2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 14-17, 2024

ENHANCING RFI ANALYSIS IN CONSTRUCTION PROJECTS: A COMPARATIVE STUDY OF TEXT CLUSTERING METHODS AND VISUALIZATION TECHNIQUES

Neziha Yilmaz[1] and Esin Ergen[1]
[1]Istanbul Technical University, Istanbul, Turkey

Abstract

The Request for Information (RFI) is a vital communication tool in construction projects, aiding teams in addressing queries and navigating challenges. Unstructured RFIs hinder manual analysis for extracting hidden knowledge. Prior research in RFI analysis employed NLP and text clustering and often relied on a single clustering method. This paper performs a comparative analysis using diverse clustering methods (LDA, NMF, and K-means) and visualization techniques to determine the most suitable methods. The study offers project managers and quality engineers an effective tool for extracting hidden knowledge in RFI.

Introduction

The construction industry heavily relies on efficient information management throughout the project lifecycle, generating a substantial volume of information in various formats (Al Qady and Kandil, 2014). Approximately 80% of corporate information is embedded in project documents, and challenges arise as professionals spend considerable time searching for and reading documents (Abbaszadegan and Grau, 2015). Insufficient information or delays in information exchange can lead to significant issues at construction sites (Abdirad et al., 2022).

Manual knowledge extraction from text documents is time-consuming and error-prone (Khuzaimah and Hassan, 2012), prompting researchers to adopt advanced Natural Language Processing (NLP) and text mining techniques, including sentiment analysis, semantic analysis, and clustering methods (Afzal et al., 2023; Jallan et al., 2019; Gheeseewan and Pudaruth, 2020).

In compliance and quality assurance, the Request for Information (RFI) document is a crucial text document in construction. It serves as a performance indicator and is analyzed conventionally through manual methods or text mining to extract insights and lessons learned (Herrera et al., 2019; Kim et al., 2022). Extracting insights and lessons learned from RFIs facilitates the early identification of potential issues, project uncertainties, and risks (Lee and Yi, 2017), allowing for timely corrective actions to be implemented in the early stages. Moreover, uncovering hidden information within RFIs enables the identification of the root cause of nonconformance. In the literature, Koc et al. (2024) have investigated the impact of non-conformities on project costs using machine learning methodologies. Similarly, the envisioned approach in this study is to extract insights from RFIs and utilize them to predict potential non-conformities for corrective and preventive action development. This approach fosters a culture of quality and compliance while mitigating reworks over time. Clustering the RFI texts to determine topics is the first step for utilizing RFI to predict unconformity and uncover hidden information within RFI texts. Exploring the effectiveness of various text topic modeling algorithms is crucial due to the unique characteristics of RFI texts. For instance, RFIs and responses, typically one to seven pages long and containing 20 to 5000 words, may be stored in unstructured format within a CDE. These texts often contain a mix of uppercase and lowercase letters, spelling errors, abbreviations, and inconsistent terminology (such as "frame" versus "beam" or "drawing" versus "CAD"), necessitating preprocessing techniques. Given their intricate details, RFI texts require micro-level clustering focus.

While some publications focus on text clustering and visualization techniques for RFIs (Lee and Yi, 2017; Afzal et al., 2023), they have utilized a single clustering method, lacking a comparison with other methods and determining the preferred visualization technique for experts. To address these gaps, this paper performs a comparative term analysis of RFI text using various clustering methods, namely Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and K-means, and visualization techniques (i.e., pyLDAvis, scatter plots, word cloud, and bar charts). The study involves expert interviews to determine the most suitable clustering and visualization methods for RFI clustering analysis. The research contributes to the limited comparative studies on text clustering and visualization of RFIs, offering a valuable tool for project managers to grasp RFI status.

Literature Review

Text-based data clustering methods are important for diverse applications in the construction industry. Moreover, text mining plays a crucial role in the quality management of construction engineering, involving tasks such as text cleaning, text segmentation, and semantic network analysis on unstructured data within text collections (Wang et al., 2018). The construction sector benefits from the interactive utilization of user-centric dictionaries, sentiment analysis based on word clusters, and the automated extraction and clustering of relevant terms for dictionary generation (Kohita et al., 2020; Alshari et al., 2020). These approaches ensure the construction industry's access to comprehensive and industry-specific term dictionaries, facilitating effective text analytics and knowledge management.

RFI texts, rich in information about project history and design functionality, offer detailed insights into potential quality issues in the construction site, allowing for a

comprehensive focus on field operations. Traditionally, extracting insights from RFIs involved manual data mining and content analysis (Bhat et al., 2017; Afzal et al., 2023). Several studies in the literature focus on classifying RFIs. Dantas Filho et al. (2016) revised constructability concerns for evaluating RFIs in residential projects, introducing categories such as Design Correction, Divergence of Information, Design Change, Design Failure, Validation Information, and Design Verification. Morales et al. (2022) analyzed 2690 RFIs from 17 projects using statistical classification methods to associate RFIs with BIM usage. The RFIs were classified, such as alternative design solutions, approvals, clarifications of information, and other categories. Subclassifications include conflict RFIs, incorrect RFIs, insufficient RFIs, and questionable RFIs. Ham and Yuh (2023) used 872 BIM RFIs to classify them based on their purpose across four completed projects and 12 construction sites.

However, existing studies generally focus on the manual classification of RFIs, and a limited number of studies have concentrated on extracting important terms from RFIs via text clustering methods (Lee and Yi, 2017; Afzal et al., 2023). Lee and Yi (2017) used topic modeling to predict uncertainty in RFIs, focusing on the pre-bidding stage. Yet, this study did not consider the construction phase, although numerous RFIs are typically created as projects progress. Afzal et al.'s (2023) study proposes an innovative approach using the LDA algorithm to extract patterns from RFIs, offering valuable insights for the construction industry but lacking a critical comparison of clustering and visualization techniques.

LDA is an unsupervised generative probabilistic model commonly used for text topic modeling for uncovering hidden topics within a collection of text documents (Buntine et al., 2004; Lee and Yi, 2017; Afzal et al., 2023). LDA was used for text clustering in annual reports of construction companies (Jagannathan et al., 2022), focusing on macro-level clustering to identify key concerns related to understanding corporate and industry strategies. Lai and Konstanta (2019) also investigated job descriptions, which are relatively short, for building permits using LDA. In another study, LDA was used for clustering on-site inspection text, and this data is entered within constrained timeframes of on-site inspections, thus including typos and variations in upper/lower case (Lin et al., 2020). Lastly, construction-defect litigation cases were clustered using LDA, and in this type of text, many words are used interchangeably, for example, grading versus slope (Jallan et al., 2019).

On the other hand, NMF is a linear algebraic model. Unlike other topic modeling approaches, NMF utilizes matrix factorization and multivariate analysis to generate coefficients rather than probabilities for each word, associating them with specific topics (Naik, 2016; Jagannathan et al., 2022). NMF effectively reduces the dimensionality of large datasets by finding a low-rank approximation of the original data matrix, capturing essential features while discarding noise and irrelevant information (Lee et al., 2012). However, NMF works better with shorter texts, such as tweets or titles, and smaller datasets (Egger and Yu, 2022). Previous studies in non-construction contexts, particularly analyzing large text streams and review data, have demonstrated varying performance between LDA and NMF, with one algorithm often outperforming the other (George and Vasudevan, 2020).

K-Means is another unsupervised learning algorithm that groups data into n clusters with equal variance by minimizing the inertia or sum-of-squares within each cluster (Wu, 2012). K-means efficiently process large datasets by grouping data points into clusters based on similarities (Farhang, 2017). However, the initial selection of cluster centers in advance can impact the outcome (Rana et al., 2011). Additionally, K-means are sensitive to noise and outliers, potentially compromising clustering accuracy (Kaur & Aggarwal, 2017). In the literature, Liu et al. (2021) investigated the reasons for pipeline incidents and contributory factors using the K-means clustering method.

Application of text mining and visualization techniques, including sentiment analysis, semantic analysis, content analysis, and clustering methods, is crucial for unlocking insights and analyzing construction issues in RFI documents. These techniques extract meaningful knowledge from unstructured data, providing valuable insights for addressing construction challenges. However, previous studies have not determined which of these methods is the most suitable for analyzing RFI texts.

Methodology

Data Collection and Pre-Processing

The project's unique characteristics and phases (i.e., pre-bid, design, or under construction) can significantly alter the terms and topics within RFI texts. Using a common RFI dataset from various project types might lead to overlooking project-specific hidden information. Therefore, this study analyzed the RFI documents of a single project, which is an airport project. The RFIs were retrieved from a common data environment (CDE), which has streamlined the RFI handling process by enabling online communication and ensuring tracked information (Afzal et al., 2023).

During the data preparation phase, if an RFI conveyed multiple topics, these were treated as separate data instances (i.e., RFI texts); for example, a clarification topic might be observed at multiple points in a structure but communicated through a single RFI. Each clarification point is treated as an individual RFI text in this case. Additionally, an RFI should be evaluated in conjunction with its responses, as submitting an RFI triggers opinions and additional inquiries from multiple parties, and the responses should provide information about the RFI topic. A subject stored as 4-5 distinct RFI documents in CDE was considered a single RFI text to maintain the RFI context. Consequently, 288 RFI texts were extracted, representing different disciplines and locations within the structure. A similar number of RFI documents (243 RFI data) were analyzed in another study

(Lee and Yi, 2017). To ensure the extraction of pertinent and meaningful information from the RFI dataset while minimizing irrelevant or confounding data, the following pre-processing steps were employed: (1) lowering case, (2) removing punctuation, (3) stop words removal, (4) filtering out alphanumerical patterns, (5) removing currency symbols, (6) copyright symbols removal, (7) tokenization, and (8) lemmatization.

Text Topic Clustering

Despite preprocessing efforts, not all obtained words (terms) hold equal significance on the RFI topic. Therefore, to prioritize terms based on their weights, the "Term Frequency-Inverse Document Frequency" (TF-IDF) stands out as the most representative weighting method for word prioritization (McArthur et al., 2018). TF-IDF considers the frequency of words across different documents and their frequency within each document. If a word occurs frequently in a document but is rare in the general collection, its importance within the document increases (Christopher et al., 2008). This situation forms the basis for using TF-IDF to reduce common words, focus on significant terms that differentiate documents from others, and extract key terms based on TF-IDF scores. Then, RFI texts were transformed into a TF-IDF matrix. Subsequently, this TF-IDF matrix is processed using four different clustering algorithms, and distinct key terms are identified.

Evaluation of Topic Clusters and Visualization Techniques

Identifying themes for each topic involves subjective human interpretations, necessitating domain knowledge from the building and construction industry (Lai and Kontokosta, 2019). To highlight the importance of domain experts, Lai and Kontokosta (2019) stated that they plan to gather a panel of experts, including architects and contractors, to further validate the topic model. Therefore, the efficacy of clustering methods and visualization techniques for complex RFI data analysis was assessed through expert evaluation by three quality engineers with over five years of experience. The following question was posed to the experts, "How successful are text clustering algorithms' outputs in understanding an RFI text's content (topic) and determining the keywords that reveal its content?". Similar questions are included in measurement tools, such as The Usability, Satisfaction, and Ease of Use (USE) and The System Usability Scale (SUS), which are commonly used in assessment of prototype, process, and model evaluation studies (Lund, 2001; Brooke, 1996). Results from each algorithm were evaluated on a 5-point Likert scale, ranging from most to least useful, practical, and applicable (i.e., successful).

Additionally, engineers were presented with visualization techniques tailored to each algorithm's output, rated on a 5-point Likert scale. This comprehensive approach identified the most advantageous clustering methods and visualization techniques for analyzing RFI texts.

Findings

All clustering algorithms were applied to the same dataset, forming 10 clusters with 20 distinct key terms. The researchers systematically named the topics to enhance clarity and facilitate interpretation. Tables 1-3 briefly list these keywords into topic clusters.

Within the framework of these three distinct clustering methods, certain pivotal keywords have undergone encryption due to privacy issues. This encryption is necessary because it includes project names, places, people involved, and contract details. For reference, this encryption scheme adheres to the following mapping: A*: (5 characters), B*: (5 characters), C*: (5 characters), D*: (6 characters), E*: (7 characters), F*: (8 characters), G*: (5 characters). Additionally, words with two or three characters have been included in the analysis. These short words hold significance for experts involved in the project and dealing with RFIs, serving as a distinctive feature of a known phenomenon within the project.

Evaluation of LDA Clusters

LDA is a probabilistic model where each word in a specific document can be associated with multiple topics. LDA considers each document a mixture of different topics and assumes that each word has a probability distribution among the topics in the document. Therefore, a word can be associated with multiple topics. The list of the 10 topic sets was obtained from the LDA model in Table 1 and the top 20 keywords with the highest weights that characterize each topic.

Table 1: Partial List of RFI Topics in LDA Clustering Method

Topic	Keywords
0	detail, attach, beam, see, column, connection, level, information, lift, steel
1	subrfi, rfi, A*, date, doc, response, damper, fire, room, rficir
2	ceiling, type, bracket, shutter, roller, camera, height, case, width, custom
3	lock, door, cut, provision, button, require, room, water, out, contact
4	door, schedule, use, cylinder, requirement, fire, confirm, attach, lock, handle
5	fire, report, foundation, contractor, stair, document, structural, provide, wall, include
6	change, rfi, frame, relate, B*, instruct, prab, draw, acm, core
7	lamella, sensor, dual, bliptrack, wifisensor, gate, strip, lift, type, installation
8	comment, mount, request, draw, area, equipment, revise, elevation, update, change
9	prab, show, acm, connection, sounder, design, accord, heating, floor, draw

The explanation of each topic is provided as follows:
Topic 0_Structural Details: Discusses details, attachments, beams, columns, and steel lift information.
Topic 1_RFI and Document Handling: Involves submissions, responses, dates, documents, and fire related topics.
Topic 2_Ceiling and Custom Design: Focuses on ceiling types, brackets, shutters, roller cameras, and custom designs.

Topic 3_Door and Water Requirements: Covers door related discussions, schedules, cylinders, and water requirements.
Topic 4_Door Scheduling and Confirmation: Focusses door scheduling, usage requirements, and confirmation processes.
Topic 5_Fire Safety and Foundation: Relates to fire reports, foundations, contractors, stairs, and structural documents.
Topic 6_Change Requests: Centers on changes, RFIs, frames, instructions, and core design.
Topic 7_Sensor Installations: Discusses lamellas, sensors, dual systems, WiFi sensors, gates, and lift installations.
Topic 8_Comments and Revisions: Involves comments, mounting requests, drawing areas, equipment revisions, and elevation updates.
Topic 9_Prab Design and Heating: Focuses on designs, connections, sounder designs, heating, and floor drawings.

Evaluation of NMF Clusters

NMF decomposes the given data matrix into the product of two lower-dimensional non-negative matrices. In text analysis, one matrix represents the distribution of topics across documents, and the other represents the distribution of words across topics. Due to this decomposition, a word in NMF can possess non-zero weights in multiple topics, resulting in the occurrence of the same keyword across different topics.

The list of the 10 topic sets was obtained from the NMF model in Table 2 and the top 20 keywords with the highest weights that characterize each topic.

Table 2: Partial List of RFI Topics in NMF Clustering Method

Topic	Keywords
0	subrfi, A*, doc, date, response, rfi, C*, mep, pdf, sprinkler
1	door, lock, leaf, cylinder, hardware, bk, closer, handle, schedule, D*
2	rficir, A*, shaft, doc, fcu, date, group, air, return, near
3	drawing, prab, td, acm, wall, structural, str, floor, foundation, design
4	opening, beam, existing, fit, slope, routing, pipe, designer, existing, solution
5	subrfi, A*, damper, control, supply, power, sys, date, doc, td
6	bracket, color, ral, camera, drawing, custom, shop, colour, commented, schedule
7	beam, level, connection, lift, steel, attached, E*, load, column, imd
8	total, gh, lcp, dcl, belongs, area, toilet, principle, room, follow
9	ceiling, shutter, roller, casing, height, mm, large, lamella, type, leeuw

The explanation of each topic is provided as follows:
Topic 0_RFI and Sprinkler Systems: Addresses RFIs, documents, responses, MEP, PDFs, and sprinkler.
Topic 1_Door Hardware and Schedules: Focuses on doors, locks, leaves, cylinders, hardware, and schedules.
Topic 2_RFICIR and Air Return: Discusses RFICIR (indicated the RFI which is in circulation among different stakeholders), shafts, documents, FCUs, dates, groups, air, and near returns.
Topic 3_Drawing and Structural Design: Involves drawings, walls, structural elements, floors, foundations, and designs.
Topic 4_Opening Solutions: Addresses openings, beams, existing structures, fits, slopes, routing, pipes, and design.
Topic 5_RFI and Damper Control: Covers RFIs, dampers, control, supplies, power systems, dates, and documents.
Topic 6_Brackets and Color Design: Focuses on brackets, colors, cameras, drawings, custom designs, shops, and schedules.
Topic 7_Beam Connections and Steel: Discusses beams, levels, connections, lifts, steel attachments, and columns.
Topic 8_Total Area and Ceiling Design: Addresses total areas, belongings, toilets, principles, rooms, and designs.
Topic 9_Design and Casing: Discusses specific design elements, ceiling shutters, roller casings, large lamellas, and leeuw types.

Evaluation of K-means Clusters

In addition to LDA and NMF, in the K-means model, each cluster is defined by the similarity of documents in the feature space. If a particular keyword is relevant to documents in multiple clusters, it can appear in the top keywords for each cluster.

The list of the 10 topic sets was obtained from the K-means model in Table 3 and the top 20 keywords with the highest weights that characterize each topic.

Table 3: Partial List of RFI Topics in K-means Clustering Method

Topic	Keywords
0	door, schedule, switch, bar, push, frame, ral, alarm, need, F*
1	door, lock, cylinder, leaf, bk, hardware, closer, D*, provision, handle
2	drawing, wall, attached, prab, td, acm, document, information, design, rfi
3	fit, opening, routing, existing, slope, pipe, changed, possible, beam, order
4	convector, size, card, reader, cad, lift, send, signal, riser, inspection
5	subrfi, A*, doc, date, rfi, damper, control, response, td, supply
6	beam, opening, existing, solution, ipe, steel, level, connection, attached, E*
7	ceiling, total, shutter, mep, roller, height, lighting, final, casing, type
8	rficir, A*, shaft, doc, fcu, date, group, rfi, air, near
9	lamella, mm, panel, cps, plate, new, G*, circuit, solution, damper

The explanation of each topic is provided as follows:
Topic 0_Door Schedules and Bar: Involves door schedules, switches, bars, pushes, frames, and alarms.
Topic 1_Door and Hardware: Focuses on doors, locks, cylinders, leaves, hardware, closers, and provisions.
Topic 2_Drawing and Design: Discusses drawings, attached documents, information, and design-related topics.

Topic 3_Fit and Routing Solutions: Addresses fits, openings, routings, existing structures, slopes, pipes, changes, and possible orders.
Topic 4_Convector Size and Lifts: Involves convector sizes, cards, readers, CAD, lifts, signals, and risers.
Topic 5_RFI and Supply Control: Covers RFIs, documents, dates, responses, and damper control.
Topic 6_Beam Solutions and Steel: Discusses beams, openings, existing structures, solutions, IPE, steel, levels, and connections.
Topic 7_Ceiling and Lighting Design: Focuses on ceilings, total areas, shutters, MEP, rollers, heights, lighting, final casings, and types.
Topic 8_RFICIR and Lamella Design: Addresses RFICIR, shafts, documents, FCUs, dates, groups, air, near returns, and lamella designs.
Topic 9_Circuit and New Solutions: Discusses specific elements like circuits, new panels, plates, and innovative solutions.

Comparison of All Clusters

Upon reviewing the results, the observed behavior can be attributed to the inherent differences in the assumptions and structures of the three clustering algorithms (LDA, NMF, and K-means). Each algorithm assigns unique topics based on its approach to modeling the connections between RFI documents and words, albeit with some word overlaps. For instance:

- LDA Topic_0, NMF Topic_7, and K-means Topic_6 share five common keywords (beam, level, connection, steel, attach (LDA) / attached (NMF&K-means)).
- LDA Topic_1, NMF Topic_0, and K-means Topic_5, with different weights, share 6 common keywords (subrfi, doc, date, response, rfi).
- LDA Topic_2, NMF Topic_9, and K-means Topic_7, with different weights, share 6 common keywords (ceiling, shutter, roller, case (LDA) / casing (NMF&K-means), height, type).
- LDA Topic_4, NMF Topic_1, and K-means Topic_1, with different weights, share 4 common keywords (door, lock, cylinder, handle). Additionally, leaf, hardware, bk, and closer are common only in NMF Topic_1 and K-means Topic_1.
- LDA Topic_9, NMF Topic_3, and K-means Topic_2, with different weights, share 4 common keywords (draw (LDA) / drawing (NMF&K-means), prab, acm, design).
- Although LDA exhibits similar topics, as seen in the examples above, it diverges with some topics (e.g., LDA Topic_8).

Table 4: Expert Evaluation of RFI Clustering Methods

	Expert #1	Expert #2	Expert #3	Mean
LDA	5	4	5	4,67
NMF	3	5	4	4,00
K-means	3	4	4	3,67
Mean	3,67	4,33	4,33	

The scale for understanding an RFI text's content (topic) and determining the keywords was defined as 1: very unsuccessful and 5: very successful (Table 4).

From the experts' perspective, LDA is more successful than the other two clustering algorithms in clustering RFIs. LDA can create more distinctive topics and provides a focus on micro-level clustering rather than generating general topics. When considering the characteristic features of RFIs compared to other text data used in the construction sector, such as their lengthiness, noisiness, inclusion of typos, combinations of upper- and lowercase letters, and utilization of variable terms for a single concept, the obtained results are consistent with the literature.

Ratings assigned to NMF and K-means models, except for Expert #2, are identical. This topic similarity has led to experts giving similar ratings, as these models create quite identical topics. None of the experts gave a rating below three because clustering algorithms are generally not used in current RFI analyses. Therefore, a clustering study of this kind is perceived to positively impact on understanding the RFI text's content (topic) and determining the keywords that reveal its content. In summary, even the model with the lowest score (K-means) demonstrates average success in the clustering RFI text data.

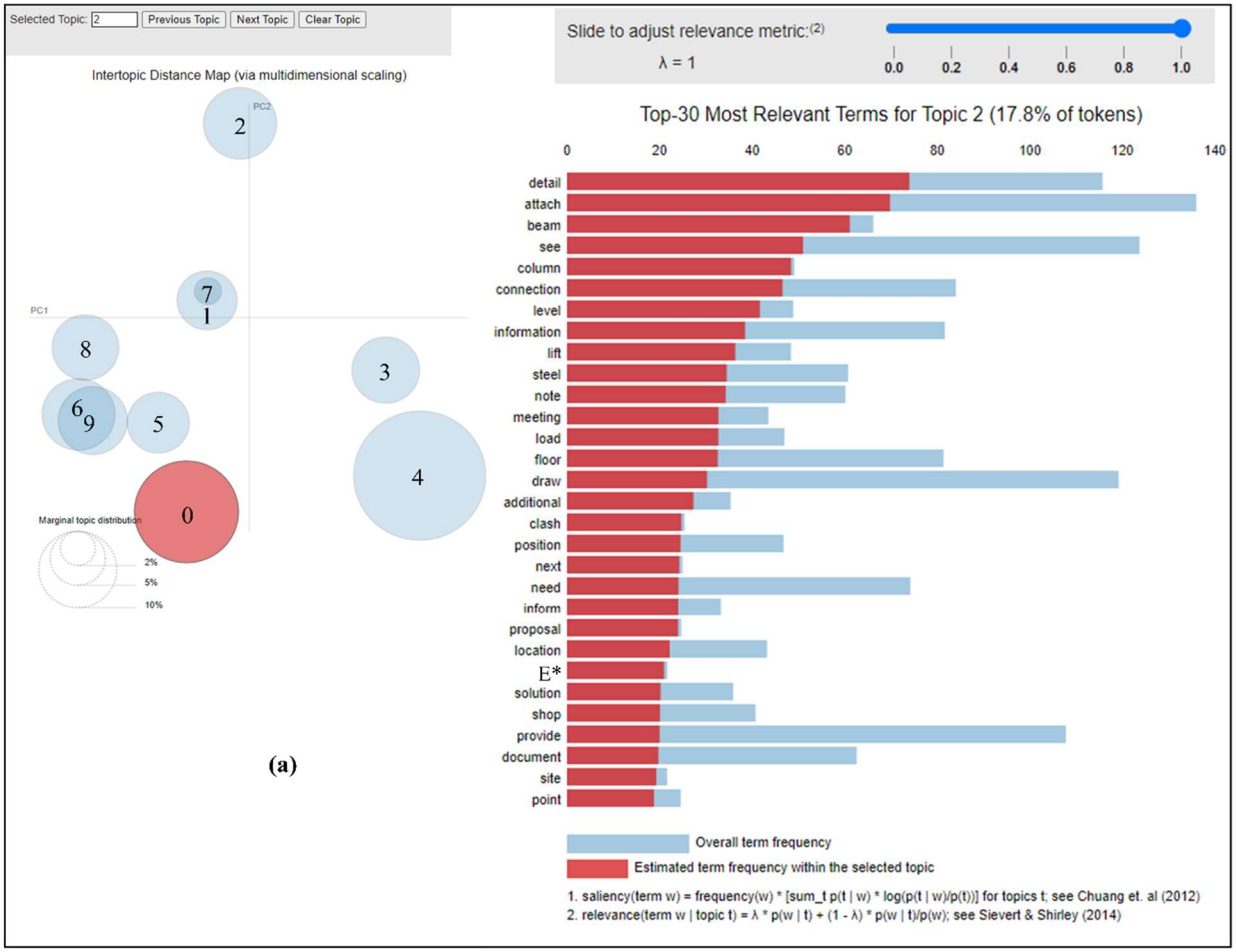
Comparison of Visualization Techniques

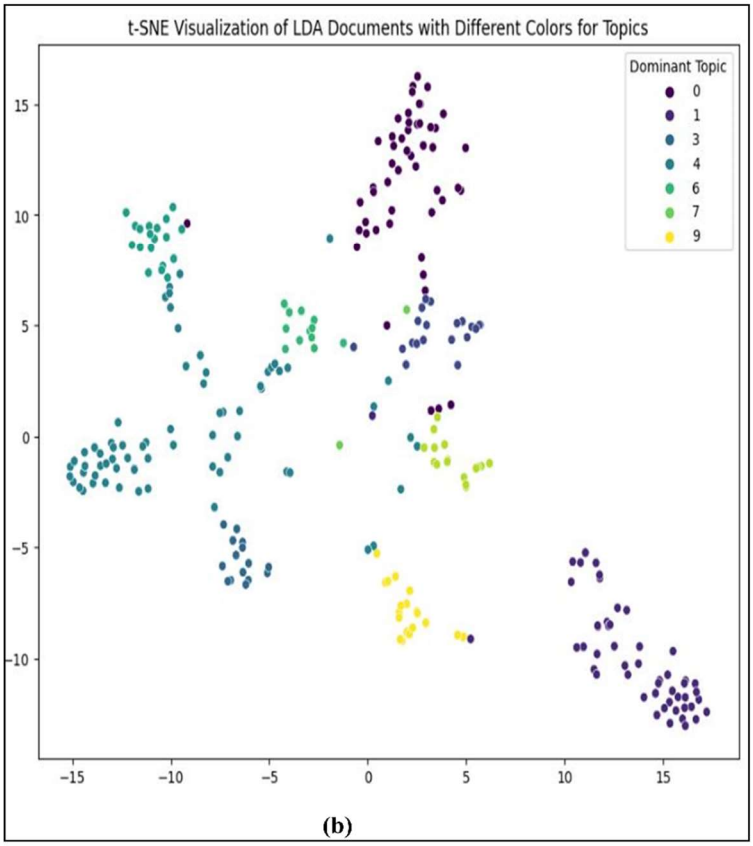
Multiple visualization techniques are employed in clustering problems. The pyLDAvis, specific to LDA models (See Figure 1(a)), is a Python package that empowers users to visualize LDA models effectively. This interactive tool provides features like hovering and clicking for in-depth data exploration. Essentially, pyLDAvis is a critical tool for gaining insights into the topics generated by LDA models and their distribution across documents.

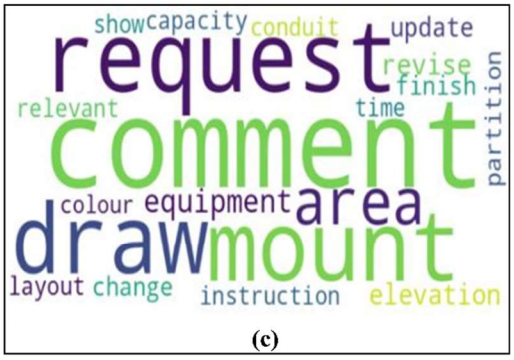
The second method involves scatter plot visualization. For this method, the t-SNE (t-Distributed Stochastic Neighbor Embedding) algorithm was applied first (See Figure 1(b)). t-SNE is a dimensionality reduction technique designed to visualize high-dimensional data compactly by preserving pairwise distances between data points as much as possible in a lower-dimensional space. In essence, t-SNE rearranges data points from a high-dimensional space to a lower-dimensional one, emphasizing the preservation of relationships among similar points.

The third method visualizes topics with word clouds (See Figure 1(c)). Word clouds are highly preferred for identifying and highlighting key themes and subjects within a textual corpus.

They facilitate a quick and effortless analysis of crucial keywords in a dataset by visually displaying frequently occurring words, where the font size is directly proportional to their frequency of occurrence. These word clouds represent a set of words visually, with the most frequent words appearing larger in size and color. They are typically arranged in a circular or elliptical pattern.







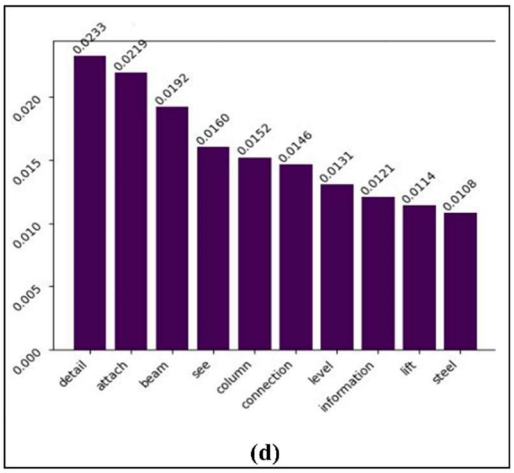


Figure 1: RFI Clustering Visualization Methods (a) pyLDAvis; (b) scatter plot; (c) word cloud; (d) bar charts

The final method utilizes bar charts (See Figure 1(d)). Bar charts are employed in statistical studies for simplicity, interpretability, distribution insight, representation of topic data, and space efficiency. This study applied visualization techniques, including scatter plots, word clouds, and bar charts, to LDA, NMF, and K-means, while pyLDAvis was exclusively used for LDA.

The experts evaluated the outputs of each model with each visualization method and were asked to rank their preferences for visualization tools used for analyzing topics and keywords. Table 5 shows the experts' preferences for visualization techniques regarding RFI topics and keywords.

Table 5: Expert Evaluation of RFI Clustering Visualization Methods

	pyLDAvis	scatter plots with t-SNE	word clouds	bar charts
Expert #1	1	4	3	2
Expert #2	1	4	2	3
Expert #3	1	4	3	2

The most preferred visualization method was identified as pyLDAvis. The experts deemed its ability to filter topic clusters with their keywords rapidly and provide interactive access to be the most useful, practical, and applicable.

Two experts ranked word clouds and bar charts second, providing data on keyword weights. While word clouds offered insights into keyword weights through font sizes, experts found the provision of numeric data more practical.

The scatter plot was evaluated as the least preferred visualization method because of its perceived lack of user-friendliness for quickly analyzing topics despite providing insights into the proximity or overlap of topic clusters.

Conclusions

In this study, 288 RFI texts retrieved from an airport project were analyzed with three different clustering algorithms (LDA, NMF, and K-means), and the results of these algorithms were compared. The generated clusters and keywords were presented using four visualization methods (pyLDAvis, scatter plots, word cloud, and bar charts). In the final stage of the study, with expert opinions, the most preferred algorithm (LDA) and visualization technique (pyLDAvis) were determined. The results of this study provide crucial insights into determining the most effective clustering algorithm and visualization technique for analyzing RFI texts in the literature. Moreover, the findings of this study can benefit project managers and quality engineers for rapidly and effectively analyzing RFI topics to develop proactive measurements for non-conformities before they arise on the construction site.

In the future, comparative analyses can be conducted by increasing the number of RFIs and including different types of projects. Furthermore, the topics in this study's results can be extended by processing them as metadata in RFI texts and analyzing them with machine learning algorithms. While the results of this study play a role in the prior work for visualizing RFIs on the BIM model, visualization and 3D querying practices within the scope of BIM are left for future studies.

References

Abbaszadegan, A., & Grau, D. (2015). Assessing the influence of automated data analytics on cost and schedule performance. *Procedia Engineering*, *123*, 3-6.

Abdirad, H. (2022). Managing digital integration routines in engineering firms: Cases of disruptive BIM cloud collaboration protocols. *Journal of Management in Engineering*, *38*(1), 05021012.

Afzal, M., Wong, J. K. W., & Fini, A. A. F. (2023, August). Unlocking Insights: Analysing Construction Issues in Request for Information (RFI) Documents with Text Mining and Visualisation. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)* (pp. 1-6). IEEE.

Al Qady, M., & Kandil, A. (2014). Automatic clustering of construction project documents based on textual similarity. Automation in construction, 42, 36-49.

Alshari, E., Azman, A., Doraisamy, S., Mustapha, N., & Alksher, M. (2020). Senti2vec: an effective feature extraction technique for sentiment analysis based on word2vec. Malaysian Journal of Computer Science, 33(3), 240-251.

Bhat, A. S. (2017). *Data visualization of requests for information to support construction decision-making* (Doctoral dissertation, University of British Columbia).

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Buntine, W., Lofstrom, J., Perkio, J., Perttu, S., Poroshin, V., Silander, T., ... & Tuulos, V. (2004, September). A scalable topic-based open source search engine. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)* (pp. 228-234). IEEE.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). Introduction to information retrieval. *An Introduction To Information Retrieval*, *151*(177), 5.

Dantas Filho, J. B. P., Angelim, B. M., Guedes, J. P., & Neto, J. D. P. B. (2016). BIM based Request For Information classification and distribution: two residential tower cases. *PARC Pesquisa Em Arquitetura E Construção*, *7*(2), 75-88.

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498.

Farhang, Y. (2017). Face extraction from image based on K-means clustering algorithms. *International Journal of Advanced Computer Science and Applications*, *8*(9).

George, S., & Vasudevan, S. (2020). Comparison of LDA and NMF topic modeling techniques for restaurant reviews. *Indian J. Nat. Sci*, *10*.

Gheeseewan, H., & Pudaruth, S. (2020). Categorisation of Computer Science Research Papers using Supervised Machine Learning Techniques. *International Journal of Computing and Digital Systems*, *9*(6), 1165-1175.

Ham, N. H., & Yuh, O. K. (2023). Performance Analysis and Assessment of BIM-Based Construction Support with Priority Queuing Policy. *Buildings*, *13*(1), 153.

Herrera, R. F., Mourgues, C., Alarcón, L. F., & Pellicer, E. (2019). Assessing design process performance of construction projects. In *Proceedings of the CIB World Building Congress* (pp. 1-10).

Jagannathan, M., Roy, D., & Delhi, V. S. K. (2022). Application of NLP-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies. *CSI Transactions on ICT*, *10*(2), 97-106.

Jallan, Y., Brogan, E., Ashuri, B., & Clevenger, C. M. (2019). Application of natural language processing and text mining to identify patterns in construction-defect litigation cases. *Journal of legal affairs and dispute resolution in engineering and construction*, *11*(4), 04519024.

Kaur, N., & Aggarwal, S. (2017). Comparative analysis of hybrid k-mean algorithms on data clustering. *International Journal of Computer Applications Technology and Research,* 6(8), 384-390.

Khuzaimah, K. H. M., & Hassan, F. (2012). Uncovering tacit knowledge in construction industry: Communities of practice approach. *Procedia-Social and Behavioral Sciences*, *50*, 343-349.

Kim, J. J., Petrov, A. L., Lim, J., & Kim, S. (2022). Comparing cost performance of project delivery methods using quantifiable RFIs: cases in California heavy civil construction projects. *International journal of civil engineering*, *20*(3), 323-335.

Koc, K., Budayan, C., Ekmekcioğlu, Ö., & Tokdemir, O. B. (2024). Predicting Cost Impacts of Nonconformances in Construction Projects Using Interpretable Machine Learning. *Journal of Construction Engineering and Management*, *150*(1), 04023143.

Kohita, R., Yoshida, I., Kitamura, H., & Nasukawa, T. (2020). Interactive construction of user-centric dictionary for text analytics. https://doi.org/10.18653/v1/2020.acl-main.72.

Lai, Y., & Kontokosta, C. E. (2019). Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems*, *78*, 101383.

Lee, J., & Yi, J. S. (2017). Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Applied Sciences*, *7*(11), 1141.

Lee, S. Y., Song, H. A., & Amari, S. I. (2012). A new discriminant NMF algorithm and its application to the extraction of subtle emotional differences in speech. *Cognitive neurodynamics*, *6*, 525-535.

Lin, J. R., Hu, Z. Z., Li, J. L., & Chen, L. M. (2020). Understanding on-site inspection of construction projects based on keyword extraction and topic modeling. *IEEE Access*, *8*, 198503-198517.

Liu, G., Boyd, M., Yu, M., Halim, S. Z., & Quddus, N. (2021). Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Safety and Environmental Protection*, *152*, 37-46.

Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability interface*, *8*(2), 3-6.

McArthur, J. J., Shahbazi, N., Fok, R., Raghubar, C., Bortoluzzi, B., & An, A. (2018). Machine learning and BIM visualization for maintenance issue classification and enhanced data collection. *Advanced Engineering Informatics*, 38, 101-112.

Morales, F., Herrera, R. F., Rivera, F. M. L., Atencio, E., & Nuñez, M. (2022). Potential Application of BIM in RFI in Building Projects. *Buildings*, *12*(2), 145.

Morchid, M., Bouallegue, M., Dufour, R., Linarès, G., Matrouf, D., & De Mori, R. (2015). Compact multiview representation of documents based on the total variability space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(8), 1295-1308.

Naik, G. R. (2016). *Non-negative matrix factorization techniques*. Heidelberg: Springer.

Prihatini, P. M., Putra, I. K. G. D., Giriantari, I. A. D., & Sudarma, M. (2017). Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents. *Contemporary Engineering Sciences*, *10*(9), 403-421.

Rana, S., Jasola, S., & Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. *Artificial Intelligence Review*, *35*, 211-222.

Wang, D., Fan, J., Fu, H., & Zhang, B. (2018). Research on optimization of big data construction engineering quality management based on rnn-lstm. Complexity, 2018, 1-16.

Wu, J. (2012). *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.