Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 1177–1188
DOI : 10.47974/JIOS-1701

A comprehensive survey and review of machine learning techniques in document processing : Industry applications and future directions

Manisha Tiwari *
Department of Computer Engineering
Mukesh Patel School of Technology Management & Engineering
NMIMS (Deemed-to-be-University)
Mumbai
Maharashtra
India

Padmanabhan Aital †
Department of Technology & Management
Mukesh Patel School of Technology Management & Engineering
NMIMS (Deemed-to-be-University)
Mumbai
Maharashtra
India

Padmaja Joshi §
Centre for Development of Advanced Computing (CDAC)
Mumbai
Maharashtra
India

Abstract

The use of deep learning, machine learning, and natural language-based approaches in the artificial intelligence (AI) field has rapidly advanced document processing efficiency across a wide range of business areas. This research provides a comprehensive survey by providing approaches, models and datasets used for document processing especially for the project proposal documents. The objective of this study is to review the literature on AI-based methods for the aforementioned application, which involves the automated processing of project proposal documents with deep learning, machine learning, and natural

* *E-mail:* **manisha.tiwari@nmims.edu (Corresponding Author)**
† *E-mail:* **padmanabha.aital@nmims.edu**
§ *E-mail:* **padmaja@cdac.in**

© Taru Publications

language processing. It also investigates a case study, which discusses the application of the methodologies in one of the e-governance document-processing task.

Subject Classification: *68T10, 68U10.*

Keywords: *Text mining, Natural language processing, Machine learning, Deep learning.*

1. Introduction

Proposal and contract documents are essential for agreements between contractors and owners, outlining key details such as functionality, roles, delivery schedules, deliverables, and cost. These intricate documents demand expertise for thorough comprehension. [1], [2],[3].

Document processing in Natural Language Processing (NLP) involves analyzing human language text for tasks like information extraction, summarization, and document classification. This survey focuses on project proposal documents, addressing challenges and advantages of using NLP, ML, and DL for information extraction.

We review the application NLP techniques like tokenization, named entity recognition (NER), and sentiment analysis for extracting information from project proposals. ML algorithms aid in document classification and information extraction. DL models, including deep neural networks, offer advantages for advanced tasks like summarization and document understanding.

2. Document Processing Tasks

This survey delves into document processing tasks in Natural Language Processing (NLP), exploring various techniques for data collection and handling, as detailed below:

2.1. *Text Extraction*

Text extraction (TE) is a pivotal NLP technique for automatically retrieving structured information from unstructured text, such as tender documents [4],[5]. In the context of tender documents, TE becomes crucial for extracting key information, facilitating effective analysis, and supporting decision-making in procurement processes. Recent advancements in TE techniques, particularly tailored for tender documents, demonstrate the application's versatility and significance in extracting valuable insights from this specific type of textual data.

2.2 *Text Summarization*

Text summarization is the process of distilling the most important information from a source text to create a concise summary. Latest research in automatic text summarization, particularly in the context of text extraction from tender documents, has focused on single-document, multi-document, and multimodal summarization (text, images, and video). A parallel development is the creation of a multi-document extractive text summarizer, enhancing information synthesis using datasets like ROUGE and DUC-2007. This summarizer incorporates semantic analysis to ensure coherent and representative summaries of tender-related content. [6].

2.3 *Text Sentiment Analysis*

In our research, we explore the multifaceted realm of natural language processing, specifically focusing on sentiment analysis, document classification, document clustering, and Named Entity Recognition (NER). In sentiment analysis, we determine text sentiment using ML and DL techniques, addressing complexities like ambiguity and sarcasm. Our exploration extends to Graph Neural Networks (GNNs)[6][7] to enhance our understanding of sentiment analysis, addressing challenges like domain adaptation and handling noisy data. Transitioning to document classification, we leverage this technique to categorize documents, aiding us in efficient data organization, content analysis, and automation.

In document clustering, we group documents based on similarity without prior knowledge of categories, utilizing algorithms like K-means and hierarchical clustering. This proves valuable in organizing, exploring, and extracting insights from large document collections. Lastly, with NER, we identify and classify named entities, contributing significantly to advancing NLP by extracting structured information and analyzing trends, patterns, and relationships in extensive textual data. Our research strives to overcome challenges and contribute to the evolution of these essential NLP techniques [8],[9],[10].

3. Decoding Project Proposal Documents

Any project, in general, must specify goals and conditions that must be met before it can move on with development. These needs are often collected in the form of text documents from multiple stakeholders and put in one location. In a variety of commercial contexts, such as construction and litigation, documentation is an essential tool.

Numerous businesses get project contracts as a result of their submissions to the agency's solicitation. The company submitting the tender must comprehend the specifications included in the call for bids, including the pre-qualification standards. It must then prepare a response in the form of a proposal that details the technical solution, the estimated time of completion for the project, the cost, the roles of various stakeholders, etc. The team considers a number of factors when preparing such a proposal, including the company's prior work in that field, current components, Proposal documents typically contain a variety of information crucial for project procurement. These documents serve as a communication tool between the company issuing the proposal and potential vendors, outlining the requirements and expectations for the project. The following are key components commonly found in proposal documents:

Project Overview: A detailed description of the project's objectives, scope, and desired outcomes. This section sets the foundation for understanding the project's purpose and goals, **Technical Requirements**: Detailed specifications for the technology, tools, and infrastructure required for the project are outlined. This information helps vendors understand the project's technical complexities and requirements, **Timeline**: Includes a proposed timeline for the project, including key milestones and deadlines. This helps vendors assess the project's feasibility and align their resources accordingly, **Budget**: Budgetary constraints or guidelines, providing vendors with an understanding of the financial parameters of the project. **Evaluation Criteria**: The criteria used to evaluate vendor proposals, including technical expertise, experience, and cost. This information helps vendors tailor their proposals to meet the buyer's expectations, **Terms and Conditions**: Legal and contractual terms governing the project, such as intellectual property rights and termination clauses.

4. Challenges in Processing Proposal Documents Manually

In our research, we identify challenges inherent in the manual processing of proposal documents. Manual reading proves time-consuming, especially when dealing with large volumes, and is subject to interpretation variations among individuals, leading to subjective analysis. The process demands significant human resources, making it less scalable for a growing workload. There is a risk of oversight, potentially missing crucial details during manual reading. The approach incurs labor costs and inefficiencies in collaboration, hindering seamless sharing and

cooperation among stakeholders. Our investigation focuses on these challenges [48] and explores the application of AI techniques to address them.1) Machine learning and its proposal document processing.

Quantifying Efficiency: Metrics for Evaluating Proposal Document Processing: Metrics are quantitative measures used to evaluate the performance of a model or system. In the context of proposal document processing, metrics used to assess the effectiveness of algorithms or techniques in automating various tasks such as classification, summarization, and analysis of proposal documents which includes Accuracy, Precision, Recall, F1-Score, Mean Absolute Error, Cross validation scores.

The literature on proposal document processing highlights various approaches and techniques used in different domains. For instance, in the legal domain, there is a focus on extracting technical terms and identifying emerging trends from US government RFPs using machine learning and NLP, with an emphasis on transfer learning to improve performance. In the construction domain, research on classifying requirements using rule-based and machine learning approaches, with SVM showing the best accuracy. Additionally, in the technology domain, RFPs for IT projects are analyzed using NLP and ML, extracting information on emerging technologies. Other works focus on text classification for proposal documents, using methods like the Maximum Entropy classifier and BERT for requirement extraction and categorization. Some studies also delve into generating answers for questions extracted from RFP documents, aiming to automate the RFP response drafting process. Overall, the literature demonstrates a wide range of NLP and ML techniques being

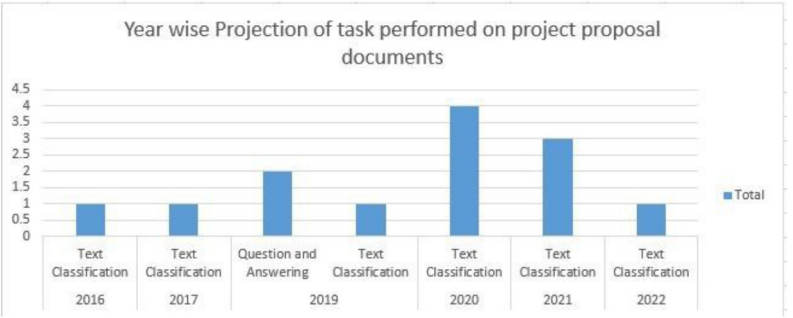


Figure 1

Artificial Intelligence Usage for document processing

applied to enhance proposal document processing, from sentence embedding techniques like Word2Vec and Doc2Vec to advanced neural models like LSTM and BERT for better semantic understanding and context extraction.

Figure 1 depicts the year wise application of AI on project proposal documents, focusing on the specific domains where AI techniques applied. The graph shows the distribution of AI applications across different domains, highlighting the areas where AI used and areas that have seen relatively less application.

5. Open Challenges in automated classification

In automated text classification using NLP, challenges persist in handling ambiguous and context-dependent language, adapting to new domains, managing imbalanced datasets, integrating multimodal information, ensuring model explainability, addressing adversarial attacks, and navigating the dynamic evolution of language. Achieving computational efficiency while balancing model complexity is crucial. Moreover, there is a need for effective integration of user feedback, and interpreting the results accurately in evolving linguistic landscapes remains a priority. Overcoming these challenges demands ongoing research to enhance the adaptability, robustness, and interpretability of text classification models.

6. Text Processing Techniques for Project Proposal Documents: Datasets and Evaluation Metrics

This section outlines the key resources utilized for evaluating text-processing techniques tailored to project proposal documents. These resources encompass datasets used and evaluation metrics used. In this section, domain wise classification of work presented with respect to proposal documents at the early project management stage.

In the legal domain, the extraction of meaningful information from complex documents is time-consuming, with sequential classification used to identify paragraph boundaries and extract clauses from legal contracts. Information retrieval, extraction, NLP, and ontology employed to identify entities in legal documents for data privacy protection. Text summarization using context-free grammar rules used to extract clause information. In the railway domain, the tendering process for bid-related documents is human-dependent, but studies focus on understanding

patterns and classifying requirements automatically, with an expert system proposed for requirement extraction and assignment. In the construction domain, NLP and ML extensively used for understanding document requirements. In the technology domain, analysis of RFPs for IT projects done using NLP and ML to extract information on emerging technologies. Healthcare insurance proposal documents analyzed using a machine learning Bi-LSTM regression model to score RFP responses based on sections affecting proposal responses. The Table 1 discusses the details related to document processing techniques used.

Table 1

Techniques used for text processing for an industry-specific Proposal Document

Citations	Tasks	Domain	Metric Used
[10]	Information Retrieval and NLP for Query Formation	Proposal document of 6 business unit	Precision and Recall
[11]	Text Classification of 7 proposal documents each of approx. 2000 statements	Proposal document of construction domain	Precision, Recall and F1 Score
[12]	An Ontology module system to extract information from the text with Project Proposal document with approx 6000 requirements	Proposal document of Railwaysdomain	Precision and Recall
[13]	Maximum Entropy classifier to extract and categorize information from RFP documents, 16 RFP of IT Services domain	Proposal document of IT Service domain	Accuracy
[19]	Capture semantic relationships between words and documents using text word embedding's and soft cosine measure of 4600 requirements in the form of question and answers	Extraction of information from proposal document of generic domain	Similarity Metrics like Cosine Similarity is used to assess how SCM performed better than WMD

Contd...

[15]	Various ML algorithms are used to classify governance and architecture specific requirements like NB: Naive Bayes SVM: Support Vector Machine LR: Logistic Regression KNN: k-Nearest Neighbors DT: Decision Tree CNN: Convolutional Neural Network RNN: Recurrent Neural Network LSTM: Long Short-Term Memory	Proposal documents with around 3000 clauses and the bid document containing valuable information regarding the specific scope of subcontractors and the classification of design-build needs into three categories: design, construction, and operation and maintenance of construction domain	Precision, Recall, F1-Score
[14]	ABi-LSTM is bidirectional long short-term memory neural network architecture. A regression model based on transformers refers to us- ing transformer architecture, like BERT or GPT, for regression tasks. and 1300 proposal document responses and 1300 proposal document responses	To analyse and score proposal document responses of Insurance Industry Domain	Precision, Recall Accuracy
[20]	PABRE meta model is proposed to classify requirements and prose and Proposal documents with approx 18000 requirement from 6 proposal documents	classification of railway domain proposal documents	NA
[17][18]	Classify documents using logistic Regression model	Proposal documents with approx 2634 statements and	Accuracy
[16]	Classification model RFPcog used to classify requirements using uni- gram features and approx. 300 requirements in the documents	Classification of IT based software proposal documents	Precision

7. Case Study: Proposal document Processing

E-governance proposal documents play a crucial role in modern governance, outlining plans and initiatives for leveraging technology to improve government services. However, the processing of these documents can be complex and time-consuming. This paper proposes a model for the efficient processing of e-governance proposal documents using AI and NLP techniques. The proposed model involves document collection from various sources, text preprocessing, document classification, entity recognition, summarization, sentiment analysis, topic modeling, and insights generation. Automation of this process enhances efficiency, significantly reducing time and effort compared to manual processing. The utilization of AI and NLP ensures accuracy and consistency in information extraction. The insights generated from the processed documents serve as valuable inputs for decision-makers, aiding in the formulation of effective e-governance strategies and policies.

8. Discussion and Conclusion

In conclusion, our survey provides a comprehensive overview of various tasks in document processing within the NLP field. By delving into each task and offering pertinent references, we aim to enhance the understanding and appreciation of advancements in these domains. This survey stands as a valuable resource for fellow researchers and practitioners in NLP, fostering exploration and further advancements in document processing.

References

[1] Kanakaris, Nikos et al. "On the Advancement of Project Management through a Flexible Integration of Machine Learning and Operations Research Tools." International Conference on Operations Research and Enterprise Systems (2019).

[2] Kam K.H. Ng, Chen et al. "A systematic literature review on intelligent automation: Aligning concepts from theory, practice, and future perspectives," *Advanced Engineering Informatics*, Volume 47, 101246, (2021) ISSN 1474-0346.

[3] Bhatnagar, Vaibhav, et al. "Descriptive analysis of COVID-19 patients in the context of India." *Journal of Interdisciplinary Mathematics* 24.3 : 489-504 (2021).

[4] Spasic I, Nenadic G. "Clinical Text Data in Machine Learning: Systematic Review.' *JMIR Med Inform Med Inform*; 8(3) : e17984 (2020).

[5] Pandey et al. "Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images." *Soft Comput* 25, 1563–1580 (2021).

[6] Wenxiong Liao, Bi Zeng et al. "Multi-level graph neural net- work for text sentiment analysis", *Computers and Electrical Engineering*, Volume 92, 107096, (2021) ISSN 0045-7906.

[7] Wankhade,Rao et al."A survey on sentiment analysis methods, applications, and challenges." *Artif Intell Rev* 55, 5731–5780 (2022).

[8] Shervin Minaee, Kalchbrenner et al. "Deep Learning–based Text Classification: A Comprehensive Review" *ACM Computing Surveys*, Volume 54 Issue 3 Article No. : 62 pp 1–40

[9] P. Semberecki and H. Maciejewski, "Deep learning methods for subject text classification of articles", Federated Conference on Computer Science and Information Systems (FedCSIS) pp. 357–360 (2017) ISSN 2300-5963.

[10] Paech, B.et al. "Answering a Request for Proposal – Challenges and Proposed Solutions." In: Regnell, B., Damian, D. (eds) Requirements Engineering: Foundation for Software Quality. REFSQ 2012. Lecture Notes in Computer Science, vol 7195. Springer, Berlin, Heidelberg (2012).

[11] Fahad ul Hassan and Tuyen Le, " Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing", *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*.Volume 12, Issue 2

[12] G. Fantoni, E. Coli et al." Text mining tool for translating terms of the contract into technical specifications: Development and application in the railway sector", *Computers in Industry.* Volume 124, 103357, ISSN 0166-3615 (2021).

[13] Beason, Sterling; Hinton et al."Automated Analysis of RFPs using Natural Language Processing (NLP) for the Technology Domain," SMU Data Science Review Vol. 5 (2021)

[14] Asha Rajbhoj, Padmalata Nistala, Vinay Kulkarni, Pulkit Batra, "AI enabled Project Initiation: An approach based on RFP Response Doc-

ument ", ISEC : 15th Innovations in Software Engineering, Article No.: 22 (2022).

[15] Fahad ul Hassan; Tuyen Le et al. "Multi-Class Categorization of Design-Build Contract Requirements Using Text Mining and Natural Language Processing Techniques", Construction Research Congress (2020).

[16] H. R. Motahari-Nezhad et al. "RFPCog : Linguistic-Based Identification and Mapping of Service Requirements in Request for Proposals (RFPs) to IT Service Solutions, ," 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA (2016),

[17] Fahad ul Hassan, Tuyen Le, "Computer-assisted separation of design-build contract requirements to support subcontract drafting," Automation in Construction. Volume 122, 103479, ISSN 0926-5805 (2021).

[18] A. Sainani, Anish et al."Extracting and Classifying Requirements from Software Engineering Contracts," 2020 IEEE 28th International Requirements Engineering Conference (RE), Zurich, Switzerland, pp. 147-157 (2020).

[19] S. Maji, Appe et al. "An Interpretable Deep Learning System for Automatically Scoring Request for Proposals," 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, pp. 851-855 (2021).

[20] Costal, D., Franch, et al. " On the Use of Requirement Patterns to Analyse Request for Proposal Documents." In: Laender, A., Pernici, B., Lim, EP., de Oliveira, J. (eds) Conceptual Modeling. ER 2019. Lecture Notes in Computer Science(), vol 11788. Springer, Cham

[21] Altameem, Ayman, et al. " P-ROCK: a sustainable clustering algorithm for large categorical datasets." *Intell. Autom. Soft Comput* 35.1 : 553-566 (2023).

[22] Ningwei Liu, Charles Feng et al. "Automate RFP Response Generation Process Using FastText Word Embeddings and Soft Cosine Measure," AICS 2019, Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science.

[23] Costal, D., Franch et al. " On the Use of Requirement Patterns to Analyse Request for Proposal Documents." In: Laender, A., Pernici, B., Lim, EP., de Oliveira, J. (eds) Conceptual Modeling. ER 2019. Lecture Notes in Computer Science(), vol 11788. Springer, Cham.

[24] Pandey, Trilok Nath et al. " Empirical analysis of machine learning techniques for prediction of indian exchange rate," *Journal of Statistics and Management Systems*, 26:1, 13-22.

[25] Bhoyar, Sanjay et al." A machine learning-based predictive approach in evaluating consumer behavior," Journal of Statistics and Management Systems , 26:8 (1955–1963), DOI: 10.47974/JSMS-1131.

[26] Sharma, Ankita & Ghose, Udayan. " Deep learning based bi-polar sentiment classification of movie reviews in Hindi," *Journal of Statistics and Management Systems*, 27:1, 59–86 (2024), DOI: 10.47974/JSMS-1030.