Journal of Engineering, Project, and Production Management
2024, 14(2), 0014

Enhancing Construction Site Safety: Natural Language Processing for Hazards Identification and Prevention

Shrutika Ballal[1], K. A. Patel[2], and D. A. Patel[3]

[1]Former Post Graduate Student, Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology (SV-NIT), Ichchhanath, Dumas road, Surat, Gujarat, India-395007, E-mail: shrutikaballal@gmail.com

[2]Assistant Professor, Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology (SV-NIT), Ichchhanath, Dumas road, Surat, Gujarat, India-395007, E-mail: kapatel@amd.svnit.ac.in (corresponding author).

[3]Associate Professor, Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology (SV-NIT), Ichchhanath, Dumas road, Surat, Gujarat, India-395007, E-mail: dap@ced.svnit.ac.in

Project Management

Received May 31, 2023; revised July 26, 2023; accepted October 9, 2023

Available online November 5, 2023

Abstract: Construction sites are well known for the inherent risks that negatively impact the safety and well-being of workers. Identifying and minimising these hazards is critical for preventing accidents and creating a safe working environment. Traditional techniques of hazards identification in construction rely on visual assessments and professional expertise, which can be time-consuming and subjective. The goal of this research is to identify traits that indicate potential dangers in the construction industry by extracting meaningful information from accident narratives. This will be achieved through the application of a rule-based iteration approach, using the Natural Language Toolkit (NLTK) for keyword extraction and text tokenization. It is a branch of artificial intelligence and computational linguistics concerned with the interaction of computers and human language. The research methodology involves the utilization of NLTK and the application of a rule-based iteration approach to extract hazards from construction-related accident narratives. The proposed approach includes gathering accident narratives, pre-processing data, and textual analysis with NLP tool for information extraction and training the algorithm with identified attributes. The textual analysis eventually leads to the extraction of significant sources of dangers that cause accidents. The study contributes to the developing subject of construction safety management by utilizing the capabilities of NLP to enhance hazard detection, resulting in safer construction practices and lower occupational hazards. The findings emphasise the accuracy with which NLP approaches detect dangers, allowing construction professionals to proactively decrease risks and enhance overall safety on construction sites.

Keywords: Keyword extraction, NLP, Risk, Safety, Text mining.

DOI 10.32738/JEPPM-2024-0014

1. Introduction

Construction sites are dynamic places with several hazards that endanger workers' safety and well-being. Identifying and minimizing these risk factors is critical for ensuring a safe working environment and preventing accidents. Traditional building danger identification methods rely on manual inspection and specialist knowledge, which can be time-consuming, subjective, and vulnerable to human error. As technology advances, there is an increasing interest in using computational tools to improve danger identification procedures.

Despite this, there has been a rising interest in using computational techniques to improve the processes involved in hazard identification on building sites as a result of the fast growth of technology. The technology and applications in this developing sector have the potential to fundamentally alter how risks are recognized and managed. The use of machine learning and artificial intelligence (AI) algorithms is one of the most notable developments in this field. These algorithms are capable of processing enormous volumes of information gathered from sensors, cameras, and other sources installed on building sites. These data are frequently processed by AI systems to find trends, abnormalities, and possible risk issues that manual inspections could overlook. Because of the earlier detection of risks, prompt actions and preventative measures can easily be put in place.

Natural Language Processing (NLP) is a developing technology is a sub-branch of Artificial Intelligence with the potential to enhance construction hazard detection. NLP involves the study, analysis, and generation of human language using computational algorithms and models. In the research, an NLP library called NLTK (Natural Language Toolkit for Construction) will be utilized to leverage construction-related textual data, such as incident reports, safety manuals, and inspection records. (Wu et al., 2022) By employing NLP tools like NLTK, relevant information can be automatically extracted, and potential dangers can be identified. The application of NLP for construction hazard identification offers several advantages. Firstly, it significantly reduces the time and effort required for hazard identification by automating the process. Instead of relying solely on manual inspection, these tools swiftly analyze vast amounts of textual data, enabling the extraction of crucial insights. Consequently, hazards can be identified more quickly, facilitating timely and proactive risk management. Secondly, utilizing NLP-based risk assessment enhances objectivity and consistency. Human judgment is susceptible to subjectivity and influenced by various factors, leading to disparities in risk assessment among individuals or teams. In contrast, NLP tools operate on standardized standards and impartial rules, ensuring a more uniform and unbiased approach. NLP enables the examination of extensive accident narratives and historical data, unveiling hidden patterns and trends related to potential hazards. Furthermore, NLP tools can identify common issue categories by analyzing textual data, thereby improving knowledge of typical construction hazards. This information informs proactive measures for hazard reduction and management. With the availability of vast amounts of digital data and advancements in computer capacity, practical applications of NLP-based knowledge discovery have become feasible.

The current study has demonstrated the enormous potential of NLP detecting potential sources of dangers in construction. NLP approaches provide a helpful tool for construction safety management by automating the interpretation of textual data. The incorporation of NLP algorithms into hazard identification procedures has the potential to revolutionize the profession by enabling faster, more objective, and consistent danger assessment, leading to safer building practices and enhanced worker well-being. Further research and development in NLP and its application to construction safety management will surely help ongoing efforts to mitigate risks and promote a safer construction sector as technology advances. The findings of this research have the potential to revolutionize the field of construction safety management by harnessing the power of NLP. By leveraging automated analysis of textual data, we can enhance the efficiency, accuracy, and objectivity of hazard identification, leading to improved safety practices, reduced occupational risks, and ultimately safer construction environments.

2. Literature Review

Accidents in a variety of sectors pose a serious threat to worker safety and negatively affect production. For safety measures to be put in place and future occurrences to be avoided, it is essential to recognize and comprehend the risks that led to these accidents. Natural language processing (NLP) approaches have been explored by researchers as an alternative way of automating the extraction of risks from accident narratives, facilitating effective analysis and proactive risk management. This study of the literature attempts to give a general overview of the research and approaches used for hazard extraction using NLP, with an emphasis on the strategy used in the given study.

The construction industry is associated with numerous risk factors, prompting extensive studies focused on mitigating injuries resulting from these variables. However, the process of examining reports to identify these risks is both time-consuming and labor-intensive. (Zhang et al., 2020) introduced a C-BiLSTM (Convolutional-Bidirectional Long Short-Term Memory) approach for automated categorization of construction-related incidents. This method encompasses several steps, including data collection, case labeling, model design, model training, and model performance evaluation. Notably, the chosen methodology outperformed traditional techniques like SVM (Support Vector Machine), LB (Logistic Regression), and LR (Linear Regression). Another study, (Zhang, 2022) employed the Word2Vec skip-gram model to classify accident causes. The process involved two key steps: training the Word2Vec skip-gram model and constructing a hybrid structural deep neural network. The results provided empirical evidence of the algorithm's effectiveness in accurately categorizing accident causes.

Xu et al. (2021) utilized text mining techniques to analyze a dataset of 221 reports on metro construction accidents. Their methodology encompassed text screening, text segmentation, computation of entropy-weighted words, selection of high-frequency phrases, and pattern formation. These steps collectively facilitated the effective identification of safety risk factors associated with construction accidents. Baker et al. (2020) evaluated various approaches, including TF-IDF (Term Frequency - Inverse Document Frequency) + SVM, CNN (Convolutional Neural Networks), and HAN (Hierarchical Attention Networks), to identify reliable injury precursors. Their findings demonstrated the effectiveness of these techniques in recognizing key [indicators of potential accidents. Additionally, Chi et al. (2014) utilized ontology-based text categorization to investigate critical construction job roles and discover probable hazards. Their research underscored the importance of Job Hazard Analysis (JHA) in assessing and mitigating risks in construction projects. Finally, Goldberg (2022) leveraged large datasets from prior studies to develop high-performing classification algorithms. By employing machine learning, the study automated the labeling of accident narratives, enhancing the efficiency and accuracy of risk assessment in the construction industry. Rupasinghe and Panuwatwanich (2020) describes a method for extracting risks from injury data using a Natural Language Processing (NLP) and Text Mining (TM) methodology. The Case-Based Reasoning (CBR) methodology, which combines the Vector Space Model and Semantic Quiry Expansion NLP methodologies, was used by Zou et al. (2017). The risk case dataset extraction prototype was made using the Python programming language.

Despite the recognition of risks and dangers on construction sites, there is a research gap in accurately and comprehensively identifying and reducing these risks. Therefore, there is a need to explore and implement Natural Language Processing (NLP) techniques to automate hazard identification, injury analysis, and the extraction of specific details such as injury causes, body parts affected, nature of hospitalization, and severity of incidents. By leveraging NLP on textual data from various sources, such as incident reports, safety manuals, and inspection records, a more efficient and comprehensive understanding of hazards and injuries can be achieved. Therefore, the authors employed cutting-edge machine learning methods to extract information from accident instances and safety data. The current study adopts a "Rule-Based" technique that makes use of a basic NPL library and simple syntax to rapidly and easily categorise accident facts. The approach simply uses pre-identified injury characteristics to train the algorithm and multiple rounds to extract accurate accident data. The broad objectives of the present study are (i) To develop an algorithm for knowledge discovery of construction accidents, (ii) To evaluate possible hazards, cause of injury, body parts injured, fatal and nonfatal and other informative attributes based on past accident reports and narratives using the developed algorithm, and (iii) To examine the identified attributes related to injury characteristics.

3. Methodology

The principal objective of this study is to discern specific characteristics that can act as indicators of hazards within the construction industry. In pursuit of this goal, accident narratives pertaining to construction incidents are meticulously considered as valuable data sources. To achieve efficient information extraction from these narratives, the "Rule-based Iteration" approach is adopted, facilitating the retrieval of relevant and meaningful insights. By identifying and isolating fundamental injury precursors through this iterative process, accident stories can be subjected to a more comprehensive and insightful analysis. This approach not only enhances the understanding of construction-related hazards but also contributes to the development of effective strategies for risk mitigation and accident prevention.

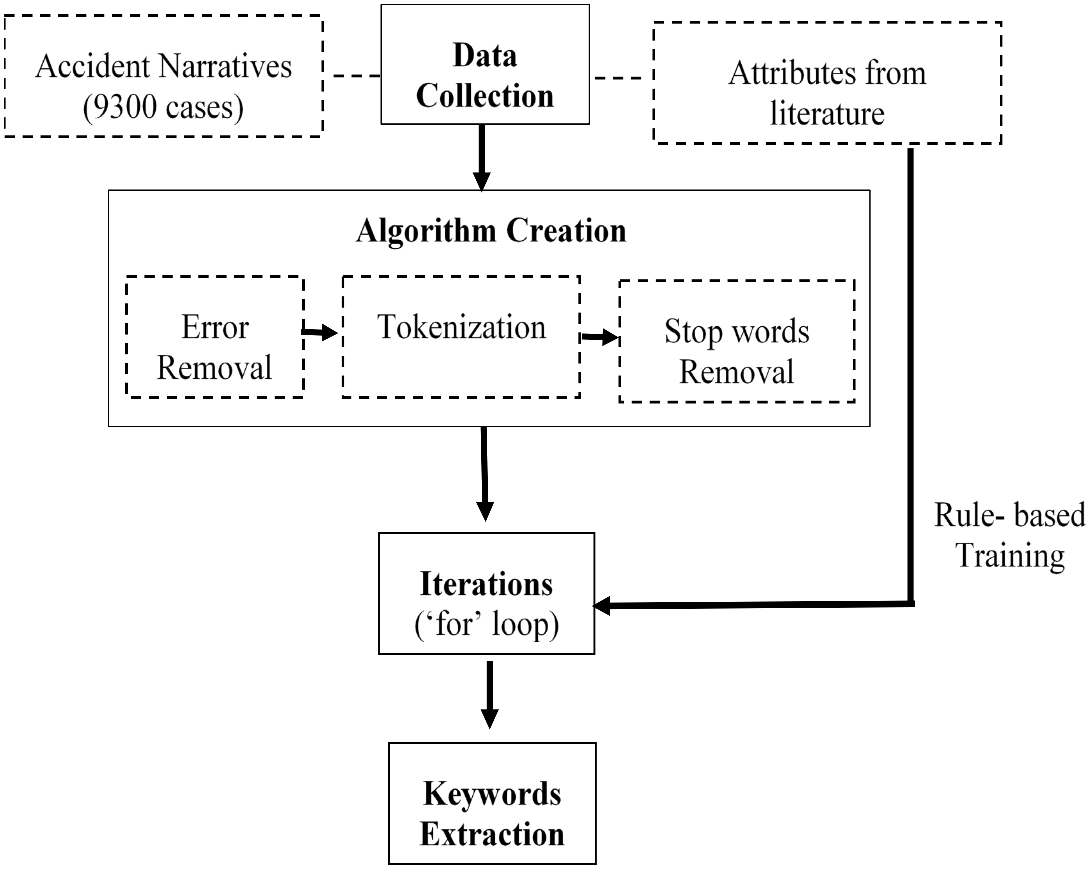

Fig. 1. Methodology for hazard extraction

Fig. 1 depicts the Hazard Extraction methodology, a crucial process that leverages the power of the Natural Language Toolkit (NLTK) for proficient keyword extraction. The NLTK provides a diverse set of tokenization techniques, allowing for the effective segmentation of text into smaller, more manageable units, such as sentences or words. For researchers and practitioners interested in replicating or examining the methods used for hazard extraction, the complete code developed for this purpose can be found in Appendix A, providing a valuable reference for future investigations in the field of construction safety and risk management.

3.1. Data Collection

As part of the data collection process, event narratives and construction accident reports are collected. An example of an accident narrative can be found in Table 1. The gathered information constitutes an unstructured collection of exceptional events. To create a comprehensive input library, around 9300 incident reports are utilized. These reports had been made/filled by the human; therefore, any errors arising from human involvement in writing these reports are ignored. Initially, the necessary attributes for training are sourced from literature and internet databases. Based on the type of information extraction, relevant characteristics are acquired. To handle the input database, a function called 'extract_hazards (narrative)' is defined.

3.2. Error Removal

The data is prepared for information extraction by removing unnecessary spacing between words and converting each phrase to lowercase. When it comes to accident narratives or input data, the NLTK library does not immediately discard errors. Instead, it provides users with various tools and functionalities to analyze and assess textual data. Tokenization, particularly sentence and word tokenization, is a primary feature of NLTK. By employing these tokenization methods, accident narratives are segmented into sentences and words, respectively, facilitating further analysis.

NLTK offers tokenization tools like 'sent_tokenize()' and 'word_tokenize()' designed to handle diverse text input formats and address common issues such as punctuation, whitespace, and formatting inconsistencies. By utilizing these capabilities, NLTK ensures that accident narratives are effectively divided into meaningful units, enabling accurate hazard extraction.

Table 1. Illustration of an accident narrative

Accident Narrative	Hazards
"at 11:30 a.m. on July 20, 2016, two employees were working together to remove a finished product from a mold. one co-worker, a forklift operator, was operating the forklift in an attempt to position the finished product from the mold machine on the forklift's forks. the other employee, a 23-year-old male, was located between the raised forks of the forklift and the rotational mold machine. the employee was attempting to slide the product in place on the forks of the forklift when the forklift operator accidentally pressed on a pedal of the forklift causing the forklift to pin the employee between the forklift and the rotational mold machine. the employee died as a result of crushing injuries to the head."	machine, forklift

3.3. Tokenization

A variety of tokenization techniques are provided by NLTK, effectively dividing text into smaller units such as sentences or words. The significance of these methods lies in their role in language processing tasks, as they furnish a structured representation of the text's elements. NLTK's tokenization capabilities are harnessed by researchers and developers, enabling the extraction of valuable information from textual data and facilitating enhanced analysis and processing of individual units. As a result, a more precise and comprehensive understanding of natural language is achieved (Manning *et al.*, 2014).

3.3.1. Sentence Tokenization

The 'sent_tokenize()' function of the NLTK is frequently used for sentence tokenization. To locate sentence boundaries in the text, it uses algorithms and models that have already been trained. To detect sentence breaks, the function reads the input text and looks for specified patterns, such as punctuation (such as periods, question marks, and exclamation points). To accomplish precise sentence segmentation, it takes into account a variety of language-specific rules and patterns. For example, if you have an accident narrative like: "An employee slipped and fell on the wet floor. He suffered a minor injury." NLTK's 'sent_tokenize()' function would tokenize it into two sentences:

1) "An employee slipped and fell on the wet floor".
2) "He suffered a minor injury".

This exemplifies the proficiency and accuracy of NLTK's 'sent_tokenize()' function in accomplishing effective sentence segmentation, thereby aiding in more detailed and comprehensive language processing tasks.

3.3.2. Word Tokenization

Word tokenization is accomplished using NLTK's word_tokenize() function. It divides a sentence or a paragraph into separate words or tokens. To establish word boundaries, NLTK takes into account a variety of linguistic norms and patterns, including whitespace, punctuation, and special characters. For example, if you have a sentence: "An employee slipped and fell on the wet floor." NLTK's word_tokenize() function would tokenize it into a list of words: ["An", "employee", "slipped", "and", "fell", "on", "the", "wet", "floor."]

In the process of segmenting the text into meaningful units, NLTK's tokenization techniques consider essential factors such as punctuation, other linguistic elements, and language-specific constraints. These tokenization functions play a pivotal role in preparing accident narratives for further analysis, notably for tasks like hazard extraction. Their efficacy ensures the data is suitably structured and ready for comprehensive examination, enabling researchers and developers to extract valuable insights and valuable information from the textual data.

3.4. Stopwords Removal

A list of frequently used stopwords for many languages, including English, is provided by NLTK. Stopwords are words that, in the context of text analysis, are regarded as unimportant or have little meaning. Stopwords like "a," "an," "the," "is," "and," "in," etc. are examples of stopwords in English. By removing these widespread terms from the text data with NLTK's stopwords, one may concentrate on words with more meaning and increase analytical precision. The process of removing stopwords with NLTK involves the following steps:

A. *Stopword Initialization:* The NLTK code initializes the stopwords by downloading them from the NLTK data source. For English-specific stopwords, the nltk.download('stopwords') method facilitates easy download.

B. *Creating Stopword Set:* After downloading the stopwords, NLTK constructs a stopword set using the stopwords.words('english') function, encompassing all the English stopwords available in the NLTK data.

C. *Stopword Filtering:* During the hazard extraction procedure, NLTK's stopword list is employed to filter out stopwords from the accident narratives. The code iterates through the accident stories, examining each word against the stopword list. If a word is identified as a stopword, it is excluded from the analysis, ensuring a more focused and meaningful examination of the data.

3.5. Rule-based Training

The 'Rule-based hazard extraction' method in the code operates by utilizing a predefined list of construction-related hazard phrases (Le and David Jeong, 2017). In this algorithm, each sentence within the accident narratives is repeatedly examined to identify potential matches with hazard keywords. Whenever a match is found, it is duly appended to the list of hazards.

Unlike the use of labeled data to train a machine learning algorithm to establish a connection between accident narratives and risks, this particular code does not follow that approach. Instead, it relies on a pre-defined set of hazard terms to diligently search for hazards within the accident narratives. This rule-based approach provides a practical and effective means of hazard extraction, enabling the identification of hazards without requiring extensive training data or complex machine learning models.

3.6. Iterations

In the developed code, there are two main iterations performed. The first iteration is performed using the sentences that were taken directly from the accident narratives. This is accomplished by utilizing a ‘for loop’ that iterates through the sentences list that is created by subjecting the narrative text to NLTK's ‘sent_tokenize()’ function. The loop iterates through each sentence sequentially, allowing for further processing and sentence-level analysis. The second iteration is performed over the words within each sentence. This is done using another for loop that iterates over the words list, obtained by applying NLTK's ‘word_tokenize()’ function to each sentence. The loop iterates through each word within a sentence, allowing for further analysis and checks on individual words. Within this loop, the code checks if each word is a hazard keyword and appends it to the hazards list if it matches any of the predefined hazard keywords. These iterations allow the code to process the accident narratives at a more granular level, sentence by sentence and word by word, enabling the extraction of hazards based on specific criteria.

3.7. Hazards Extraction

First, the code imports the necessary libraries. It uses the NLTK's library for natural language processing tasks and specifically imports functions for tokenization and stop words. It downloads necessary resources like 'punkt', 'averaged_perceptron_tagger', 'omw-1.4', 'stopwords' from the NLTK package. The NLTK stopwords for the English language are downloaded and stored in a variable named stop_words. These stopwords are common words like "a," "the," "and," etc. that do not carry significant meaning in the context. It creates a set of stop words in English language using NLTK's stopwords corpus.

Next, a function called extract_hazards(narrative) is defined. This function takes a narrative as input and aims to extract hazards from it. The narrative is divided into sentences using the sent_tokenize() function from NLTK. A list of hazard keywords related to construction accidents has been created. These keywords represent potential hazards that could be present in the accident narratives. Examples of hazard keywords include "machine", "fire," "electrocution," "chemical," and many others. For each sentence in the narrative, the sentence is tokenized into individual words using the word_tokenize() function. These words are then converted to lowercase and filtered to remove non-alphabetic words and stopwords. The code iterates over each word in the sentence and checks if it matches any of the hazard keywords. If a match is found, the word is considered a hazard and added to a list of hazards. Finally, the list of hazards is converted to a set to remove duplicates and returned as the output of the extract_hazards() function. To apply the hazard extraction function to the accident narratives, the code uses the apply() function on the 'Narrative' column of the DataFrame. It assigns the extracted hazards to a new column named 'Hazards'. Lastly, the resulting DataFrame with the extracted hazards is saved to an Excel file named 'sources_hazards.xlsx' using the to_excel() function. For example, the extracted hazards are stored in the 'Hazards' column of the DataFrame as shown in Table 1. Each row in the 'Hazards' column contains a list of hazards associated with the corresponding accident narrative.

To determine the number of various types of hazards through graphs, the Python programming language was employed, specifically utilizing the 'matplotlib.pyplot' tool, which is a part of the matplotlib library. This tool provides a user-friendly interface for creating diverse types of visualizations and plots. 'Matplotlib.pyplot' is widely recognized and extensively used within the Python community as a powerful plotting library. It provides an extensive range of options for generating high-quality graphs and charts, making it a popular choice for data visualization tasks. Its versatility allows users to create various types of plots, including line graphs, bar charts, scatter plots, and more. By utilizing 'matplotlib.pyplot,' researchers and data analysts can effectively convey complex information through visually appealing and easily interpretable graphs. The tool's capabilities enable the representation of trends, patterns, and relationships within the accident reports, aiding in the identification of common hazardous scenarios and facilitating targeted preventive measures. The code for graph generation by using 'matplotlib.pyplot' tool is given below in Appendix B.

In summary, the code uses NLTK's tokenization and stopwords functionalities to extract hazards from accident narratives. It matches the words in the narratives against a predefined list of hazard keywords and stores the extracted hazards in a new column of the DataFrame.

4. Findings and Discussion

As mentioned in the objectives, the code/algorithm has been developed to evaluate hazards, causes of injury, injured body parts, fatal and nonfatal incidents, and other informative attributes through past construction accident reports and narratives. Additionally, the attributes can be used to identify injury-related characteristics. In this study, the information from the analysis can prove beneficial for safety managers in pinpointing the exact source of accidents and enhancing the safety of workers.

Fig. 2 illustrates the graph representing the most common hazardous scenario derived from around 9300 accident reports. As stated earlier, the graph was generated using the Python programming language, specifically employing the 'matplotlib.pyplot' tool. The depicted graph highlights the prevalence of various hazards in the construction context, with "Truck" being identified as the highest common hazard. This signifies that trucks pose a significant risk factor in construction environments. Trucks are extensively utilized for transporting materials, equipment, and personnel within construction sites, but their presence introduces potential dangers that must be effectively managed to ensure the safety of workers.

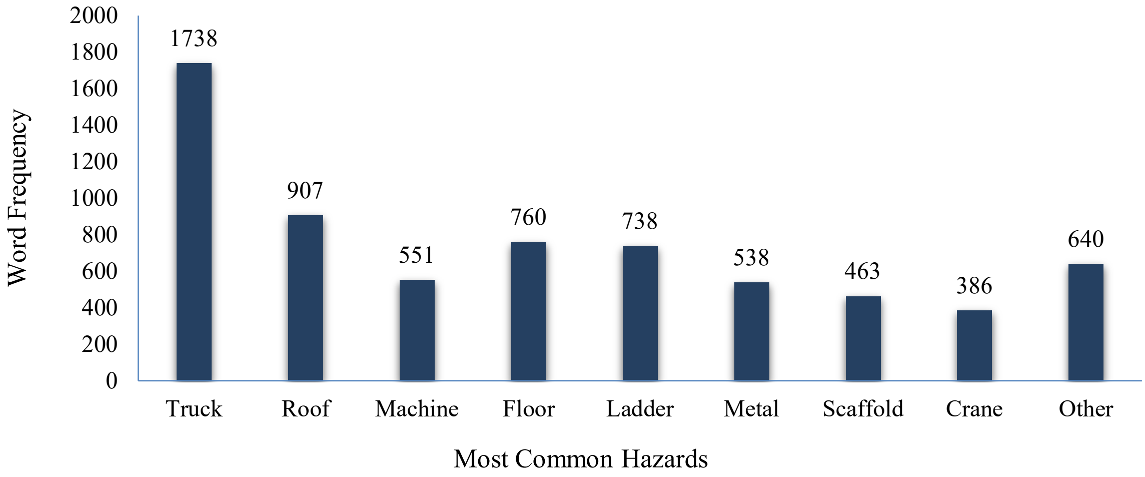

Fig. 2. Graph of most common hazards

The second most common hazard identified in the graph is "Roof." Construction workers often perform tasks on rooftops, which exposes them to hazards such as falls, unstable surfaces, or collapsing structures. Proper safety measures, such as fall protection systems and regular inspections, are essential to mitigate these risks. The third common hazard identified is a combination of "Floor" and "Ladder". The floor may have hazards such as slippery surfaces or uneven terrain, while ladders can present risks of falls if not used correctly. The fourth common hazard grouping consists of "Machine," "Metal," "Scaffold," and "Crane." Construction sites often involve the use of heavy machinery and equipment, which can pose hazards like entanglement, crushing, or mechanical failures. Metal structures, such as beams or sharp edges, can cause cuts or other injuries. Scaffolds and cranes, if not properly erected or operated, can lead to falls or falling objects, endangering workers. These identified hazards emphasize the importance of implementing robust safety protocols, comprehensive training programs, and diligent site inspections in construction environments. By addressing and mitigating these common hazards, construction companies can significantly reduce the risk of accidents, injuries, and fatalities, fostering a safer working environment for all personnel involved. The ‘Other’ category encompasses causes such as illness, trauma, fire, heart attack, burning due to chemical, electrical shock and other incidents where limited information is available. Although the specific details may vary, these accidents highlight the diverse range of factors that can contribute to workplace injuries.

The various hazards fall in the 'Other' category have been listed in Fig. 3 along with their percentages. The category in workplace injuries is a catch-all classification that includes causes not explicitly specified under more specific headings. It encompasses a broad spectrum of incidents, such as illnesses resulting from exposure to hazardous substances, infectious agents, or workplace-related stress. Traumatic events like slips, falls, being struck by objects, or physical altercations are also included, as are fires caused by electrical faults, overheated equipment, or flammable materials. Moreover, heart attacks that occur at work due to stress, physical exertion, or underlying health conditions are considered within this category, along with burns resulting from chemical exposure and injuries caused by electrical shocks. This comprehensive category also accounts for cases where information is insufficient to classify incidents under more specific headings. These could involve unique or rare accidents or injuries, or situations where data collection and reporting might be incomplete or inconclusive. The category serves as a vital reminder that workplace injuries can arise from a myriad of factors, some of which may be unforeseen or less common. It underscores the complexity of ensuring workplace safety and the need for comprehensive risk assessment and mitigation strategies. Employers must be proactive in identifying potential hazards and implementing appropriate safety measures to protect their employees. Adequate safety training and education should be provided to raise awareness among workers about potential risks, irrespective of their rarity. By thoroughly investigating incidents in this category and identifying trends or patterns, organizations can continually improve their safety protocols and standards, reducing the likelihood of workplace injuries and promoting a culture of safety and well-being for their workforce.

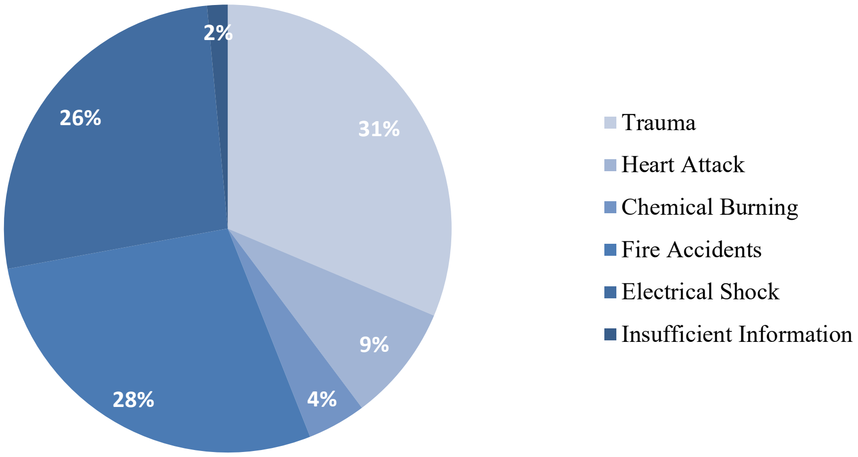

Fig. 3. Pie chart of ‘Other’ category

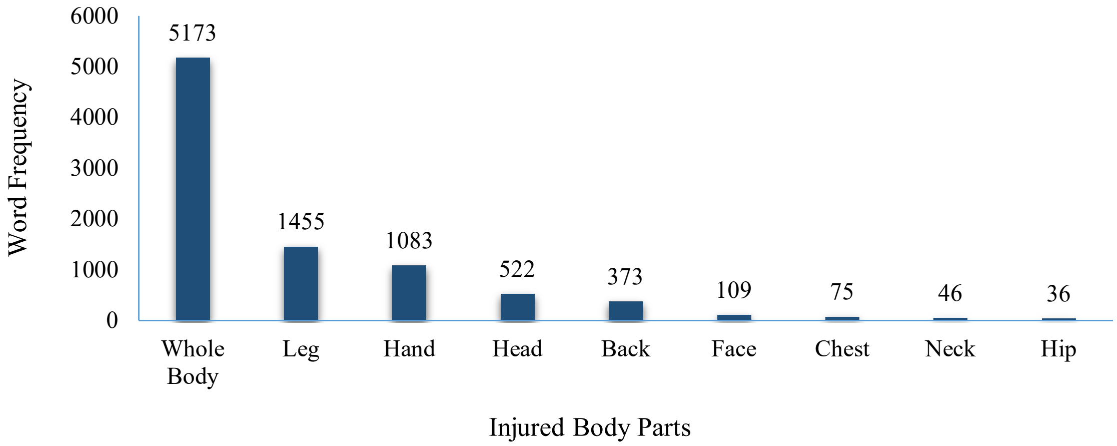

Fig. 4. Graph of injured body parts

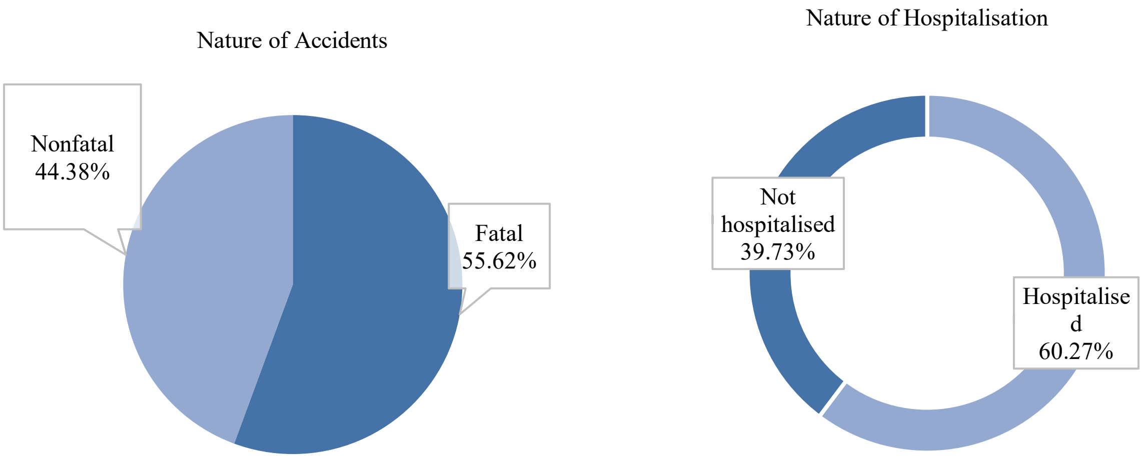

Fig. 5. Pie charts of nature of accidents and nature of hospitalization

In most accidents, the graph represents the employees' 'Whole Body' is the most seriously injured as shown in Fig. 4. Workers who are confined due to heat stroke or who are traumatized are also included in the Whole Body group. The 'Leg' is the second bodily component group, which contains feet, ankles, knees, leg fingers, and so on. Leg accidents are often caused by burning, crushing, or being hit between items. Similarly, under the 'Hand' category, fingers, arms, nails, right and left hands, and so on are all grouped.

Nature of accidents and nature of hospitalization is depicted in terms of Pie chart in Fig. 5. Out of a total of 9300 instances, 5182 are fatal, while the rest are nonfatal. The code categorises the information based on terms such as "death," "killed," and so on. This implies that if these terms are discovered, the instances are classified as fatal. As previously stated, the majority of catastrophic instances are caused by 'Fall'. Workers are hospitalized in 60.27% of instances and the remaining instances omitted hospital-related terms. Because of the low severity of the injury, employees may be treated on-site.

5. Conclusions and Further Research

The study addresses the need for accurate and comprehensive identification and reduction of risks on construction sites. By leveraging Natural Language Processing (NLP) techniques, the research aims to automate hazard identification, injury analysis, and extraction of specific details from accident reports and narratives. Using machine learning methods and a "Rule-Based" approach, the algorithm effectively categorizes accident facts, leading to precise hazard evaluation. Herein, the Hazard Extraction methodology has been proposed utilizing NLTK's tokenization techniques and accordingly the code/algorithm has been developed. The methodology evaluates hazards, causes of injury, and other attributes using past accident reports and narratives. The information aids safety managers in identifying accident sources and enhancing worker safety. The study reveals the most common hazardous scenario with "Truck" as the highest common hazard. "Roof," "Floor and Ladder," and "Machine, Metal, Scaffold, and Crane" are also identified as frequent hazards, emphasizing the need for robust safety protocols. The "Other" category includes various incidents with limited information. Workplace injuries encompass a wide range of factors, necessitating comprehensive risk assessment and mitigation strategies. Thorough investigation and trend analysis lead to continuous safety improvement. The presented work shows that the 'Whole Body' is most frequently injured in accidents. The 'Leg' and 'Hand' categories also account for significant injury instances. The nature of accidents and hospitalization is depicted, with a majority being nonfatal, and 60.27% of cases requiring hospitalization.

The identified attributes offer crucial insights into accidents, aiding understanding of causes and contributing factors. Companies can implement targeted safety measures to mitigate risks and protect workers' well-being. Knowledge of commonly affected body parts helps implement ergonomic improvements. Identifying specific hazards enables effective risk assessment and control measures. The nature of hospitalization guides safety measure evaluations and future prevention strategies. The findings have far-reaching implications for safety planners and engineers, enhancing hazard identification and risk mitigation practices. The extracted features serve as valuable tools, empowering safety planners and engineers to create safer working environments. Additionally, they contribute to refining workplace safety regulations and industry-wide best practices, fostering a culture of safety across construction. Eventually, this research significantly contributes to construction safety knowledge and promotes a safer working environment industry-wide.

One of the key advantages of this research is its ability to analyze large volumes of data more efficiently and accurately. This capability enables safety planners and engineers to process and interpret vast amounts of information quickly, allowing for more comprehensive hazard identification and risk assessment. Furthermore, the study can be expanded to further segment and analyze the data to reveal additional insights and literary characteristics. By delving deeper into the dataset, researchers can uncover patterns, correlations, and hidden factors that may contribute to accidents and safety incidents. The expanded analysis can provide a more nuanced understanding of the factors influencing safety in the workplace, leading to more effective preventive strategies. Moreover, as part of the future scope, a careful evaluation of the highlighted features can be conducted. This evaluation can involve a thorough examination of the extracted features to assess their relevance, reliability, and potential for implementation in real-world safety planning and engineering practices. By critically evaluating the highlighted features, researchers can refine and improve the methodology and explore potential avenues for practical application.

Author Contributions

Shrutika Bhallal contributes to conceptualization, data collection, methodology, computer program, analysis, and investigation. K. A. Patel contributes to draft preparation, manuscript editing, revision as per comments from the reviewers, and supervision. D. A. Patel contributes to manuscript editing, visualization, supervision, and project administration.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

Baker, H., Hallowell, M. R., and Tixier, A. J. P. (2020). Automatically learning construction injury precursors from text. *Automation in Construction*, 118, 106145. doi.org/10.1016/j.autcon.2020.103145

Chi, N. W., Lin, K. Y., and Hsieh, S.H. (2014). Using ontology-based text classification to assist job hazard analysis. *Advanced Engineering Informatics*, 28(4), 381-394. doi.org/10.1016/j.aei.2014.05.001

Goldberg, D. M. (2022). Characterizing accident narratives with word embeddings: Improving accuracy, richness, and generalizability. *Journal of Safety Research*, 80, 441-455. doi.org/10.1016/j.jsr.2021.12.024

Le, T., and David Jeong, H. (2017). NLP-based approach to semantic classification of heterogeneous transportation asset data terminology. *Journal of Computing in Civil Engineering*, 31(6), 04017057. doi.org/10.1061/(asce)cp.1943-5487.0000701.

Manning, C .D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D. (2014). The Stanford coreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55-60. Retrieved from https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf on July 24, 2023.

Rupasinghe, N. K. A. H., and Panuwatwanich, K. (2020). Extraction and analysis of construction safety hazard factors from open data. *IOP Conference Series: Materials Science and Engineering*, 849(1), 012008. doi.org/10.1088/1757-899X/849/1/012008.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., and Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, 134, 104059. doi.org/10.1016/j.autcon.2021.104059.

Xu, N., Ling, M. A., Liu, Q., Wang, L., and Deng, Y. (2021). An improved text mining approach to extract safety risk factors from construction accident reports. *Safety Science*, 138, 105216. doi.org/10.1016/j.ssci.2021.105216.

Zhang, F. (2022). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *International Journal of Construction Management*, 22(6), 1120-1140. doi.org/10.1080/15623599.2019.1683692.

Zhang, J., Zi, L., Hou, Y., Deng, D., Jiang, W., and Wang, M. (2020). A C-BiLSTM approach to classify construction accident reports. *Applied Sciences,* 10(17), 5754. doi.org/10.3390/app10175754.

Zou, Y., Kiviniemi, A., and Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in Construction*, 80, 66–76. doi.org/10.1016/j.autcon.2017.04.003.

Shrutika Ballal is presently working as a Post Graduate Trainee at Larson and Tubro company, STT Data Centre, Kolkata, West Bengal, India. She has completed her BTech from KIT's College of Engineering, Kolhapur in 2021; MTech in Construction Technology and Management from Sardar Vallabhbhai National Institute of Technology (SV-NIT), Surat in 2023. His research interests include construction technology & management, and construction safety.

K. A. Patel is presently working as an Assistant Professor at Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology (SV-NIT), Surat, Gujarat, India. He received his BE from L. D. college of engineering, Ahmedabad in 2007; MTech from Malaviya National Institute of Technology Jaipur in 2009; and PhD from Indian Institute of Technology Delhi, New Delhi in 2016. His research specialisations include construction technology & management, structural engineering, and earthquake engineering.

D. A. Patel is presently working as an Associate Professor at Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology (SV-NIT), Surat, Gujarat, India. He received his BE from Government Engineering College, Modasa in 2001; MTech from Centre for Environmental Planning and Technology (CEPT) university, Ahmedabad in 2003; and PhD from Indian Institute of Technology Delhi, New Delhi in 2015. His research interests include construction technology & management, construction safety, valuation, and construction laws.

Appendix A: Code for extraction of Hazards from Accident Narratives:

```
import nltk

nltk.download('punkt')

nltk.download('stopwords')

import pandas as pd

import nltk

from nltk.tokenize import sent_tokenize, word_tokenize

from nltk.corpus import stopwords

# Load the accident narratives from the Excel file

df = pd.read_excel('accident_narratives.xlsx')

# Initialize NLTK's stopwords

nltk.download('stopwords')

stop_words = set(stopwords.words('english'))
```

```
# Function to extract hazards from a given narrative
def extract_hazards(narrative):
    hazards = []
    sentences = sent_tokenize(narrative)
    # List of hazard keywords related to construction
    hazard_keywords = ['mechanical power press', 'fireworks', 'roof', 'load', 'metal', 'machine', 'ladders', 'scaffold', 'cranes', 'excavation', 'electricity', 'welding', 'cutting', 'demolition', 'confined spaces', 'equipment', 'tools', 'nail guns', 'saws', 'grinders', 'compressors', 'forklifts', 'explosive', 'chemical', 'asbestos', 'noise', 'vibration', 'pressure', 'objects', 'materials', 'power lines', 'trenches', 'roofing', 'energy', 'flammable', 'heat', 'cold', 'radiation', 'ergonomics', 'biological agents', 'sharp objects', 'crushing', 'motion', 'fire', 'lighting', 'fumes', 'ventilation', 'surface', 'guarding', 'lifting', 'collision', 'signage', 'disposal', 'floors', 'openings', 'protocols', 'training', 'maintenance', 'collapse', 'ground', 'edges', 'fall protection', 'noise levels', 'substances', 'scaffolding', 'security', 'workloads', 'dehydration', 'electrocution', 'stroke', 'disorders', 'eye injuries', 'respiratory', 'fatigue', 'communication', 'heavy', 'ergonomics', 'entrapment', 'cave-ins', 'carbon monoxide', 'wiring', 'malfunctions', 'ventilation', 'failure', 'guards', 'operators', 'modifications', 'traffic', 'barricades', 'debris', 'weather', 'machinery parts', 'chemical', 'defective', 'signage', 'lighting', 'ladder use', 'floor', 'weather', 'chemical', 'temperature', 'noise', 'vibration', 'manual handling', 'machinery', 'equipment', 'maintenance', 'supervision', 'signage', 'methyl chloride', 'structural collapse', 'procedures', 'overexertion', 'heat stress', 'cold stress', 'drowning', 'musculoskeletal disorders', 'respiratory disorders', 'vision impairment', 'insect', 'animal bites', 'skin irritation', 'chemical burns', 'electrical', 'radiation', 'explosive blasts', 'debris', 'heavy machinery',  'suffocation', 'poisoning', 'hazardous waste', 'truck', 'hook', 'bar', 'drill', 'toxic fumes', 'ventilation', 'emergency exits', 'tools', 'traffic', 'inclement weather', 'roofing', 'scaffolding', 'ladder', 'trenching', 'excavation', 'energy sources', 'barricading', 'power lines', 'flammable liquids', 'welding', 'cutting', 'lighting', 'ergonomic', 'temperature', 'stairs', 'sharp objects', 'handling heavy loads', 'noise levels', 'strains', 'sprains', 'wire', 'chain', 'motor', 'vehicle', 'trailer', 'trolley', 'moving objects', 'moving machinery', 'entrapment', 'equipment failure', 'overhead cranes', 'rope', 'log', 'breath', 'skidder', 'forklift', 'mower', 'safety equipment', 'unsafe work practices', 'inspection', 'work sites', 'lighting', 'confined spaces', 'tractor', 'crusher', 'scraper', 'crane', 'tower', 'chemical spills', 'biological contamination', 'structural defects', 'security', 'training', 'risk assessment', 'tools', 'hazardous materials', 'confined space entry', 'collapsing structures', 'debris', 'soil conditions', 'power tool']
    for sentence in sentences:
        words = word_tokenize(sentence)
        words = [word.lower() for word in words if word.isalpha() and word.lower() not in stop_words]
        for word in words:
            if word in hazard_keywords:
                hazards.append(word)
    return list(set(hazards))
# Apply the extract_hazards function to the 'Narrative' column
df['Hazards'] = df['Narrative'].apply(extract_hazards)
# Save the results to an Excel file
df.to_excel('sources_hazards.xlsx', index=False)
```

Appendix B: Code for Graph Generation of results using 'matplotlib.pyplot' tool:

```
pip install pandas matplotlib openpyxl
import pandas as pd
import matplotlib.pyplot as plt
# Read the Excel file
df = pd.read_excel('sources_hazards1.xlsx')
```

```
# Extract the 'Hazards' column as a Series
hazards = df['Hazards']
# Concatenate all words from the 'Hazards' column into a single string
all_words = ','.join(hazards)
# Split the string into individual words and clean them
words = [word.strip() for word in all_words.split(',')]
# Calculate the word frequencies
word_freq = pd.Series(words).value_counts()
# Plot the bar graph
word_freq[:10].plot(kind='bar')
plt.xlabel('Most common hazards')
plt.ylabel('Word frequency')
plt.title('Frequency of Words in Hazards Column')
plt.show()
```