Progress in Computing Applications

DLINE JOURNALS

Print ISSN: 2278 – 6465
Online ISSN: 2278 – 6473

PCA 2026: 15 (1)
https://doi.org/10.6025/pca/2026/15/1/18-29

Unsupervised Semantic Analysis of Python FAQ Corpora: Topic Modeling, Similarity Detection, and Structural Optimization

Maleerat Maliyaem
Faculty of Information Technology, King Mongkut's University
of Technology North Bangkok, Bangkok. Thailand
maleerat.m@itd.kmutnb.ac.th

ABSTRACT

This study demonstrates how unsupervised machine learning techniques can audit and optimize technical knowledge repositories through systematic semantic analysis of a curated Python FAQ corpus comprising 163 question answer pairs. Addressing the challenge that over 80% of organizational data exists in unstructured text form, we implement a transparent, eight stage natural language processing pipeline encompassing preprocessing, TF-IDF feature extraction, cosine similarity analysis, Latent Dirichlet Allocation (LDA) topic modeling, Principal Component Analysis (PCA) dimensionality reduction, and network-based co-occurrence analysis. Our methodology emphasizes contextualized preprocessing decisions, responding to contemporary gaps in methodological transparency within organizational text mining research. Results reveal five coherent thematic clusters corresponding to core Python programming concepts: language fundamentals, data structures, syntax and control flow, function operations, and advanced mechanisms. Cosine similarity analysis identifies non trivial semantic overlap among FAQ entries, highlighting actionable opportunities for content consolidation to reduce redundancy and improve retrieval efficiency. Network analysis establishes "python," "function," "list," and "dictionary" as high centrality conceptual anchors within the corpus topology. These findings translate into practical recommendations for technical documentation management, including evidence based FAQ merging, hierarchically organized navigation schemas, and gap identification for underrepresented subject areas. By bridging unstructured text data with structured organizational intelligence, this reproducible framework supports educational chatbot development, curriculum design, and knowledge base optimization. The study underscores that rigorous, contextually justified preprocessing combined with multi perspective unsupervised analytics enables researchers and practitioners to unlock significant value from complex text corpora while ensuring methodological transparency and analytical validity in organizational research applications.

Keywords: Unsupervised Machine Learning, Topic Modeling, Semantic Similarity Analysis, Python FAQ corpus, Natural Language Processing, Text Preprocessing, Knowledge Base Optimization, Latent Dirichlet Allocation, Cosine Similarity

Received: 28 September 2025, Revised 13 December 2025, Accepted 19 December 2025

1. Introduction and Background

It has been estimated that over 80% of all data available to organisations is in unstructured text form [1,2]. Organisations that employ this data in their decision making have been shown to be more successful than those that do not [3, 4], which makes exploiting the vast quantities of available text data tempting. As a potential means of leveraging this data, machine learning (ML) has attracted increasing attention from both industry and academia. In organisational research, ML offers vast potential through its ability to identify patterns that researchers can apply to develop hypotheses grounded in data, to support inductive or abductive exploratory research, and to detect previously undetected patterns in post hoc analyses of regression results, among other applications [5]. However, the task of turning text into data that computers can analyse is not straightforward, as machines understand only numbers, and transforming language into numbers involves many potential pitfalls [6].

1.1 Unsupervised Data Preprocessing

Preprocessing is the application of various techniques to reduce data complexity and size [7,8]. Data preprocessing decisions significantly affect the interpretability and validity of UML results, and what is applicable in one research setting may not be applicable to another. Thus, choices need to be justified specifically within the research discipline and setting [9].Contemporary organizational research using UML lacks transparency for preprocessing choices and does not consider the theoretical and contextual factors pertinent to making these choices e.g.,[10, 11, 12, 14]. Merely copying the preprocessing used in previous literature without providing contextual justification may lead to unsuitable methodological decisions and unwarranted inferences from the analysis.

1.2 Advances in Code Analytics and Translation

Beyond general organizational text, these methodological challenges extend to specialized domains such as software engineering. Semi-supervised learning models are trained by combining both the unlabeled and labelled data, specifically utilizing a small proportion of labelled data with a great deal of unlabeled data. Some event detection efforts based on semi-supervised learning are now presented. In recent years, researchers have begun exploring automatic program translation using supervised deep learning techniques trained on large-scale parallel code corpora. However, parallel resources are scarce in the programming language domain, and collecting bilingual data manually is costly. To address this issue, several unsupervised programming translation systems are proposed [14].

Further advancing code analytics, Liang, Lu [15] [2022] presented KG4Py, a toolkit for generating a knowledge graph from Python source code and enabling semantic search over it. The system analyses Python files on GitHub using static analysis (via concrete syntax trees) to extract functions, imports, and other code entities, building a structured Python code knowledge graph. A hybrid model combining pre-trained and unsupervised embeddings enables natural language queries to retrieve relevant Python code snippets based on semantic similarity rather than simple keyword matching. Experiments show effective graph construction and improved semantic search performance for Python code.

1.3 Social Media Streams and Event Detection

Similarly, social media streams present unique challenges characterized by short messages; utilization of progressively advancing sporadic, casual, abridged words, syntactic and spelling blunders; blended dialects; vagueness; and inappropriate sentence structure [16]. These characteristics make it difficult for methods that rely on them for computational purposes to perform adequately and effectively [17 -21]. Likewise, many current event-detection methodologies have focused primarily on the use of trivial keywords or theme retrieval. Still, they have not fully considered the valuable semantics embedded in social media streams, which hinders the accuracy of event detection [22, 23].

To mitigate these issues, Kolajo, T et al. [24] improved the semantic analysis of noisy terms in social media streams, the representation/embedding of social media stream content, and the summarisation of event clusters in social media streams. These advancements highlight the need to move beyond simple keyword matching toward robust semantic understanding across both code and social media domains.

2. Dataset Description

This study utilizes a curated Python Frequently Asked Questions (FAQs) dataset, designed to support research in natural language processing, information retrieval, and educational chatbot development. The dataset comprises a structured collection of commonly asked Python programming questions, sourced primarily from official Python documentation and reputable community resources. To enhance usability and accessibility, the dataset is further supplemented with concise, self written FAQs aimed at beginner-level understanding while preserving technical accuracy.

Each entry in the dataset consists of a question answer pair accompanied by a subject label that categorizes the content into thematic domains such as *Basics*, *Errors*, *Libraries*, and *Advanced Concepts*. This categorical annotation enables both supervised and unsupervised analytical workflows, including topic discovery and semantic similarity analysis.

The dataset contains 163 FAQ entries, with an average question length of approximately 40 characters and an average answer length of 50 characters, ensuring concise yet informative responses. The data is encoded in Latin-1 and stored in CSV format, facilitating easy integration with standard data analysis and machine learning pipelines.

2.1 Dataset Structure

- question: Textual representation of the Python-related query
- answer: Official or simplified explanation addressing the query
- subject: High-level topical category of the FAQ

This structured design makes the dataset suitable for tasks such as FAQ retrieval, semantic clustering, topic modeling, and chatbot response generation.

3. Methodology

The methodological framework adopted in this study follows a systematic text analytics pipeline encompassing preprocessing, feature extraction, similarity analysis, topic modeling, and visualization, as summarized below.

3.1 Text Preprocessing

All textual data underwent standard preprocessing steps to ensure consistency and reduce noise. These steps included:

1. Conversion of text to lowercase

2. Removal of punctuation and special characters

3. Whitespace normalization

These operations improve downstream feature extraction and similarity computations by minimizing lexical variations.

3.2 Feature Extraction

Textual features were extracted using the Term Frequency Inverse Document Frequency (TF-IDF) representation. A vocabulary size of 50-100 features was used to balance expressiveness and computational efficiency. TF-IDF weighting emphasizes discriminative terms while suppressing commonly occurring but less informative tokens.

3.3 Similarity Analysis

Semantic similarity between FAQ entries was computed using cosine similarity over the TF-IDF vector space. This approach enables the identification of redundant or closely related questions, supporting FAQ optimization and deduplication tasks.

3.4 Topic Modeling

To uncover latent thematic structures, Latent Dirichlet Allocation (LDA) was applied with five topic components. LDA facilitates unsupervised discovery of dominant topics within the dataset and provides insights into topic distribution across questions.

3.5 Dimensionality Reduction

For visualization and exploratory analysis, Principal Component Analysis (PCA) was employed to reduce the high dimensional TF-IDF feature space to two principal components. This transformation enables intuitive visualization of question diversity and topic separation.

3.6 Network Analysis

A co-occurrence network was constructed by thresholding similarity scores to retain only meaningful relationships between FAQs. The resulting network graph highlights structural relationships among questions and reveals clusters of semantically related content.

3.7 Software Environment

All analyses were conducted using Python 3.12, with the following libraries:

- *pandas* for data manipulation
- *NumPy* for numerical computation
- *scikit-learn* for NLP and machine learning tasks
- *matplotlib* and *seaborn* for visualization
- *networkx* for graph-based analysis

This reproducible pipeline ensures methodological transparency and supports extensibility for future research.

4. Analysis

A comprehensive descriptive analysis was conducted to examine the structural, lexical, and semantic characteristics of the Python FAQ dataset. The analysis was implemented using Python 3.12 with standard natural language processing and machine learning libraries, providing both quantitative summaries and visual insights into the dataset composition and thematic organization

4.1 Lexical and Statistical Characteristics

The lexical distribution of the dataset is illustrated in Figure 1, which presents the top 30 most frequently occurring terms in both questions and answers, along with the corresponding length distributions. The word frequency analysis reveals dominant terminology closely aligned with Python programming fundamentals, reflecting the instructional nature of the dataset.

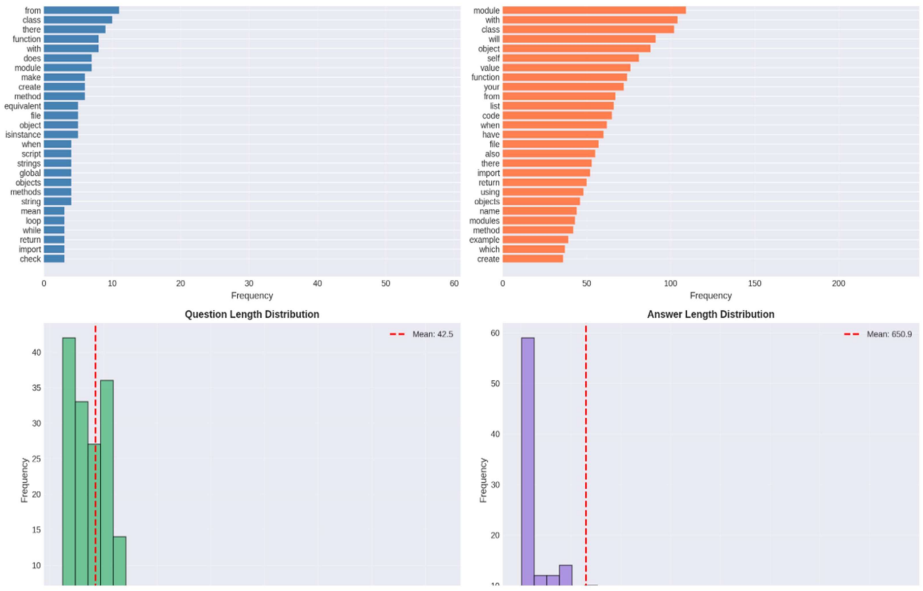

Figure 1. Topic Analysis showing top 30 words in questions/answers and length distributions

Histograms of question and answer lengths indicate a relatively narrow distribution, confirming that most entries are concise and focused. This consistency supports the dataset's suitability for FAQ retrieval systems and conversational agents, where brevity and clarity are essential. Together, these statistics provide an overview of common linguistic patterns and content density within the dataset

4.2 Topic Network and Semantic Relationships

To explore semantic relationships among frequently occurring terms, a keyword co-occurrence network was constructed and visualized in Figure 2. In this network, nodes represent key terms extracted from the dataset, while edges denote their co-occurrence relationships within questions and answers. Node size corresponds to term importance, measured by degree centrality, and edge thickness reflects the strength of association between terms.

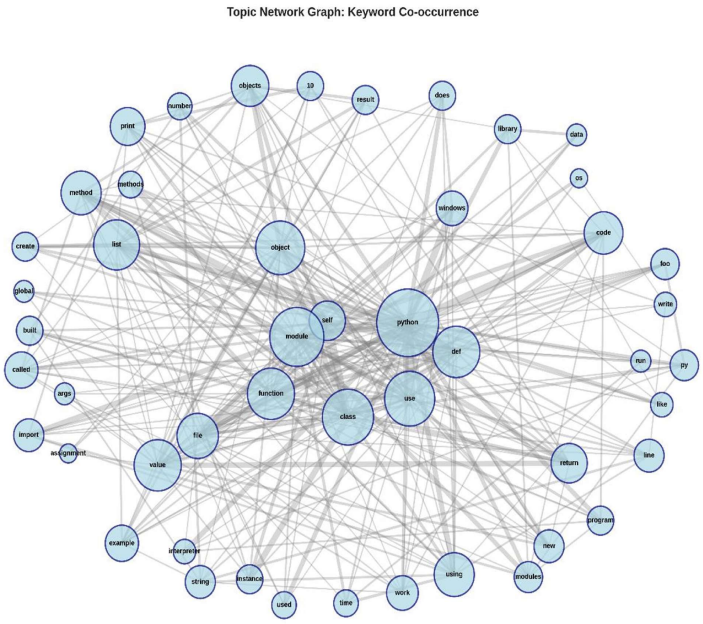

Figure 2. Topic Network Graph visualizing keyword co-occurrences and semantic relationships

The resulting network highlights several central concepts, demonstrating how core Python topics are interconnected. This analysis provides insight into the conceptual structure of the FAQ corpus and identifies key thematic anchors around which other topics are organized

4.3 Similarity Analysis of FAQ Entries

Semantic similarity between questions was quantified using cosine similarity and visualized as a heatmap in Figure 3. The symmetric similarity matrix, computed for a subset of questions, uses a color gradient to indicate degrees of similarity, with higher similarity values represented by warmer colors.

The heatmap reveals clusters of highly similar or near-duplicate questions, indicating redundancy within the dataset. Such patterns suggest opportunities for FAQ consolidation and optimization, particularly in applications aimed at reducing repetitive user queries and improving information retrieval efficiency

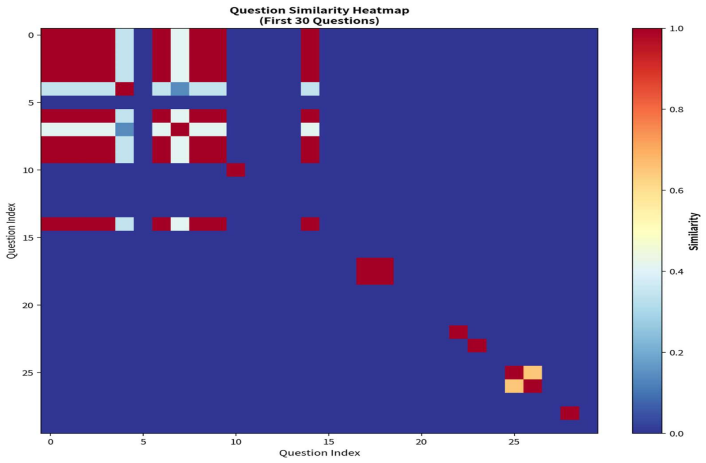

Figure 3. Similarity Heatmap displaying cosine similarity matrix for questions (Red=High, Blue=Low)

4.4 Dimensionality Reduction and Global Structure

To better understand the dataset's global structure, Principal Component Analysis (PCA) was applied to project high-dimensional textual features into a two-dimensional space, as shown in Figure 4. Each point in the projection represents a question, with color encoding used to indicate question length.

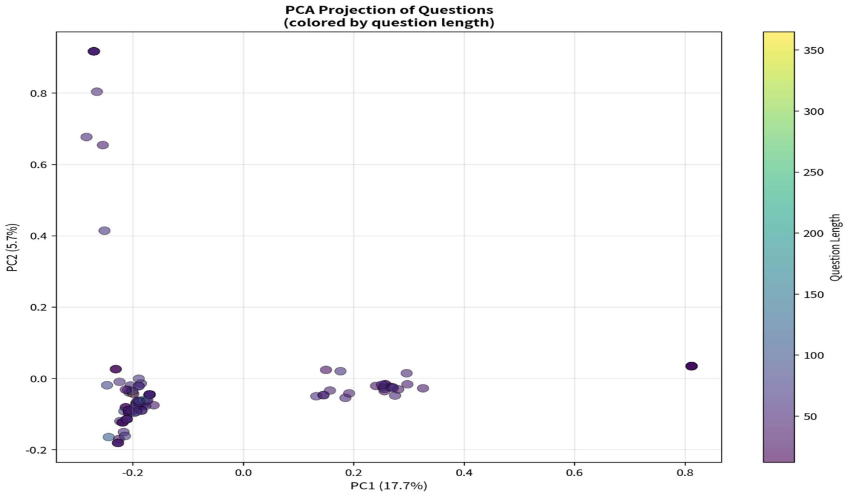

Figure 4. PCA Analysis projecting questions into 2D space, color-coded by length

The PCA visualization reveals natural groupings of questions, suggesting latent similarities in content and structure. This reduced dimensional representation aids in identifying clusters and outliers, providing an intuitive overview of question diversity and dataset organization

4.5 Analysis Workflow and Methodological Transparency

The complete analytical workflow is summarized in Figure 5, which presents an eight-stage pipeline encompassing data input, preprocessing, feature extraction, topic modeling, similarity analysis, dimensionality reduction, visualization, and insight generation. This pipeline ensures methodological transparency and reproducibility while integrating multiple complementary analytical techniques.

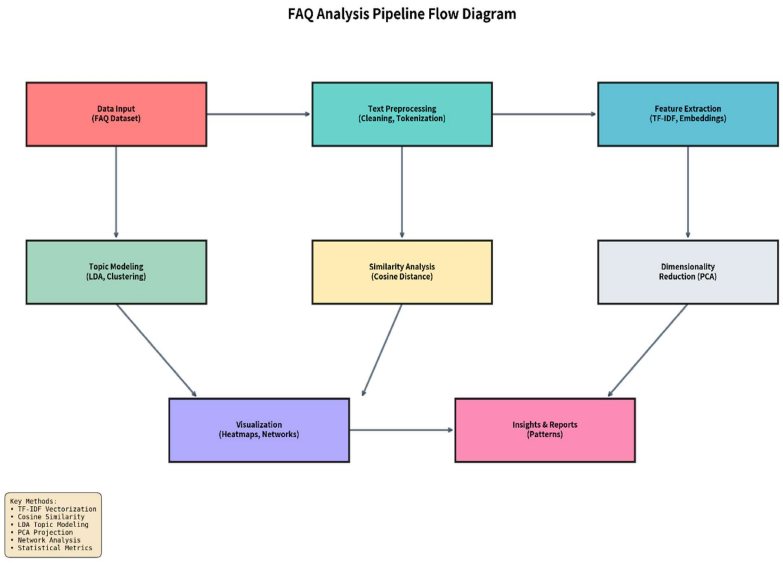

Figure 5. Analysis Pipeline diagram showing the 8-stage workflow from Data Input to Insights

By documenting each stage and the associated methods such as TF-IDF vectorization, cosine similarity, LDA topic modeling, PCA projection, and network analysis the pipeline serves as a structured framework for future extensions of the study

4.6 Topic Modeling and Thematic Distribution

Latent thematic patterns within the dataset were identified using Latent Dirichlet Allocation (LDA) with five topics, as illustrated in Figure 6. Each topic is characterized by its top-ranked terms and relative contribution across the corpus.

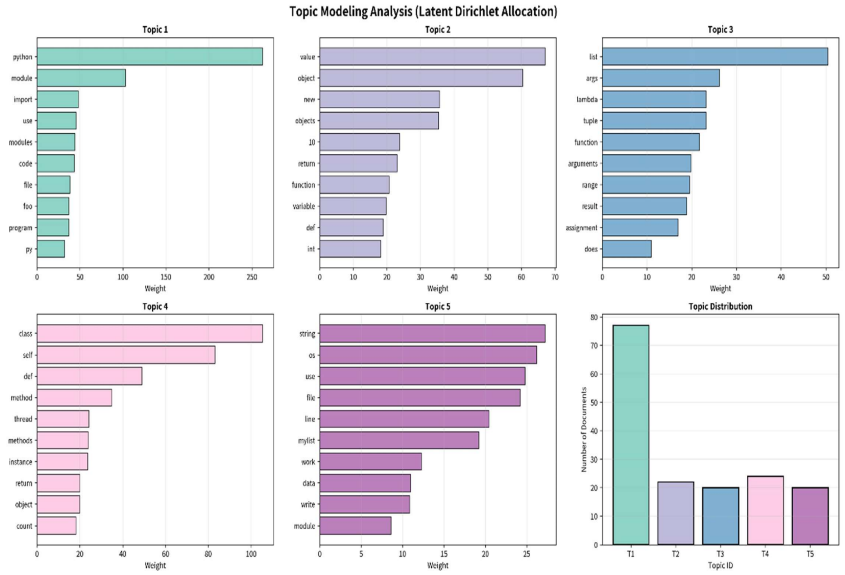

Figure 6. Topic Modeling (LDA) results showing 5 discovered topics with top 10 words each

The discovered topics correspond to key areas of Python programming, including core language features, data structures, syntax, functional operations, and advanced concepts. This thematic decomposition demonstrates the dataset's broad topical coverage and supports its use for content organization, curriculum design, and intelligent FAQ categorization

5. Summary

This study presents a systematic computational analysis of a curated Python programming FAQ dataset comprising 163 question answer pairs. Employing a natural language processing pipeline implemented in Python 3.12 with scikit learn, we generated a suite of complementary visualizations including topic coherence maps, pairwise similarity heatmaps, principal component analysis (PCA) projections, and a reproducible workflow diagram to characterize the semantic structure, thematic organization, and informational redundancy within the corpus.

Our analysis yields four principal findings. First, exploratory lexical statistics confirm that the dataset comprehensively addresses foundational Python programming concepts. Second, Latent Dirichlet Allocation (LDA) topic modeling identifies five coherent thematic clusters: (i) core language features, (ii) data structures and collections, (iii) programming syntax and control flow, (iv) function definition and invocation, and (v) advanced language mechanisms. Third, cosine similarity analysis reveals non trivial semantic overlap among a subset of questions, indicating opportunities for content consolidation to enhance knowledge base efficiency. Fourth, network-based co-occurrence analysis establishes "python," "function," "list," and "dictionary" as high-centrality nodes, underscoring their foundational role in the conceptual topology of introductory Python discourse.

These insights translate into actionable recommendations for technical documentation management. The similarity matrix provides an empirical basis for merging semantically redundant entries, thereby reducing maintenance overhead. Topic modeling outputs support the development of a hierarchically organized, user-centric navigation schema. Word frequency distributions facilitate gap analysis by highlighting underrepresented subject areas warranting expanded coverage. Finally, the explicitly documented eight-stage analytical pipeline from raw data ingestion through preprocessing, feature extraction, modeling, and visualization ensures methodological transparency and supports reproducibility across similar documentation analytics tasks.

Collectively, this multi-perspective analytical framework demonstrates how unsupervised machine learning techniques can be leveraged to audit, optimize, and structurally refine community maintained technical knowledge resources.

6. Conclusion

This study demonstrates the efficacy of unsupervised machine learning techniques in auditing and optimizing technical knowledge repositories. By applying a systematic text analytics pipeline to a curated Python FAQ dataset, we successfully uncovered latent thematic structures and semantic redundancies often obscured in manual curation. The Latent Dirichlet Allocation (LDA) model identified five coherent topics ranging from core language features to advanced mechanisms, validating the dataset's comprehensive coverage of

foundational programming concepts. Furthermore, cosine similarity analysis revealed significant semantic overlap among specific entries, highlighting actionable opportunities for content consolidation to reduce maintenance overhead and improve retrieval efficiency.

Beyond specific dataset insights, this research underscores the critical importance of contextualized preprocessing in organizational text mining. Contemporary research often lacks transparency regarding preprocessing choices, potentially leading to unsuitable methodological decisions. Our transparent, eight-stage pipeline offers a reproducible framework to mitigate these risks. The network analysis further established key conceptual anchors like “function” and “list,” while PCA visualizations aided in identifying outliers and question diversity.

These findings translate into practical recommendations for technical documentation management and educational chatbot development. The similarity matrix supports merging redundant entries, while topic modeling outputs facilitate hierarchically organized navigation. Ultimately, this multi perspective analytical framework illustrates how unsupervised learning can transform unstructured text into actionable organizational intelligence. Given that over 80% of organizational data exists in unstructured form, leveraging such methods is crucial for competitive success. Future research should explore scaling this methodology to larger, multilingual corpora and integrating supervised learning for automated FAQ updating. By bridging the gap between raw text data and structured knowledge, this approach empowers organizations to leverage their unstructured assets more effectively, enhancing both decision making, information retrieval, and user support systems. This work confirms that with rigorous methodological transparency, machine learning can unlock significant value from complex text corpora, ensuring that data driven hypotheses are grounded in robust, contextually justified analysis.

References

[1] Gandomi A., Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35 (2) 137-144.

[2] Robinson T. J., Giles R. C., Rajapakshage R. U. (2020). Discussion of “experiences with big data: Accounts from a data scientist’s perspective. *Quality Engineering*, 32 (4) 543-549.

[3] Cao G., Duan Y. (2017). How do top and bottom performing companies differ in using business analytics? *Journal of Enterprise Information Management*, 30 (6) 874-892.

[4] McAfee A., Brynjolfsson E., Davenport T. H., Patil D. J., Barton D. (2012). Big data: The management revolution. *Harvard Business Review*, 90 (10) 60-68.

[5] Choudhury P., Allen R. T., Endres M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal,* 42 (1) 30-57.

[6] Cambria E., White B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine,* 9 (2) 48-57.

[7] Denny M., Spirling A. (2017, September 27). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *SSRN*. https://doi.org/10.2139/ssrn.2849145.

[8] Hardeniya N., Perkins J., Chopra D., Joshi N., Mathur I. (2016). Natural language processing: Python and NLTK. Packt.

[9] Hickman L., Thapa S., Tay L., Cao M., Srinivasan P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25 (1) 114-146.

[10] Jeong Y., Park I., Yoon B. (2019). Identifying emerging research and business development (R&BD) areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change*, 146, 655-672.

]11] Kim J. H., Chen W. (2018). Research topic analysis in engineering management using a latent Dirichlet allocation model. *Journal of Industrial Integration and Management*, 3 (4) Article 1850016.

[12] Westerlund M., Leminen S., Rajahonka M. (2018). A topic modelling analysis of living labs research. *Technology Innovation Management Review*, 8 (7) 40-51.

[13] White G. O., Guldiken O., Hemphill T. A., He W., Khoobdeh M. S. (2016). Trends in international strategic management research from 2000 to 2013: Text mining and bibliometric analyses. *Management International Review,* 56 (1) 35-65.

[14] Liu, F., Li, J., Zhang, L. (2023). Syntax and Domain Aware Model for Unsupervised Program Translation, 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), *Melbourne, Australia,* 2023, p. 755-767.

[15] Liang, Lu., Li, Yong., Wen, Ming., Liu, Ying. (2022). KG4Py: A toolkit for generating Python knowledge graphs and code semantic search. Connection Science *Connection Science*, 34 (1) 1384 - 1400.

[16] Kolajo, T., Daramola, O., Adebiyi, A., Seth, A. (2020). A framework for pre-processing of social media feeds based on integrated local knowledge base. *Inf Process Manag*. 2020;57(6):102348. https://doi.org/10.1016/j.ipm.2020.102348.

[17] Atefeh, F., Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Comput Intell*. 31 (1) 132–64.

[18] Jain, V. K., Kumar, S., Fernandes, S. L. (2017). Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *J Comput Sci*. 2017;21:316–26. https://doi.org/10.1016/j.jocs.2017.01.010.

[19] Rao, D., M c, Namee, P., Dredze M. (2013). Entity linking: finding extracted entities in a knowledge base. *In*: Poibeau T, Saggion H, Piskorski J, Yangarber R, editors. Multi-source, Multilingual information extraction and summarization. Theory and Applications of Natural Language Processing. Heidelberg *Springer* p. 93–115.

[20] Singh, T., Kumari, M. (2016). Role of text pre-processing in Twitter sentiment analysis. *Procedia Comput Sci*.2016;89:549–54. https://doi.org/10.1016/j.procs.2016.06.095.

[21] Zhan, J., Dahal, B. (2017). Using deep learning for short text understanding. *Journal of Big Data*. 4:34. https://doi.org/10.1186/s40537-017-0095-2.

[22] Katragadda, S., Benton, R., Raghavan, V. (2017). Framework for real-time event detection using multiple social media sources. *In*: *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS). Waikoloa, Hawaii,* p. 1716–1725 https://doi.org/10.24251/HICSS.2017.208.

[23] Xia, C., Schwartz, R., Xie, K., Krebs, A., Langdon, A., Ting, J., Naaman, M. CityBeat: Real-time social media visualisation of hyper-local city data. *In: Proceedings of the 23rd International World Wide Web Conference Committee (IW3C2). Seoul, South Korea*. p. 167–170. https://doi.org/10.1145/2567948.2577020.

[24] Kolajo, T., Daramola, O. Adebiyi, A.A. (2022). Real-time event detection in social media streams through semantic analysis of noisy terms. *J Big Data* 9, 90.