A Survey on Analyzing Construction and Accident Reports Using Deep Learning and Natural Language Processing

www.surveyx.cn

Abstract

This survey paper explores the transformative role of deep learning and natural language processing (NLP) in analyzing construction and accident reports, highlighting their significance in enhancing safety management practices. By automating information extraction and improving data analysis precision and efficiency, these technologies enable proactive safety measures. The integration of external knowledge into large language models (LLMs) further enhances their capability to address safety engineering challenges, providing actionable insights. Despite these advancements, challenges remain, particularly concerning user trust and understanding in AI systems, which are critical in decision-critical domains like construction safety management. The development of user-centric explanation processes and the refinement of neural information retrieval systems are essential for overcoming these barriers. The survey also underscores the potential of innovative methods, such as the RCCN approach, in enhancing text classification tasks, and highlights the successful integration of crowdsourced data with structured optimization approaches, exemplified by CROME, in balancing accuracy and localization in emergency response. Overall, the findings emphasize the transformative impact of deep learning and NLP technologies in the construction industry, providing a foundation for future research and innovation in safety management, with a focus on addressing existing challenges and exploring new methodologies to offer more sophisticated, reliable, and efficient solutions for analyzing construction and accident reports.

1 Introduction

1.1 Significance of Advanced Computational Techniques

The integration of advanced computational techniques, particularly deep learning and natural language processing (NLP), is crucial for analyzing construction and accident reports, significantly enhancing information extraction precision and efficiency. These methodologies automate the understanding of complex textual data, essential for meticulous information retrieval in high-stakes environments. Taye et al. highlight deep learning's transformative role in addressing challenges across various fields, including NLP [1].

In the construction sector, such techniques improve safety measures through the analysis of textual data, as demonstrated by near-miss management systems [2]. Semantic search methods that examine workplace documents for potential hazards further exemplify their utility in enhancing safety in new construction projects [3].

Autonomous vehicles also benefit from these advancements, where accident report analysis is vital to understanding dynamics, collision types, and contributing factors, thus addressing safety and regulatory concerns [4]. Additionally, the systematic identification and synthesis of deep learning

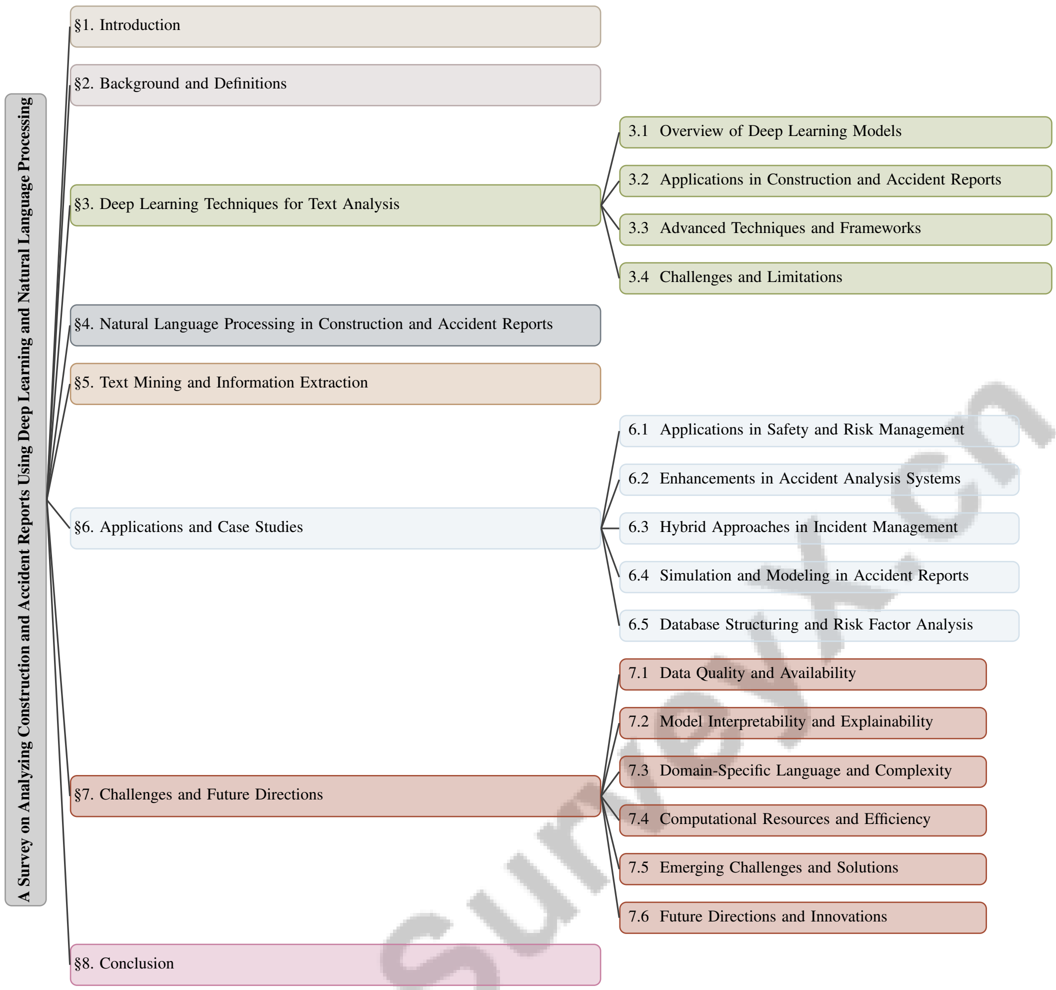


Figure 1: chapter structure

algorithms for software defect prediction underscore the broader applicability of these techniques in improving software reliability and performance [5].

In radiology, deep learning approaches to report generation reflect the rapid growth and potential of these methods in revolutionizing textual data analysis across sectors [6]. Notably, LSTM-based models that avoid manual features or external resources showcase innovative advancements in this field [7].

Furthermore, Terven et al. provide an overview of loss functions and performance metrics used in deep learning tasks, emphasizing their definitions, strengths, limitations, and applications, which are essential for optimizing model performance in textual data analysis [8].

Implementing sophisticated computational methods is vital for transforming the analysis of construction and accident reports. By leveraging advanced techniques like deep learning architectures and semantic search engines, organizations can extract valuable insights that enhance safety protocols and operational outcomes. These methods facilitate the identification of injury precursors, streamline document retrieval, and enable accurate severity classification of incidents, ultimately fostering a safer construction environment [9, 2, 10, 3, 11].

1.2 Role of Deep Learning and Natural Language Processing

Deep learning and natural language processing (NLP) play a pivotal role in analyzing construction and accident reports, providing advanced mechanisms for automated extraction and interpretation of information from extensive textual datasets. These technologies enhance the processing of complex

AI-generated, for reference only.

data, which is essential for improving safety and operational decision-making in the construction sector and related fields [12]. The core focus is on advancing deep learning techniques, particularly neural networks, which significantly enhance the handling of unstructured data prevalent in construction and accident reports [13].

Deploying deep learning architectures enables the identification of safety risks through automated understanding and analysis of incident data. For instance, structured databases of crane-related incidents, integrating qualitative and quantitative analyses, exemplify the role of deep learning and NLP in automating safety risk assessments [2]. This integration is supported by a systematic evaluation of loss functions and performance metrics, crucial for optimizing model performance in both computer vision and NLP tasks [8].

Deep learning has notably advanced NLP by facilitating automatic feature learning from large datasets, a critical capability for analyzing the unstructured data typical of construction and accident reports [1]. This advancement allows for sophisticated model development capable of extracting actionable insights, thereby aiding in proactive safety measures. The ability to learn hierarchical representations of language data enhances the understanding of intricate textual content, reinforcing the transformative impact of these technologies.

The integration of deep learning and NLP techniques in analyzing construction and accident reports significantly enhances the accuracy and efficiency of information extraction, while advancing safety management practices in the industry. State-of-the-art models such as Convolutional Neural Networks (CNN) and Hierarchical Attention Networks (HAN) have successfully identified predictive textual patterns linked to safety outcomes, improving our understanding of injury precursors. Additionally, large language models (LLMs) have been employed to develop sophisticated QA systems and semantic search engines, facilitating the retrieval of relevant information from historical incident reports and enhancing safety data comprehension. This innovative approach streamlines the analysis of unstructured text and provides actionable insights, contributing to effective safety management strategies across the construction sector [10, 3, 9, 11].

1.3 Structure of the Survey

This survey is meticulously structured to explore the application of deep learning and natural language processing (NLP) in analyzing construction and accident reports. It begins with an introduction that emphasizes the significance of advanced computational techniques, particularly deep learning and NLP, in automating the extraction and understanding of information from extensive textual datasets. The survey examines foundational elements and terminology relevant to the study, including discussions on construction and accident reports within the context of road traffic incidents, as well as methodologies in deep learning, NLP, text mining, and information extraction. It highlights the unique characteristics of insurance claim reports, produced under specific constraints that guide both writing and interpretation, thereby enhancing clarity and succinctness. The emerging role of neural networks in information retrieval is also noted, showcasing their potential to transform NLP applications [14, 15].

Subsequent sections focus on specific methodologies and applications. The section on deep learning techniques for text analysis examines various algorithms and models, their applications, and challenges encountered in processing construction and accident reports. This is followed by an in-depth analysis of NLP's role in extracting meaningful information from these reports, encompassing named entity recognition, event extraction, sentiment analysis, and topic modeling.

The survey provides a thorough examination of text mining and information extraction techniques, focusing on advanced methodologies for detecting patterns and trends within textual data. It highlights the integration of neural network approaches, particularly deep learning, which enhances relevant information identification by leveraging learned representations of text. Additionally, it discusses challenges such as the influence of stopwords and the need for improved topic quality measures, offering insights into the evolving landscape of NLP and its applications [14, 16, 15, 17, 18]. Real-world applications and case studies illustrate the successful implementation of these technologies in enhancing safety and risk management, improving accident analysis systems, and supporting hybrid approaches in incident management.

The paper concludes by identifying challenges in applying deep learning and NLP to construction and accident reports and suggesting future research directions. These include issues related to data

AI-generated, for reference only.

quality, model interpretability, domain-specific language complexity, and computational efficiency. By systematically addressing key topics in deep learning and NLP, this survey aims to deliver a comprehensive overview of their current applications and future potential within the construction and safety management sectors. It explores innovative methods for analyzing digitally recorded safety reports, highlights advancements in neural network approaches for information retrieval, and discusses the transformative impact of deep learning algorithms across various domains, particularly their implications for enhancing safety incident analysis and predictive modeling in construction environments [14, 9, 16].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions of Core Concepts

The analysis of construction and accident reports through advanced computational techniques requires a foundational grasp of key concepts. Construction reports are instrumental in tracking project progress, challenges, and safety incidents, serving as essential tools for project management and safety oversight [2]. They document near-miss incidents and other safety-related events, crucial for effective safety management.

Accident reports, especially in traffic contexts, elucidate incident causes and consequences, necessitating the use of natural language processing (NLP) to infer responsibilities and extract insights from linguistic cues [19]. Their effective analysis is critical for improving safety measures and ensuring regulatory compliance.

Deep learning, a subset of machine learning, employs multi-layer neural networks to autonomously learn data representations, proving effective for unstructured data like text [13]. Architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are pivotal in recognizing complex patterns within textual data [1].

NLP facilitates the interaction between computers and human language, enabling the automated processing and analysis of large textual datasets [20]. It is essential for extracting structured representations of events from unstructured text, aiding in the understanding and categorization of information in construction and accident reports.

Text mining derives meaningful information from text, employing techniques like topic modeling and sentiment analysis to identify patterns within extensive datasets [21]. Methods such as Latent Dirichlet Allocation (LDA) help discover abstract topics within document collections, enhancing information organization and retrieval [17].

Information extraction automates the retrieval of structured data from unstructured text, identifying entities, relationships, and events within reports. This process is vital for converting raw data into actionable insights that inform safety management and operational strategies [21]. Knowledge base construction (KBC) closely relates to information extraction, involving the derivation of relations from richly formatted data, crucial for developing comprehensive databases of construction and accident reports [22].

Integrating these core concepts enables the development of advanced systems for analyzing construction and accident reports, using multi-task learning to enhance the efficiency and accuracy of information extraction across related tasks [23]. Speculation cue detection and scope resolution further refine the management of uncertainty within textual data, providing a nuanced understanding of the information contained in these reports [24].

2.2 Challenges in Traditional Analysis Methods

Traditional methods for analyzing construction and accident reports encounter significant challenges in managing complex textual data. Manual feature extraction is labor-intensive and prone to errors, particularly with high-dimensional data, often failing to capture the semantic richness needed for effective analysis [6]. These methods struggle to integrate non-numerical, unstructured data, like textual incident descriptions, into cohesive analytical frameworks, a critical need in domains like traffic incident management where diverse data types must be synthesized [2].

The inadequacy of existing benchmarks, which often lack flexibility for different deep learning frameworks and datasets, further complicates these challenges, restricting the implementation and comparison of adversarial attacks [20]. Additionally, the noisy and uncertain nature of crowdsourced data complicates precise incident localization, increasing the risk of false alerts [19].

Traditional predictive analytics often inadequately predict accident occurrences and rely on historical data for preventive strategies. This limitation arises from dependence on labeled data from known domains, which does not generalize well to unseen data, leading to poor performance [1]. Ensuring the accuracy and consistency of generated reports, along with dependence on high-quality datasets, further compounds the limitations of traditional methods [13].

The substantial computational power needed to train models on high-dimensional data presents a significant challenge [8]. Designing models that effectively share knowledge across tasks risks overfitting due to increased complexity [6]. Existing benchmarks also exhibit generalizability and performance limitations, particularly when applied to varied datasets or tasks, restricting their applicability in diverse analytical contexts [25].

These limitations highlight the need for sophisticated computational techniques capable of addressing the complexities inherent in analyzing construction and accident reports. Advanced approaches, including state-of-the-art deep learning architectures for NLP and semantic search systems, can enhance the accuracy, efficiency, and scalability of analyses. Techniques like CNNs and Hierarchical Attention Networks (HAN) show promise in identifying injury precursors from raw construction accident reports, while tools such as HSEarch facilitate semantically enhanced searches within workplace accident documentation. The integration of Large Language Models into machine learning workflows has demonstrated substantial improvements in incident severity classification, underscoring the potential of these methodologies to yield deeper insights and more effective solutions across diverse datasets and contexts [15, 3, 9, 10].

2.3 Importance in Safety Management

The deployment of advanced computational techniques, including deep learning and natural language processing, is essential for enhancing safety management in the construction industry by facilitating efficient analysis of construction and accident reports. These reports are vital for identifying patterns and trends that inform proactive safety measures. The survey by Cayo et al. underscores the importance of time series analysis in examining trends within occupational accident reports, crucial for understanding the temporal dynamics of safety incidents and implementing timely interventions [26].

Learning from near misses is another critical aspect of safety management. Raviv et al. emphasize the need for systematic analysis of near-miss incidents to enhance safety practices and prevent future accidents [2]. By leveraging deep learning and NLP techniques, organizations can automate insight extraction from near-miss reports, thereby improving hazard identification and corrective action implementation.

These computational techniques also allow for the integration of diverse data sources, fostering a comprehensive understanding of safety risks. This holistic approach supports the development of effective safety management strategies that utilize data-driven insights from advanced technologies, such as semantic search engines and NLP. By employing tools like HSEarch for enhanced document retrieval and deep learning models to identify injury precursors from accident reports, the construction industry can systematically analyze safety incidents. This methodology aids in identifying risk factors, particularly in crane operations, and promotes a safer working environment by reducing accidents through informed decision-making and proactive risk management. The integration of these innovative solutions is crucial for advancing safety protocols and ensuring a more secure construction landscape [9, 10, 2, 3, 11].

3 Deep Learning Techniques for Text Analysis

Deep learning has significantly transformed text analysis by revolutionizing the processing and interpretation of textual data. Table 1 provides a detailed overview of the various deep learning methods and their applications in the analysis of construction and accident reports, highlighting the key features and techniques employed in these models. This section delves into foundational

AI-generated, for reference only.

Category	Feature	Method
Overview of Deep Learning Models	Data Integration Approaches	TCIAM[2]
Applications in Construction and Accident Reports	Predictive Analysis Multimodal Data Integration	DL-IPE[9], HMLA-SC[10], CS[11] FDR[22]
Advanced Techniques and Frameworks	Hierarchical Structures Dynamic Simulation	DRNN[12], HABERT[27] CS[25]
Challenges and Limitations	Computational Challenges	QA-LSTM[7]

Table 1: This table presents a comprehensive summary of deep learning methods applied to text analysis, specifically focusing on their applications in construction and accident reports. It categorizes the methods into four primary areas: an overview of deep learning models, applications in construction and accident reports, advanced techniques and frameworks, and challenges and limitations. Each category highlights specific features and the corresponding methods, providing insights into the current state of deep learning applications in these domains.

models that have driven these advancements, highlighting key architectures and their applications in construction and accident reports to enhance safety management practices.

3.1 Overview of Deep Learning Models

Deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers have fundamentally changed text analysis by offering advanced methods for understanding complex data. As illustrated in Figure 2, the primary categories of deep learning models used in text analysis highlight the distinct roles of these architectures: CNNs are adept at identifying local patterns, making them ideal for tasks like text classification and sentiment analysis by capturing spatial hierarchies [1]. Their application in analyzing construction-related injuries demonstrates their capacity to learn and visualize predictive patterns from safety reports [2].

RNNs, especially Long Short-Term Memory (LSTM) networks, excel in handling sequential data and modeling time-step dependencies, significantly enhancing tasks such as language modeling and sequence prediction [8]. Their integration into deep learning frameworks enables automatic feature extraction from large datasets, crucial for analyzing unstructured data typical of construction and accident reports [1].

Transformers have advanced text analysis by modeling long-range dependencies through self-attention mechanisms. This architecture underpins state-of-the-art large language models, excelling in semantic feature extraction and context understanding [8]. Their capability to perform complex tasks with minimal task-specific training sets new benchmarks in text analysis [19].

The Tower Crane Incident Analysis Method (TCIAM) exemplifies advanced deep learning for text analysis by combining qualitative content analysis with quantitative clustering techniques [2]. The evolution of these models highlights the need to integrate diverse algorithms to enhance information extraction and interpretation across various domains.

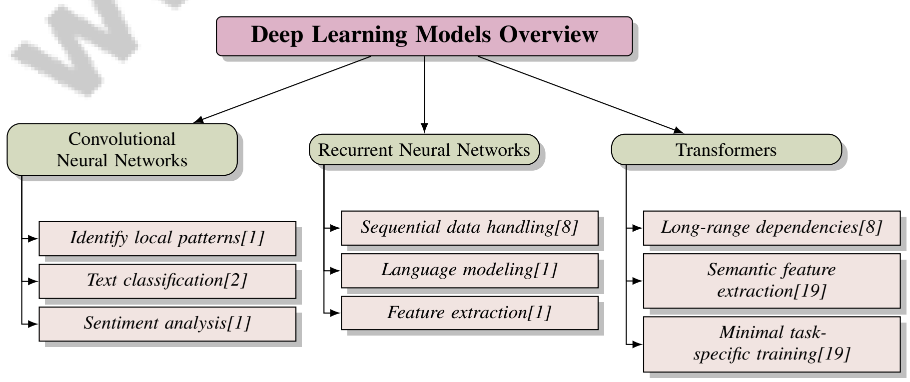


Figure 2: This figure illustrates the primary categories of deep learning models used in text analysis, highlighting Convolutional Neural Networks for local pattern identification, Recurrent Neural Networks for sequential data handling, and Transformers for modeling long-range dependencies.

AI-generated, for reference only.

3.2 Applications in Construction and Accident Reports

Method Name	Analytical Techniques	Model Integration	Application Scenarios
DL-IPE[9]	Cnn And Han	-	Construction Safety
HMLA-SC[10]	Deep Learning	Large Language Models	Traffic Incident Management
FDR[22]	Deep Learning	Multimodal Lstm	Construction Safety
CS[11]	Prompt Engineering	Regulatorybert	Traffic Incident Management

Table 2: Comparison of Deep Learning Methods and Their Applications in Construction Safety and Traffic Incident Management. This table summarizes various analytical techniques, model integrations, and specific application scenarios for different deep learning methods, highlighting their roles in enhancing safety and management practices.

Deep learning models offer substantial advancements in analyzing construction and accident reports, surpassing traditional methods in accuracy and efficiency. Automated extraction of precursors from construction injury reports aids in identifying potential safety risks and outcomes, using deep learning architectures to process raw text data and enhance incident prediction and prevention [9].

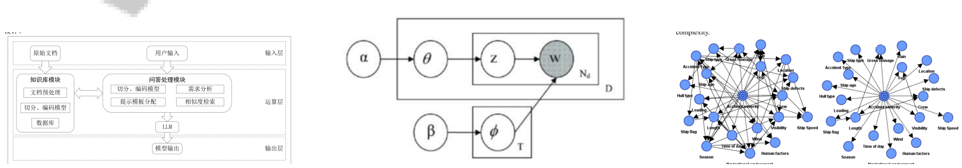
In traffic incident management, hybrid approaches combining baseline and language-based features from deep learning models have improved severity classification accuracy, demonstrating deep learning's potential to refine incident analysis precision for better emergency response and management [10].

Large language models (LLMs), such as GPT-4, highlight the transformative impact of deep learning, outperforming traditional machine learning algorithms in classification tasks and demonstrating superior capabilities in analyzing complex textual data from construction and accident reports [28].

Topic modeling techniques like Latent Dirichlet Allocation (LDA) effectively extract meaningful topics from large datasets, achieving higher coherence scores than Non-negative Matrix Factorization (NMF) and aiding information organization and retrieval from construction and accident reports [21].

Specialized models like RegulatoryBERT enhance domain-specific language processing, significantly improving performance in text classification and named entity recognition tasks, which is critical for compliance and improving safety protocols in construction projects [20]. Systems like Fonduer, utilizing deep learning for knowledge base construction from richly formatted documents, exemplify the integration of deep learning models in extracting relations and building comprehensive databases from construction and accident reports [22].

Integrating deep learning models in construction and accident report analysis enhances information extraction accuracy and efficiency, leading to improved safety management practices. Advanced NLP architectures allow researchers to identify predictive textual patterns associated with safety outcomes, facilitating a deeper understanding of injury precursors. Incorporating LLMs into machine learning workflows for traffic incident management refines incident severity classification by merging traditional data with language-based features, promoting proactive, data-driven strategies for global risk mitigation [9, 10]. Table 2 provides a comprehensive comparison of deep learning methods used in construction safety and traffic incident management, detailing the analytical techniques, model integrations, and application scenarios relevant to each method.



(a) A diagram illustrating a machine learning model architecture[11]

(b) A neural network model with multiple layers and nodes[21]

(c) Navigational Environment Complexity[29]

Figure 3: Examples of Applications in Construction and Accident Reports

As illustrated in Figure 3, deep learning techniques have become powerful tools for extracting and interpreting complex data in text analysis, especially in construction and accident reports. The figures demonstrate various applications of these techniques, showcasing their versatility and effectiveness.

AI-generated, for reference only.

The first figure presents a machine learning model architecture, emphasizing the flow from input to output through a structured knowledge base, essential for effective question-answering systems. The second figure highlights a multi-layered neural network model, facilitating intricate input processing through interconnected nodes. This model exemplifies how deep learning manages diverse data inputs to yield meaningful outputs. Lastly, the third figure illustrates navigational environment complexity, demonstrating how factors like ship attributes and external conditions interact to influence navigation outcomes. This network diagram emphasizes the capability of deep learning models to map and analyze multifaceted relationships, providing valuable insights into accident prevention and construction safety. Collectively, these examples underscore the transformative potential of deep learning in enhancing the understanding and management of text-based information in construction and accident reporting contexts [11, 21, 29].

3.3 Advanced Techniques and Frameworks

Method Name	Modeling Techniques	Hierarchical Structures	Multi-task Learning
DRNN[12]	Bidirectional Lstm	Hierarchical Representations	Relationship Learning
HABERT[27]	Hierarchical Attention	Hierarchical Attention	Recursive Regularization
CS[25]	-	-	-

Table 3: This table provides a comparative overview of various advanced deep learning methodologies applied to textual data analysis, particularly in the context of construction and accident reports. It highlights the specific modeling techniques, hierarchical structures, and multi-task learning strategies employed by each method, showcasing their contributions to improved accuracy and efficiency in processing complex datasets.

The evolution of deep learning frameworks has led to advanced techniques that enhance textual data analysis, particularly in construction and accident reports. Notable advancements include bidirectional Long Short-Term Memory (LSTM) cells and RNN transducers, which improve accuracy and efficiency for tasks like speech recognition [12]. These innovations are crucial for processing sequential data, allowing effective modeling of dependencies and contextual information within text.

Hierarchical attention mechanisms and recursive regularization represent another key innovation, enabling models to capture hierarchical structures within text and improve classification and extraction from complex datasets [27]. This approach benefits multi-label classification tasks, where understanding nuanced relationships between different labels is crucial.

Multi-task learning (MTL) frameworks exemplify advancements in deep learning for text analysis. MTL methods are categorized into regularization, relationship learning, feature propagation, optimization, and pre-training [23]. These strategies enhance model performance across multiple tasks, facilitating knowledge sharing and improving generalization capabilities, especially in domains like construction and accident report analysis, where data may be sparse or heterogeneous.

The mathematical foundations of deep learning are critical for developing these advanced frameworks, providing the theoretical underpinnings necessary for effective model design and implementation [13]. Understanding these foundations is essential for applying deep learning techniques across diverse AI applications, including textual data analysis in safety management and risk assessment.

Frameworks focusing on dynamic object interactions, such as CarSim, offer insights into real-time animation and simulation of dynamic systems [25]. While primarily used for dynamic simulations, the principles underlying such frameworks can inform the development of models requiring real-time processing and analysis of construction and accident reports, where dynamic interactions and temporal aspects are critical.

The integration of advanced deep learning techniques and frameworks has significantly improved the analysis and interpretation of complex textual data, enabling more accurate and efficient solutions for information extraction and decision-making in safety management and related fields. Table 3 presents a comparative analysis of advanced deep learning techniques, illustrating their application in enhancing textual data analysis through diverse modeling approaches and multi-task learning frameworks. This enhancement is largely attributed to the hierarchical representation capabilities of deep neural networks, successfully applied in various domains, including natural language processing and information retrieval. Recent studies demonstrate that architectures such as Convolutional Neural Networks (CNN) and Hierarchical Attention Networks (HAN) effectively identify predictive textual patterns in safety reports, facilitating a deeper understanding of injury precursors and contributing to enhanced safety outcomes. The growing computational power available for high-dimensional tensor calculations further boosts the efficiency and accuracy of these applications, underscoring the transformative impact of deep learning on traditional safety management practices [14, 9, 12, 1].

3.4 Challenges and Limitations

Applying deep learning to text analysis in construction and accident reports presents several challenges and limitations affecting its effectiveness. A primary challenge is the computational complexity inherent in deep learning models, requiring substantial resources for training and inference [7]. This complexity is compounded by the need for extensive model tuning, which can be resource-intensive and may not yield optimal performance if the model architecture does not align with specific task requirements [19].

Reliance on large, annotated datasets for training poses a significant limitation, especially in niche domains like construction safety management, where such datasets are often scarce. This scarcity hampers the development of models capable of generalizing effectively to new, unseen data, impacting their real-world applicability [7]. Additionally, the presence of noisy and unrelated information complicates the selection process, as correct insights may not directly correspond to lexical units in the input text [7].

The evaluation scope of many deep learning methods is often restricted to specific tasks, limiting their generalizability across different applications and datasets. This limitation is evident where assumptions about normal and anomalous data distributions lead to potential biases and inaccuracies in anomaly detection outcomes [19]. Variability in results necessitates averaging across multiple model runs to ensure statistical significance, which can be resource-intensive and time-consuming [24].

Regulatory constraints on sharing annotated datasets and models further complicate these challenges, potentially affecting the performance and development of deep learning models for text analysis. The high variance or numerical instability of mutual information estimation methods in high-dimensional settings poses additional challenges, impacting the reliability of model outputs [7].

Despite these challenges, recent advancements in integrating semantic and syntactic features demonstrate significant potential for improving extraction performance. A novel framework utilizing BERT for semantic features and Graph Convolutional Networks for syntactic features has outperformed traditional methods in Chinese event extraction, often hindered by the language's unique characteristics. This suggests that leveraging the complementary strengths of different feature types could pave the way for overcoming existing limitations in various domains, including specialized texts like insurance claim reports and biomedical literature, where precise event representation is critical [24, 30, 15, 17]. Addressing these challenges is essential for enhancing the accuracy, interpretability, and scalability of deep learning models in analyzing construction and accident reports, ultimately improving safety management practices.

4 Natural Language Processing in Construction and Accident Reports

Natural Language Processing (NLP) is integral in analyzing construction and accident reports, offering significant enhancements to safety management. Essential NLP techniques, such as Named Entity Recognition (NER) and event extraction, are crucial for identifying key entities and events within textual data. This section explores these techniques' importance in construction and accident reports and their implications for safety management.

Figure 4 illustrates the hierarchical structure of NLP techniques applied in these reports, highlighting the primary categories of Named Entity Recognition and Event Extraction, Sentiment Analysis and Contextual Understanding, Topic Modeling and Information Retrieval, and Large Language Models and Safety Engineering. Each of these categories is further divided into subcategories and specific techniques or applications, demonstrating the interconnectedness and comprehensive approach to enhancing safety management in the construction industry.

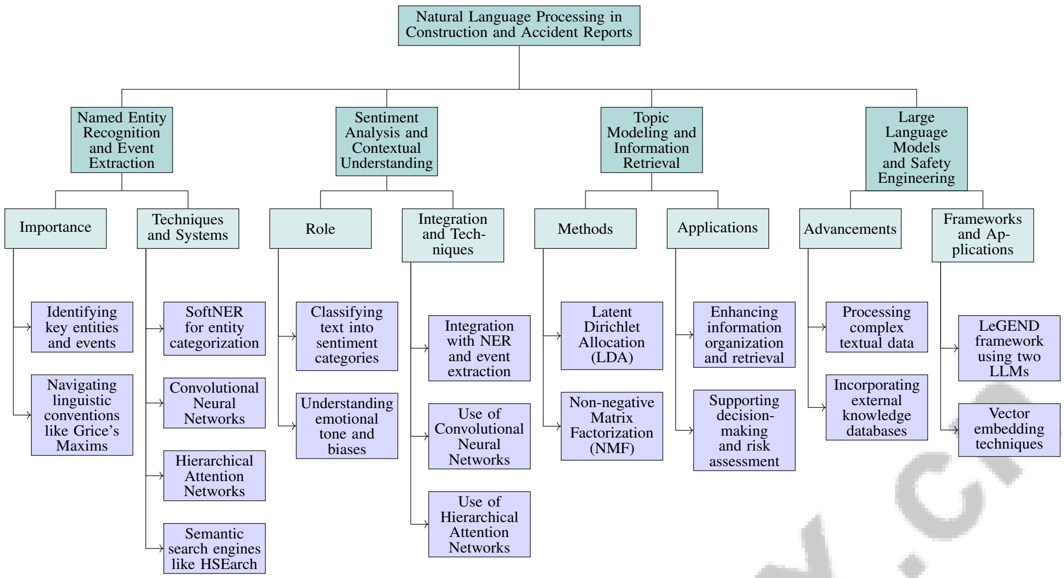


Figure 4: This figure illustrates the hierarchical structure of Natural Language Processing (NLP) techniques applied in construction and accident reports, highlighting the primary categories of Named Entity Recognition and Event Extraction, Sentiment Analysis and Contextual Understanding, Topic Modeling and Information Retrieval, and Large Language Models and Safety Engineering. Each category is further divided into subcategories and specific techniques or applications, demonstrating the interconnectedness and comprehensive approach to enhancing safety management in the construction industry.

4.1 Named Entity Recognition and Event Extraction

NER and event extraction are pivotal in analyzing construction and accident reports, enabling the identification of key entities and events. Effective NLP systems must navigate linguistic conventions, such as Grice's Maxims, to interpret reports accurately [15]. Advanced systems like SoftNER, which integrate NER with relation extraction, enhance entity categorization within complex texts, crucial for constructing knowledge graphs that organize and retrieve report information [31]. These advancements facilitate accurate entity recognition and critical event extraction, providing context and insights into safety incidents. Techniques such as Convolutional Neural Networks and Hierarchical Attention Networks, alongside semantic search engines like HSEarch, streamline information retrieval and enhance workplace safety measures [3, 9]. Leveraging these methodologies allows for automated insights extraction, supporting proactive safety management.

4.2 Sentiment Analysis and Contextual Understanding

Sentiment analysis is vital for understanding the nuances of construction and accident reports by classifying text into sentiment categories—positive, negative, or neutral. This process provides insights into the emotional tone and potential biases, crucial for evaluating attitudes toward safety incidents. Advanced NLP techniques, including Convolutional Neural Networks and Hierarchical Attention Networks, enhance sentiment extraction from complex data, improving the understanding of expressions and contextual cues [9, 2]. Integrating sentiment analysis with NER and event extraction supports comprehensive report analysis, uncovering deeper issues and trends [32, 17]. This integration refines safety management by highlighting improvement areas, fostering continuous improvement and proactive risk management [9, 2].

Implementing sentiment analysis underscores the need for advanced computational techniques to extract insights from unstructured data. State-of-the-art models effectively identify injury precursors from accident reports, while semantic search systems like HSEarch enhance information retrieval efficiency, improving safety outcomes [3, 9]. As NLP evolves, integrating sentiment analysis with other frameworks will enhance the ability to derive actionable intelligence, contributing to safer construction practices.

4.3 Topic Modeling and Information Retrieval

Topic modeling techniques are crucial for extracting and organizing information from construction and accident reports, identifying latent topics within extensive datasets. Latent Dirichlet Allocation (LDA) is a prevalent method for generating semantically coherent topics, demonstrating superior coherence scores in aviation incident reports compared to Non-negative Matrix Factorization (NMF) [21]. Integrating topic modeling with information retrieval systems enhances the organization and retrieval of relevant information, supporting informed decision-making and risk assessment in safety management [17].

Advanced topic modeling techniques that incorporate prior knowledge refine topic extraction from complex datasets, improving relevance and specificity for construction and safety reports [17]. Tailoring topic models to specific contexts enhances precision, enabling more targeted analyses. The application of these techniques highlights the importance of leveraging computational methods to manage large volumes of textual data. As these techniques advance, their integration with emerging NLP and deep learning frameworks will significantly enhance actionable insights extraction from unstructured data, improving safety management and operational efficiency in construction [14, 9, 3, 11, 20].

4.4 Large Language Models and Safety Engineering

Large language models (LLMs) represent a significant advancement in the automated analysis and interpretation of accident reports in safety engineering within the construction industry. These models, known for their ability to process complex textual data, improve accuracy and efficiency over traditional methods. Incorporating external knowledge databases into LLMs enhances their capacity to address safety-related challenges [11].

The LeGEND framework exemplifies an innovative approach using two LLMs for scenario transformation: LLM1 extracts interactive patterns, while LLM2 converts these into formal scenarios [33]. This dual-phase approach highlights LLMs' potential to transform raw data into structured insights for safety management. LLMs enhance critical insights extraction, identifying potential risks and facilitating proactive measures. Advancements like integrating external knowledge databases and vector embedding techniques improve LLMs' reliability in analyzing historical reports, allowing effective summarization and recommendations [10, 3, 11, 28, 34]. This capability is crucial in construction, where timely safety issue resolution is vital.

The impact of LLMs on safety engineering is profound, offering new avenues for enhancing accident report analysis. As these models evolve, their integration with external knowledge sources and analytical frameworks will enhance safety management effectiveness, improving incident report analysis, injury precursor identification, and risk assessments, contributing to safer construction practices [9, 10, 2, 3, 11].

5 Text Mining and Information Extraction

5.1 Datasets and Preprocessing

The effectiveness of text mining and information extraction in analyzing construction and accident reports is contingent upon the quality and diversity of the datasets employed. Structured incident reports in construction are systematically compiled and analyzed to inform safety strategies, providing comprehensive datasets that mirror real-world scenarios [2]. Similarly, traffic incident datasets from multiple US states serve as a foundation for understanding traffic patterns and developing predictive models [19]. Preprocessing techniques such as tokenization, stop-word removal, and stemming are essential for enhancing data quality, enabling the extraction of relevant information for tasks like 3D scene simulation of accidents [25]. Robust preprocessing ensures datasets are well-structured, allowing advanced NLP techniques, including CNN and HAN, to effectively analyze complex textual data, identify predictive injury precursors, and facilitate semantic searches [9, 2, 10, 3, 15]. Table 4 provides a detailed overview of the representative benchmarks employed in text mining and information extraction, illustrating their applicability across different domains and tasks.

Benchmark	Size	Domain	Task Format	Metric
GNN-RoadSafety[19]	9,000,000	Traffic Accident Analysis	Edge-level Regression/classification	MAE, AUROC
ZSL-LLM[28]	4,000	Text Classification	Zero-shot Classification	Accuracy, F1 Score
ChatGPT-Text-Distinction[18]	10,000	Text Classification	Binary Classification	Accuracy, MCC
AUTONLU[35]	1,000,000	Image Editing	Intent Detection	F1-score
COLIEE[36]	725	Legal Text Processing	Legal Information Retrieval	F1-score
ChatSOS[34]	117	Safety Engineering	Question Answering	Accuracy, F1-score
SpecNeg[24]	13,541	Biomedical Text Analysis	Speculation Detection	F1-score

Table 4: This table presents a comprehensive overview of various benchmarks utilized in diverse domains, highlighting their respective sizes, task formats, and evaluation metrics. The benchmarks span multiple applications including traffic accident analysis, text classification, and legal information retrieval, providing a broad spectrum of data for evaluating model performance. Each benchmark is associated with specific metrics such as MAE, AUROC, and F1-score, which are critical for assessing the effectiveness of the respective tasks.

5.2 Methods for Topic and Pattern Identification

Topic and pattern identification within textual data is crucial for analyzing construction and accident reports. Utilizing methods like LDA, trained on Document-Term Frequency Matrices, uncovers hidden thematic structures within datasets [21]. This is complemented by integrating diverse data sources, such as traffic records and weather data, to reveal patterns that inform predictive models [19]. Evaluation metrics like accuracy and the Matthews correlation coefficient ensure reliability in topic and pattern identification [18]. These methodologies enable the extraction of critical insights, enhancing incident severity classification and systematic exploration of safety documents [3, 11, 10].

5.3 Entity and Relation Extraction

Entity and relation extraction are pivotal for converting unstructured text into structured data, enhancing the understanding of safety incidents. Techniques like CNN and HAN automatically identify injury precursors from accident reports, while semantic search systems like HSEarch facilitate efficient document retrieval [15, 3, 9]. Methods such as SoftNER automate entity and relation extraction, constructing knowledge graphs that support systematic analysis [31]. The LeGEND framework further integrates these techniques with scenario generation, transforming natural language into formal logical scenarios, crucial for proactive safety measures [33].

5.4 Temporal and Semantic Analysis

Temporal and semantic analysis are vital for understanding patterns and trends in construction and accident reports. Temporal analysis, using methods like time series analysis, identifies evolving risk patterns [10, 26, 21, 2]. Semantic analysis employs NLP techniques to interpret linguistic and contextual information, enhancing the understanding of complex scenarios [30, 15]. Integrating these analyses improves correlation identification between incidents and causes, supporting effective safety management strategies. This comprehensive approach enables organizations to anticipate safety issues and implement timely interventions, improving operational outcomes by addressing hazardous risk factors [9, 2].

6 Applications and Case Studies

6.1 Applications in Safety and Risk Management

The integration of deep learning and natural language processing (NLP) has transformed safety and risk management by enhancing the analysis and interpretation of complex datasets. These technologies support informed decision-making through tools like LLM-based knowledge QA systems, which utilize external databases for precise information retrieval, and machine learning workflows that merge traditional data with language model-generated features to classify traffic incident severity. In construction, deep learning applied to accident reports facilitates the automatic identification of injury precursors, thereby improving predictive capabilities and understanding of safety incidents

[10, 15, 9, 11]. Systems like ChatSOS exemplify these advancements by providing contextually relevant information that enhances decision-making processes in safety engineering.

The Tower Crane Incident Analysis Method (TCIAM) demonstrates the practical application of deep learning and NLP in safety management by analyzing accidents and near misses to guide targeted interventions, thereby reducing future incidents [2]. Cloud-based natural language understanding (NLU) systems, such as AutoNLU and ChatSOS, further illustrate the transformative potential of advanced technologies in safety management by automating incident reporting and improving decision-making through efficient information retrieval [15, 35, 31, 11].

In traffic incident management, integrating features from large language models (LLMs) with traditional accident report data has significantly increased classification accuracy of incident severity. This integration highlights the benefits of combining advanced NLP techniques with traditional data sources, enhancing safety analyses in specialized domains [15, 32]. Hierarchical classification methods have also improved fine-level classification accuracy, particularly for rare events, by merging hierarchical classification techniques with advanced NLP methods [32, 9, 27, 11].

The incorporation of advanced technologies, especially LLMs and machine learning algorithms, has significantly enhanced the capacity to analyze intricate datasets, accurately identify risk factors, and implement effective preventive measures. Recent studies underscore the successful application of LLMs in traffic incident management, improving incident severity classification by merging traditional data with language-based features. Moreover, innovations such as LLM-based QA systems in safety engineering have demonstrated their capability to process historical incident reports, providing actionable insights through improved comprehension and response accuracy [9, 2, 10, 11, 34].

6.2 Enhancements in Accident Analysis Systems

The integration of deep learning and NLP into accident analysis systems has significantly advanced automation and accuracy in incident handling. SoftNER's incorporation into Microsoft's incident management platform exemplifies this progress by efficiently extracting and categorizing entities from complex reports [31]. Deep learning's effectiveness in high-dimensional function estimation further underscores its transformative impact on accident analysis systems, aiding in precise estimation and prediction of incident outcomes [37].

In autonomous vehicles (AVs), deep learning and NLP have been crucial in analyzing accident reports to identify common patterns and areas for improvement. Research indicates that AVs are more prone to rear-end collisions than conventional vehicles, highlighting specific areas for technological and regulatory enhancements [4]. Systems like CarSim, which convert accident reports into 3D simulations, demonstrate the potential for further improvements in information extraction and modeling, despite current limitations in accuracy and comprehensiveness [25].

These advancements in accident analysis systems, driven by deep learning and NLP technologies, enhance automation, accuracy, and comprehensiveness of safety analyses. By integrating LLMs with traditional machine learning techniques, researchers have achieved significant improvements in incident severity assessment accuracy, while vector databases enrich LLM analytical capabilities, enabling timely, reliable insights from historical data. This comprehensive approach optimizes global incident management processes and opens avenues for future automation and intelligent systems in transportation safety [34, 11, 10].

6.3 Hybrid Approaches in Incident Management

Hybrid approaches in incident management combine diverse technologies and analytical frameworks, such as LLMs and traditional machine learning algorithms, to enhance safety systems' efficiency and effectiveness. Recent research illustrates how LLMs improve incident severity classification by integrating features from unstructured accident reports with conventional data, leading to enhanced accuracy across various models. Systems like ChatSOS leverage LLMs for real-time knowledge retrieval and historical report analysis, enriching decision-making in safety engineering [11, 10]. These approaches often involve the synergistic application of deep learning, NLP, and other computational techniques to address complex safety incident management challenges.

The combination of deep learning with NLP techniques automates information extraction and interpretation from incident reports, enabling hybrid systems to process extensive textual data and accurately identify key entities and events [31]. Moreover, hybrid approaches frequently incorporate additional data sources, such as sensor data, to provide a comprehensive understanding of incident scenarios. This integration enables predictive model development that anticipates potential safety risks and informs proactive mitigation measures [19].

Multi-task learning frameworks further exemplify the potential of hybrid approaches by enhancing model generalization capabilities across diverse datasets and scenarios [23]. This adaptability is vital for managing the dynamic and unpredictable nature of safety incidents. By integrating advanced technologies such as LLMs and vector databases, these innovative methods significantly improve incident analysis and response capabilities, contributing to safer and more efficient operational environments [10, 3, 11].

6.4 Simulation and Modeling in Accident Reports

Simulation and modeling play a crucial role in accident report analysis by providing dynamic visual representations that enhance understanding of accident dynamics and inform safety management strategies. Technologies like CarSim transform accident reports into 3D simulations, enabling visualization of temporal and spatial aspects of accidents [25]. Integrating simulation with advanced analytical techniques, such as NLP and deep learning, significantly improves information extraction and interpretation from accident reports, facilitating a nuanced understanding of safety incidents [15, 9].

These technologies enable the generation of accurate, comprehensive simulations that reflect real-world incident complexities, aiding accident investigation and supporting preventive measures by highlighting areas for safety protocol improvement. Moreover, the application of simulation and modeling extends beyond individual accident analysis to inform broader safety management strategies by aggregating data from multiple reports to identify recurring patterns and trends [4, 11, 21].

By integrating LLMs and machine learning techniques, recent studies have improved incident severity classification accuracy and extracted valuable insights from unstructured data. This not only deepens understanding of accident dynamics but also contributes to developing safer, more efficient global transportation systems [10].

6.5 Database Structuring and Risk Factor Analysis

Database structuring and risk factor analysis are critical components in construction and accident report analysis, serving as foundational elements for effective safety management and decision-making. Proper database structuring enables systematic information organization and retrieval, facilitating efficient analysis of large textual data volumes. This capability is essential for identifying significant patterns, trends, and anomalies that inform targeted safety strategies and interventions [3, 9].

The Rocas framework exemplifies effective database structuring by identifying triggering entities and misconfigurations in Automated Driving Systems (ADS) accidents through robust root cause analysis [38]. Risk factor analysis involves identifying and assessing variables that may increase accident likelihood, essential for prioritizing safety interventions and allocating resources. Advanced topic modeling techniques extract meaningful insights from aviation accident reports, identifying underlying causes that inform targeted safety improvements [17, 21].

The integration of advanced computational techniques, such as machine learning and NLP, enhances database structuring and risk factor analysis capabilities by facilitating automated extraction and interpretation of data from unstructured text. This holistic approach supports developing more accurate and predictive safety models, ultimately contributing to accident reduction and enhanced safety management practices [32, 18].

By establishing a comprehensive framework for data organization and risk assessment, these processes empower organizations to extract actionable insights that guide proactive safety strategy development, enhancing operational outcomes. This is particularly evident in high-hazard industries like construction, where systematic near-miss and accident report analysis has identified critical risk factors, such as technical failures in tower cranes. Furthermore, integrating advanced machine learning techniques, including LLMs, into incident management workflows has led to significant improvements in incident severity classification, providing organizations with precise data to inform safety measures and operational practices [10, 2].

7 Challenges and Future Directions

7.1 Data Quality and Availability

Data quality and availability are critical in the analysis of construction and accident reports, particularly when employing advanced computational techniques like deep learning and NLP. The necessity for extensive, high-quality datasets to train models that generalize effectively across scenarios remains a primary challenge [1]. Variability in document formats complicates accurate information extraction, necessitating advanced preprocessing methods. In low-resource settings, limited labeled data leads to overfitting and reduced model generalizability, resulting in biased outputs that do not accurately reflect real-world conditions [1]. Data biases and model interpretability are also significant concerns, as inadequate attention to these issues undermines output reliability. Crowdsourced data quality varies considerably, especially in regions with low engagement, affecting incident analysis reliability [1]. Moreover, evaluation benchmarks often lack comprehensive coverage, requiring refinement [25].

Many studies overlook critical accident variables, leading to incomplete insights, particularly in the architecture, engineering, and construction (AEC) sector, where labeled training data is scarce [33]. Addressing these challenges involves developing training methods that generalize across diverse domains and data types. Improving data quality and access to well-structured datasets can significantly enhance predictive model accuracy and reliability, crucial for refining safety management practices. Techniques such as Convolutional Neural Networks and Hierarchical Attention Networks facilitate automatic extraction of injury precursors from reports. Semantic search engines like HSEarch enhance retrieval of relevant information, aiding systematic accident report reviews. Structured analyses of near-miss incidents using clustering methodologies can identify critical risk factors, leading to more effective safety interventions [3, 9, 2].

7.2 Model Interpretability and Explainability

Deep learning model interpretability and explainability present significant challenges in construction and accident report analyses. The complexity of these architectures often results in "black box" models, obscuring prediction rationales [1]. This lack of transparency hinders adoption in practical scenarios, where stakeholders require clear explanations to ensure trust and accountability [39]. User-centric explanation methods aim to bridge this gap, facilitating AI integration into decision-making [40]. However, their effectiveness depends on data and algorithm quality, which can limit real-world applicability [3]. Reliance on external knowledge databases for model explanations complicates the issue, as database completeness and accuracy directly impact model performance and explanation reliability [11].

Integrating symbolic reasoning with neural networks in neurosymbolic AI offers promising avenues for enhancing interpretability. However, applying these approaches to NLP tasks within construction and accident reporting is challenging due to language complexity [41]. Future research should focus on improving training algorithms and exploring new model architectures that inherently support interpretability and generalization [37]. Practical examples and exercises should be developed to enhance deep learning education, equipping practitioners to create interpretable models [13]. Addressing these challenges is crucial for advancing deep learning applications in safety management and ensuring responsible AI use in critical domains.

7.3 Domain-Specific Language and Complexity

Analyzing construction and accident reports is complicated by domain-specific language complexity, which challenges effective communication and understanding. In Automated Driving Systems (ADS), domain-specific language intricacies hinder root cause analysis, necessitating precise interpretation of technical details [38]. This complexity requires careful selection and tuning of priors in modeling, introducing additional complexity and potential bias [17]. User-centric explanation processes must consider diverse stakeholder backgrounds, complicating accessible and meaningful explanation creation [40]. This is evident in the construction industry, where NLP application is hindered by specialized report language [42].

Dataset quality and model biases affect result interpretability, highlighting the need for robust data management practices [21]. De-identification in clinical settings illustrates adapting to evolving domain-specific language, requiring regular updates [32]. In languages like Chinese, domain-specific language complexity impacts event extraction performance, complicating NLP applications [30]. Future research should explore hybrid models combining shallow and deep methods, alongside techniques to manage high-dimensional, noisy, and imbalanced datasets [43]. Enhancing memory modules and jury mechanisms could improve performance in ambiguous scenarios, providing robust solutions for handling domain-specific language complexity [44]. Advanced methodologies and innovative NLP solutions can enhance interpretability and applicability, improving safety management practices across domains. For instance, NLP effectively analyzes car accident reports, where linguistic constraints aid in producing concise texts for clearer interpretation. An LLM-based QA system for safety engineering illustrates how integrating external knowledge databases enriches model responses and improves safety assessment accuracy. Research into automatically learning injury precursors from construction accident reports highlights deep learning architectures' potential to identify predictive patterns, contributing to a deeper understanding of safety incidents and enhancing proactive safety measures [15, 9, 11].

7.4 Computational Resources and Efficiency

Implementing deep learning and NLP in construction and accident report analysis requires substantial computational resources. These resources are vital for training complex models capable of processing large textual data volumes and extracting meaningful insights. Integrating Knowledge Graphs with Deep Learning models highlights computational demands associated with hybrid systems enhancing information extraction and interpretation [45]. Efficiency challenges related to these computational requirements are significant, particularly regarding processing speed and resource allocation. High data dimensionality necessitates robust computational infrastructure for timely and accurate analysis. Model optimization and hyperparameter tuning further intensify resource demands, especially in low-resource settings where effective utilization of limited labeled data is essential for improving performance. Strategies like early stopping, weight initialization, and innovative training-validation splits can mitigate resource demands, enhancing outcomes without additional data or extensive redesigns [14, 46, 1, 16].

Scalability of deep learning models poses additional challenges, as their complexity often leads to increased computational costs. Efficient resource management is essential for optimizing the balance between model accuracy and computational efficiency, enabling effective deployment in real-world applications while adhering to constraints. This involves leveraging limited labeled data through techniques like early stopping and weight initialization, significantly enhancing performance without additional resources or complex redesigns. Prioritizing effective training strategies maximizes available data utility, ensuring models are accurate and resource-efficient, facilitating broader adoption across fields, including NLP and computer vision [14, 46, 1, 16, 23].

Addressing challenges of deploying advanced analytical techniques in resource-constrained environments requires developing more efficient algorithms and computational frameworks. These advancements will enhance deep learning models' performance, recognized for their transformative impact across domains, including NLP and computer vision, while facilitating optimal utilization of limited labeled data [47, 46, 16]. Optimizing computational processes and leveraging innovative approaches like parallel processing and distributed computing can enhance deep learning and NLP application efficiency in construction and accident report analysis, contributing to improved safety management and operational outcomes.

7.5 Emerging Challenges and Solutions

The deep learning and NLP field in construction and accident report analysis is continually evolving, presenting emerging challenges requiring innovative solutions. One challenge is the need for better-defined reasoning theories and robust benchmarks to evaluate Neuro-Symbolic AI's effectiveness in real-world applications. Integrating symbolic reasoning with neural networks offers promising avenues for enhancing interpretability and decision-making capabilities, yet it requires comprehensive evaluation frameworks for practical applicability [41].

Another pressing issue is deep learning resource accessibility, which remains a barrier for many researchers and practitioners. Improving resource availability and comprehensibility is crucial for fostering innovation and collaboration. Future research should focus on developing educational materials and tools simplifying deep learning technique understanding and application, broadening participation and accelerating advancements [13]. The challenge of overfitting persists, particularly as models grow more complex and datasets expand. Hybrid models combining deep learning with evolutionary algorithms represent a promising solution, optimizing performance while mitigating overfitting risks [13]. Addressing these challenges requires a multifaceted approach incorporating theoretical advancements, practical solutions, and collaborative efforts across the research community. By focusing on developing advanced methodologies for analyzing construction accident reports and near-miss incidents, and leveraging semantic search technologies, the field can significantly enhance safety incident understanding. This evolution will lead to creating more effective analytical tools identifying injury precursors through state-of-the-art NLP techniques and facilitating structured incident data investigations, resulting in improved safety management practices and operational efficiency in the construction industry [3, 9, 2].

7.6 Future Directions and Innovations

The future of analyzing construction and accident reports using deep learning and NLP is poised for significant advancements, with several promising research directions and innovations emerging. A key focus is refining Knowledge Graphs integrated with deep learning models to enhance information retrieval accuracy and efficiency from complex datasets. This includes improving query conversion mechanisms and exploring additional features integration, like real-time traffic data and advanced modeling techniques, to enhance accident prediction accuracy [19].

Fine-tuning large language models (LLMs) for specific accident reporting contexts and exploring their applicability to other domains is another promising direction. Reducing computational complexity in these models will be crucial for broader adoption across various sectors, including traffic incident management. Future research should investigate applying functional scenarios from additional sources and refining LLM capabilities in transforming natural language descriptions into formal scenarios [33].

Enhancing scalability and extraction quality of systems like Fonduer is a critical research direction, focusing on exploring additional modalities and techniques to improve knowledge base construction robustness [22]. Similarly, expanding datasets with more domain-specific texts and exploring additional transfer learning techniques will strengthen models like RegulatoryBERT in safety engineering applications [20].

Further research should aim to enhance training datasets and explore Deep Recurrent Neural Networks (DRNNs) applications in sentiment analysis and other NLP tasks [12]. This involves developing structured methodologies for model building and assessing data quality to improve machine learning applications' usability and trustworthiness in business contexts [48]. Exploring transfer learning techniques and methods to automatically refine generated reports for clinical use will enhance model robustness [6].

Refining models to improve robustness against data noise and exploring additional data sources to enhance detection capabilities are critical areas for future research [49]. Enhancing AV detection systems and examining human factors' impact on AV safety are crucial for developing more robust safety protocols [4]. Future research should focus on developing more flexible multi-task learning (MTL) models that can adapt to changing task requirements, exploring emerging techniques like zero-shot learning and task-agnostic training [23].

Expanding text mining algorithms and datasets to include more diverse workplace accident information sources will also be a focus [3]. Improving benchmark capabilities to encompass diverse datasets and attack methods will enhance integration with emerging deep learning technologies [47]. Pursuing these research directions can develop more sophisticated, reliable, and versatile systems for analyzing construction and accident reports, ultimately enhancing safety management practices and operational efficiency.

Future research should also prioritize developing more efficient training techniques and enhancing deep learning models' interpretability and adaptability [1]. This involves exploring new model architectures and training algorithms that generalize across diverse domains and data types, ensuring models are robust and adaptable to a wide range of applications.

8 Conclusion

Deep learning and natural language processing (NLP) have significantly transformed the analysis of construction and accident reports, enhancing safety management practices by automating information extraction and improving analytical precision. These technologies enable proactive safety measures by integrating external knowledge into large language models, thereby addressing complex safety engineering challenges. Despite these advancements, challenges remain, particularly in fostering user trust and understanding of AI systems, which are crucial for decision-making in construction safety management. Enhancing user-centric explanation processes and improving neural information retrieval systems are essential steps toward ensuring the effectiveness and reliability of AI solutions in practical applications.

Innovative methods, such as the RCCN approach, have demonstrated their efficacy in text classification tasks, proving their applicability in real-world scenarios. Additionally, the integration of crowdsourced data with structured optimization techniques, exemplified by CROME, highlights the importance of balancing accuracy and localization in emergency response efforts. The transformative impact of deep learning and NLP within the construction industry underscores the need for ongoing research and innovation in safety management. By addressing current challenges and exploring new methodologies, the field can advance towards more sophisticated, reliable, and efficient solutions for the analysis of construction and accident reports.

References

[1] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.

[2] Gabriel Raviv, Barak Fishbain, and Aviad Shapira. Analyzing risk factors in crane-related near-miss and accident reports. *Safety science*, 91:192–205, 2017.

[3] Emrah Inan, Paul Thompson, Tim Yates, and Sophia Ananiadou. Hsearch: semantic search system for workplace accident reports, 2021.

[4] Francesca M Favarò, Nazanin Nader, Sky O Eurich, Michelle Tripp, and Naresh Varadaraju. Examining accident reports involving autonomous vehicles in california. *PLoS one*, 12(9):e0184952, 2017.

[5] Görkem Giray, Kwabena Ebo Bennin, Ömer Köksal, Önder Babur, and Bedir Tekinerdogan. On the use of deep learning in software defect prediction, 2022.

[6] Chang Liu, Yuanhe Tian, and Yan Song. A systematic review of deep learning-based research on radiology report generation, 2024.

[7] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Lstm-based deep learning models for non-factoid answer selection, 2016.

[8] Juan Terven, Diana M. Cordova-Esparza, Alfonso Ramirez-Pedraza, Edgar A. Chavez-Urbiola, and Julio A. Romero-Gonzalez. Loss functions and metrics in deep learning, 2024.

[9] Henrietta Baker, Matthew R. Hallowell, and Antoine J. P. Tixier. Automatically learning construction injury precursors from text, 2020.

[10] Artur Grigorev, Khaled Saleh, Yuming Ou, and Adriana-Simona Mihaita. Enhancing traffic incident management with large language models: A hybrid machine learning approach for severity classification, 2024.

[11] Haiyang Tang, Zhenyi Liu, Dongping Chen, and Qingzhao Chu. Chatsos: Llm-based knowledge qa system for safety engineering, 2023.

[12] Sarvesh Patil. Deep learning based natural language processing for end to end speech translation, 2018.

[13] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic programming and evolvable machines*, 19(1):305–307, 2018.

[14] Ye Zhang, Md Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. Neural information retrieval: A literature review, 2017.

[15] Dominique Estival and Francoise Gayral. An nlp approach to a specific type of texts: Car accident reports, 1995.

[16] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[17] Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling, 2017.

[18] Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. Distinguishing human generated text from chatgpt generated text using machine learning, 2023.

[19] Abhinav Nippani, Dongyue Li, Haotian Ju, Haris N. Koutsopoulos, and Hongyang R. Zhang. Graph neural networks for road safety modeling: Datasets and evaluations for accident analysis, 2024.

[20] Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. Pretrained domain-specific language model for general information retrieval tasks in the aec domain, 2022.

[21] Aziida Nanyonga, Hassan Wasswa, and Graham Wild. Topic modeling analysis of aviation accident reports: A comparative study between lda and nmf models, 2024.

[22] Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. Fonduer: Knowledge base construction from richly formatted data, 2018.

[23] Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Namboodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras, 2024.

[24] Benita Kathleen Britto and Aditya Khandelwal. Resolving the scope of speculation and negation using transformer-based architectures, 2020.

[25] Sylvain Dupuy, Arjan Egges, Vincent Legendre, and Pierre Nugues. Generating a 3d simulation of a car accident from a written description in natural language: the carsim system, 2001.

[26] Tony Gabriel-Fernandez Cayo and Fred Torres-Cruz. Time series analysis applied to notifications of work accidents, 2023.

[27] Xinyu Zhao, Hao Yan, and Yongming Liu. Hierarchical multi-label classification for fine-level event extraction from aviation accident reports, 2024.

[28] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. Large language models are zero-shot text classifiers, 2023.

[29] Likun Wang and Zaili Yang. Bayesian network modelling and analysis of accident severity in waterborne transportation: A case study in china. *Reliability Engineering & System Safety*, 180:277–289, 2018.

[30] Jiangwei Liu, Jingshu Zhang, Xiaohong Huang, and Liangyu Min. Syntactic-gcn bert based chinese event extraction, 2021.

[31] Manish Shetty, Chetan Bansal, Sumit Kumar, Nikitha Rao, and Nachiappan Nagappan. Softner: Mining knowledge graphs from cloud incidents, 2021.

[32] Xavier Tannier, Perceval Wajsbürt, Alice Calliger, Basile Dura, Alexandre Mouchet, Martin Hilka, and Romain Bey. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse, 2023.

[33] Shuncheng Tang, Zhenya Zhang, Jixiang Zhou, Lei Lei, Yuan Zhou, and Yinxing Xue. Legend: A top-down approach to scenario generation of autonomous driving systems assisted by large language models, 2024.

[34] Haiyang Tang, Dongping Chen, and Qingzhao Chu. Chatsos: Vector database augmented generative question answering assistant in safety engineering, 2024.

[35] Nham Le, Tuan Lai, Trung Bui, and Doo Soon Kim. Autonlu: An on-demand cloud-based natural language understanding system for enterprises, 2020.

[36] Chau Nguyen, Phuong Nguyen, Thanh Tran, Dat Nguyen, An Trieu, Tin Pham, Anh Dang, and Le-Minh Nguyen. Captain at coliee 2023: Efficient methods for legal information retrieval and entailment tasks, 2024.

[37] Nicholas G. Polson and Vadim O. Sokolov. Deep learning, 2018.

[38] Shiwei Feng, Yapeng Ye, Qingkai Shi, Zhiyuan Cheng, Xiangzhe Xu, Siyuan Cheng, Hongjun Choi, and Xiangyu Zhang. Rocas: Root cause analysis of autonomous driving accidents via cyber-physical co-mutation, 2024.

[39] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: What, why, and where we are?, 2022.

[40] Bettina Finzel, David E. Tafler, Stephan Scheele, and Ute Schmid. Explanation as a process: user-centric construction of multi-level and multi-modal explanations, 2021.

[41] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promise in natural language processing? a structured review, 2022.

[42] Siwei Lai. Word and document embeddings based on neural network approaches, 2016.

[43] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection, 2021.

[44] Yuxuan Hu, Chenwei Zhang, Min Yang, Xiaodan Liang, Chengming Li, and Xiping Hu. Learning to generalize unseen domains via multi-source meta learning for text classification, 2024.

[45] Ankush Agarwal, Raj Gite, Shreya Laddha, Pushpak Bhattacharyya, Satyanarayan Kar, Asif Ekbal, Prabhjit Thind, Rajesh Zele, and Ravi Shankar. Knowledge graph - deep learning: A case study in question answering in aviation safety domain, 2022.

[46] Hongseok Choi and Hyunju Lee. Exploiting all samples in low-resource sentence classification: Early stopping and initialization parameters, 2024.

[47] John X. Morris, Jin Yong Yoo, and Yanjun Qi. Textattack: Lessons learned in designing python frameworks for nlp, 2020.

[48] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[49] Yasas Senarath, Ayan Mukhopadhyay, Sayyed Mohsen Vazirizade, Hemant Purohit, Saideep Nannapaneni, and Abhishek Dubey. Practitioner-centric approach for early incident detection using crowdsourced data for emergency services, 2021.

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.