A corpus database for cybersecurity topic modeling in the construction industry

Dongchi Yao[1,2] and Borja García de Soto[1,2]

[1]S.M.A.R.T. Construction Research Group, Division of Engineering, New York University Abu Dhabi (NYUAD), Experimental Research Building, Saadiyat Island, P.O. Box 129188, Abu Dhabi, United Arab Emirates

[2]Tandon School of Engineering, New York University (NYU), United States
dy2037@nyu.edu, garcia.de.soto@nyu.edu

Abstract –

In the digitalized era of Construction 4.0, ensuring the confidentiality, availability, and integrity of digital assets through cybersecurity is crucial for the construction industry. Although more than 75% of respondents from a Forrester survey who are in the construction, engineering, and infrastructure industries reported to have experienced a cyber-incident in the past 12 months, only 0.25% of cybersecurity publications focus on the construction industry until Jan 2023. Considering the significance of ensuring cybersecurity in construction, this study uses Latent Dirichlet Allocation (LDA) Topic Modeling technique to identify potential research directions in cybersecurity in the construction industry, based on various text sources collected, including news, articles & blogs, academic publications, books, standards, and company reports. The results of the study identify eight topics for future research: Perform Risk Analysis, Prevent the Increasing Cyber Incidents, Detect Ransomware, Strengthen Management Process, Protect Network Devices, Regulate Information Storage and Sharing, Protect Privacy, and Improve Authentication Process. Additionally, the corresponding action is proposed for addressing each topic. These findings can be used by researchers, practitioners, and policymakers in the construction industry to address the challenges and opportunities in cybersecurity.

Keywords –

Cybersecurity; Topic Modeling; Deep Learning; Natural Language Processing; Construction Industry

1 Introduction

The construction industry is shifting towards digitalization, known as Construction 4.0, which involves the use of digital technologies such as digital twins, drones, robotics, cloud computing, and virtual reality [1]. These technologies have the potential to improve the speed, cost, and quality of construction, operation, and maintenance of assets [2]. Protecting the data generated in construction projects, such as blueprints, bidding documents, sensor data, and maintenance parameters, is critical to ensure the smooth implementation of these technologies and prevent data breaches that can cause damage to data and physical spaces, disrupting business operations and compromising safety [3].

Several cybersecurity incidents have been reported in the construction industry recently, such as the theft of floorplan files from the Australian intelligence headquarters [4], identity theft resulting in a loss of 17.2 million euros for a Finnish crane maker [5], exposure of private information of employees in a data breach at Turner Construction [6], theft of trade secrets from ThyssenKrupp (a construction escalator manufacturer) due to a data breach [7], ransomware attack at the federal construction firm called Bird Construction [8]. The construction industry was the third largest target of ransomware attacks in 2020, following government agencies and manufacturing, with AEC companies being more than twice as likely to be targeted. Many businesses in the industry fail to develop a protection plan for their digital assets, which can lead to significant losses in the event of a cyberattack [9]. These incidents highlight the importance of prioritizing cybersecurity in the construction industry. However, cybersecurity in the construction industry presents unique challenges compared to general cybersecurity. This is due to the dynamic nature of construction projects, which involves frequent changes in teams, varying levels of cybersecurity knowledge among employees, scattered and frequent communications within the supply chain, frequent exchange of digital information among stakeholders, and the overlapping of different projects. As a result, standard frameworks and methodologies for general cybersecurity designed for more static industries, like the manufacturing industry, may not be adequate for addressing the specific needs of the construction industry. Instead, tailored and industry-specific approaches are

required to effectively address the cybersecurity challenges in the construction sector.

In the last few years, some studies have addressed cybersecurity in the construction industry, the purpose of which can be classified into general discussions, review papers and specific solutions. General discussions on cybersecurity topics in construction include works by Bello and Maurushat [10], EI-Sayegh et al. [11], García de Soto et al. [1], Mantha and García de Soto [4], to name a few. Review papers include works by Pärn and Edwards [12], Sonkor and García de Soto [13]. Specific technical solutions cover blockchain technology [12] [14], machine learning and deep learning algorithms [15][16], threat modeling [7][17], framework proposal [5], and CVSS score [3]. Despite these efforts, the search in the Web of Science database with query words "cybersecurity" and "construction industry" only resulted in a total number of no more than 62 in the past 5 years, among which there is no unified framework to help practitioners who are not computer science experts to implement cybersecurity measures handily. In this context, this study proposes to synthesize existing content that has been published online, screen and organize it into a corpus database (a corpus is simply a list of sentences), and implement a topic modeling technique on the corpus that can be used to identify future research directions in the field of cybersecurity in the construction industry. Compared to previous works [2] [4] [18], this study is the first to analyze content from such extensive sources, including news, articles & blogs, LexisNexis database, academic publications, books, standards, and company reports, which provides a more concrete foundation and a more comprehensive information integration for topic identification.

The rest of the paper is organized as follows: Section 2 introduces the steps and methods for collecting, processing, and screening the text and illustrates how the topic modeling is implemented. Section 3 demonstrates the experiments of semantic screening, lists the statistical information of all the corpora established and demonstrates how to decide the input variable for topic modeling. Section 4 discusses and summarizes the topics. Finally, Section 5 concludes the paper, identifies the existing limitations of the paper and proposes future work.

2 Steps and methods

Figure 1 illustrates the overall process of this study. It can be seen that there are 6 main steps leading to the final purpose of topic modeling, during which 4 intermediate corpora are formed and stored. The steps are explained in detail in Sections 2.1 to 2.6.

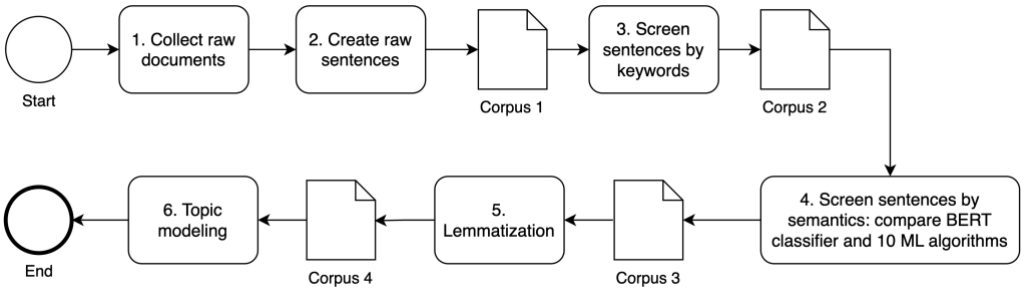

Figure 1. Overview of the main steps used in this study

2.1 Collect raw documents

As much relevant text as possible was collected to perform topic modeling in the field of cybersecurity in the construction industry. The text was collected from six sources to cover most of the necessary data.

(1) News articles and blogs are useful sources of domain-specific text as they provide detailed information about contemporary cyber incidents, including their causes, procedures, and impacts. In this study, we used the Octoparse software to crawl over 70 websites. From each website, we extracted the title, URL, and content of each piece of text. Search criteria: "cyber*" and "construction industry".

(2) LexisNexis is a reputable provider of online legal, business, news, and public records information, which includes documents dating back decades and adding about 54 million new documents each month. This study utilized the free API service provided by New York University (NYU) to request relevant documents using the Python query code. Through this process, we obtained and saved nearly 4,000 pieces of text in CSV form. Search criteria: "cyber*" and "construction industry".

(3) Academic peer-reviewed publications are important data sources on cybersecurity in the construction industry due to the high quality and relevance of their professional knowledge. We conducted

searches using the same query words as above in the Google Scholar, Web of Science, and Scopus databases to identify the most relevant publications on this topic. Search criteria: "cyber*" and "construction industry".

(4) Books and book chapters related to our topic. Although we could only find a small number of relevant books and book chapters, we could use relevant information from those sources. Search criteria: "cyber*" and "construction industry".

(5) Specifications and standards are important sources as they contain complete statements of legal or industry requirements related to cybersecurity. In this study, we collected 37 documents for text extraction, including PAS 1192-5:2015, ISO/IEC 27001 etc. Search criteria: "cyber*" and "construction industry".

(6) Reports from companies, such as slides and manuals, can be useful sources of text, although the text may be scattered throughout the document and require manual compilation. We used Google Search to locate PDF files of company reports. Search criteria: "cyber*" and "construction industry".

The collected raw documents are open-sourced on the GitHub page [19].

2.2 Create raw sentences (Corpus 1)

In this step, we organized the raw documents into a list of sentences. There are four types of raw documents: CSV, PDF, JSON, and TXT. We used open-sourced or self-created Python scripts and libraries to extract the texts from these documents. This resulted in Corpus 1 (summarized in Section 3.1), which is a list of sentences. The Python scripts and libraries used are summarized as follows: (1) The Slate3k library was used to extract text from one-column PDF files of books, standards, and reports; (2) The Pytesseract library is used to extract text from PDF files of papers that have two columns; (3) The Pandas library was used for processing CSV and JSON files and text handling; (4) Mysentences, a self-created Python script, was used to preliminarily clean and split the texts into sentences.

2.3 Screen sentences by keywords (Corpus 2)

Not all sentences were relevant to this study. Sentences containing noisy or irrelevant information were removed. To do this, the authors have created a list of 76 keywords to screen out related sentences, including "cyber", "threat", "vulnerab", "internet", "secur", "safe", "information", "cloud", etc. Some of the keywords are partial words, as there are different variations of the word, and some are abbreviations commonly used in the cybersecurity field. This process resulted in Corpus 2 (summarized in Section 3.1).

2.4 Screen sentences by semantics (Corpus 3)

To ensure that all sentences in our study were meaningful and relevant, we used a semantic screening process that involved training a deep learning model. We randomly selected and manually labeled 2,000 pieces of sentences with the label "0" indicating "Exclusion" and the label "1" indicating "Inclusion." The criteria for inclusion of sentences in the training corpus are as follows: (1) the sentence should have a good format and structure, and (2) the sentence should be directly relevant to the construction industry or cybersecurity. These two criteria are clear and not ambiguous and should effectively screen out useful texts for training the model.

The 2,000 sentences were then randomly split into a training set with 1,660 samples and a test set with 340 samples (~80:20 data split). The training set was fed into a deep learning model for classification tasks, which was composed of a pre-trained BERT model [20] and a linear layer. We experimented with different sets of hyperparameters and configurations for fine-tuning the model and ultimately selected the set of hyperparameters that achieved the highest performance on the test set. This model was trained using the Hugging Face's Transformers library [21].

For comparison to the BERT classifier, we also trained and tested 10 traditional statistics-based machine learning algorithms [22] as baselines: Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting Decision Tree (GDBT), eXtreme Gradient Boosting (XGBoost), Stochastic Gradient Descent (SDG) Classifier, Decision Tree (DT), and Light Gradient Boosting Machine (LGBM) Classifier. To do this, we used the lemmatized version of Corpus 2 (see Section 2.5 for the explanation of lemmatization) to create Term Frequency-Inverse Document Frequency (TF-IDF) features and fed them into these machine learning algorithms. The training of these 10 algorithms was implemented using the scikit-learn library [23].

After comparing the performance of the 11 models (1 BERT classifier and 10 ML algorithms), we selected the model with the highest test performance to screen all sentences in Corpus 2. The resulting sentences labeled with "Inclusion" formed Corpus 3 (summarized in Section 3.1).

2.5 Lemmatization (Corpus 4)

Once the samples in Corpus 3 were cleaned in terms of semantics, we established a corpus that integrates information and knowledge about cybersecurity in the construction industry. In order to perform topic modeling, we need to create a lemmatized version of Corpus 3. Lemmatization involves removing inflectional endings from a word and returning it to its base or dictionary form,

known as the lemma. Lemmatization can reduce the number of tokens while maintaining the same level of information, improving the performance of NLP tasks. The lemmatization process consists of the following steps: lowercasing each sentence, tokenizing each sentence into words using the NLTK library, removing stop words and punctuations and lemmatizing the words.

A Python script named "myRaw2Lemmatized.py" [24] was developed to perform the entire lemmatization process in an integrated way, resulting in Corpus 4 (summarized in Section 3.1).

2.6 Topic modeling

Topic modeling is a technique for identifying the key themes or topics in a collection of texts. It is performed using the Latent Dirichlet Allocation (LDA) method [25]. To optimize the results of the LDA model, we must first determine the input variable, which represents the number of topics to be generated. To achieve this, we experimented with input variables ranging from 5 to 30 and evaluated the models based on their perplexity and coherence scores. Perplexity scoring measures the accuracy of a probabilistic model in predicting a sample, while coherence scoring measures the relatedness of the top words in a topic. A low perplexity score and high coherence score indicate that the model is more accurate and the topics are more coherent, respectively. A Python script called "topic_modeling.py" [26] was developed to run the entire process. Section 3.3 describes the experimentation process and indicates how we ultimately chose 8 as the input variable (i.e., the number of topics), which means that eight topics will be summarized. The eight topics identified and summarized can provide insights into future directions for research on cybersecurity in the construction industry.

3 Experiments and Results

In Section 3.1, we present the experiments and results of screening sentences by semantics (corresponding to Section 2.4) and then present an overview of the statistical information for the four corpora established. Section 3.2 describes the experiment process for deciding the input variable for topic modeling and presents the preliminary results of the eight topics (corresponding to Section 2.6).

3.1 Corpora establishment

Sections 2.1 and 2.2 are for collecting related text or documents and tokenizing them into individual sentences, which result in Corpus 1. Table 1 shows statistical information on Corpus 1. Sections 2.1 through 2.3 outline the process of collecting raw sentences and filtering them using keywords, resulting in the creation of Corpora 1 and 2. Although these steps require significant effort, they are relatively straightforward to implement. In contrast, Section 2.4, which involves screening sentences based on semantics, requires extensive experimentation to determine the most effective model. Therefore, in this section, we provide a detailed explanation of the experimentation process.

Table 1. Statistics of Corpus 1 (Raw Documents)

Text source	Tools	# of document	# of sentence	Total number
News articles and blogs	Octoparse crawling tool	75 websites	71 K	
LexisNexis databases	LexisNexis API through NYU	3,968 pieces of news	596 K	
Academic papers	Google Scholar, Web of Science, and Scopus	78 pieces	26 K	802 K
Books (chapters)	Google search	13 files	73 K	
Specifications/ Standards	Google search	37 files	22 K	
Company reports	Google search	46 files	14 K	

3.1.1 Screen sentences by semantics

In Section 2.4, we used Corpus 2 as input for the BERT classifier; at the same time, we compared the BERT classifier with 10 ML algorithms. A total of 1,660 examples were used for training and 340 for testing. The training and experimenting of the BERT classifier were conducted on the Google Colab. A total of 10 experiments using the BERT classified were conducted. The hyperparameter settings and training results are presented in Table 2. In these experiments, the main variables were the batch size for gradient accumulation, learning rate, class weights, and dropout rate. We set the total number of training epochs to 6, the optimizer with scheduler to Adam with decoupled weight decay (AdamW), and the loss function to CrossEntropy. Note that in the 2,000 sentences, the number of label 'Exclusion' is 1,381 while that of 'Inclusion' is 619. To make the training dataset more balanced for training, we set different class weights when computing the loss, respectively 1:3, 1:2 and 1:1.4. Any ratio between 1:3 and 1:1 should be fine for experimentation.

We found that both Setting 7 and Setting 9 achieved the highest testing accuracy of 0.903, demonstrating the BERT model's strong ability to fit training examples and its superiority in semantic understanding. This validates our labeling process on the 2,000 random samples and subsequent training on the entire Corpus 2. In comparison, we also applied 10 statistic-based ML algorithms (in Section 2.4) to predict labels on the 340 test samples. The results indicate that the performance of

statistic-based ML algorithms is generally lower than that of the BERT model, with the highest accuracy being 0.797 from the XGBoost and GBDT models.

Table 2. BERT classifier experiments

Setting number	Batch size	Learning rate	Class weights	Dropout rate
1	4	5e-5	1:3	0.2
2	8	5e-5	1:3	0.2
3	4	5e-5	1:2	0.4
4	8	5e-5	1:2	0.4
5	4	5e-5	1:1.4	0.2
6	8	5e-5	1:1.4	0.2
7	4	4e-5	1:3	0.2
8	8	4e-5	1:3	0.2
9	4	4e-5	1:1.4	0.4
10	8	4e-5	1:1.4	0.4

Therefore, we selected the BERT classifier as our final model, which was configured with Setting 7 in Table 2. We then used this trained model to automatically label all sentences in Corpus 2, resulting in the creation of Corpus 3. For example, the 2nd sentence in Corpus 3 is "*For some construction companies, recent ransomware attacks have led to the loss of confidential data or a systems shutdown*", and the BERT classifier labels it as "Inclusion" with probability 0.998, meaning that this sentence should be included in our corpus with a high confidence score. The 189th sentence is "*Innovative building firms employ Building Information Modeling (BIM) as a central database for blueprints, designs, and other assets*", and the BERT classifier labels it as "Inclusion" with a probability of 0.687, meaning that this sentence can only be included in our corpus with a low confidence score. Sentences with probabilities lower than 0.5 should be excluded from our corpus. From our result, more than 90% of the final sentences have a probability of over 0.85, which demonstrates the effectiveness of our labeling process. These high probabilities also indicate that the BERT classifier is highly confident in its judgments.

3.1.2 Statistical information of the four corpora

The statistical information (number of sentences, number of words, number of unique words, and storage size) for Corpora 1 through 4 are summarized below.

- Corpus 1: 802K sentences, 23,753K words, 905K vocabulary items, 163 MB
- Corpus 2: 238K sentences, 9,224K words, 447K vocabulary items, 66 MB
- Corpus 3: 66K sentences, 2,227K words, 121K vocabulary items, 15 MB
- Corpus 4: 66K sentences, 1,305K words, 34K vocabulary items, 11 MB

Each step significantly reduces the number of sentences, words, vocab size, and storage size, with some reductions exceeding 50%. This results in a more content-rich and semantically useful corpus for our study. We will use Corpus 4 to perform topic modeling.

3.2 Decide the input variable for topic modeling

In Section 2.6, we used Corpus 4 for topic modeling through the LDA method. To determine the optimal input variable for the LDA model, we evaluated different numbers (5 to 30) using perplexity and coherence scores as metrics [25]. Figure 2 shows how the perplexity score decreases with an increase in the input variable while the coherence score fluctuates, generally decreasing. Out of the three candidate input variables (8, 15, and 30) that reach a peak respectively in the fluctuating line, we chose 8 as it results in more focused topics. Therefore, the final number of topics derived is 8, with a perplexity score of -8.68 and a coherence score of 0.48.

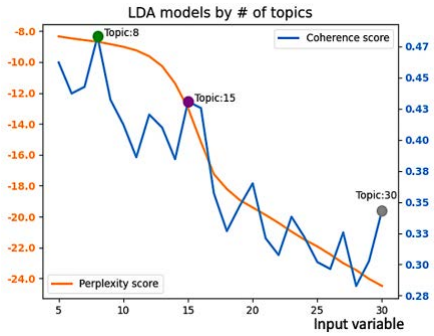

Figure 2. Perplexity and coherence scores

The LDA model then uses 8 as the input variable for topic modeling. The results are visualized in Figure 3 as a 2D mapping of the 8 topics grouped into clusters. The distances between the clusters indicate their correlation, with topics 1, 2, 3, and 8 being closely related and topics 4, 5, 6, and 7 being more independent. The size of the bubble represents the prevalence of each topic, with topics 1, 2, and 3 being the most prevalent and thus to be prioritized.

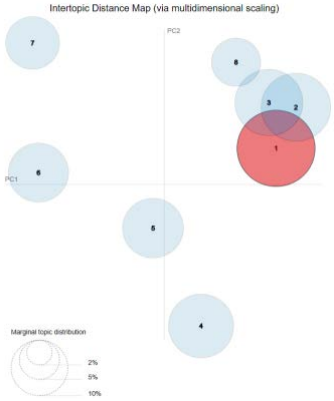

Figure 3. 2D mapping of eight topics

4 Discussions on summarizing topics

Once the number of topics was selected, in this case, eight, they were summarized based on two elements: (1) keywords associated with each topic, which were identified by the LDA model as the 10 keywords with the highest weights, and (2) representative original sentences for each topic, which were selected from the 3 most meaningful sentences out of a pool of 10 sentences with the highest contribution percentage, calculated using the Gensim library [27]. The 3 most meaningful sentences from a pool of 10 sentences with the highest percentage of contribution for each topic were selected.

For example, for the 1st topic, the top 10 keywords with the highest weights were: 0.05: security; 0.04: cybersecurity; 0.03: safety; 0.02: risk, training, may; 0.01: area, provide, private, ensure. The chosen three representative sentences are as follows: (1) Construction companies need to conduct comprehensive and frequent third-party cyber security assessments so they can identify and remediate vulnerabilities; (2) Cyber security is not just an IT issue. Cyber risk should not be seen as an issue solely for your IT department or provider; (3) Many construction firms also lack awareness regarding cyber security. Their percentages of contribution are 0.96, 0.92, and 0.92, respectively. By combining the 10 keywords and the 3 representative sentences, we can infer and summarize the topic: *Increase awareness of construction firms by providing training and contact periodic risk analysis*.

A similar process for Topic 1 was used to determine the remaining topics. The results are summarized in Table 3. In general, the 8 topics can be described as Perform Risk Analysis, Prevent the Increasing Cyber Incidents, Detect Ransomware, Strengthen Management Process, Protect Network Devices, Regulate Information Storage and Sharing, Protect Privacy, and Improve Authentication Process. The action to take to address these topics is also proposed in Table 3. These 8 identified topics and corresponding suggested actions can serve as a reference for researchers interested in studying future research directions in cybersecurity within the construction industry. Based on the topic prevalence (the size of the bubble) shown in Figure 3, we recommend prioritizing research on risk analysis and specific solutions to threats, such as ransomware, corresponding to Topics 1, 2, and 3.

5 Conclusions and outlooks

This study employs topic modeling to uncover potential research directions in cybersecurity in the construction industry. We gathered 6 types of raw text sources, which were transformed into 802K raw sentences. Through a process of keyword and semantic screening, the data was reduced to 66K sentences, forming the corpus (Corpus 4) for our topic modeling analysis. We utilized the LDA method and tested input variables ranging from 5 to 30, evaluating and selecting the model based on perplexity and coherence scores. Our analysis revealed that 8 was the most suitable number of topics. The results of our analysis identified 8 potential topics for future research, and corresponding actions to take are also proposed.

Some of the limitations of this study include (1) the corpus was not as clean as it could have been, as some samples contained irregular characters and signs from the original crawled websites due to semantic screening only considering the meaning of sentences and ignoring their format, and (2) the number of specific studies on cybersecurity in the construction industry is limited, which inevitably impacted the comprehensiveness of our final corpus.

To address these limitations, the authors are adding a new corpus and comprehensively cleaning up selected texts. Additionally, the resulting corpus can be used for other purposes, such as training large language models to create question and answer systems (like the ChatGPT dialogue system released in Nov 2022 [28]), which can provide relevant people with preliminary consulting advice on cybersecurity management and will also be the focus of our future research.

Acknowledgments

The authors want to thank the Center for Cyber Security at New York University Abu Dhabi (CCS-AD) for the support provided.

References

[1] García de Soto B., Agustí-Juan I., Joss S., and Hunhevicz J. Implications of Construction 4.0 to the workforce and organizational structures, International Journal of Construction Management, 22(2): 205-217, 2022.

[2] Kayan H., Nunes M., Rana O., Burnap P., and Perera C. Cybersecurity of Industrial Cyber-Physical Systems: A Review, [Online]. Available: http://arxiv.org/abs/2101.03564, 2021.

[3] Mantha B. R. K., and García de Soto B. Assessment of the cybersecurity vulnerability of construction networks, Engineering, Construction and Architectural Management, 28(10): 3078–3105, 2021.

[4] Mantha B. R. K., and García de Soto B. Cyber security challenges and vulnerability assessment in the construction industry, in Proceedings of the Creative Construction Conference 2019, pages 29–37, Budapest, Hungary, 2019.

Table 3. Summarized topics and actions to take

Topic No.	Keywords	Three representative sentences for each topic	Topic summarized	Some actions to take
1	security, cybersecurity, safety, risk, may, training, area, provide, private, ensure	• Construction companies need to conduct comprehensive and frequent third-party cyber security assessments so they can identify and remediate vulnerabilities. • Cyber security is not just an IT issue. Cyber risk should not be seen as an issue solely for your IT department or provider. • Many construction firms also lack awareness regarding cyber security.	Perform Risk Analysis	Researchers should take the lead to develop frameworks first.
2	cyber, security, standard, sector, industry, technology, development, dimension, report, impact	• Regardless of the industry, there is a worrying shift in the mindsets of security professionals. • Rising cyber-attacks 75% of respondents reported an increase in cyber-attacks in the last 12 months. • We are undoubtedly in a changing world, and cyber security is an ever-changing industry.	Prevent the Increasing Cyber Incidents	Industry practitioners should share the incident data.
3	threat, infrastructure, national, critical, attack, cyber, government, public, measure, communication	• You don't want to get ransomware, and the government also doesn't want you to accept ransomware. • Phishing emails have increased by 20%, while malware soared by 423%. • For example, supply-chain-based attacks, such as the SolarWinds SUNBURST attack, are not simple system vulnerabilities.	Detect Ransomware and Malware	Software providers must constantly update their algorithms.
4	construction, risk, project, building, design, system, management, build, personnel, may	• These risks can be physical and directly affect the insurance business, or they may be more transitional and affect insurers. • Stakeholder-associated life cycle risks in the construction supply chain. • Due to known and unknown risks, the Company's results may differ materially from its expectations and projections.	Strengthen Management Process	Companies should adopt strict policies and train employees effectively.
5	system, http, use, control, security, network, secure, access, internet, device	• The Internet of Things (IoT) is a system where devices are connected wirelessly with the help of unique identifiers. • Access to the software is not properly controlled. • Centralized visibility allows organizations to control access for any device, connecting from any network.	Protect Network Devices	Hardware manufacturers should integrate advanced security features.
6	information, data, asset, personal, include, use, level, access, specific, person	• Besides proprietary employee data, other potentially vulnerable information includes sensitive client data, and tenant personally identifiable information (PII). • Are any assets (e.g., web server, web application) Internet-facing? • What information is being shared, and what is the purpose of sharing it?	Regulate Information Storage and Sharing	Governments must establish strict data protection laws.
7	data, protection, activity, policy, breach, require, state, privacy, key, manage	• The attacked data was encrypted, and ransom payments were demanded in the Bitcoin cryptocurrency. • After that, both processing and privacy should be determined before clarifying the requirement of granularity of trajectory data. • They are triggered if: (1) there is a breach of contract, and (2) the penalty for breach is stated in the contract.	Protect Privacy	Users should be educated on best practices, and developers should prioritize privacy by design.
8	cybersecurity, develop, strategy, response, incident, framework, security, number, practice, plan	• Here are some strategic approaches: Deploy Multi-Factor Authentication. • Implementing cyber security strategies to fortify endpoints and IT/OT integration while establishing a robust incident response plan will go a long way toward delivering peace of mind. • Choose a good, strong password: passwords are the weakest cybersecurity link.	Improve Authentication Process	Companies should implement multi-factor authentication or biometrics.

[5] Turk Ž., García de Soto B., Mantha B. R. K., Maciel A., and Georgescu A. A systemic framework for addressing cybersecurity in construction, Automation in Construction, 133: 103988, 2022.
[6] Infosec, Phishing Attacks in the Construction Industry, Infosec, On-line: https://resources.infosecinstitute.com/topic/phishing-attacks-construction-industry/, Accessed: 14/03/2022.
[7] Mantha B., García de Soto B., and Karri R. Cyber security threat modeling in the AEC industry: An example for the commissioning of the built environment, Sustainable Cities and Society, 66: 102682, 2021.
[8] Glamoslija Katarina. Ransomware Facts, Trends & Statistics for 2021, Safety Detective, On-line: https://www.safetydetectives.com/blog/ransomware-statistics/, Accessed: 08/01/2023.
[9] Salami Pargoo N., and Ilbeigi M. A Scoping Review for Cybersecurity in the Construction Industry, Journal of Management in Engineering, 39(2), 2023.
[10] Bello A., and Maurushat A. Technical and Behavioural Training and Awareness Solutions for Mitigating Ransomware Attacks, in Advances in Intelligent Systems and Computing, 1226(AISC): 164-176, 2020.
[11] El-Sayegh S., Romdhane L., and Manjikian S. A critical review of 3D printing in construction: benefits, challenges, and risks, Archives of Civil and Mechanical Engineering, 20(2), 2020.
[12] Pärn E. A., and Edwards D. Cyber threats confronting the digital built environment: Common data environment vulnerabilities and block chain deterrence, Engineering, Construction and Architectural Management, 26(2), 2019.
[13] Sonkor M. S., and García de Soto B. Is Your Construction Site Secure? A View from the Cybersecurity Perspective, in Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC), pages 864-871, Dubai, United Arab Emirates, 2021.
[14] Shemov G., Garcia de Soto B., and Alkhzaimi H. Blockchain applied to the construction supply chain: A case study with threat model, Frontiers of Engineering Management, 7(4): 564-577, 2020.
[15] Goh G. D., Sing S. L., and Yeong W. Y. A review on machine learning in 3D printing: applications, potential, and challenges, Artificial Intelligence Review, 54(1): 63-94, 2021.
[16] Yao D., and García de Soto B. A preliminary SWOT evaluation for the applications of ML to Cyber Risk Analysis in the Construction Industry, IOP Conference Series: Materials Science and Engineering, 1218(1): 012017, 2022.
[17] Mohamed Shibly M. U. R., and García de Soto B. Threat Modeling in Construction: An Example of a 3D Concrete Printing System, in 37th International Symposium on Automation and Robotics in Construction, pages 625-632, Kitakyushu, Japan, 2020.
[18] Sonkor M. S., and García de Soto B. Operational Technology on Construction Sites: A Review from the Cybersecurity Perspective, Journal of Construction Engineering and Management, 147(12), 2021.
[19] Yao D. Data for topic modeling, GitHub repository, On-line: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/Data%20for%20topic%20modeling, Accessed: 16/03/2023.
[20] Devlin J., Chang M.-W., Lee K., and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, [Online]. Available: http://arxiv.org/abs/1810.04805, 2018.
[21] Jain S. M. Hugging Face, Introduction to Transformers for NLP, pages 51–67. CA: Apress, Berkeley, 2022.
[22] Giuseppe B. Machine Learning Algorithms. Packt Publishing Ltd, Birmingham, 2017.
[23] Kramer O. Scikit-Learn, Machine Learning for Evolution Strategies, pages 45–53. Springer, Cham, Oldenburg, 2016.
[24] Yao D. myRaw2Lemmatized, GitHub repository, On-line: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/myRaw2Lemmatized.py, Accessed 07/01/23.
[25] Blei D. M., Ng A. Y., and Edu J. B. Latent Dirichlet Allocation, Journal of Machine Learning Research, 3: 993-1022, 2003.
[26] Yao D. Topic Modeling, GitHub repository, On-line: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/Topic%20modelling/topic_modeling.py, Accessed 07/01/2023.
[27] Rehurek R., and Sojka P. Gensim-python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2), 2011.
[28] OpenAI ChatGPT: Optimizing Language Models for Dialogue, OpenAI, On-line: https://openai.com/blog/chatgpt/, Accessed: 07/01/2023.