A Survey of Text Embedding Information Retrieval and Domain Adaptation in Construction Technology

www.surveyx.cn

Abstract

This survey explores the multidisciplinary intersection of text embedding, information retrieval, project management, domain adaptation, natural language processing (NLP), construction technology, and semantic analysis. By transforming text into numerical vectors, these technologies enhance computational methodologies across diverse domains. The integration of NLP with systematic reviews and construction technology exemplifies the synergy between machine learning and NLP, improving information retrieval and decision-making processes. The survey highlights the evolution of text embedding techniques, from traditional models to sophisticated neural network-based architectures, and their role in semantic search and content generation. Information retrieval systems are crucial for managing large datasets, with applications in construction technology demonstrating significant advancements in project management and execution. Domain adaptation ensures models are optimized for specific fields, enhancing their performance and security. Semantic analysis facilitates meaning extraction, crucial for effective data management and communication in construction. Challenges such as data quality, integration, scalability, and computational costs are addressed, emphasizing the need for innovative solutions. The survey concludes by underscoring the transformative impact of these technologies, fostering advancements in computational methodologies and enhancing system capabilities across various domains.

1 Introduction

1.1 Multidisciplinary Intersection

The convergence of natural language processing (NLP), construction technology, and project management is critical for advancing computational methodologies across various domains. This intersection is exemplified by the integration of NLP techniques with literary analysis, enhancing character identification and relationship extraction in narratives, thereby improving information retrieval capabilities [1]. The synergy between NLP and systematic reviews in social sciences underscores the importance of machine learning (ML) in synthesizing large datasets for informed decision-making [2]. Furthermore, addressing the challenge of obtaining high-quality text and code embeddings enhances semantic search and text similarity applications, highlighting the role of embeddings in improving information retrieval and content generation [3]. Efficient text summarization techniques are essential for distilling large datasets into actionable insights across various domains [4]. In Mathematical Information Retrieval (MIR), the intersection of mathematics, ML, and information retrieval emphasizes the interconnectedness of these fields in managing complex mathematical data [5]. Additionally, developing benchmarks for lay language generation that incorporate external knowledge illustrates NLP's expanding role in producing accessible content across sectors [6]. Collectively, these intersections highlight the potential for leveraging modern computational techniques to optimize processes across diverse sectors, advancing technological and methodological innovations.

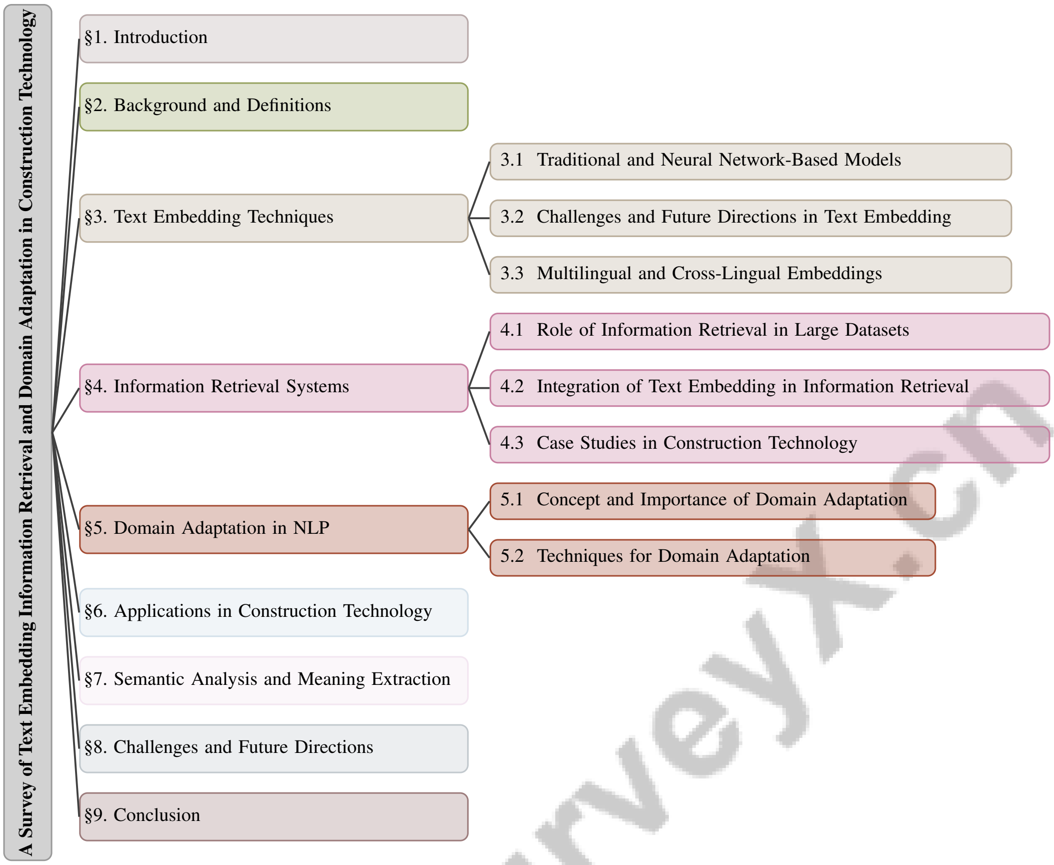

A Survey of Text Embedding Information Retrieval and Domain Adaptation in Construction Technology
§1. Introduction
§2. Background and Definitions
§3. Text Embedding Techniques
3.1 Traditional and Neural Network-Based Models
3.2 Challenges and Future Directions in Text Embedding
3.3 Multilingual and Cross-Lingual Embeddings
§4. Information Retrieval Systems
4.1 Role of Information Retrieval in Large Datasets
4.2 Integration of Text Embedding in Information Retrieval
4.3 Case Studies in Construction Technology
§5. Domain Adaptation in NLP
5.1 Concept and Importance of Domain Adaptation
5.2 Techniques for Domain Adaptation
§6. Applications in Construction Technology
§7. Semantic Analysis and Meaning Extraction
§8. Challenges and Future Directions
§9. Conclusion

Figure 1: chapter structure

1.2 Significance of Text Transformation

Transforming text into numerical vectors is fundamental in computational analysis, facilitating the application of NLP across various domains. This transformation enhances AI-assisted writing tools, significantly improving writing quality and efficiency, particularly in business contexts [7]. In educational settings, this process is essential for thematic analysis using large language models (LLMs), enabling theme identification in student writing and demonstrating the profound impact of text vectorization [8].

In sentiment analysis, effective text representation is pivotal, as exemplified by a novel LSTM-based approach for emotion detection [9]. The healthcare sector also benefits, with NLP techniques extracting embedded information from unstructured clinical data, thereby enhancing system efficiency [10]. The application of WordNet-based semantic search systems, which improve information retrieval through word sense disambiguation, further underscores the significance of text embeddings in overcoming keyword search limitations [11].

In systematic reviews, automating time-intensive stages through ML and NLP transforms traditional processes, emphasizing the importance of text vectorization in streamlining information synthesis [2]. Text summarization is another critical application, where the ability to distill lengthy documents into concise, coherent summaries is paramount while preserving core meaning and essential information [4]. These examples collectively illustrate the pivotal role of converting text into numerical vectors, driving advancements in computational technologies and enhancing system capabilities.

1.3 Structure of the Survey

The survey is organized into nine sections, each exploring distinct yet interconnected areas of study. The introductory section examines the multidisciplinary intersection of text embedding, information retrieval, project management, domain adaptation, NLP, construction technology, and semantic analysis, emphasizing the importance of transforming text into numerical vectors for computational processing. The background and definitions section provides a comprehensive overview of core concepts, defining key terms and elucidating their interrelationships.

The section on text embedding techniques investigates both traditional and neural network-based models, addressing challenges and future directions, and highlighting multilingual and cross-lingual embeddings. The subsequent section on information retrieval systems focuses on their role in managing large datasets, the integration of text embedding, and practical applications illustrated through case studies in construction technology.

Domain adaptation in NLP is scrutinized next, discussing its concept, importance, and various techniques. The applications of innovative NLP techniques within construction technology are analyzed, highlighting their transformative effects on project management and execution processes. By optimizing communication, documentation, and data retrieval, these advancements enhance efficiency, accuracy, and decision-making in construction projects, ultimately leading to improved outcomes [12, 13].

The exploration of semantic analysis and meaning extraction focuses on its role within the construction domain and methods for addressing polysemy and word sense representation. The penultimate section identifies challenges and future directions, addressing issues related to data quality, integration, scalability, and computational costs. The survey concludes with a summary of key findings and contributions, underscoring the significance of a multidisciplinary approach in advancing technological innovations.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts and Definitions

This survey employs a multidisciplinary framework that integrates key concepts such as text embedding, information retrieval (IR), project management, domain adaptation, natural language processing (NLP), construction technology, and semantic analysis. Text embedding involves converting text into dense vector representations, which is crucial for NLP tasks like semantic textual similarity and understanding complex textual structures. The PhysBERT model exemplifies domain-specific embeddings, enhancing information retrieval and semantic analysis in physics, showcasing the utility of specialized embeddings [14]. The ruMTEB benchmark supports evaluating Russian text embedding models, fostering research through comparative analyses [15].

Information retrieval focuses on extracting relevant information from large datasets, addressing vocabulary mismatch through query expansion [16]. Neural networks integrated with traditional IR methods improve document ranking and retrieval efficiency [17]. Understanding multilingual embeddings and inversion attacks is vital for addressing security issues related to semantic content recovery [18].

Project management involves strategic planning and execution to achieve specific objectives, particularly in computational technology [3]. Domain adaptation customizes NLP models for specific fields, ensuring effective knowledge transfer from general-purpose to specialized contexts [16].

Natural Language Processing (NLP) uses computational techniques to understand and generate human language, facilitating tasks like Named Entity Recognition (NER) and Extractive Question Answering (QA) [3]. Modeling textual relationships across modalities presents challenges in representation and retrieval [16]. The application of ML and NLP in synthesizing large data volumes for decision-making is illustrated in systematic reviews and outcomes-based contracting [5].

Semantic analysis is crucial for evaluating machine translations and capturing deep semantic relationships [19]. It involves analyzing language to discern meaning and relationships between concepts, essential for tasks like sentence classification and patent analysis [16]. In construction technology, advancements in NLP and IR optimize project outcomes, emphasizing scalability and efficient data processing [3]. Challenges in tagging and retrieving articles with similar topics but differing terminologies across disciplines are addressed through these advancements [16]. These definitions collectively provide a comprehensive understanding of the interconnected concepts explored in this survey, highlighting their significance in advancing computational methodologies across various domains.

In recent years, the evolution of text embedding techniques has gained considerable attention within the field of natural language processing. This review aims to elucidate the various methodologies employed in this domain, particularly focusing on both traditional and neural network-based models. To enhance the understanding of this complex landscape, Figure 2 illustrates the hierarchical categorization of text embedding techniques. This figure encompasses traditional and neural network-based models, challenges and future directions, as well as advancements in multilingual and cross-lingual embeddings. Each category within the figure explores significant methodologies, current challenges, and innovative solutions, thereby highlighting the diverse landscape and ongoing progress in text embedding technologies. By analyzing these categories, we can better appreciate the intricacies involved in the development of effective text embeddings and the implications for future research.

3 Text Embedding Techniques

3.1 Traditional and Neural Network-Based Models

Text embedding methodologies have transitioned from traditional statistical frameworks to sophisticated neural network architectures, significantly advancing textual data representation and processing. Foundational models such as the Vector Space Model (VSM) and Language Models (LM) employ term weighting and relevance feedback for effective information retrieval, with VSM particularly adept in health information contexts [20]. The integration of latent Dirichlet allocation with keyphrase extraction enhances topic interpretability and computational efficiency, underscoring the enduring relevance of traditional models [21].

As illustrated in Figure 3, the categorization of text embedding methodologies into traditional models, neural network models, and benchmarks/tests highlights key methods and their applications in text representation. Neural network-based models capture intricate semantic relationships within text. The PhysBERT model, built on the BERT architecture and pre-trained on an extensive corpus of arXiv physics papers, exemplifies the efficacy of domain-specific embeddings in task performance enhancement [14]. Similarly, the Semantic Dependency and Keyword Evaluation Method (SDKEM) merges semantic dependency analysis with keyword information to improve semantic comprehension [19].

Large language models (LLMs) like Llama 2 and GPT-4 have further enriched text embedding capabilities, excelling in tasks such as text summarization and retrieval augmentation [6]. The application of Generalized Log-Likelihood Ratio Tests (GLRT) in text embedding provides a unique approach to maximizing information from rare events [22].

Recent advancements are exemplified by benchmarks such as ruMTEB, which introduces 23 tasks tailored for the Russian language, addressing significant gaps in existing benchmarks and emphasizing multilingual and cross-domain applications [15]. These methodologies illustrate a diverse landscape of text embedding, where both traditional statistical techniques and advanced neural network models contribute to the effective representation and processing of textual data across various applications.

3.2 Challenges and Future Directions in Text Embedding

The progression of text embedding technologies faces challenges that necessitate innovative solutions to enhance efficacy and applicability. A key issue is the integration of domain-specific knowledge into embedding models, as traditional methods often struggle with mathematical symbol ambiguity and diverse sequence types, especially in languages lacking clear word boundaries. Rare event handling remains a challenge due to underrepresentation in datasets, potentially leading to biased models [22].

Developing privacy-preserving methods is hindered by limited understanding of how various embedding strategies affect models like Vec2Text [23], complicating robust privacy solutions. Adapting retrieval technologies to accommodate diverse languages and sequence types also presents ongoing challenges [24].

Future research should explore strategies to enhance semantic structure and reduce processing time, such as employing pre-trained models fine-tuned for specific tasks [16]. The introduction of benchmarks assessing retrieval systems against LLM-generated relevance judgments highlights the need for statistically significant evaluations to improve retrieval accuracy [25]. Incorporating adversarial training frameworks, like the Adversarial Context-aware Network Embedding (ACNE),

AI-generated, for reference only.

offers promising avenues for learning embeddings for unseen nodes, addressing current limitations in model generalization [26].

Addressing these challenges and exploring innovative future directions, such as integrating advanced graph techniques for keyphrase extraction and utilizing large language models for text enrichment, can significantly improve text embedding accuracy and efficiency. These enhancements may lead to superior document similarity assessments, with potential performance boosts of up to 27

3.3 Multilingual and Cross-Lingual Embeddings

Advancements in multilingual and cross-lingual embeddings have been pivotal in overcoming the limitations of traditional text embedding models, which are predominantly trained on monolingual corpora. These embeddings facilitate text processing across multiple languages by projecting them into a shared semantic space, thereby enhancing cross-lingual information retrieval and semantic understanding [27]. The Multilingual Embedding for Cross-Lingual Information Retrieval (MECIR) method exemplifies this approach by constructing a multilingual embedding space from topically aligned corpora, enabling effective cross-lingual retrieval [27].

Recent developments have produced top-performing models such as E5, GTE, and BGE, which significantly enhance the robustness and accuracy of multilingual embeddings [28]. These models employ innovative architectures and training strategies to align text representations across languages, ensuring semantic nuances are preserved.

The G2P2 model introduces three graph interaction-based contrastive strategies during pre-training, effectively aligning text and graph representations to improve low-resource text classification across languages, highlighting the integration of graph-based methodologies in enriching multilingual embeddings' semantic representation capabilities [29].

Moreover, the development of novel evaluation metrics and defense mechanisms tailored for multilingual and cross-lingual settings addresses the unique challenges posed by these models, ensuring embeddings maintain their integrity and security across diverse linguistic environments, as demonstrated by benchmarks introducing new evaluation criteria [18].

Advancements in multilingual and cross-lingual embedding techniques, along with innovations such as large language model-based text enrichment, signify substantial progress in enhancing the functionality of text embedding models. These developments enable effective navigation of diverse linguistic environments, improving cross-lingual information retrieval and broadening applicability across various natural language processing applications. The integration of high-quality multilingual datasets and the use of instruction-tuned models further enhance embedding systems' performance, addressing challenges posed by varying linguistic contexts [27, 30, 31, 28].

4 Information Retrieval Systems

4.1 Role of Information Retrieval in Large Datasets

Information retrieval (IR) systems play a crucial role in managing and extracting insights from large datasets, a necessity heightened by the rapid proliferation of digital information. The challenges of processing vast data volumes are notably pronounced in Cross-Language Information Retrieval (CLIR) systems, where translating extensive document collections imposes significant computational demands [32]. Dynamic frameworks like the Dynamic Data Partitioning Framework (DDPF) enhance the adaptability of IR systems, offering more effective real-time applications than static methods [33].

In the biomedical domain, the PubMed dataset, with its 198,366 documents and structured Medline citation fields, serves as a benchmark for advanced indexing and retrieval methodologies, underscoring the necessity for IR systems capable of handling large-scale, structured data [34]. This capability is essential for improving access to medical literature and supporting research initiatives.

Incorporating entity type recognition into IR processes enhances retrieval accuracy, particularly for short search queries often lacking context [35]. IR systems are also vital for systematic reviews and academic literature management, addressing the complexities of synthesizing extensive data and managing large document collections [2]. In software tools, IR systems are crucial for classifying and retrieving metagenomics software, where precise classification based on descriptions is vital for effective data management [36].

The importance of IR in large datasets is further illustrated by the need for standardized metadata and efficient data processing algorithms, as seen in projects like the Gutenberg Corpus, which includes comprehensive metadata such as author information and genre [37]. Recent advancements in IR methodologies, including retrieval-augmented generation techniques, highlight the necessity for innovative approaches that address traditional limitations and enhance processing efficiency [38]. The current landscape of IR systems emphasizes natural language queries and the need for improvements in code retrieval, reflecting the growing complexity of modern IR applications [17]. Additionally, the construction of test collections for IR has become increasingly labor-intensive due to expanding document sizes, necessitating efficient relevance assessment methods to maintain IR efficacy [25].

These developments collectively underscore the critical role of information retrieval systems in managing large datasets across diverse applications, driving innovations that enhance data accessibility and utility while addressing challenges posed by the expanding digital information landscape [24]. As illustrated in Figure 4, the hierarchical structure of key challenges, applications, and advancements in information retrieval systems within large datasets highlights innovations in cross-language retrieval, biomedical applications, and advancements in metadata standardization.

4.2 Integration of Text Embedding in Information Retrieval

The integration of text embedding techniques into information retrieval (IR) systems has markedly enhanced their accuracy and efficiency by transforming text into numerical vectors that capture semantic similarities often overlooked by traditional keyword-based methods. This transformation addresses the limitations associated with natural language nuances, thereby improving the retrieval process [16]. In specialized domains such as physics, models like PhysBERT illustrate how domain-specific embeddings can capture unique semantics and contextual relationships within literature, significantly boosting retrieval accuracy [14].

Text embeddings have also demonstrated substantial improvements in multilingual contexts, where methods leveraging topical relevance between documents in different languages enhance cross-lingual information retrieval [27]. The synergy between embedding techniques and retrieval-augmented models improves the quality and simplicity of summaries while maintaining factual correctness, showcasing the potential of embeddings to refine IR outcomes [6].

Moreover, the combination of IR-based and neural-based comment generation approaches dynamically enhances comment generation accuracy, illustrating the versatility of embedding techniques in augmenting IR systems [39]. The incorporation of Generalized Log-Likelihood Ratio Tests (GLRT) into IR systems further enhances accuracy and efficiency, particularly in analyzing structured symbolic sequences [22].

Advanced embedding transformation methods that ensure privacy protection while maintaining retrieval effectiveness offer practical solutions for service providers, addressing security concerns associated with text embeddings [23]. Additionally, integrating semantic dependencies and keyword relevance in machine translation evaluation has been shown to enhance IR accuracy, as demonstrated by methods incorporating these elements into the evaluation process [19].

The continuous advancement and integration of text embedding techniques, such as hybrid document embedding approaches and domain-specific vector space representations, are essential for enhancing the effectiveness of information retrieval systems. Recent studies indicate that these methods, utilizing graph techniques for keyphrase extraction and incorporating semantic similarity measures, can significantly improve the retrieval of relevant documents, outperforming traditional baselines by up to 27

4.3 Case Studies in Construction Technology

The application of information retrieval and text embedding technologies in construction technology has been pivotal in advancing project management and data handling. A notable example is the deployment of retrieval-augmented large language models (LLMs), which have shown significant improvements over traditional models in generating accurate and contextually relevant responses, especially in the complex linguistic landscape of construction documents [40]. This capability is crucial given the specificity and technicality of language in construction, where precise information retrieval is essential for effective decision-making.

The integration of advanced embedding techniques, such as those employed in Arctic-embed models, demonstrates their potential in real-world information retrieval scenarios. These models have achieved state-of-the-art retrieval accuracy, facilitating efficient processing of large datasets, a common requirement in construction projects that necessitate extensive documentation analysis [41]. Furthermore, contextualization and organization algorithms like ITR, combined with Word2Vec, have outperformed traditional methods such as IS-TFIDF in clustering quality, highlighting the importance of efficient document management in enhancing project execution [42].

Moreover, enriching metadata in multidisciplinary digital libraries has significantly improved retrieval performance, as evidenced by pipelines that enhance F1 measures compared to baseline methods. This approach is particularly relevant in construction technology, where integrating diverse data sources is critical for comprehensive project analysis [43]. The application of these technologies is further illustrated by experiments on large collections of scientific articles, such as those from the NASA SciX digital library, demonstrating the capability to efficiently retrieve and process vast amounts of data [44].

BIM-GPT, a prompt-based virtual assistant, has exhibited remarkable NLP capabilities, achieving high accuracy rates in classifying user intent and building object categories in natural language queries. This showcases the potential of NLP and text embedding technologies in enhancing data efficiency and accuracy in construction-related tasks [45].

Collectively, these case studies reveal the transformative effects of information retrieval and text embedding technologies within the construction industry. They illustrate how these technologies enhance data management processes, improve information retrieval accuracy, and ultimately optimize project outcomes. Advancements such as hybrid document embedding approaches, data-driven strategies for combining word embeddings, and the benchmarking of pre-trained text embedding models exemplify their effectiveness in aligning complex built asset information with domain-specific technical terminology. These innovations facilitate automated cross-mapping of data and significantly increase the relevance and precision of retrieved documents, addressing long-standing challenges in managing and integrating technical content in construction projects [46, 47, 48, 49, 50].

5 Domain Adaptation in NLP

5.1 Concept and Importance of Domain Adaptation

Domain adaptation is vital in natural language processing (NLP) for tailoring models to specific fields, thereby enhancing semantic representation and performance. General models often fail to capture domain-specific nuances, necessitating adaptation methods such as Domain Adaptation with Description (DAD), which allows for effective adaptation without direct access to target documents [51]. Models like PhysBERT underscore the significance of domain-specific NLP customization, adeptly handling the specialized language of physics literature [14]. Techniques like Generalized Log-Likelihood Ratio Tests (GLRT) enhance analytical accuracy in genomic and text analysis across diverse applications [22]. Additionally, adapting models to counter multilingual vulnerabilities, such as inversion attacks, ensures security and effectiveness across languages [18]. Refining evaluation methods, such as the Semantic Dependency and Keyword Evaluation Method (SDKEM) for machine translation, further emphasizes the need for specialized performance metrics [19].

Domain adaptation overcomes global model limitations by employing feature generation and domain-specific insights, significantly improving reliability and effectiveness in various applications. This dual approach enhances classification tasks, embedding performance, and keyword extraction in fields like healthcare, finance, and enterprise search [31, 52, 53, 54].

5.2 Techniques for Domain Adaptation

Domain adaptation techniques are crucial for enhancing NLP model performance by aligning them with the specific characteristics of various domains. Maximum Mean Discrepancy (MMD) is a prominent technique that fine-tunes global models for enterprise settings, reducing domain shifts and improving accuracy [54]. This approach is particularly beneficial in enterprise email systems, where domain-specific language necessitates tailored adjustments.

Adapting document retrieval algorithms to accommodate diverse sequence types and languages exemplifies the versatility of domain adaptation methods. Novel frameworks ensure effective processing of various linguistic structures and content, maintaining retrieval systems' relevance and accuracy across applications [24]. Cross-lingual retrieval challenges, especially without aligned corpora, are addressed by leveraging multilingual embeddings, aligning text representations across languages in a shared semantic space to enhance effectiveness [27].

Dense Retrieval Adaptation with Description (DAD) represents an innovative domain adaptation technique. By constructing synthetic target collections and generating queries from domain descriptions, DAD facilitates effective model adaptation without extensive domain data [51].

These techniques illustrate strategies designed to adapt NLP models to various domains, addressing specific challenges to improve the models' capacity for accurate and contextually relevant outputs. Incremental methods enhance thematic context characterization by refining vocabulary through iterative querying, improving retrieval effectiveness. In engineering education research, NLP applications like clustering, summarization, and prompting techniques analyze student essays, enabling educators to identify key themes and patterns efficiently. These advancements enhance NLP model adaptability across fields, facilitating more effective analysis and insights [8, 55].

As depicted in Figure 5, domain adaptation in NLP enhances model performance in new domains or datasets divergent from initial training. The "Optimal Loss vs Computational Budget" technique compares methods like full fine-tuning and bias tuning, illustrating the trade-off between computational resources and performance. The "All Enterprise Data" Venn diagram organizes enterprise data into overlapping categories, highlighting data type complexity within larger datasets. The "MatchZoo Studio and Library" image outlines a platform for automated machine learning, featuring components that facilitate domain-adapted model development and application. These examples collectively provide a comprehensive overview of strategies and tools for adapting NLP models to diverse and evolving data environments [56, 54, 40].

6 Applications in Construction Technology

6.1 Innovative NLP Techniques in Construction

NLP techniques are revolutionizing construction technology by enhancing efficiency and accessibility. Semantic-aware representation through contrastive learning improves multimodal data embedding, crucial for interpreting diverse data in complex construction environments [57]. AI-assisted tools like Google Smart Compose and GPT-3 streamline communication, reduce errors, and enhance collaboration in construction project management [7]. Meta-textual embeddings advance information retrieval, facilitating quick access to extensive textual data and supporting informed decision-making [58]. The development of domain-specific NLP models, akin to ChemTEB in chemistry, suggests potential advancements in construction technology, improving data processing quality and efficiency [59].

Recent NLP advancements are transforming construction by boosting data processing, communication, and project management. For example, NLP has been applied to analyze drilling reports in the oil and gas sector for automatic classification of operational data. Additionally, NLP techniques predict citation impacts, analyze large datasets, and extract insights. Weak supervision-based NLP assesses climate change impacts on infrastructure, efficiently labeling vast scientific literature. These innovations streamline construction processes, optimizing decision-making and operational efficiency across industries [12, 60, 13].

6.2 Impact on Project Management and Execution

Advanced technologies like text embedding and information retrieval are optimizing project management and execution in construction by enhancing workflows and decision-making. For instance, sequence mining and pattern analysis significantly impact project management in the oil and gas industry by facilitating informed decisions and efficient operations [13]. These technologies also

AI-generated, for reference only.

accelerate access to information crucial for climate adaptation strategies, supporting effective policy development [60].

Interpretable embeddings, such as Supervised Explicit Semantic Analysis (SESA), align project tasks with human-understandable concepts, enhancing execution efficiency by clearly defining roles [61]. The Retrieval-Augmented Feature Generation (TIFG) method excels in generating meaningful features, critical for improved construction project outcomes [53]. Additionally, integrating deep semantic representations into recommendation systems improves material relevance, influencing project management by ensuring access to pertinent research [62].

Terminology-Based Text Embedding (TDE) focuses on keyphrases and contextual relationships, providing accurate technical content representations vital for effective project management [49]. The integration of machine learning, NLP, and generative pre-trained transformers is reshaping construction project management, enhancing operational efficiency through streamlined information retrieval from complex building information models (BIM), improving decision-making accuracy via automated dataset analysis, and leading to favorable project outcomes through improved communication throughout the project lifecycle [62, 13, 2, 12, 43].

7 Semantic Analysis and Meaning Extraction

Semantic analysis plays a crucial role in deciphering the complex meanings of words, especially in specialized domains such as construction. This involves addressing language complexities, such as polysemy, where a single term may have multiple interpretations depending on context. Tackling these challenges is essential for enhancing the precision and effectiveness of semantic analysis. The following subsection explores polysemy and the methods used to represent word senses accurately, highlighting their significance in construction semantics.

7.1 Semantic Analysis in Construction

In the construction industry, semantic analysis enhances the understanding and interpretation of complex textual data by examining language to uncover meanings and relationships among concepts. This is vital for managing the vast information generated in construction projects. Techniques like Latent Semantic Analysis (LSA) and advanced Natural Language Processing (NLP) methods extract nuanced insights from unstructured data, identifying topic structures and improving semantic annotation. These capabilities optimize information retrieval and extraction, enriching decision-making processes and project outcomes by providing contextually relevant information tailored to specific domains [63, 64, 65].

Semantic analysis primarily enhances information retrieval systems by leveraging semantic relationships to retrieve pertinent documents accurately, crucial for project managers needing quick access to specific information. Advanced models, including those integrating deep learning techniques, automate the extraction of semantic patterns from construction documents, improving retrieval efficiency and accuracy [19].

Furthermore, semantic analysis aids in developing intelligent systems that automatically classify and organize construction-related data, beneficial in large projects where data volume can be overwhelming. By implementing semantic analysis, construction firms can systematically categorize project documentation, enhancing accessibility and overall project management [16].

Moreover, semantic analysis contributes to creating effective communication tools within the construction industry. By employing advanced NLP techniques to analyze the semantic context of language in construction documents, these tools enhance clarity and precision among stakeholders, reducing misunderstandings and errors, leading to more efficient project execution and better outcomes [66, 67, 63, 68, 55].

Semantic analysis is integral to modern construction technology, utilizing methodologies like LSA and Supervised Explicit Semantic Analysis (SESA) to reveal hidden structures in textual data, improve information retrieval, and extract actionable insights from complex documents such as drilling reports. This transformation of unstructured data into operational intelligence enhances project management efficiency and promotes continuous improvement in construction practices through data-driven decision-making [63, 13, 61, 65].

7.2 Polysemy and Word Sense Representation

Managing polysemy and accurately representing word senses are significant challenges in semantic analysis, particularly in construction, where technical terms often have multiple meanings. Sparse coding techniques effectively extract distinct sense vectors from the combined embeddings of polysemous words, facilitating nuanced meaning representation and enhancing semantic analysis precision [68].

Sparse coding decomposes word embeddings into basis vectors, each representing a potential word sense, allowing researchers to identify and separate meanings associated with polysemous words, improving the clarity and precision of semantic representations. Leveraging linear algebraic structures to recover individual word senses from embeddings and employing adversarial neural networks to disentangle denotation from connotation significantly enhances interpretability and performance in tasks like information retrieval and semantic analysis. These advancements are particularly beneficial in construction technology, where recognizing subtle terminology differences can impact project outcomes and communication efficacy [61, 69, 48, 65, 68].

Integrating advanced NLP techniques, such as part-of-speech tagging and word sense disambiguation, further enhances word meaning disambiguation, improving performance in critical tasks like information retrieval and text classification. Semantic search methods utilizing resources like WordNet to understand polysemous and synonymous words achieve substantial performance improvements, evidenced by a 17.7

Research into managing polysemy and representing word senses is critical in semantic analysis. Approaches like linear algebraic structures in word embeddings demonstrate that multiple word senses can coexist in linear superposition. Methods such as LSA modulate word meanings based on context, effectively addressing polysemy and synonymy issues. These advancements enhance our understanding of word sense representation and contribute to developing robust models for natural language processing and information retrieval [68, 65]. Techniques like sparse coding deepen our understanding of complex language structures and support the development of reliable NLP tools, advancing construction technology capabilities.

8 Challenges and Future Directions

8.1 Challenges in Data Quality and Integration

The integration of NLP and text embedding technologies faces significant challenges concerning data quality and integration, crucial for their effectiveness. General-purpose text embedding models often fall short in specialized domains like physics, adversely impacting data quality [14]. Additionally, the computational complexity of methods such as Generalized Log-Likelihood Ratio Tests (GLRT) complicates their application in large datasets, hindering integration efforts [22].

Biases in training data and inconsistent data quality across tasks, particularly in multilingual contexts vulnerable to inversion attacks, necessitate tailored defenses to enhance multilingual embeddings' robustness. Traditional evaluation methods often fail to capture semantic correctness, highlighting the need for improved methodologies to assess data quality and integration accurately [19].

The benchmark for retrieval augmentation with large language models (LLMs) reveals inadequacies in addressing complexities of generating high-quality background explanations, indicating a need for refinement in integration strategies [6]. Furthermore, false positives in LLM-generated judgments raise concerns about fairness and accuracy, essential for reliable data integration [25].

These challenges underscore the necessity for innovative methodologies to enhance data quality and integration across diverse systems. Addressing these issues is vital for advancing NLP and text embedding technologies, improving model accuracy and utility in applications like document similarity assessment, citation impact prediction, and domain-specific tasks. Leveraging advancements in LLMs can help overcome vocabulary, context, and grammatical accuracy limitations, driving innovation in performance metrics [12, 28, 31, 49].

AI-generated, for reference only.

8.2 Scalability and Computational Costs

Scalability and computational cost challenges significantly hinder the deployment of advanced NLP and text embedding technologies, especially for large datasets and diverse applications. Evaluating multilingual embeddings requires substantial computational resources, complicating the scaling of NLP research across multiple languages [18]. The intensive processing power needed for analytical methods like GLRT further limits their applicability in large-scale contexts [22].

Document retrieval algorithms encounter scalability challenges when handling diverse sequences, necessitating adaptive computational strategies to maintain performance [24]. Although methods for learning multilingual embeddings have improved retrieval times, they still face scalability issues that require innovative solutions to manage computational costs effectively [27].

Recent advancements in comment generation methods, which integrate information retrieval with neural-based approaches, show significant progress over traditional techniques; however, ensuring scalability remains a critical area for future research [39]. Addressing these challenges involves refining embedding transformation methods to enhance privacy and retrieval effectiveness while managing computational expenses [23].

Exploring domain-specific information and implicit descriptions in dense retrieval adaptation methods is crucial for overcoming scalability barriers and optimizing resource allocation [51]. Future research should focus on improving the reliability and statistical validity of LLM-generated judgments, essential for effective relevance assessment in expansive applications [25].

9 Conclusion

The survey underscores the transformative role of integrating text embedding, information retrieval, and domain adaptation within construction technology and NLP. This multidisciplinary approach has substantially enhanced efficiency and adaptability, as demonstrated by models such as BERT, which improve the accuracy of retrieving similar bug reports. The application of ML and NLP in systematic reviews has streamlined the management of complex social science literature, highlighting a fruitful avenue for future exploration. The advancement of retrieval-augmented methods enhances the interpretability of lay language summaries, paving the way for innovations in automated background explanation generation. The incorporation of convolutional residual retrieval models has markedly improved document retrieval performance, underscoring the need for sophisticated retrieval techniques in information processing. Additionally, the survey emphasizes the significance of enhanced semantic understanding in mathematical information retrieval, with the introduction of Mathematical Object Identifiers (MOIs) offering a promising research direction. Techniques like the Generalized Log-Likelihood Ratio Tests (GLRT) illustrate the broad applicability of these methodologies in enhancing accuracy and efficiency across text and genomic analyses. The successful identification and classification of metagenomics tools provide valuable guidance for researchers in selecting suitable software for analysis, thereby expanding the utility of NLP-based classification systems. Furthermore, advancements in machine translation evaluation methods, which bolster accuracy through semantic analysis, highlight the pivotal role of semantic techniques in refining language processing applications. This survey highlights the importance of a multidisciplinary approach in advancing computational methodologies, thereby enhancing system capabilities and fostering innovative solutions to complex challenges in information processing and retrieval.

AI-generated, for reference only.

References

[1] Adrian Groza and Lidia Corde. Information retrieval in folktales using natural language processing, 2015.

[2] Iman Munire Bilal, Zheng Fang, Miguel Arana-Catania, Felix-Anselm van Lier, Juliana Outes Velarde, Harry Bregazzi, Eleanor Carter, Mara Airoldi, and Rob Procter. Machine learning information retrieval and summarisation to support systematic review on outcomes based contracting, 2024.

[3] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training, 2022.

[4] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023.

[5] André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William Grosky, and Bela Gipp. Why machines cannot learn mathematics, yet, 2019.

[6] Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. Retrieval augmentation of large language models for lay language generation, 2024.

[7] Carlos Alves Pereira, Tanay Komarlu, and Wael Mobeirek. The future of ai-assisted writing, 2023.

[8] Andrew Katz, Umair Shakir, and Ben Chambers. The utility of large language models and generative ai for education research, 2023.

[9] Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. A sentiment-and-semantics-based approach for emotion detection in textual conversations, 2018.

[10] Sijia Liu, Yanshan Wang, Andrew Wen, Liwei Wang, Na Hong, Feichen Shen, Steven Bedrick, William Hersh, and Hongfang Liu. Create: Cohort retrieval enhanced by analysis of text from electronic health records using omop common data model, 2019.

[11] Vuong M. Ngo, Tru H. Cao, and Tuan M. V. Le. Wordnet-based information retrieval using common hypernyms and combined features, 2018.

[12] Adilson Vital Jr. au2, Filipi N. Silva, Osvaldo N. Oliveira Jr. au2, and Diego R. Amancio. Predicting citation impact of research papers using gpt and other text embeddings, 2024.

[13] Júlio Hoffimann, Youli Mao, Avinash Wesley, and Aimee Taylor. Sequence mining and pattern analysis in drilling reports with deep natural language processing, 2017.

[14] Thorsten Hellert, João Montenegro, and Andrea Pollastro. Physbert: A text embedding model for physics scientific literature, 2024.

[15] Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design, 2025.

[16] Haoming Jiang, Tianyu Cao, Zheng Li, Chen Luo, Xianfeng Tang, Qingyu Yin, Danqing Zhang, Rahul Goutam, and Bing Yin. Short text pre-training with extended token classification for e-commerce query understanding, 2022.

[17] Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Yichun Yin, Hao Zhang, Yong Liu, Yasheng Wang, and Ruiming Tang. Coir: A comprehensive benchmark for code information retrieval models, 2024.

[18] Yiyi Chen, Heather Lent, and Johannes Bjerva. Text embedding inversion security for multilingual language models, 2024.

[19] Kewei Yuan, Qiurong Zhao, Yang Xu, Xiao Zhang, and Huansheng Ning. Evaluation of machine translation based on semantic dependencies and keywords, 2024.

[20] Harsh Thakkar, Ganesh Iyer, and Prasenjit Majumder. A comparative study of approaches in user-centered health information retrieval, 2015.

[21] Michał Łopuszyński. Application of topic models to judgments from public procurement domain, 2014.

[22] Ted Dunning. Finding structure in text, genome and other symbolic sequences, 2012.

[23] Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. Understanding and mitigating the threat of vec2text to dense retrieval systems, 2024.

[24] Gonzalo Navarro. Spaces, trees and colors: The algorithmic landscape of document retrieval on sequences, 2013.

[25] David Otero, Javier Parapar, and Álvaro Barreiro. On the statistical significance with relevance assessments of large language models, 2024.

[26] Tony Gracious and Ambedkar Dukkipati. Adversarial context aware network embeddings for textual networks, 2020.

[27] Mitodru Niyogi, Kripabandhu Ghosh, and Arnab Bhattacharya. Learning multilingual embeddings for cross-lingual information retrieval in the presence of topically aligned corpora, 2018.

[28] Hongliu Cao. Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark, 2024.

[29] Zhihao Wen and Yuan Fang. Augmenting low-resource text classification with graph-grounded pre-training and prompting, 2023.

[30] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[31] Nicholas Harris, Anand Butani, and Syed Hashmy. Enhancing embedding performance through large language model-based text enrichment and rewriting, 2024.

[32] Atsushi Fujii and Tetsuya Ishikawa. Applying machine translation to two-stage cross-language information retrieval, 2000.

[33] Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen. Exploring a multidimensional representation of documents and queries (extended version), 2010.

[34] Alexandros Ioannidis. An analysis of indexing and querying strategies on a technologically assisted review task, 2021.

[35] Walid Shalaby, Khalifeh Al Jadda, Mohammed Korayem, and Trey Grainger. Entity type recognition using an ensemble of distributional semantic models to enhance query understanding, 2016.

[36] Kaoutar Daoud Hiri, Matjaž Hren, and Tomaž Curk. Nlp-based classification of software tools for metagenomics sequencing data analysis into edam semantic annotation, 2022.

[37] Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics, 2018.

[38] Mohamed Morchid, Juan-Manuel Torres-Moreno, Richard Dufour, Javier Ramírez-Rodríguez, and Georges Linarès. Automatic text summarization approaches to speed up topic model learning process, 2017.

[39] Huang Yuchao, Wei Moshi, Wang Song, Wang Junjie, and Wang Qing. Yet another combination of ir- and neural-based comment generation, 2021.

[40] Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. Matchzoo: A learning, practicing, and developing system for neural text matching, 2019.

[41] Gulshan Saleem, Nisar Ahmed, and Usman Qamar. Text mining through label induction grouping algorithm based method, 2021.

[42] Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping pubmed queries to documents, 2017.

[43] Junwen Zheng and Martin Fischer. Bim-gpt: a prompt-based virtual assistant framework for bim information retrieval, 2023.

[44] Yan Xiao, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Beyond precision: A study on recall of initial retrieval with neural representations, 2018.

[45] Rahul Ragesh, Sundararajan Sellamanickam, Arun Iyer, Ram Bairi, and Vijay Lingam. Hetegcn: Heterogeneous graph convolutional networks for text classification, 2020.

[46] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. Legal document retrieval using document vector embeddings and deep learning, 2018.

[47] Mehrzad Shahinmoghadam and Ali Motamedi. Benchmarking pre-trained text embedding models in aligning built asset information, 2024.

[48] Alfredo Silva and Marcelo Mendoza. A data-driven strategy to combine word embeddings in information retrieval, 2021.

[49] Hamid Mirisaee, Eric Gaussier, Cedric Lagnier, and Agnes Guerraz. Terminology-based text embedding for computing document similarities on technical content, 2019.

[50] Martin Semmann and Mahei Manhei Li. New kids on the block: On the impact of information retrieval on contextual resource integration patterns, 2023.

[51] Helia Hashemi, Yong Zhuang, Sachith Sri Ram Kothur, Srivas Prasad, Edgar Meij, and W. Bruce Croft. Dense retrieval adaptation using target domain description, 2023.

[52] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[53] Xinhao Zhang, Jinghan Zhang, Fengran Mo, Yuzhong Chen, and Kunpeng Liu. Retrieval-augmented feature generation for domain-specific classification, 2024.

[54] Brandon Tran, Maryam Karimzadehgan, Rama Kumar Pasumarthi, Michael Bendersky, and Donald Metzler. Domain adaptation for enterprise email search, 2019.

[55] Carlos M. Lorenzetti and Ana G. Maguitman. Learning better context characterizations: An intelligent information retrieval approach, 2010.

[56] Alicja Ziarko, Albert Q. Jiang, Bartosz Piotrowski, Wenda Li, Mateja Jamnik, and Piotr Miłoś. Repurposing language models into embedding models: Finding the compute-optimal recipe, 2024.

[57] Pierre Lamart, Yinan Yu, and Christian Berger. Semantic-aware representation of multi-modal data for data ingress: A literature review, 2024.

[58] Toshitaka Kuwa, Shigehiko Schamoni, and Stefan Riezler. Embedding meta-textual information for improved learning to rank, 2020.

[59] Ali Shiraee Kasmaee, Mohammad Khodadad, Mohammad Arshi Saloot, Nick Sherck, Stephen Dokas, Hamidreza Mahyar, and Soheila Samiee. Chemteb: Chemical text embedding benchmark, an overview of embedding models performance efficiency on a specific domain, 2024.

[60] Tanwi Mallick, Joshua David Bergerson, Duane R. Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. Analyzing the impact of climate change on critical infrastructure from the scientific literature: A weakly supervised nlp approach, 2023.

[61] Dasha Bogdanova and Majid Yazdani. Sesa: Supervised explicit semantic analysis, 2017.

[62] Hebatallah A. Mohamed, Giuseppe Sansonetti, and Alessandro Micarelli. Tag-aware document representation for research paper recommendation, 2022.

[63] Thierry Hamon, Adeline Nazarenko, Thierry Poibeau, Sophie Aubin, and Julien Derivière. A robust linguistic platform for efficient and domain specific web content analysis, 2007.

[64] Gagnon Michel, Zouaq Amal, Aranha Francisco, Ensan Faezeh, and Jean-Louis Ludovic. An analysis of the semantic annotation task on the linked data cloud, 2018.

[65] Juan C. Valle-Lisboa and Eduardo Mizraji. The uncovering of hidden structures by latent semantic analysis, 2006.

[66] Javad Rafiei Asl and Juan M. Banda. Gleake: Global and local embedding automatic keyphrase extraction, 2020.

[67] Enes Altuncu, Jason R. C. Nurse, Yang Xu, Jie Guo, and Shujun Li. Improving performance of automatic keyword extraction (ake) methods using pos-tagging and enhanced semantic-awareness, 2025.

[68] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy, 2018.

[69] Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. Are "undocumented workers" the same as "illegal aliens"? disentangling denotation and connotation in vector spaces, 2020.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

AI-generated, for reference only.

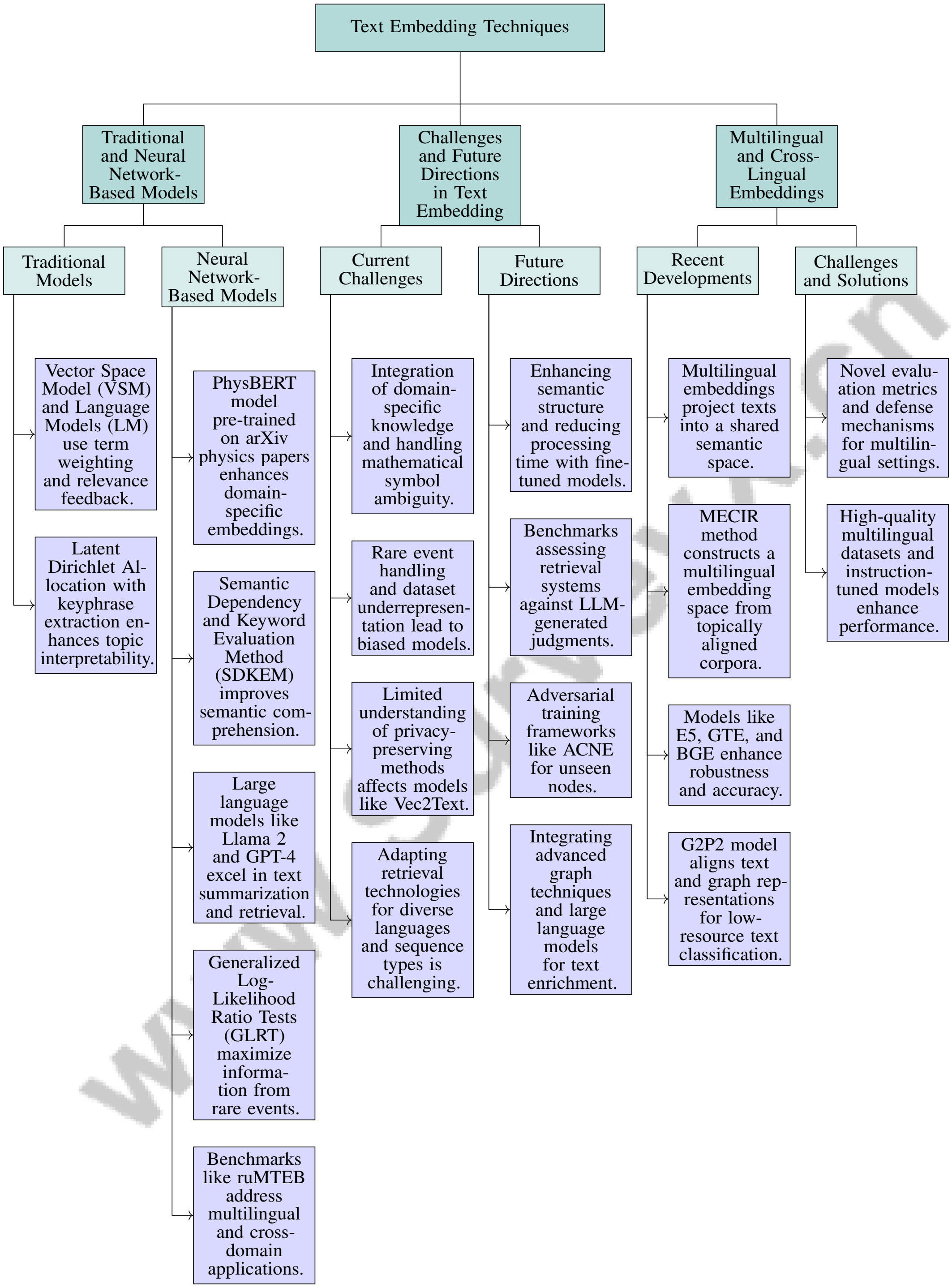

Text Embedding Techniques
Traditional and Neural Network-Based Models
Challenges and Future Directions in Text Embedding
Multilingual and Cross-Lingual Embeddings
Traditional Models
Neural Network-Based Models
Current Challenges
Future Directions
Recent Developments
Challenges and Solutions
Vector Space Model (VSM) and Language Models (LM) use term weighting and relevance feedback.
Latent Dirichlet Allocation with keyphrase extraction enhances topic interpretability.
PhysBERT model pre-trained on arXiv physics papers enhances domain-specific embeddings.
Semantic Dependency and Keyword Evaluation Method (SDKEM) improves semantic comprehension.
Large language models like Llama 2 and GPT-4 excel in text summarization and retrieval.
Generalized Log-Likelihood Ratio Tests (GLRT) maximize information from rare events.
Benchmarks like ruMTEB address multilingual and cross-domain applications.
Integration of domain-specific knowledge and handling mathematical symbol ambiguity.
Rare event handling and dataset underrepresentation lead to biased models.
Limited understanding of privacy-preserving methods affects models like Vec2Text.
Adapting retrieval technologies for diverse languages and sequence types is challenging.
Enhancing semantic structure and reducing processing time with fine-tuned models.
Benchmarks assessing retrieval systems against LLM-generated judgments.
Adversarial training frameworks like ACNE for unseen nodes.
Integrating advanced graph techniques and large language models for text enrichment.
Multilingual embeddings project texts into a shared semantic space.
MECIR method constructs a multilingual embedding space from topically aligned corpora.
Models like E5, GTE, and BGE enhance robustness and accuracy.
G2P2 model aligns text and graph representations for low-resource text classification.
Novel evaluation metrics and defense mechanisms for multilingual settings.
High-quality multilingual datasets and instruction-tuned models enhance performance.

Figure 2: This figure illustrates the hierarchical categorization of text embedding techniques, encompassing traditional and neural network-based models, challenges and future directions, and advancements in multilingual and cross-lingual embeddings. Each category explores significant methodologies, current challenges, and innovative solutions, highlighting the diverse landscape and ongoing progress in text embedding technologies.

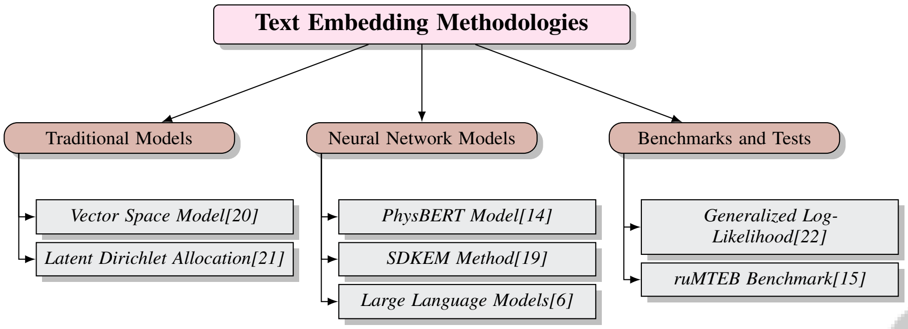


Figure 3: This figure illustrates the categorization of text embedding methodologies into traditional models, neural network models, and benchmarks/tests, highlighting key methods and their applications in text representation.

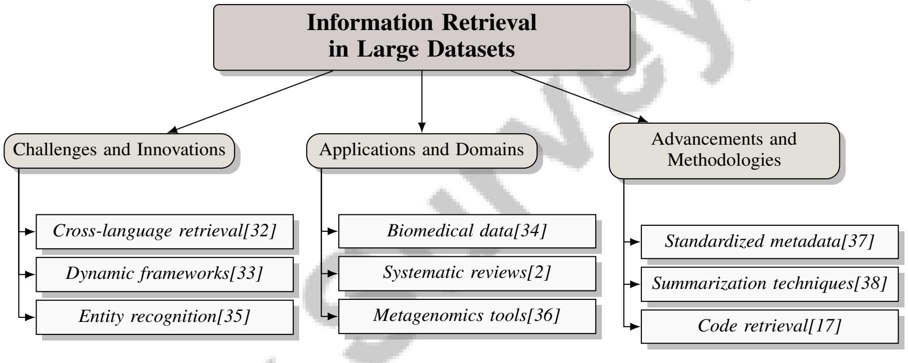
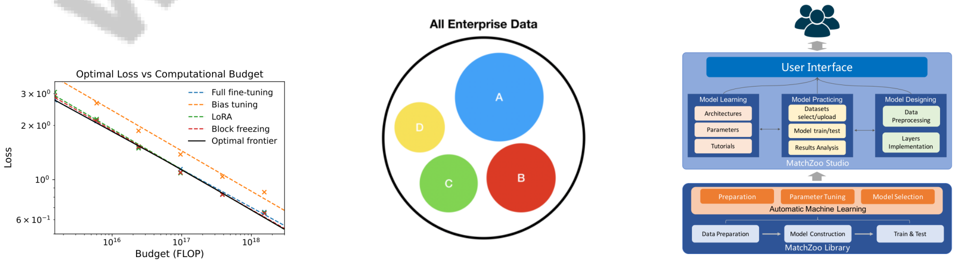

Information Retrieval in Large Datasets
Challenges and Innovations
Applications and Domains
Advancements and Methodologies
Cross-language retrieval[32]
Dynamic frameworks[33]
Entity recognition[35]
Biomedical data[34]
Systematic reviews[2]
Metagenomics tools[36]
Standardized metadata[37]
Summarization techniques[38]
Code retrieval[17]

Figure 4: This figure illustrates the hierarchical structure of key challenges, applications, and advancements in information retrieval systems within large datasets, highlighting cross-language retrieval innovations, biomedical applications, and advancements in metadata standardization.



(a) Optimal Loss vs Computational Budget[56]

(b) All Enterprise Data[54]

(c) MatchZoo Studio and Library[40]

Figure 5: Examples of Techniques for Domain Adaptation

AI-generated, for reference only.