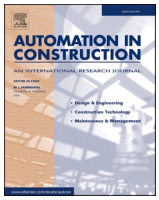
Automation in Construction 164 (2024) 105466

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Construction risk identification using a multi-sentence context-aware method

Nan Gao *, Ali Touran, Qi Wang, Nicholas Beauchamp

Northeastern University, 360 Huntington Ave, Boston, MA 02115, United States

ARTICLE INFO

Keywords:
Project-level risk
Risk identification
Context-aware text classification
Natural language processing

ABSTRACT

Knowledge of risk events with potentially negative consequences from previous projects is essential for risk identification in early stages of new infrastructure projects. However, historical risk events are usually scattered in various sources and reports, rendering collecting such risk information time-consuming and expensive. To expand the current risk data sources and facilitate risk events' extraction, the paper presents a synthetic approach that utilizes Natural Language Processing (NLP) techniques to automatically identify and extract risk-related sentences from news articles. A supervised Multi-sentence Context-aware Risk Identification (MCRI) model is devised to exploit both sentence-level and multi-sentence level context to boost the sentence classification performance. The MCRI model outperformed several baseline models with a risk-class F1-score of 87.1% and an accuracy of 86.7%. This paper provides a baseline for future studies aimed at automating the extraction of project-level risk information within the construction domain.

1. Introduction

Infrastructure projects are capital-intensive, of long durations and fraught with uncertainty, cost overruns and delays. According to Gao and Touran [15], the average cost of the US urban rail projects completed in the last four decades was $826.5 million with an average development duration of 10 years with an average cost overrun of 31.2%. Risk assessment can help avoid or reduce the detrimental consequences on project performance caused by adverse future events. Though broad risk assessment can incorporate both positive uncertainties (opportunities) and negative uncertainties (risk), the risk in this paper is refined as negative uncertainty only. It involves establishing procedures for quantifying, mitigating, and monitoring risks. The results of risk quantification can help determine the appropriate level of contingency needed. Therefore, risk assessment is a critical part in project management to ensure the successful delivery of construction projects.

Risk identification is the first step in the risk assessment process, typically performed during the early phase of a project life cycle [45]. The identified risks serve as crucial input for later steps and affect/determine the effectiveness of risk assessment. Current risk identification approaches include literature review, questionnaire survey, brainstorming, and use of risk checklists [42]. While the last two methods are commonly employed in practice [31,46], they heavily rely on project managers' experience. Thus, the outcome of these approaches can be influenced by individual attitudes and perceptions [33]. Furthermore, since the expertise is stored in the individual's mind rather than a centralized corporate database, it can be lost when staff members depart or retire [37].

To mitigate human bias and complement manual efforts, an effective approach for risk identification involves extracting risk events from past similar projects. However, due to lack of an effective knowledge management system and learning culture, the practice of capturing and storing such knowledge is less common among construction companies [37,43]. To supplement the scarce data of risk-related knowledge, non-technical data such as news articles can be utilized. News articles record issues and risk events causing project performance issues as projects progress, making it a valuable additional resource. Indeed, several studies have employed news data for analysis. As an example, Bhadani et al. [4] extracted financial risk events from news and assessed their impacts on various stock indices. Lu et al. [30] developed an automated framework to extract firm-specific risks from the Wall Street Journal. Chu et al. [7] conducted sentiment analysis on online news articles to recognize the pattern of risk variation in the supply chain area. And in the construction domain, Ninan [36] manually reviewed safety news from Google news repository to investigate the world perception of construction safety and uncover the underrepresented safety concerns.

* Corresponding author.
E-mail address: gao.n@northeastern.edu (N. Gao).

https://doi.org/10.1016/j.autcon.2024.105466
Received 26 January 2023; Received in revised form 1 May 2024; Accepted 8 May 2024

Identifying project-level risk events from text poses unique challenges, however, including information sparsity, domain specialty, and context dependencies. Firstly, unlike financial news, there is no dedicated publisher or news source focused on reporting the development process of ongoing individual construction projects. As a result, the potential documents containing risk information are scattered in various reports and news articles, making indexing and collection more difficult. Additionally, many risk events are specific to the construction domain due to their exposure to the external environment, such as unforeseeable underground conditions and inclement weather. Consequently, the model trained in the general management domain would not be suitable for extracting project-level risk event in the construction area. Furthermore, identifying construction risk events usually requires the co-occurrence of information about "uncertainty" and "potential negative impact on cost or schedule", which can span multiple sentences. This characteristic requires the designed classification model to account for the interaction between sentences, instead of ignoring the context and taking an individual sentence as an isolated input.

Overall, intelligent systems that can automatically detect project-level risk events from a broad range of textual data such as news articles carry significant potential to make up for the scarcity of explicit risk data and complement the current practices. To the best of the authors' knowledge, there is no such established system. In light of this, the paper proposes a synthetic approach to collect and cleanse the corpus from news articles, to extract potential risk paragraphs using domain knowledge, and to identify construction risk events scattered in a large number of documents. Specifically, the Multi-sentence Context-aware Risk Identification (MCRI) model is proposed which integrates RoBERTa and Bidirectional Gated Recurrent Unit (BiGRU). The contributions of this study lie in three areas: 1) the MCRI model advances existing classification models in the construction domain by exploiting both sentence-level and multi-sentence level context. Instead of taking an individual sentence as the input, the MCRI model takes a block of sentences as inputs and assigns labels to each sentence; 2) the study demonstrates that by adding higher level of context, incremental performance gains can be achieved in the construction risk identification; 3) non-technical data such as news articles are less studied in the risk identification domain and this study fills that gap and improves the efficiency of collecting risk information from news and other textual documents.

2. Related work

2.1. Intelligent risk management

Text mining is the process of extracting meaningful information such as patterns and trends from text. It has been adopted in the risk management domain for various tasks including risk identification, risk categorization, and risk list generation etc. [12,28,48]. The main goal is to replace labor-intensive manual works with automatic models, enable large-scale analysis, and present more structured data for project managers, which are achieved through three broad approaches: 1) text classification, 2) clustering, and 3) information extraction. Text classification is the task of assigning predefined classes to text documents. For instance, Hassan and Le [18] classified contractual clauses into three types of requirements to prevent ignoring requirements and thus assist contract risk management. The study compared various classification models including both shallow learning methods such as logistic regression and deep learning neural networks and showcased the capability of NLP in classifying general contract requirements. In addition to machine learning-based classification approaches, rule-based approach can also be used for classification. For example, Lee et al. [25] built a domain ontology and defined patterns based on sentence structure such as subject-verb-object (SVO) to identify poisonous contract clauses and map them to 11 risk types. The rule-based method allows the integration of domain knowledge, is transparent and interpretable, and can achieve high accuracy. However, crafting and updating the rules over time is labor-intensive.

Clustering refers to finding groups of similar documents from a collection of unstructured data and can assist in document retrieval. One of the common methods used for text clustering is to calculate the similarity between text representations. Text representations can be a vector of word counts (BOW) or more refined representations such as Word2Vec. For example, Jallan and Ashuri [21] grouped risk disclosures from Securities and Exchange Commission (SEC) Financial Filings under 18 risk types by calculating the cosine similarity between the disclosures and risk types defined by a set of keywords. Similarly, Erfani and Cui [13] grouped risk items in project risk registers based on semantic similarity and generated the risk register template for a given new project defined by a set of characteristics such as project type and size in dollar values. Information extraction relates to automatically extracting structured data from unstructured texts, including Name Entity Recognition (NER) and relation extraction [2]. NER locates and extracts a sequence of words that is a single entity such as organizations, while relation extraction identifies semantic relationships between two or more entities in the text documents. For example, Jeon et al. [23] proposed a method based on defect thesaurus and pre-trained language models to extract 23 classes of defect named entities from building quality complaints. To enable quick access and efficient use of knowledge contained in construction contracts, Al Qady and Kandil [1] extracted concepts (i.e., noun phrases and prepositional phrases) and relations (verb phrases) from construction contract documents using shallow parsing.

Existing studies showcase the successful application of a variety of text mining approaches in the risk management domain. However, these studies focus on identifying and categorizing risks in technical documents such as contracts and specifications which are of limited availability and uniform writing. There is less research on extracting risk information from open source and noisy text such as tweets and news articles [11,55]. Also, most existing research focuses on specific types of risk including contractual risk and safety risks. There is a lack of studies identifying the wide variety of risks that can occur during the full lifecycle of project development. Therefore, it is necessary to establish an intelligent system that can automatically detect project-level risk events from a broad range of textual data such as news articles. Since the goal is to extract risk-related information, a text classification approach that differentiates risk from non-risk sentences in the news text will be developed. Specifically, this paper chooses a machine learning based classification approach over the rule-based approach for three reasons: firstly, there are various types of risks and each of these risks can be expressed in numerous ways, which makes it difficult and cumbersome to speculate rules to differentiate risk narratives from non-risk narratives. Secondly, the rule-based method is highly dependent on the predefined lexicon size and quality. It may not be able to identify new types of risks if their vocabulary is not included in the lexicon. Thirdly, the machine learning based method is more scalable. In other words, it can be easily updated and its accuracy may further increase when new data becomes available.

2.2. Text classification techniques and strategies

Machine-learning based text classification methods can be grouped into two categories: shallow learning and deep learning methods. Shallow learning models such as Naive Bayes (NB), Logistic Regression (LR), and support vector machine (SVM) have dominated in the earlier text classification studies due to the limitations of computation and data [27]. These approaches are also easy to interpret and implement. They usually follow two steps. Firstly, text data are transformed into feature vectors using methods such as Bag of words (BOW) and their variations. Feature engineering and analysis are usually performed to truncate the vocabulary, reduce the dimension of feature vector, and obtain better performance. Secondly, the extracted features are input into a classifier

to map between the features and classes. Several studies have investigated such approaches for text classification in the construction domain [17,32,40]. Hassan and Le [17], for example, compared four machine learning algorithms (i.e., NB, SVM, LR, and feedforward neural network) to classify the contractual text into requirement and nonrequirement. Likewise, Williams and Gong [47] developed a classification model combining text and numerical data to predict the level of cost overruns. The shallow learning approach has several limitations. For instance, it is mainly based on the word/term occurrence and loses the word order information. In addition, the feature engineering process is time-consuming and the selected features may not generalize well to new data or tasks [35]. Although some risk sentences can be identified based on the occurrence of explicit keywords such as delay and utility relocation, text sequence information plays an important role in the classification of more complex sentences, which renders the shallow learning methods insufficient.

Compared with shallow learning, deep learning methods can preserve the sentence structure and capture contextual information. These methods are also an end-to-end procedure where representations are learned from an extensive training database without domain expertise for feature engineering. Deep learning methods have become more popular in recent years due to the availability of a large amount of training data in general domain and the increased computational power. For example, word2vec is trained on 6 billion words and outperforms shallow learning models for many NLP tasks [34]. More recently, pretrained large language models such as Bidirectional Encoder Representations from Transformers (BERT) [10] are developed to effectively capture the semantics of words in context and allow further fine-tuning in downstream classification tasks. For example, Tian et al. [44] improved upon the pretrained BERT model by adding a BiGRU to enhance the global feature information and a self-attention mechanism to strengthen local feature information. Zhang et al. [52] proposed a multi-feature channels Convolutional Neural Network (CNN) model which inputs the BERT embeddings and word2vec embedding into a multi channels CNN to automatically classify construction quality records.

Models that can capture context at the word-level and at the sentence-level have been explored extensively in various domains. However, most existing deep learning classification models in the construction domain overlook the importance of context in modeling text or can only exploit contextual information within a sentence. Such models are not ideal for construction risk identification because without context, differentiation between risk vs. non-risk can be ambiguous and challenging even for humans. Specifically, identifying risk requires considering both a potentially adverse circumstance or event and its impact on cost or schedule, which are often spanned over multiple sentences. While an individual sentence could include both elements, the surrounding sentences play a crucial role in risk classification in the target sentence.

To address these challenges, this research introduces MCRI, a novel model that integrates a pretrained language model RoBERTa and a BiGRU encoder. RoBERTa is pretrained on large amounts of data and the knowledge gained during the pretraining process is transferred to the construction domain. To prevent overfitting, only the final layer of RoBERTa is fine-tuned. The addition of the BiGRU structure on top of the sentence-level representation infuses multi-sentence level context information and compensates for the inadequate information of an individual sentence for risk classification. Furthermore, by experimenting with models that capture context at various levels, this research showcases the incremental performance gains achieved through enhanced contextual understanding. Given that context is also crucial for modeling other construction-related texts, such as contract clauses or project reports where the surrounding sentences are vital for comprehension, this research sheds light on the effective use of context information. The model developed in this study can be adopted by other text classification tasks.

3. Methodology

Fig. 1 shows the overall research framework of this paper. There are mainly three steps involved: 1) develop an automatic risk paragraph extraction program to collect potential risks scattered in numerous news articles and speed up the data labeling process; 2) preprocess raw texts and build MCRI to capture multi-sentence level context; and 3) train the model and validate its effectiveness.

3.1. Data preparation and preprocessing

Labeled datasets for automatic identification of project risks from public data such as news are not readily available. Furthermore, based on the definition of risks mentioned earlier, the labeling requires risk management expertise in the construction domain. This requirement ruled out the option of using crowdsourcing to build large datasets [3]. Compared to research in financial or corporate news, there are no such news categories dedicated to construction, not to mention a category for construction projects. Furthermore, risk information is often scattered in numerous articles and the number of risk sentences per article is small, rendering the necessity of a filtering program to boost the density of risks. Therefore, this study created a three-step process to develop the training dataset: 1) target and download project risk-relevant news articles using the searching string; 2) automatically extract potential risk paragraphs from the downloaded news articles; 3) assign labels (risk or non-risk) to each sentence in the extracted paragraphs. The following sections 3.1.1–3.1.3 explain each step in detail. Fig. 2 demonstrates the three steps just mentioned using an example news article from Nexis Uni published by the Engineering News-Record (ENR).

3.1.1. Collecting news articles

To target the project risk-relevant news articles from a large number of miscellaneous news articles, different search methods are experimented using two of the largest news databases, namely Nexis Uni (the educational version of LexisNexis) and ProQuest. The research used projects' names along with the word "construction" and any of the following terms: delay, schedule, budget, cost, estimate, and contingency as search strings in both databases. A set of 2445 news articles, ranging in date of publication from 1969 to 2021, were collected after manually screening of the search results to keep project risk-relevant news. These news articles originated from a wide range of publishers including both national news publishers such as the New York Times and local news publishers such as Pittsburgh Tribune-Review.

The news articles are downloaded in PDF format. Duplicated news articles are deleted first. Except for the news title and body text, which are of interest for this study, each news article contains unnecessary information (noise) such as the name of publisher, publishing date, and URL (Fig. 2). Regular expression (RE) patterns are defined to extract news sentences and remove noise. The body text is then parsed into individual sentences. Usually, a combination of specific punctuations including a period (“.”), an exclamation mark (“!”), or a question mark (“?”) along with a blank space character indicates the ending of a sentence. However, special instances in the news article such as unit “3-in.” or abbreviations such as “dept. of transportation” and “Rep. Mike Doyle” could lead to false parsing. In order to cope with this issue, a list of extra abbreviations summarized from construction news is supplemented to the default abbreviation list in the sentence tokenizer of the Natural Language Tool Kit (NLTK) package.

3.1.2. Extracting potential risk paragraphs

In construction engineering and management, risk is defined as a potentially adverse circumstance or event that can cause undesired cost growth or time delays. The definition reveals that risk has two basic dimensions including negative impact on project performance and uncertainty [30]. Since news articles are written to inform the public about the project progress rather than specifically reviewing the project risks,

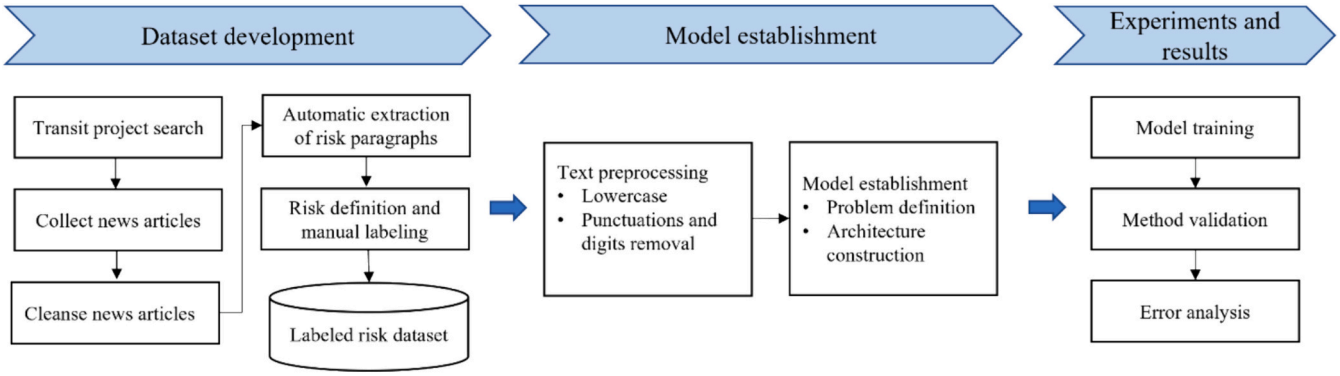


Fig. 1. Research steps.

the proportion of risk sentences in each document is low. While there are a few sentences discussing project risks and their impacts on the performance, most sentences are non-risk and give background information about the project. Furthermore, there are many types of risks, requiring a relatively large dataset to represent/cover the features of each category. As a result, it is impractical to go through each individual sentence in the news articles to build the training dataset. Thus, a filtering program is devised to extract possible risk paragraphs first. The filtering program helps to boost the density of risks and balance the number of risk and non-risk sentences in the labeling dataset. Through manually reviewing a sample of news articles, it is observed that when a risk event is reported, its impact on the project's performance is also discussed. In other words, performance indicating words (cost and schedule descriptions) occur either in the risk narratives or in their surrounding sentences. Given this, the lexicon describing cost overruns and schedule delays was built for the extraction of relevant paragraphs.

The filtering program includes two rules. The first rule depends on the appearance of a single item; this list of terms contains performance indicating words and phrases such as "cost overrun", "delay", "on schedule" (Fig. 3). The authors expected that rule #1 would be insufficient because the writing in news articles is less uniform and often purposefully colorful to make them more readable compared to technical documents. Therefore, rule #2 was created to expand the extraction power by allowing a wide range of term combinations. The second rule depends on the simultaneous appearance of any item from two lists: one list consists of time and cost specific words such as "time", "progress", "milestone", "budget", "forecast". Another list consists of generic words including trending words such as "increase", "insufficient", "escalate", negative words such as "challenge", "complexity", and sources of risk such as "utility", "inflation". The initial lists under both rules are developed based on reviewing news articles, brainstorming, and literature review. Wordnet from NLTK library is used to enhance the lists by supplementing synonyms to the terms in the list [5]. Through running the extraction in a random sample of articles and inspecting the extracted paragraphs, the lexicon was finalized, consisting of 140 unigrams and 7 bigrams.

Considering that the risks may occur in the surrounding sentences and the context of the sentence is important for interpretation, the sentence containing performance indicating terms is extracted, along with i sentence(s) before and after. The selection of i is a tradeoff. If the number is too low, important discussions on risks could be omitted. Table 1 compares two scenarios: $i = 1$ (3-sentence group) and $i = 2$ (5-sentence group). It can be seen that risk sentences are missed out in the 3-sentence group. However, if the number is too high, the study may suffer from the extraction and labeling inefficiency and reaches a highly unbalanced dataset where non-risk sentences overwhelm risk sentences. For example, $i = 3$ (7-sentence group) would result in around 40% more sentences being extracted compared to $i = 2$. After reviewing extracted paragraphs from a sample of 27 articles using different i values, it was determined that 2 surrounding sentences, i.e., 5 sentences as a group, can satisfactorily capture the context and cover the associated risks leading to cost overrun and schedule delay. Adjacent sentence groups (as shown in Fig. 2) and overlapped sentence groups are merged into bigger groups while duplicated sentences were removed. Out of 2445 news articles, 1122 articles had paragraphs extracted.

3.1.3. Labeling

The researchers went through the extracted paragraphs and annotated each sentence as risk or non-risk based on the risk definitions mentioned in the previous section (as shown in Fig. 2). Though the labeling unit is an individual sentence, the annotators were asked to assign the most appropriate label with consideration of the surrounding sentences from the same news article. For example, "Although the numbers of executive staff declined to eight from nine, executive office salaries increased 56 percent to $1.2 million from $790,000." Without the context information for this sentence, it is hard to link the staff salary increase to the rail project's cost overrun and decide the corresponding label category. However, given the previous sentence "The budget calls for a $2.1 million increase to pay 137 full-time HART employees, including 33 design and construction workers …", the causal relation between the current sentence and project's cost overrun becomes clear and the corresponding label should be "risk".

Two rounds of pilot labeling were conducted before labeling the whole dataset. The purpose is to examine if the annotators have the same understanding of the definition and iteratively improve the definition. Cohen's kappa score is used to measure inter-annotator agreement [9] calculated by Eq. (1). Simple agreement is the proportion of sentences in which all annotators assigned the same label whereas chance agreement is the proportion of sentences in which agreement is expected by chance.

$$\text{Cohen's kappa} = \frac{\textit{Simple agreement} - \textit{Chance agreement}}{1 - \textit{Chance agreement}} \quad (1)$$

After the first round of pilot labeling, the kappa score was calculated and a score of 0.66 (moderate agreement) was obtained. The authors discussed the problems and further clarified the labeling rules. Specifically, it was emphasized that the annotators should assign label based on the meaning of the sentence and its context and should not overinterpret or imagine the risk events separated from the sentence itself when labeling. After the clarification, the authors conducted a second-round pilot labeling and the kappa score reached 0.80 (strong agreement). Then the authors labeled a subset of the extracted paragraphs and classified 1148 sentences as risk sentences and 1064 sentences as non-risk sentences.

3.1.4. Text preprocessing

Text preprocessing is a necessary step before inputting the labeled raw text into the classification model where text is cleaned and sharpened by removing the unhelpful parts of the data. The processing

Step 1 Collect news articles

Tunneling work delayed since last fall, first by a collapse and later by contaminated soil, will resume soon on a $ 321-million Baltimore Metro line.

Frank Hoppe, project director at the Maryland Mass Transit Administration (MTA), says the delay began when sandy soil at the face of one tunnel slid into the tunneling machinery, creating a void that led to the collapse of a small section of a street.
Soon after repairs were made and work resumed, tunnelers encountered a layer of gasoline-soaked soil that threatened to ignite or to sicken workers with its fumes.

Since then, the joint venture of Kiewit Construction Co., Omaha, and J.F. Shea Co., Walnut, Calif., that won the $ 70-million tunneling contract has installed large ventilation pipes in the area. The contractor also has revamped its tunneling shield to prevent slides and refitted its tunneling equipment with explosion-proof electronics that generate no sparks.

The nine-month-plus delay may cost as much as $ 20 million but overruns will be absorbed by a $ 53-million "cushion" for contingencies, according to estimates prepared by the MTA. The federal government is providing 85% of the project funding; the rest is being paid by state and local governments.

To make up for lost time, Hoppe says MTA will begin to issue overlapping contracts with the hope of finishing the project by late 1994, rather than in mid-1995, as was initially feared. The line, which will connect the existing Charles Center station and Johns Hopkins hospital, had been scheduled to be finished by June 1994.

URL: http://www.enr.com

Step 2 Extract risk paragraph automatically

Extracted Paragraphs
Tunneling work delayed since last fall, first by a collapse and later by contaminated soil, will resume soon on a $ 321-million Baltimore Metro line. Frank Hoppe, project director at the Maryland Mass Transit Administration (MTA), says the delay began when sandy soil at the face of one tunnel slid into the tunneling machinery, creating a void that led to the collapse of a small section of a street. Soon after repairs were made and work resumed, tunnelers encountered a layer of gasoline-soaked soil that threatened to ignite or to sicken workers with its fumes. Since then, the joint venture of Kiewit Construction Co., Omaha, and J.F. Shea Co., Walnut, Calif., that won the $ 70-million tunneling contract has installed large ventilation pipes in the area.
The contractor also has revamped its tunneling shield to prevent slides and refitted its tunneling equipment with explosion-proof electronics that generate no sparks. The nine-month-plus delay may cost as much as $ 20 million but overruns will be absorbed by a $ 53-million "cushion" for contingencies, according to estimates prepared by the MTA. The federal government is providing 85% of the project funding; the rest is being paid by state and local governments. To make up for lost time, Hoppe says MTA will begin to issue overlapping contracts with the hope of finishing the project by late 1994, rather than in mid-1995, as was initially feared.

Step 3 Label sentences

Sentence	Label
Tunneling work delayed since last fall, first by a collapse and later by contaminated soil, will resume soon on a $ 321-million Baltimore Metro line.	Risk
Frank Hoppe, project director at the Maryland Mass Transit Administration (MTA), says the delay began when sandy soil at the face of one tunnel slid into the tunneling machinery, creating a void that led to the collapse of a small section of a street.	Risk
Soon after repairs were made and work resumed, tunnelers encountered a layer of gasoline-soaked soil that threatened to ignite or to sicken workers with its fumes.	Risk
Since then, the joint venture of Kiewit Construction Co., Omaha, and J.F. Shea Co., Walnut, Calif., that won the $ 70-million tunneling contract has installed large ventilation pipes in the area.	Non-Risk

Fig. 2. Data preparation process.

```
delay_lexicon=['delay', 'hold_up', 'holdup', 'holdup', 'postponement', 'timeextension', 'time_lag', 'timely', 'behindschedule', 'prolong', 'lengthen', 'ontime', 'staywithin', 'stayon', 'push back']
```

Fig. 3. Sample list of indicating keywords and phrases of delay.

Table 1
Comparison of different extracted sentence group.

News Title	Extracted sentences (3-sentence group)	Extracted sentences (5-sentence group)
Legal fight threatens to slow light rail construction	(1) Qwest Corp. is suing Sound Transit and the City of Tacoma, demanding that the tri-county transit agency pay an estimated $5 million to $7 million to move thousands of phone lines from under Commerce Street. (2) Sound Transit is countersuing, asking for an injunction requiring the Denver-based phone company to move its equipment immediately and pay for any construction *delays*. (3) "We're not going to cave into them," said Pierce County Executive John Ladenburg, also a Sound Transit board member.	(1) A **federal lawsuit** over who should **pay to move buried phone lines** could slow construction of a downtown light-rail line here - the first segment of Sound Transit's light-rail network to break ground. (2) Qwest Corp. is suing Sound Transit and the City of Tacoma, demanding that the tri-county transit agency pay an estimated $5 million to $7 million to move thousands of phone lines from under Commerce Street. (3) Sound Transit is countersuing, asking for an injunction requiring the Denver-based phone company to move its equipment immediately and pay for any construction *delays*. (4) "We're not going to cave into them," said Pierce County Executive John Ladenburg, also a Sound Transit board member. (5) "We're going to fight them all the way."
And, as always, obstacles in the transit path	(1) Corridor planners and university leaders have been oh-so-slow to come to an understanding about how best to **keep vibration and electromagnetic interference from disrupting research in buildings** on Washington Avenue, adjacent to the proposed rail line. (2) They're just about out of time if the project is going to stay *on schedule* and *within budget*. (3) The $940 million project counts on receiving half of its funding from the federal government.	(1) The most recent source of Bell's aggravation has been a **dispute with the University of Minnesota**. (2) Corridor planners and university leaders have been oh-so-slow to come to an understanding about how best to **keep vibration and electromagnetic interference from disrupting research in buildings** on Washington Avenue, adjacent to the proposed rail line. (3) They're just about out of time if the project is going to stay *on schedule* and *within budget*. (4) The $940 million project counts on receiving half of its funding from the federal government. (5) **Unless that money is in President Obama's 2010 budget, this railroad won't be running as scheduled in 2014.**

Note: the italic font refers to the performance lexicon and the center sentence. Bold fonts indicate risk sentences.

approach depends on the adopted language model and downstream task thus varies from study to study. This research implemented two data cleaning techniques. Firstly, since the risk identification task is case insensitive, all the input text is converted into lowercase. Similarly, punctuation and digits are removed since they are not contributing to the text classification.

3.2. MCRI model

In order to automatically identify risks from massive news articles, the research must abstract the task into a text classification problem and builds a novel model architecture that can address unique challenges associated with construction risk identification. The following section articulates the problem definition and model architecture in detail.

3.2.1. Problem definition

The input corpus $D = \{(P_n, Y_n)\}_{n=1}^{N}$ consists of N text segments, where $P_n = \langle s_t^n \rangle_{t=1}^{T}$ is a segment instance containing a sequence of T sentences, $Y_n = \langle y_t^n \rangle_{t=1}^{T}$ is the corresponding risk labels. The input corpus is a collection of text documents such as news articles or post-project reviews. The text segment can be an individual paragraph or an arbitrary number of consecutive sentences. The goal is to learn a classification model from the corpus D, such that given an unseen text segment P_i, the model can predict the risk label Y_i of sentences in P_i.

3.2.2. Model architecture

Fig. 4 shows the overall architecture of the MCRI model, which involves two main components: 1) a sentence-level embedding model that converts the preprocessed text into numerical vectors; and 2) a multi-sentence level encoding model that informs each sentence of its context to generate context-aware embeddings. The embeddings are then fed into a fully connected layer to predict sentence labels. Each component is explained in detail below.

3.2.3. Sentence embedding

The first step in text classification model is to convert the text into machine readable information, i.e., numerical vectors. BERT (Bidirectional Encoder Representations from Transformers) is currently the state-of-the-art language model used in various NLP tasks including text classification [10]. It is firstly trained on a large cross-domain corpus and then followed by a specific fine-tuning task. BERT model can encode the order of word presence in a sentence and retrain the contextual meaning. By using an attention mechanism, BERT is able to read the text input as an entire sequence at once rather than reading it sequentially. It can generate different embeddings for the same word in different contexts to address the polysemy issue. BERT models have evolved over time and several variants have been created. RoBERTa, known as a 'Robustly Optimized BERT Pretraining Approach', is one of the most popular BERT variants and was developed to improve the pretraining phase by longer model training, more training data points and larger batch size [29]. Thus, this research directly employs the pretrained RoBERTa base model with 12 encoder layers/transformer blocks to leverage the power of transfer learning. Given a sentence s, the embedding of $[CLS]$ token from the last layer is used as the sentence embedding $e(s)$. $[CLS]$ stands for classification and is a dummy token added at the start of each sentence to represent the meaning of the entire sentence. Researchers find that fine tuning only a few of the final

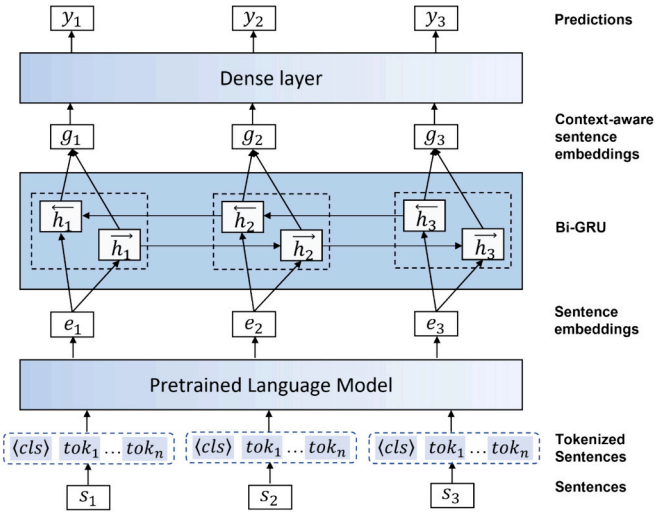


Fig. 4. MCRI Model Architecture.

encoder layers of pretrained language models is sufficient for achieving high quality on domain specific tasks [26]. Therefore, to fine tune RoBERTa for the risk identification task, this study updates the weights in the final encoder layer and freezes the bottom 11 layers. This practice can avoid model overfitting resulting from updating all 12 layers, 125 million parameters. It also helps boost the training speed and reduce memory usages.

3.2.4. Context-aware sentence embedding

The risk definition mentioned before indicates two important elements: a potentially adverse circumstance or event and its impact on cost or schedule. In other words, the appearance of both elements (event and its impact on cost/schedule) is needed to qualify a sentence as risk sentence. Though an individual sentence could include both elements, they are commonly spanned over multiple sentences. Thus, the surrounding sentences play an important role when performing the risk classification in the target sentence. For example, "Thirteen utility lines under Stanwix Street are not where utility maps show them to be, and finding them will add up to $90,000 to the project's cost. The increases represent the first serious jumps in the cost of the North Shore Connector since work started in February." The second sentence talks about the adverse impacts on cost but the cause or the potential risk factor is missing. The first sentence is required for identifying the risks. Another example is "A tentative agreement was negotiated in December under which the federal government would pay two-thirds of the bonds and the local governments one-third, but the details have never been worked out." This sentence could be talking about funding risk though it is ambiguous by itself. Specifically, it is a general discussion that happened in the early project phase and without contextual information, it is hard to judge if the circumstances affected any ongoing project.

To solve the ambiguity, the multi-sentence level context needs to be accounted for. However, BERT and RoBERTa only encode the contextual information at a sentence level. To capture the interactions between sentences within a document, a BiGRU is needed to extract additional features from the surrounding sentences and integrate them into the final feature representation of the target sentence [19]. As a variant of recurrent neural networks (RNN), BiGRU consists of two GRUs, one taking the input (i.e., the sentence embeddings e_t) in a forward direction, and the other in a backwards direction [8]. Both the forward and backward GRU are also connected to the same output layer (g_t) as shown in Fig. 4. The forward flow can inform the target sentence of context from preceding sentences while the backwards flow can inform the target sentence of context from succeeding sentences. Each GRU contains two gates: an input gate and a forget gate, which help control the flow of information and learn and preserve the important data in a sequence, thus overcomes the vanishing gradient problem of RNN. The hidden state (h_t) acts as the neural network memory and holds information on previous data that the network has seen. BiGRU has proved to be suitable for modeling context [19,39]. Specifically, given a sequence of independently encoded sentence embeddings e_t in a context group/text segment, $GRU(e_t)_{t=1}^{T}$ are the context-aware sentence embeddings from the hidden state of the BiGRU model.

Nearby sentences are usually talking about the same topic and contextually dependent while remote sentences don't contribute to the classification of the target sentence and could even add noises. In other words, not all sentences from the same news article are qualified to be in the same group. This study uses the sentence group/segment size T to control the amount of effective contextual information. Specifically, all the extracted potential risk paragraphs from the same news article are sorted in correct order and then sliced into fixed-length segment size T and pad all sentence group to the maximum size T by putting zeros. Fig. 5 shows an example of slicing a news article of 10 sentences into 3 sentence groups where $T = 4$. Different T values ([4, 8, 12, 16]) have been experimented to find the optimal sentence group size, i.e., context length. In the MCRI model, each sentence group would be treated as a data point.

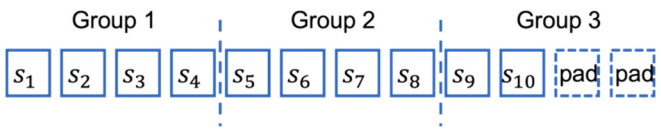


Fig. 5. Example of slicing into context group of size 4.

3.2.5. Dense layer

The context-aware sentence embeddings (g) are fed into a fully connected layer/dense layer to predict sentence labels. The dense layer learns the full set of weights in a linear function (Eq. 2) to map the high-level text features to the output (y) and make the model end-to-end trainable. Here, A is the learnable weights and b is the bias. The study uses cross-entropy loss to update and optimize the weights.

$$y = gA^T + b \tag{2}$$

3.3. Evaluation

The model is evaluated using recall, precision, F1-score, and accuracy, calculated using Eqs. (3) to (6). Since this research focused on identifying the risk sentence from news articles, the metrics except accuracy are calculated based on the risk class. Specifically, TP (True Positive) is the number of risk sentences correctly classified as risk, FP (False Positive) is the number of non-risk sentences incorrectly classified as risk, TN (True Negative) is the number of non-risk sentences correctly classified as non-risk, and FN (False Negative) is the number of risk sentences incorrectly classified as non-risk. F1-score is a comprehensive measure since it takes both FP and FN into consideration.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$\text{F1} - \text{score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

K-fold cross validation is used to calculate the average performance measures of the k times experiments' result. Compared to fixing an arbitrary testing set, k-fold validation does not waste data and is more reliable [41]. Specifically, the modeling data is randomly split into k sets of approximately equal size, with the model trained on k-1 sets and tested on the remaining set. This process repeats k times and thus k results of the metrics are obtained, which allows to investigate the mean and variance and have a deeper insight into the model's performance instead of using a single result. Also, since the dataset is small, the application of k-fold cross validation would avoid the stochasticity of results caused by the different training/test split. This study uses a common k value of 5 [22]. Therefore, 80% of the all the data is used for training while 20% is used for validation.

4. Experiments and results

This section presents the performance of the developed model and demonstrates the effectiveness of leveraging contextual information in the risk classification task. Firstly, the baseline models and experiment set up are described. Then, the study analyzes different text representations methods and classifiers to provide empirical evidence on their applicability for risk classification. These models were then used as baseline to verify the advantage of the proposed MCRI model.

4.1. Baseline models and experiment setup

Both shallow learning models and deep learning models were selected as baseline models. For shallow learning models, the combination of Term Frequency–Invert Document Frequency (TF-IDF) embedding and Logistic Regression (LR) was selected as a representation. For deep learning models, the combinations of two word embedding methods (Word2Vec and GloVe) and two classifiers (CNN and LSTM) were implemented.

- Both Word2Vec and GloVe are distributional similarity-based representations that emphasize the context as opposed to the definition. They can capture both the semantic and syntactic information of words. Semantic information mainly refers to the meaning of words and syntactic information is their grammatical functions. The basic assumption is that words appearing in similar context should share similar meaning [16]. Word2Vec was first introduced in 2013 and employs a two-layer neural network for training embedding vectors [34]. GloVe extends Word2Vec by using global word-word co-occurrence counts statistics [38]. This study uses the Word2Vec embedding (300 dimensions) pre-trained on Google News dataset (about 100 billion words) and GloVe embedding (100 dimensions) pre-trained on Wikipedia and newswire text data (about 6 billion words).
- The advancement of CNN lies in the convolution operation, achieved by a convolutional layer and a pooling layer. The CNN model can be intuitively understood as a combination of two parts: 1) the convolution + pooling layers perform feature extraction; 2) the fully connected layer performs classification using the extracted features. In the context of sentence classification, CNN can automatically detect the location of critical terms that determine whether a sentence contains a risk [24,53].
- The Long Short-Term Memory (LSTM) network [20] is a modified version of RNN. Similar to the BiGRU, LSTM can model the interdependence between inputs since it processes sequences by retaining the memory of the previous value in the sequence. Since text is sequential data, LSTM has been proved to be suitable for text classification tasks [54].

All the models were written in Python 3.8 and the main packages used included Scikit-learn and PyTorch 1.12. NVIDIA GeForce RTX 2070 GPU were used to train and run models. Most model parameters were selected through grid search while others were chosen based on the literature [19]. Table 2 shows the choices of hyperparameters.

4.2. Results

4.2.1. Analysis of text representation methods and classifiers

In this section, the research compares the results of baseline models including different combinations of text representation methods and machine learning classifiers (Table 3). The purpose is to provide empirical evidence and advice for future model design in the area of automatic risk identification and showcase the benefits of deep transfer learning models in the context of small data size. To start with, the effectiveness of shallow learning and deep learning classifiers are compared. The result shows that although Logistic Regression is a shallow learning model, its performance is comparable to CNN and LSTM. A plausible explanation is that at the data scales of 1000+ labels per class, the shallow learning models can outperform deep learning approaches since the limited dataset is not enough to train deep models from scratch and achieve superior performance [6]. Since running LR is very simple and requires minimum computing resources and time, it is advisable to use LR as a starting point in experiments. Moreover, if the task is straightforward and the achieved performance is satisfactory, there would be no need to further experiment with deep learning models given its efficiency. As for deep learning classifiers, LSTM performs slightly better than CNN. As mentioned before, LSTM can better capture the interdependence between words and semantics in the text sequence while CNN works better for tasks where feature detection is more important, for example, Part-of-speech (POS) tagging and named entities recognition (NER) [51]. Since risk identification usually requires the information of words presented in the earlier part of the sentence to understand the current terms, LSTM is more suitable for this task.

The study also investigated the most suitable methods for modeling construction risk sentences. Five most popular methods were examined, including shallow learning (TF-IDF), pretrained embeddings (GloVe, Word2Vec), and transfer learning models (BERT, RoBERTa). Comparing the best performing model in each category, it can be seen that the F1-score increased by 2.43% from TF-IDF to Word2vec and 8.46% from Word2vec to RoBERTa while the corresponding gains for accuracy are 3.04% and 7.71%, respectively (Table 3). Word2Vec in general performs better than GloVe in this research. The reason might be that Word2Vec is pretrained on Google News data which is closer to the dataset used in this study.

In addition, these results demonstrate the utility of transfer learning models such as BERT and RoBERTa. Specifically, knowledge gained from massive cross domain pretraining can be adapted and transferred to the downstream tasks in the construction domain. This aligns with the results from previous studies in both the construction [52] and general domain [27]. The performance gains of RoBERTa over BERT are expected since the former is trained on a larger dataset and uses a dynamic masking scheme to make the model more robust and leads to better downstream task performance. Overall, the results show that word embedding methods improve upon TF-IDF by reducing the dimensions of embedding vectors and encoding the context information at an individual word level. Transfer learning models further improve the word embedding methods by taking into account the word position and encoding the context information at an individual sentence level.

4.2.2. Impact of context information

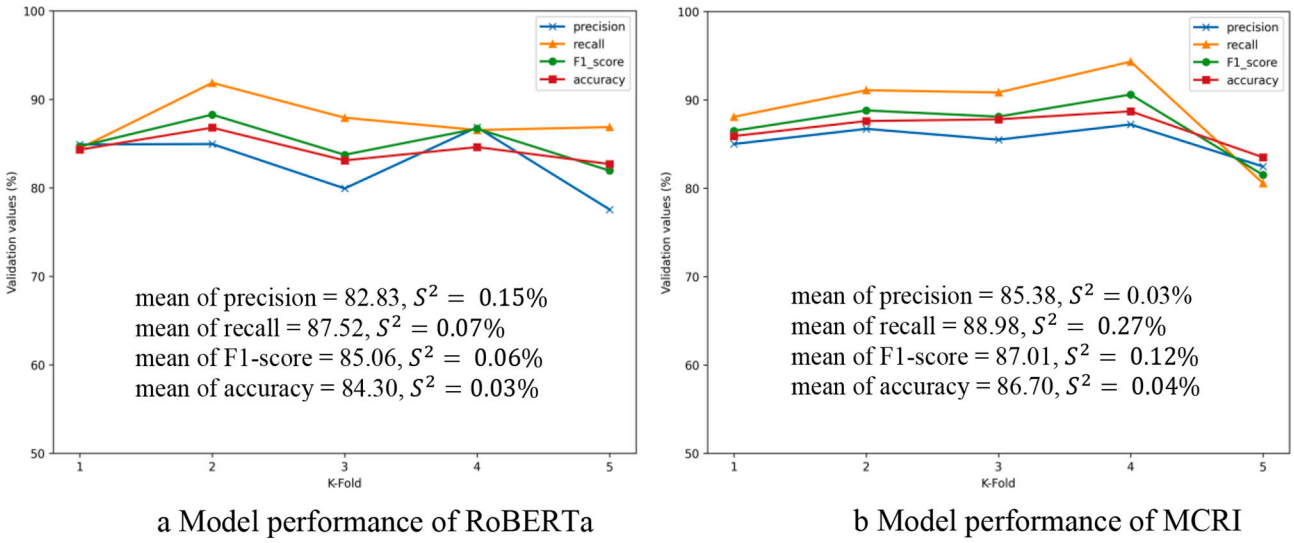
The results from the previous section have proved the importance of context in the automatic risk identification task. Though BERT and RoBERTa showed superiority over other models by taking a sentence rather than a single word as input and generating vector representation for the entire input sentence, it only exploits contextual information within a sentence. Given that in the risk identification task, relevant information from surrounding sentences is key for categorizing the current sentence, the research added the BiGRU to create the final feature representation of the sentence that encodes the multi-sentence level context information. Specifically, in this section, the study compared models with and without multi-sentence level context to inspect its impacts. Fig. 6 shows the performance obtained from each run in the 5-fold cross validation for both RoBERTa and MCRI models. It demonstrates that the mean F1-score is improved by 2.05% and the mean accuracy is improved by 2.4% through introducing the multi-sentence level context information, i.e., using the BiGRU to capture the useful information from the surrounding sentences and then adding them to the target sentence. Also, MCRI model boosts both the precision and recall, illustrating that the added context helps with reducing both false positive errors (i.e., non-risk sentences being predicted as risk sentences) and false negative errors (i.e., risk sentences being predicted as non-risk sentences). The result can be explained from two aspects:

Table 2
Choices of some hyperparameters and their experiment values.

Hyperparameter	Choice	Experiment values
Learning rate	0.0001	0.0005, 0.0001, 0.00005
Dropout	0.5	0.5, 0.6
Max seq length	64	64, 128
Batch size	16	16, 32
Number of BiGRU layers	1	1, 2

Table 3
Model performance of different text representation methods and classifiers.

	TF-IDF + LR	CNN		LSTM		BERT	RoBERTa
		GloVe	Word2Vec	GloVe	Word2Vec		
Precision (%)	74.16	75.56	79.55	78.53	78.52	80.72	**82.83**
Recall (%)	74.21	68.25	73.08	70.44	75.10	83.51	**87.52**
F1-score (%)	74.17	71.55	76.10	74.14	76.60	82.01	**85.06**
Accuracy (%)	73.55	72.22	76.53	74.78	76.59	81.25	**84.30**



a Model performance of RoBERTa b Model performance of MCRI

Fig. 6. a Model performance of RoBERTa.
Fig. 6b Model performance of MCRI.

firstly, it is necessary and effective to incorporate the context or sequence information in classifying sentences [49,50]. For example, if the current sentence is talking about cost increase or schedule delay, then the preceding or the next sentence is most likely talking about a risk event. Secondly, the BiGRU structure can effectively capture the context-related semantic information and model the sequential relations between sentences. Specifically, it can filter useless semantic information from contextual sentences while enhancing the influence of key information on the classification results, thus mitigating the problem that invalid surrounding sentence information introduces noises to the target sentence representation.

Fig. 6 also shows that the ranges of values produced from different folds are small for both models. To better understand the learning variance and the generalizability of the proposed MCRI model and the RoBERTa model, the variance (S^2) of each evaluation metric was calculated using the cross validation results. For both models, S^2 ranges from 0.03% to 0.27%, which are small and provide confidence in the model's capability of performing in new data.

4.2.3. Impact of the context length

As mentioned in section 3.2, the context length/size of sentence group controls the amount of contextual information to be accounted for in the classification task. If the size is too small, crucial information necessary for understanding the current sentence may be omitted. Also, large group size means that more context information is available for the BiGRU model to leverage [19]. However, if the size is excessively large, the model may struggle to learn long-range dependencies due to nosier and longer path lengths [39]. Specifically, gradient vanishing and gradient explosion problems could happen. Table 4 shows the results of using different context lengths. It is observed that precision and accuracy exhibit a weak trend of initially increasing and then decreasing as the group size increases. While the performance difference between group size of 4 and group size 8 is minimal, the most accurate group size of 8 is selected since error in predicting both risk and non-risk class is equally important in this study.

Table 4
MCRI model performance using different context lengths.

Group size	4 sentences	8 sentences	12 sentences	16 sentences
Precision (%)	84.13	85.38	86.19	84.30
Recall (%)	90.59	88.98	87.00	88.81
F1-score (%)	87.22	87.10	86.53	86.45
Accuracy (%)	86.54	86.70	86.33	85.89

4.3. Examples of risks from news vs project reports

This section showcases the utility of news articles as a supplementary source of project risk information through comparing risks identified from news articles versus risks reported by official project documents, i.e., Before and After studies (B&A studies) by FTA. As shown in Table 5, most risks from the B&A studies are in fact reported in the news and identified by the proposed MCRI model. Also, new articles could supplement expert evaluations through providing more granular risk information which is not covered by the B&A studies. Moreover, since B&A studies are unobtainable for most projects, risk information collected from news articles can fill the knowledge gap for these projects. It's worth noting that the scope of comparison performed here is restricted. Future studies can conduct a more comprehensive comparison between risks reported in news articles and official project documents if there is access to a sufficient number of project reports.

Table 5
Examples of risks from news articles and B&A studies.

Project names	Risks extracted from news articles	Risks reported in B&A studies
South Corridor Light Rail – Charlotte, North Carolina	• hurricanes Katrina and Rita • a shortage of skilled labor • increase in price • re-bidding and restructuring of the project	• unanticipated rapid inflation in global and regional construction costs • the later-than-anticipated opening of the project in November 2007, nearly two years later than anticipated in the MIS/AA
Sprinter Light Rail – Oceanside, California	• installation of the Sprinter's signal network • malfunctioning signaling equipment • inappropriate design • poor construction plan/ project phasing • organizational incapacity to undertake the project • federal policy, similar to "Buy America" • the federal consultants base their risk assessments on data from other projects without considering specifics of the Sprinter project • heavy rainfall • lawsuit • skyrocketing material costs • obtaining permit/ approvals • state's financial problems • federal budget crises	• high construction bids due to an active market • construction delays resulting from right-of-way access restrictions unanticipated at FFGA execution • design changes, related in part to the substitution of a longer DMU vehicle
North Shore Connector; Pittsburgh, PA	• poor contract management • increases in price • inaccurate cost estimates • uncertainty of funding • stakeholder demand • poor construction site surveys	• schedule delays (21%) • underestimated baseline unit costs • understated scope: the anticipated scope had the North Shore alignment and stations at-grade rather than the actual outcome in tunnel and on elevated structure.

4.4. Error analysis

There are two types of errors in the classification results, i.e., FP errors and FN errors. FP errors refer to the circumstances where non-risk sentences were classified as risk sentences and FN errors represent that risk sentences were identified as non-risk sentences.

The reason for FP mainly lies in the use of twisted expressions and negation. For example, the following sentence.

"The federal government has always funded such agreements, so despite uncertainties involving projects dependent on matching federal funds, concerns surrounding the nation's trust fund are a non-issue for the authority…."

is wrongly predicted as risk due to mention of funding risk and the use of "despite". Also, sentences that discuss opportunities (i.e., the opposite of risk) or project scope also share common wording as risk sentences. For example,

"Our staff at FTA has gotten really good at detecting the risk factors that could lead to significant cost overruns."

FN errors are mainly caused by two reasons in this research. Firstly, identifying some risk sentence is difficult even for human beings and require domain knowledge and interpretation. The following sentence.

"On the other hand, in 2004 voters voted on a Northwest Rail Line with a stop in downtown Louisville."

implicitly indicates the scope change risk (i.e., adding a new train stop) caused by stakeholders. Here, "voters voted" is the domain knowledge and the key indicator of stakeholders' impact. It would be too difficult, if not impossible, for the language model to learn this domain knowledge from few occurrences of the word "voters" in the whole training dataset. Secondly, risk sentences can be written without any explicit project or risk-related language or written in an ordinary style, which requires a deeper understanding of the content. For example,

"In a concentrated urban center where people have been living for more than 100 years, you start to find things that you didn't know were there."

talks about the unforeseen underground/geological conditions for a project located in the urban center. However, none of the technical terms appears, thus appearing as a non-risk sentence.

5. Discussion

Overall, this research developed an automated approach to identify and gather project-level risk events from textual documents. The study provides both theoretical and practical insights. From the theoretical perspective, this research built a labeled risk dataset that encodes risk management expertise and formalizes the engineering knowledge for digital use. The labeled dataset can be used for the training process in future risk model development. More importantly, this study represents one of the first efforts to bring the context-aware model to the construction area and showcases its effectiveness. Specifically, the research's results demonstrated incremental performance improvements produced by adding word-level context from TF-IDF to word embedding methods and sentence-level context from word embedding to BERT and RoBERTa. To overcome the limitations in capturing higher level context of existing classification models in the construction domain, this study devised MCRI, a novel model that utilizes a BiGRU encoder to infuse useful context information from surrounding sentences to the target sentence's embedding. By using BiGRU, contextual sentences and the current sentence are assigned different importance thus avoiding the dilution of the current sentence information. At the same time, it provides valuable features to help categorize the current sentence. The results showed that leveraging the multi-sentence level contextual information leads to significant improvements in the construction risk identification. Since context plays an important role in modeling other construction-related texts (such as contract clauses or other project reports where surrounding sentences are required to comprehend the target sentence), this research sheds light on the effective use of context information and the model developed in this study can be adopted by other text classification tasks.

This research also has practical implications for utilizing NLP in future construction management practices. Firstly, the MCRI model can improve the efficiency of collecting risk information from news and other textual documents. For example, in the context of identifying risks for specific construction projects, online searches can yield several documents such as news or project reports. However, manually reviewing each document to identify relevant risk information can be tedious and time-consuming. In contrast, the MCRI model developed in this study enables the pinpointing of risk sentences from these documents, reducing the volume of document review from pages to just a few paragraphs. This not only saves time and effort but also enables practitioners to focus on analyzing and mitigating risks more effectively. In addition, the MCRI model does not require supervision or domain expertise to operate. This means that users do not need to have specialized knowledge in risk identification or construction projects to utilize the model effectively. Furthermore, the model's low computation cost and insignificant processing time make it highly efficient for performing extensive searches of risk information for given projects. This capability is particularly beneficial in scenarios where comprehensive risk assessment is critical but time and resources are limited.

Secondly, the model achieved a high accuracy level of risk event recognition even given the challenge of heterogenous and largely dispersed news articles. News articles are multidomain texts where the

vocabulary is less standardized compared to technical documents [40]. For example, the standard contract clauses of 2017 FIDIC yellow book [14] are around 50,000 words long while the unique words are only about 2200 words, whereas in the dataset in this study there are about 58,134 words while the total unique words are 5037. Therefore, the research provides confidence for the automation of construction management tasks. The model comparison results can also aid construction practitioners in selecting the suitable paradigm to solve real-world problems.

6. Conclusions and future work

Risk identification is one of the most important steps in the risk assessment process, and currently, it heavily relies on practitioners' experience and individual knowledge in the construction industry. There is a lack of systematic gathering, documentation, and indexing of construction-related risks. The unfortunate absence of data and information from the past thus hinders effective and intact knowledge transfer. Collecting risk events from unstructured sources, such as news articles, provides a unique opportunity to make up for the scarcity of data and serves as a starting point for risk identification in the new project. In this study, a methodology is proposed to automatically identify and extract risk events from news articles. First, a collection of project-related news articles published between 1980 and 2020 was gathered and a rule-based filtering program was devised to extract the potential risk paragraphs to aid the labeling process. Also, a rigid definition for risk sentences was established to guide the labeling process. As a result, 2212 unique sentences were labeled, including 1148 risk sentences. Then, this study devised a novel text classification method to identify risks that fits in the construction context. The method combined a state-of-the-art pretrained language model with a BiGRU model to encode the multi-sentence level context information. The former leveraged the power of transfer learning while the latter provided each sentence with its language environment. The robustness of the method was validated through 5-fold cross validation. The proposed model outperformed several baseline approaches and achieved an F1-score of 87.1% and an accuracy of 86.7%.

The current research has successfully developed and validated an NLP pipeline for risk identification and extraction based on news articles. Because potentially more reliable sources such as technical documents and reports are difficult to obtain, confidential, and/or biased, they are not included in the training dataset. In light of this, future work will be needed to continue to validate the proposed approach. Firstly, the effectiveness and accuracy of applying the trained MCRI model in extracting risk sentences from technical documents needs to be further verified and measured. Also, the training dataset could 0be expanded by incorporating technical documents if there is access. Secondly, unsupervised methods, such as topic modeling, can be implemented to categorize the extracted risk narratives and summarize the dominant risks in specific types of projects or periods. This will enable more extensive construction risk-related empirical research. Lastly, risks' impacts on projects in terms of monetary value or delay durations are key information in the risk assessment process. Future research could develop automated models such as NER models to extract more granular risk-related information from the identified risk sentences, such as identifying specific risk causes and their potential impacts.

CRediT authorship contribution statement

Nan Gao: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Ali Touran:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Qi Wang:** Supervision, Methodology, Funding acquisition, Conceptualization. **Nicholas Beauchamp:** Writing – review & editing, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to acknowledge the financial support provided by Northeastern University (Boston) Tier 1 Grants.

References

[1] M. Al Qady, A. Kandil, Concept relation extraction from construction documents using natural language processing, J. Constr. Eng. Manag. 136 (3) (2010) 294–302, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000131.
[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, K. Kochut, A brief survey of text mining: classification, clustering and extraction techniques, arXiv preprint (2017), https://doi.org/10.48550/arXiv.1707.02919.
[3] P. Barberá, A.E. Boydstun, S. Linn, R. McMahon, J. Nagler, Automated text classification of news articles: a practical guide, Polit. Anal. 29 (1) (2021) 19–42, https://doi.org/10.1017/pan.2020.8.
[4] S. Bhadani, I. Verma, L. Dey, Mining Financial Risk Events from News and Assessing their Impact on Stocks, in: Mining Data for Financial Applications: 4th ECML PKDD Workshop, MIDAS 2019, Würzburg, Germany, September 16, 2019, Revised Selected Papers 4 (pp. 85-100), Springer International Publishing, 2020, https://doi.org/10.1007/978-3-030-37720-5_7.
[5] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, Inc., 2009 (ISBN: 978-0-596-51649-9).
[6] H. Chen, S. McKeever, S.J. Delany, A comparison of classical versus deep learning techniques for abusive content detection on social media sites, in: Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I 10, Springer International Publishing, 2018, pp. 117–133, https://doi.org/10.1007/978-3-030-01129-1_8.
[7] C.Y. Chu, K. Park, G.E. Kremer, A global supply chain risk management framework: an application of text-mining to identify region-specific supply chain risks, Adv. Eng. Inform. 45 (2020) 101053, https://doi.org/10.1016/j.aei.2020.101053.
[8] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder-decoder approaches, arXiv preprint (2014), https://doi.org/10.48550/arXiv.1409.1259.
[9] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46, https://doi.org/10.1177/001316446002000104.
[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. Doi:10.48550/arXiv.1810.04805.
[11] C. Diao, R. Liang, D. Sharma, Q. Cui, Litigation risk detection using twitter data, J. Leg. Aff. Disput. Resolut. Eng. Constr. 12 (1) (2020) 04519047, https://doi.org/10.1061/(ASCE)LA.1943-4170.000035.
[12] Y. Ding, J. Ma, X. Luo, Applications of natural language processing in construction, Autom. Constr. 136 (2022) 104169, https://doi.org/10.1016/j.autcon.2022.104169.
[13] A. Erfani, Q. Cui, Predictive risk modeling for major transportation projects using historical data, Autom. Constr. 139 (2022) 104301, https://doi.org/10.1016/j.autcon.2022.104301.
[14] FIDIC, Plant and Design-Build Contract 2nd Ed (Yellow Book), FIDIC, Lausanne, 2017. ISBN13: 978–2–88432-082-5.
[15] N. Gao, A. Touran, Cost overruns and formal risk assessment program in US rail transit projects, J. Constr. Eng. Manag. 146 (5) (2020) 05020004, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001827.
[16] Z.S. Harris, Distributional structure. WORD 10:2–3, 1954, pp. 146–162, https://doi.org/10.1080/00437956.1954.11659520.
[17] F.U. Hassan, T. Le, Automated requirements identification from construction contract documents using natural language processing, J. Leg. Aff. Disput. Resolut. Eng. Constr. 12 (2) (2020) 04520009, https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379.
[18] F. Hassan, u., & Le, T., Computer-assisted separation of design-build contract requirements to support subcontract drafting, Autom. Constr. 122 (2021) 103479, https://doi.org/10.1016/j.autcon.2020.103479.
[19] Z. He, L. Tavabi, K. Lerman, M. Soleymani, Speaker turn modeling for dialogue act classification, arXiv preprint (2021), https://doi.org/10.48550/arXiv.2109.05056.
[20] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[21] Y. Jallan, B. Ashuri, Text Mining of the Securities and Exchange Commission Financial Filings of publicly traded construction firms using deep learning to identify and assess risk, J. Constr. Eng. Manag. 146 (12) (2020) 04020137, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001932.

[22] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning (Vol. 112, p. 18), Springer, New York, 2013. ISBN: 978-1-0716-1417-4.
[23] K. Jeon, G. Lee, S. Yang, H.D. Jeong, Named entity recognition of building construction defect information from text with linguistic noise, Autom. Constr. 143 (2022) 104543, https://doi.org/10.1016/j.autcon.2022.104543.
[24] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint (2014), https://doi.org/10.48550/arXiv.1408.5882.
[25] J. Lee, J.-S. Yi, J. Son, Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP, J. Comput. Civ. Eng. 33 (3) (2019) 04019003, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000807.
[26] J. Lee, R. Tang, J. Lin, What would elsa do? Freezing layers during transformer fine-tuning, arXiv preprint (2019), https://doi.org/10.48550/arXiv.1911.03090.
[27] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S. and He, L. (2020). A survey on text classification: from shallow to deep learning. *arXiv preprint.* Doi:10.48550/arXiv.2008.00364.
[28] J. Liu, H. Luo, H. Liu, Deep learning-based data analytics for safety in construction, Autom. Constr. 140 (2022) 104302, https://doi.org/10.1016/j.autcon.2022.104302.
[29] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint.* Doi:10.48550/arXiv.1907.11692.
[30] H.-M. Lu, N.W. Huang, Z. Zhang, T.-J. Chen, Identifying Firm-Specific Risk Statements in News Articles, in: Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2009, Bangkok, Thailand, April 27, 2009. Proceedings (pp. 42–53), Springer Berlin Heidelberg, 2009, https://doi.org/10.1007/978-3-642-01393-5_6.
[31] T. Lyons, M. Skitmore, Project risk management in the Queensland engineering construction industry: a survey, Int. J. Proj. Manag. 22 (1) (2004) 51–61, https://doi.org/10.1016/S0263-7863(03)00005-X.
[32] H.R. Marucci-Wellman, H.L. Corns, M.R. Lehto, Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review, Accid. Anal. Prev. 98 (2017) 359–371, https://doi.org/10.1016/j.aap.2016.10.014.
[33] E. Maytorena, G.M. Winch, J. Freeman, T. Kiely, The influence of experience and information search styles on project risk identification performance, IEEE Trans. Eng. Manag. 54 (2) (2007) 315–326, https://doi.org/10.1109/TEM.2007.893993.
[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint (2013), https://doi.org/10.48550/arXiv.1301.3781.
[35] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep Learning–based Text Classification: A Comprehensive Review, ACM Comput. Surv. 54 (3) (2021) 1–40, https://doi.org/10.1145/3439726.
[36] J. Ninan, Construction safety in media: an overview of its interpretation and strategic use, Int. J. Constr. Manag. 23 (6) (2021) 945–953, https://doi.org/10.1080/15623599.2021.1946898.
[37] O. Okudan, C. Budayan, I. Dikmen, A knowledge-based risk management tool for construction projects using case-based reasoning, Expert Syst. Appl. 173 (2021) 114776, https://doi.org/10.1016/j.eswa.2021.114776.
[38] J. Pennington, R. Socher, C.D. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532-1543, 2014, https://doi.org/10.3115/v1/D14-1162.
[39] V. Raheja, J. Tetreault, Dialogue act classification with context-aware self-attention, arXiv preprint (2019), https://doi.org/10.48550/arXiv.1904.02594.
[40] D.M. Salama, N.M. El-Gohary, Semantic text classification for supporting automated compliance checking in construction, J. Comput. Civ. Eng. 30 (1) (2013) 04014106, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000301.
[41] G. Seni, J.F. Elder, Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions, Morgan & Claypool Publishers. ISBN, 2010, 9781608452842.
[42] N.B. Siraj, A.R. Fayek, Risk identification and common risks in construction: literature review and content analysis, J. Constr. Eng. Manag. 145 (9) (2019) 0001685, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001685.
[43] H.C. Tan, C.J. Anumba, P.M. Carrillo, D. Bouchlaghem, J. Kamara, C. Udeaja, Capture and Reuse of Project Knowledge in Construction, John Wiley & Sons. ISBN, 2009, 978-1-4051-9889-9.
[44] D. Tian, M. Li, Q. Ren, X. Zhang, S. Han, Y. Shen, Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining, Autom. Constr. 145 (2023) 104670, https://doi.org/10.1016/j.autcon.2022.104670.
[45] T.E. Uher, A.R. Toakley, Risk management in the conceptual phase of a project, Int. J. Proj. Manag. 17 (3) (1999) 161–169, https://doi.org/10.1016/S0263-7863(98)00024-6.
[46] Washington State Department of Transportation (DOT), "Project Risk Management Guide," Engineering and Regional Operations, Accessed April 27, 2024. https://wsdot.wa.gov/publications/fulltext/CEVP/ProjectRiskManagementGuide.pdf, 2018.
[47] T.P. Williams, J. Gong, Predicting construction cost overruns using text mining, numerical data and ensemble classifiers, Autom. Constr. 43 (2014) 23–29, https://doi.org/10.1016/j.autcon.2014.02.014.
[48] C. Wu, X. Li, Y. Guo, J. Wang, Z. Ren, M. Wang, Z. Yang, Natural language processing for smart construction: current status and future directions, Autom. Constr. 134 (2022) 104059, https://doi.org/10.1016/j.autcon.2021.104059.
[49] D. Yan, S. Guo, Leveraging contextual sentences for text classification by using a neural attention model, Comput. Intell. Neurosci. 2019 (2019), https://doi.org/10.1155/2019/8320316.
[50] Zahiri, S. M., & Choi, J. D. (2017). Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *arXiv preprint.* Doi:10.48550/arXiv.1708.04299.
[51] Zhang, Y., & Wallace, B. (2016). A sensitivity analysis of (and Practitioners' guide to) convolutional neural networks for sentence classification. *arXiv pre-print.* Doi:10.48550/arXiv.1510.03820.
[52] D. Zhang, M. Li, D. Tian, L. Song, Y. Shen, Intelligent text recognition based on multi-feature channels network for construction quality control, Adv. Eng. Inform. 53 (2022) 101669, https://doi.org/10.1016/j.aei.2022.101669.
[53] B. Zhong, X. Pan, P.E.D. Love, L. Ding, W. Fang, Deep learning and network analysis: classifying and visualizing accident narratives in construction, Autom. Constr. 113 (2020) 103089, https://doi.org/10.1016/j.autcon.2020.103089.
[54] C. Zhou, C. Sun, Z. Liu, F. Lau, A C-LSTM neural network for text classification, arXiv preprint (2015), https://doi.org/10.48550/arXiv.1511.08630.
[55] S. Zhou, S.T. Ng, Y. Yang, J.F. Xu, Delineating infrastructure failure interdependencies and associated stakeholders through news mining: the case of Hong Kong's water pipe bursts, J. Manag. Eng. 36 (5) (2020) 04020060, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000821.