K-Means and BERTopic results would be time-consuming. To enhance the thematic analysis, we integrate GPT-based summarization in the following way: GPT cross-references K-Means clusters with BERTopic topics to highlight thematic overlaps and missing segments, thus improving interpretability and uncovering hidden trends not explicitly labeled or easily recognizable [27].

To assess the accuracy and reliability of GPT-generated thematic summaries, we conducted a benchmarking analysis by comparing GPT's automated topic interpretations with ground truth labels derived from a manual thematic analysis [20]. In previous work, a thematic analysis was carried out based on the description and challenge specified in each case description, out of 49 cases provided by Danish industry. The analysis followed an iterative coding process between three authors (2 senior and 1 junior researchers). The qualitative data from each case were systematically analyzed via inductive coding and iterative coding cycles, until a final set of codes had been developed. The results revealed three main areas of interest: (i) technology/product, (ii) digitalization, and (ii) sustainability, along with five secondary aspects: (a) future trends, (b) customer behavior, (c) business, (d) regulations, and (e) training.

The topics derived from GPT-based summarization include employee-product interaction, digital monitoring, automation in retail and tech, gamification and robotics in urban settings, data-driven resource management, market positioning and pollution control, 3D recognition in commerce, and AI/ML-driven production optimization. After adjusting the cosine similarity threshold to 0.4, we observed the accuracy increased to 45.1%, demonstrating that GPT-based thematic analysis aligns well with human classification. A higher accuracy is likely not possible as the manual thematic analysis is also biased since the human annotation was conducted by authors that were also involved in onboarding the industry and their cases in the course. As a result, nuances in the themes developed that could not be captured by the GPT model.

Themes capturing various materials and manufacturing pro-

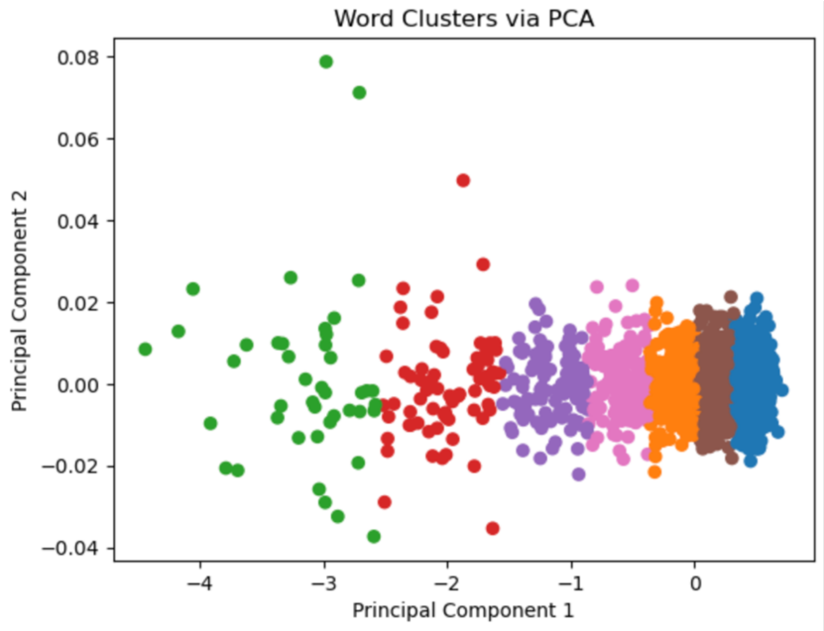

Fig. 3: Cluster Scatter Plot via K-Means Analysis

cesses overlap across the topic model results (BERTopic and K-Means) presented in figure 4. In addition, differences across the results suggest that BERTopic reveals highly specialized themes, while K-Means clusters are broader categories. BERTopic is better for meaning-based categorization which is exemplified by topic clusters on medical devices or protein research. Meanwhile K-Means is effective in discovering larger ideas, such as surface treatments and energy systems, but fails to explicate the fine-grained topics. Lastly, BERTopic provides more interpretable outputs, reducing human effort in labeling clusters. This facilitates faster association of future ideas in the pipeline with overall themes or industry needs. By merging outputs of both models, results can be interpreted in a macro- and micro-perspective, where K-Means can be used for high-level grouping, and BERTopic for detailed thematic breakdown.

Following theme interpretation, themes can be related to clusters (K-Means) or topics (BERTopic), allowing for further observations depending on the industry needs. In an industry context, when implementing IMSs, a higher variety of data sets would be made available to help re-align the idea proposals with what the company is looking to develop across its offerings and what the company has the potential to pursue. If further analysis is performed, using prompting in the GPT or other LLMs, opportunities and trends can be recognized, such as in automation and robotics or consumer and assistive technologies. Certain themes occurring across the solutions proposed by the student teams persist when compared to the manual thematic analysis [20] conducted on the industry case descriptions in 2021. This points to the conclusion that certain industry needs remained relevant over the years.

TD-IDF Analysis

To examine how ideas evolved through the different CIS innovation model stages, we applied TF-IDF to track keyword importance. Early stages of ideation such as Inputs and Ideas were characterized by problem-related and broad terms, such as “waste”, “reduction”, and “reusable”, which in the later stages of Concept and Design were replaced with more solution-oriented keywords, like “biodegradable materials”, “automation, “energy recovery”, and so on, showing conceptual maturity. Many terms disappeared between stages, suggesting refinement of ideas throughout the process as challenges arose during idea feasibility exercises. As the course intends to focus on radical innovation, ideas related to automation and AI gained prominence in the later stages, while manual labor and traditional processes declined, indicating that students recognize AI-related ideas as trendy or reflective of radicalness. Particularly, if AI-enabled ideas gain prominence in later stages of innovation, as they did in this study, workshops specifically dedicated to applications of AI may deepen the knowledge of the ideators and strengthen the feasibility of the ideas.

The dramatic reduction of keyword continuity (Fig. 5), especially from Inputs to Design (7%), highlights both critical evaluation and the maturation of ideas. The TF-IDF results narrate a common funnel of innovation, where wide and diverse ideas are proposed in the early stages in high numbers whereas narrower and high-potential ideas are proposed in lower numbers in the later stages. Tracking term evolution via TF-IDF is about providing valuable insights into the overall innovation process. These include identifying the stages where ideas drop off the most, at what part of the process are promising ideas at risk of abandonment, or whether idea

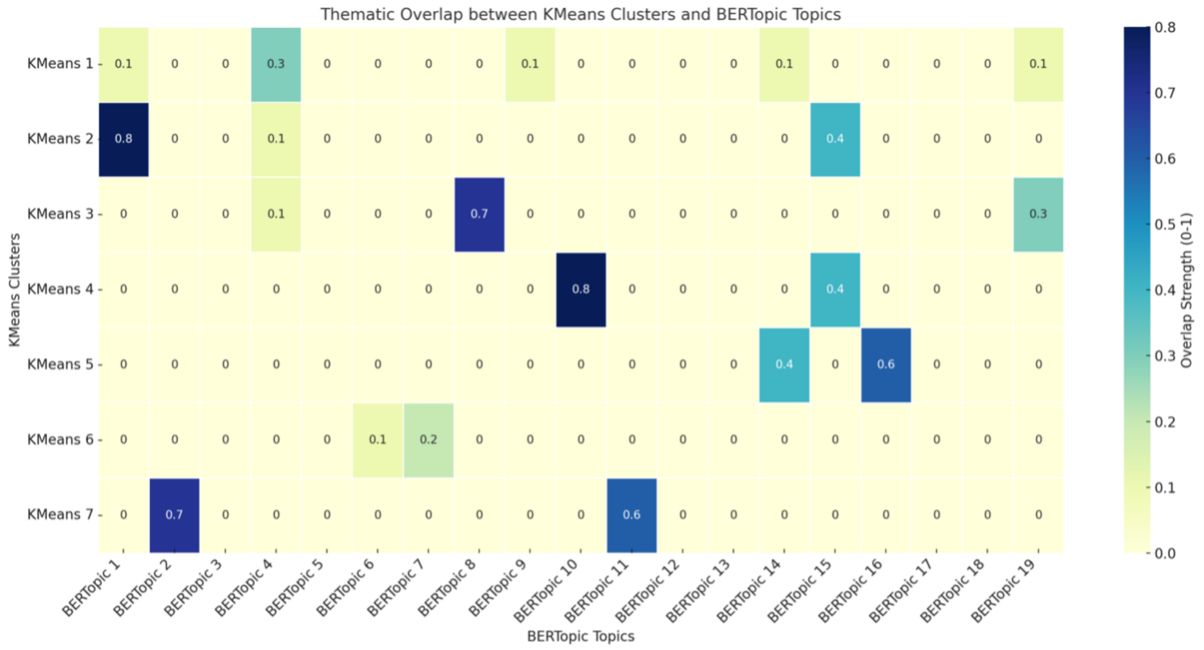

Fig. 4: Thematic Overlap between K-Means Clusters and BERTopic Topics

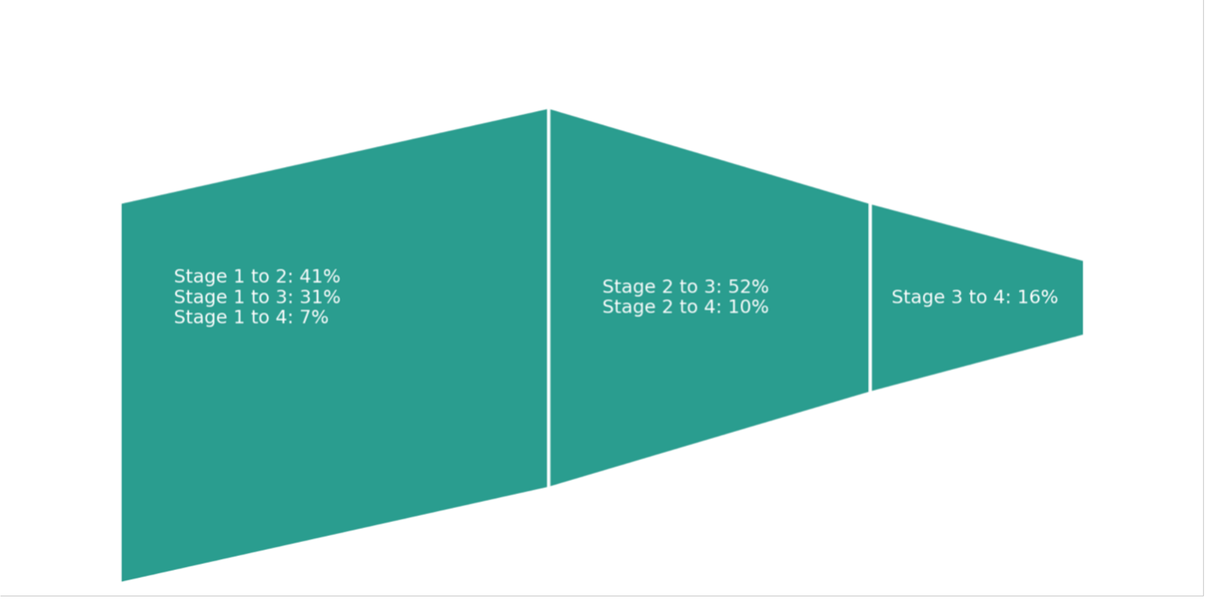

Fig. 5: Rounded Percentage of words that stayed between stages

execution via business models and go-to-market strategies are explored. Lastly, coupled with the other topic models, TF-IDF can assess the knowledge transfer in the innovation journey, from the early-stage inputs to final solutions, to ensure industry expectations are always considered.

Together, the combination of BERTopic, K-Means, GPT, and TF-IDF offers a layered understanding of ideation patterns, where broader thematic clusters are complemented by fine-grained topics. Additionally, keyword evolution reflects a dynamic process of idea refinement aligned with radical innovation intentions.

Discussion

The findings of this study suggest that TM techniques, specifically BERTopic and K-Means clustering, provide a structured framework for idea classification and trend detection in an academic IMS. While the results support the effectiveness of AI-driven classification methods, the identification of radical ideas requires further validation and other methodological and analytical limitations need to be addressed. Future work could address these limitations by analyzing company case descriptions along with student results, exploring other mainstream methods of TM such as LDA, and capturing new data from student teams who are willing to utilize the IMS across all four stages of ideation.

A wide body of research explores whether quantitative metrics of TM are reliable. Similarly, qualitative approaches require significant domain-specific knowledge to evaluate key terms for interpretability [28]. Topic interpretability is subjective, requiring manual validation to ensure meaningful insights. The innovation process presented in this educational setting requires student teams to interactively communicate with industry contacts and involve subject-matter experts towards the later stages of the CIS model. These activities facilitate more accurate interpretations of clusters provided by the topic modeling. For educators, this inspires reflections on how these AI models can improve ideation training, feedback loops, and curriculum design.

This study intends to strengthen the identification of radical innovation, which can be inferred from outlier clusters like in K-Means analysis. However, radicalness is a subjective concept and time-dependent, requiring further qualitative assessment. Student perception of radical vs. incremental innovation may be biased or inconsistent across teams. Subject-matter experts and other stakeholders can manually evaluate outliers, thus determining the potential existence of radical innovation. However, this is also an area in which to develop the model further, using deep learning that can analyze textual content for linguistic markers of novelty and disruptions [29]. However, this would require historical data on radical innovations of the companies, such as patent databases and startup funding reports to mention a few. For industry, this inspired reflections on how such analyses provided by AI models can guide early-stage innovation management in-house, focusing efforts and dedicating resources to high-potential ideas.

Conclusion

This study highlights the potential of integrating AI-driven topic modeling techniques within IMS to structure and enhance the innovation process. For academia, the findings contribute to the growing body of research on AI-assisted idea management systems, particularly in educational settings. The results provide a systematic approach for universities and research institutions to incorporate AI-based classification systems into their curriculum, allowing students to develop solutions aligned with industry needs. For industry, the study offers a structured approach to managing innovation pipelines, enabling firms to identify emerging trends, focus on promising ideas, and make data-driven RD decisions. The results indicate that integrating BERTopic and K-Means in IMS can help industry partners evaluate idea evolution at different stages, particularly in assessing radical versus incremental potential.

The results of the BERTopic indicate an effective structuring of the innovation landscape, allowing industry partners and ideators to identify common thematic trends across ideas. Simultaneously, the results allow for addressing gaps in ideation where certain industry challenges remain under-represented. In addition, BERTopic can provide an initial assessment of whether radical ideas emerge naturally, or if the process in IMS favors incremental improvements.

The output of the K-Means analysis provides actionable insight for innovation managers, ideators, and industry partners as identifying underdeveloped clusters can guide RD decisions. This is because ideas will become identifiable as radical vs incremental within the existing landscape of the company.

TF-IDF analysis provides a multi-layered perspective on how ideas evolve and how they are filtered through an educationally framed innovation management system. The observed trend of narrowing ideas from diverse early-stage concepts to fewer AI-driven, solution-focused proposals highlights the importance of balancing creative exploration with technological feasibility.

Lastly, the combination of BERTopic and K-means results with the GPT-based summarization provided a multi-layered approach to structuring idea generation and trend analysis across an innovation process. As a result, this reduces manual efforts in interpreting topic models and assists in identifying emerging and underrepresented trends.

Following these results, we can confirm Hypothesis One: topic modeling can effectively classify and cluster innovation ideas in an innovation process using IMSs. However, Hypothesis Two is partially supported, considering the progression of the themes and TF-IDF results in the innovation stages. Radical ideas can potentially be spotted, but this requires further qualitative validation from the ideators and stakeholders.

The main contributions of this study to academic literature are related to offering insights into how radical ideas emerge within structured idea management systems between industry and academia. On the other hand, contributions to practical IMS applications are towards an AI-driven framework for classifying and analyzing ideas and demonstrating the potential of topic modeling in IMS Implementation. Our findings underline the role of AI as a tool for enhancing, rather than replacing, human-centered innovation processes, with significant potential for both educational environments and industrial idea management systems.

Acknowledgments

We would like to express our sincere gratitude to Professor Rune Hylsberg Jacobsen, at the Department of Electrical and Computer Engineering, Aarhus University, for his valuable guidance throughout this work. Our thanks also go to Associate Professor Henning Sejer Jacobsen, at the Department of Biological and Chemical Engineering, Aarhus University for leading the task of collecting the industry cases and for co-running the Applied Innovation in Engineering course, from which all student data in Rosetta was gathered. We also extend our appreciation to Thomas Filyo from EasyCount for his contributions to data collection and for bringing Rosetta forward as a web application. Finally, we are grateful to DIS Study Abroad, a Danish educational institute, for facilitating the introduction between the authors, which provided a strong foundation for this collaboration.

References

[1] S. K. Taghizadeh, S. A. Rahman, M. M. Hossain, and M. M. Haque, "Characteristics of organizational culture in stimulating service innovation and performance," MIP, vol. 38, no. 2, pp. 224–238, Apr. 2020, doi: 10.1108/MIP-12-2018-0561.

[2] H. Lei, A. T. L. Ha, and P. B. Le, "How ethical leadership cultivates radical and incremental innovation: the mediating role of tacit and explicit knowledge sharing," JBIM, vol. 35, no. 5, pp. 849–862, Dec. 2019, doi: 10.1108/JBIM-05-2019-0180.

[3] J. Alammar and M. Grootendorst, Hands-on large language models: language understanding and generation, 1st edition. Beijing Boston Farnham: O'Reilly, 2024.

[4] Y. Zhang, M. (Sam) Khalilitousi, and Y. P. Park, "Unraveling dynamically encoded latent transcriptomic patterns in pancreatic cancer cells by topic modeling," Cell Genomics, vol. 3, no. 9, p. 100388, Sep. 2023, doi: 10.1016/j.xgen.2023.100388.

[5] R. Sear, N. J. Restrepo, Y. Lupu, and N. F. Johnson, "Dynamic Topic Modeling Reveals Variations in Online Hate Narratives," in Intelligent Computing, vol. 507, K. Arai, Ed., Cham: Springer International Publishing, 2022, pp. 564–578. doi: 10.1007/978-3-031-10464-0_38.

[6] S. Leka, "The Role of Artificial Intelligence in Idea Management Systems and Innovation Processes: An Integrative Review," in Proceedings of the Cognitive Models and Artificial Intelligence Conference, İstanbul Turkiye: ACM, May 2024, pp. 160–164. doi: 10.1145/3660853.3660890.

[7] A. Crijns, V. Vanhullebusch, M. Reusens, M. Reusens, and B. Baesens, "Topic modelling applied on innovation studies of Flemish companies," Journal of Business Analytics, vol. 6, no. 4, pp. 243–254, Oct. 2023, doi: 10.1080/2573234X.2023.2186274.

[8] A. W. Al-Khatib and E. M. Al-ghanem, "Radical innovation, incremental innovation, and competitive advantage, the moderating role of technological intensity: evidence from the manufacturing sector in Jordan," EBR, vol. 34, no. 3, pp. 344–369, Apr. 2022, doi: 10.1108/EBR-02-2021-0041.

[9] H. Chesbrough, Open innovation: the new imperative for creating and profiting from technology, Nachdr. Boston, Mass: Harvard Business School Press, 2011.

[10] Serena Leka and Henning Sejer Jacobson, "The role of Idea management system in Absorptive Capacity and Action-based Process of Radical innovation," Nov. 2022, pp. 1–18.

[11] H. S. Jakobsen, J. Brix, and R. S. Jakobsen, "Unraveling data from an idea management system of 11 radical innovation portfolios: key lessons and avenues for artificial intelligence integration," J Innov Entrep, vol. 13, no. 1, p. 9, Jan. 2024, doi: 10.1186/s13731-024-00368-6.

[12] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, arXiv. doi: 10.48550/ARXIV.2203.05794.

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967. [Online]. Available: https://api.semanticscholar.org/CorpusID:6278891

[14] C. Kakatkar, V. Bilgram, and J. Füller, "Innovation analytics: Leveraging artificial intelligence in the innovation process," Business Horizons, vol. 63, no. 2, pp. 171–181, Mar. 2020, doi: 10.1016/j.bushor.2019.10.006.

[15] H. Jelodar et al., "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2017, arXiv. doi: 10.48550/ARXIV.1711.04305.

[16] W. He, Z. Zhang, and W. Li, "Information technology solutions, big data and artificial intelligence in online social network-based open innovation," Enterprise Information Systems, vol. 14, no. 5, pp. 663–679, 2020.

[17] Y. Wang, W. He, and W. Zhu, "The role of artificial intelligence in the evolution of knowledge management," Information Systems Frontiers, vol. 23, no. 4, pp. 961–976, 2021.

[18] S. Bird, E. Klein, and E. Loper, Natural language processing with Python. Cambridge: O'Reilly, 2009.

[19] Kamila Kunrath, Serena Leka, Lasse Steenbock Vestergaard, Mirko Presser, and Devarajan Ramanujan, "A Digital Tool for Scaffolding Innovation Learning in Engineering Education with Local Industry Needs," International Journal of Engineering Education, vol. 40, no. 4, pp. 801–814, 2024.

[20] A. B. Pauli, M. Barrett, O. Lacroix, and R. Hvingelby, "DaNLP: An open-source toolkit for Danish Natural Language Processing," in Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), S. Dobnik and L. Øvrelid, Eds., Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 460–466. Accessed: Mar. 13, 2025. [Online]. Available: https://aclanthology.org/2021.nodalida-main.53/

[21] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, Sep. 2005, pp. 79–86. Accessed: Mar. 13, 2025. [Online]. Available: https://aclanthology.org/2005.mtsummit-papers.11/

[22] R. Řehůřek and P. Sojka, Software Framework for Topic Modelling with Large Corpora. University of Malta, 2010. Accessed: Mar. 13, 2025. [Online]. Available: https://repozitar.cz/publication/15725/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka

[23] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," 2012, doi: 10.48550/ARXIV.1201.0490.

[24] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," J. Phys.: Conf. Ser., vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[25] Kamila Kunrath, Serena Leka, Haitham Abu-Ghaida, and Devarajan Ramanujan, "IDENTIFYING INDUSTRY NEEDS FOR INNOVATION SKILLS IN ENGINEERING EDUCATION: A THEMATIC ANALYSIS OF CASES FROM DANISH INDUSTRY," 2021, p. 10.

[26] C. Ding and X. He, "K -means clustering via principal component analysis," in Twenty-first international conference on Machine learning - ICML '04, Banff, Alberta, Canada: ACM Press, 2004, p. 29. doi: 10.1145/1015330.1015408.

[27] A. Turobov, D. Coyle, and V. Harding, "Using ChatGPT for Thematic Analysis," May 13, 2024, arXiv: arXiv:2405.08828. doi: 10.48550/arXiv.2405.08828.

[28] M. Gillings and A. Hardie, "The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice," Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 530–543, May 2023, doi: 10.1093/llc/fqac075.

[29] M. Qiao and K.-W. Huang, "Correcting Measurement Error in Regression Models with Variables Constructed from Aggregated Output of Data Mining Models," MISQ, vol. 49, no. 1, pp. 29–60, Mar. 2024, doi: 10.25300/MISQ/2024/18026.