IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Understanding On-Site Inspection of Construction Projects based on Keyword Extraction and Topic Modeling

Jia-Rui Lin[1], Zhen-Zhong Hu[1], Jiu-Lin Li[2,3], and Li-Min Chen[3]

[1]Department of Civil Engineering, Tsinghua University, 100084, China; Tsinghua University-Glodon Joint Research Centre for Building Information Model (RCBIM), Tsinghua University, 100084, China.
[2]Beijing Urban Construction Group Co. Ltd., 100088, China.
[3]Beijing National Speed Staking Oval Operation CO., Ltd., 100088, China

Corresponding author: Jia-Rui Lin (e-mail: lin611@tsinghua.edu.cn, jiarui_lin@foxmail.com).

This research is supported by the Natural Science Foundation of China (No. 51908323), the Beijing Municipal Science and Technology Project (No. Z181100005918006) and the Tsinghua University Initiative Scientific Research Program (No. 2019Z02UOT).

ABSTRACT As an essential way to ensure success of construction projects, on-site inspection involves intensive paperwork, while generating large amounts of textual data. Lack of understanding of information hidden in text-based inspection records always leads to overlooking of important issues and deferred decisions. Therefore, a novel text mining approach based on keyword extraction and topic modeling is introduced to identify key concerns and their dynamics of on-site issues for better decision-making process. Then, the proposed approach was demonstrated in a real world project and tested with 7250 issue records. Results showed that the proposed method could successfully extract key concerns hidden in texts and identify their changes with time, thereby enabling a more efficient on-site inspection and data-centric decision-making process. This research contributes: (1) to the body of knowledge a new framework for discovering key concerns and their changes with time in texts, and (2) to the state of practice by providing insights on hot topics and their changes with time to reduce on-site issues and make decisions efficiently.

INDEX TERMS on-site inspection; text mining; keyword extraction; topic modeling; decision making; construction management

I. INTRODUCTION

One of the paramount efforts engineers, contractors and owners are currently facing is the need to deliver high quality construction projects without quality or safety issues. Accurate and timely information of the on-site issues is needed for a well maintained and efficient project control that will ensure cost and time efficiency of the project [1]. Since 50% to 80% of construction site problems arise from a lack of data or a delay in the receipt of information [2], it is essential to collect and analyze on-site data dynamically and make decisions in a more efficient way. That is, efficient on-site data collection, timely data analysis and communication of the data in a well interpreted way are important for construction companies [3].

The on-site inspection plays an important role in collecting timely data to take corrective actions for a high-quality and hazard-free construction project[4]. However, traditional ways of on-site inspection still depend on paper-based files like drawing and data collection forms, which are not efficient for information exchange and communication, usually leading to overlooking important issues and deferring on-site decisions [5]. McCullouch [6] noted that managers spend, on average, 30-50% of their time recording and analyzing site data due to the manual nature of monitoring and controlling methods and thus, they are distracted from other important tasks.

As information and communication technology evolves quickly, quite a few new methods have been explored to advance data collection, sensing, and visualization for on-site inspection [1]. To automate the data collection, Global Positioning System (GPS), Radio-frequency Identification (RFID) [7] as well as computer vision and 3D reconstruction [8] have been investigated by both researchers and companies. Meanwhile, mobiles and wearable computers are becoming more and more popular in collecting inspection data [9] and sharing knowledge [10 * MERGEFORMAT ,11]. Dynamic

acquisition of a large amount of unstructured data, such as, images, point cloud, videos, and texts, requires efficient data mining methods to extract valuable knowledge from them. Though image recognition [12], 3D reconstruction [8], and video processing for detection of unsafe behaviors [13] have been explored, less attention was paid to text data, another important media type that is commonly used in construction projects to record notes and share ideas of inspected on-site issues. As building information modeling (BIM) -based on-site inspection emerges and evolves quickly [14], more and more structured inspection forms and unstructured texts are collected, thus requiring efficient text mining approaches for on-site inspection.

Though the size of text data is not as large as image data or video data, manually analyzing the text and identifying key concerns is time-consuming. Also, early prediction of issues and effective response to potential challenges by extracting useful knowledge from text data is critical to ensuring the project's success [2]. Although it is argued that text analytics is vital in any construction project, it is not widely used in the construction area yet [2] and very few text and natural language processing tools nowadays appear to be suitable for the construction industry [15]. Currently, text analytics in the construction industry mainly focuses on document retrieval [15-17] to aid decision-making process, clustering[18] and classification [19,20] for easy access and management of textual data, information extraction to predict and evaluate the project's performance [21,22]. Since a construction document usually contains a few topics, it is important to model and extract topics hidden in text data for decision-making purpose. Though topic modeling is adopted for classification of BIM cases [23], trends analysis [24] of BIM research, and pattern identification [25] in the construction industry, they mainly focus on a macro-level, mining of unstructured texts from on-site inspection to discover key concerns and their changes with time is important for construction management and more exploration is needed.

To identify key concerns and their dynamics that affect efficient on-site inspection and issue resolving, an approach based on keyword extraction and topic modeling is proposed. The following section provides a brief review of relevant research and applications. Then, the methodology of the proposed approach and its four main parts, namely, data pre-processing, word segmentation and tagging, keyword extraction and topic modeling are introduced. After that, results and validations based on collected data from a real world project and their corresponding discussion are presented. Finally, contributions to the construction domain and possible improvements in the future are concluded.

II. Related Research

A. ON-SITE INSPECTION BASED ON INFORMATION AND COMMUNICATIONS TECHNOLOGY (ICT)

On the way to a high-quality and hazard-free construction project, on-site inspection plays an important role in collecting timely data that reflects the status of the project to take corrective actions, if needed [4]. Which means, on-site inspection makes it possible to monitor, audit and report the construction phases periodically so that constructed project meets all the requirements at any of the construction stages [26]. However, traditional ways for on-site inspection still depend on paper-based files like drawing and data collection forms, which are not efficient for information exchange and communication, usually leading to overlooking important issues and deferred on-site decisions [5].

Several efforts have been made to enhance data collection, sensing, and visualization for on-site inspection [1]. By integrating automated data acquisition technologies, with the scheduling system, relational database, and AutoCAD, El-Omari and Moselhi [4] proposed a method for construction progress reporting and decision making. Furthermore, computer-aided design (CAD) drawings could also be improved through building information modeling (BIM) technologies to achieve effective and efficient inspection of construction [27]. Application of GPS, RFID, barcodes, laser scanners, video and audio technologies for automated data acquisition were investigated by both researchers and companies [7]. All of them have their limitations. For example, GPS can only be used in outdoor scenes, and sometimes it is impossible to attach RFID or barcodes on construction elements. Photogrammetry, image processing [12], computer vision and 3D reconstruction [8] were all adopted for construction inspection. Efficient visualization of inspection results is essential to an intuitive understanding of construction projects, where Augmented Reality (AR) is usually utilized these days. Adoption of large and heavy equipment, marker-based AR [28], and mobile AR based on marker-less registration [29] in construction inspection was explored too. However, utilization of AR systems always means extra training and set-up efforts.

Meanwhile, as the availability of commercial mobiles and wearable computers increases, inspection data can be collected and retrieved on site using a variety of sensor systems, data management tools, and information systems [9]. Application of mobile devices for safety inspection and knowledge sharing [10,11] are investigated. As for quality inspection, mobile GIS was also adopted for information collecting and sharing to enhance the reduction of construction delays by quality compliance and efficient coordination at various stages [26]. It is reported that mobile technologies can reduce the amount of time typically needed for the file search task and transfer the time onto the general inspection task, like observing ongoing construction activities [30]. Further benefits including reduced interruptions to operations, increased safety, enhanced thoroughness, more rapid determination of results, and more accurate recording of activities are also observed [9]. Yamaura and Muench [31] also concluded that project inspectors using mobile

technology experienced productivity gains on the order of 25%, collected and shared 2.0 times as many observations, and improved the timeliness of daily reports and overall data availability. Inevitable, there is a trend to adopt various mobile technologies [32] to enable complete and consistent data, improved data accessibility and work efficiency.

Obviously, with the increasing adoption of various information technologies in the on-site inspection, a large amount of images, point cloud, and videos are generated, thus requiring efficient data mining methods to extract valuable knowledge from them. Defect detection based on images [12], 3D reconstruction based on images and point cloud [8], and detection of unsafe actions from videos [13] have been conducted. However, less attention was paid to text data, another important media type that is commonly used in construction projects to record notes and share ideas of inspected on-site issues.

B. TEXT ANALYTICS IN CONSTRUCTION

Generally, 80% of the information available in a company is in a textual format [33]. For a construction project, there are a massive amount of text documents involved, such as contracts, reports, correspondence, etc., and various text documents like specification documents are also produced before construction [21]. Therefore, effective data analytics could help in achieving successful projects with better decision-making process and improve productivity and output by 5% to 6% [2]. However, many data management problems such as complex data retrieval, difficult information reuse and inefficient interoperability between different management systems [34] may occur when dealing with textual documents.

Although it is argued that text analytics is vital in any construction project, it is not widely used in the construction area yet [2] and very few text and natural language processing tools nowadays appear to be suitable for the construction industry [15]. Text analytics in the construction area aims to discover useful information hidden within the text data [21] to provide insights and early warnings for decision makers. Currently, there are six main types of text analytics adopted in the construction industry, which are discussed in the following paragraphs.

1) Document retrieval that finds related files based on keywords, queries or the content of them. In the construction area, a content-based text mining method is proposed for retrieval of CAD documents. The proposed method could extract and index the text-based annotations embedded in CAD drawings, thus making it possible for a large collection of documents [16]. Similarly, vector space model and semantic query expansion are utilized in the retrieval of similar cases for construction risk management [15]. Project-specific term dictionary and dependency grammar parsing information of textual documents are also considered [17].

2) Classification and 3) clustering that facilitate document management [35] by dividing documents into a few groups with or without predefined labels. Caldas et al. [19] first adopted support vector machine (SVM), k-nearest neighbor, etc. for classification and retrieval of text-based construction documents. Hierarchical classification of construction documents was also reported by the same authors [20]. With content-based search for image files, Soibelman et al.[35] further extend their work to model-based management of unstructured construction data types. A method for automatic clustering of construction project documents based on textual similarity is also proposed [18].

4) Information extraction which aims to extract valuable data for further data mining application. For example, Yarmohammadi et al. applied text mining techniques to extract modeling patterns from auto-generated log files of Autodesk Revit for measurement of the modelers' performance [22]. William and Gong [21] adopt text mining to extract information for prediction of construction cost overruns. By taking the advantage of keyword extraction, a prototype for automatically extracting precursors and outcomes from unstructured injury reports is developed by Tixier et al. [36].

5) Natural language processing (NLP) is also used for regulatory documents processing for compliance checking [37], intelligent data retrieval based on natural language [38], and hidden knowledge discovery from post project reviews documents [33].

6) Topic modeling that identifies latent topics in substantial text content for analysis of public attitude and research trends of a specific field. Together with keyword extraction, topic modeling is utilized to reveal the trends of BIM research and application in Korea [24]. By extracting topics from Weibo posts, public concerns about off-site construction and their dynamic changes are explored to further promote the development of Chinese construction [39]. More research on identifying patterns hidden in construction-defect litigation cases [40], discovering thematic structures and patterns [25] based on topic modeling have been explored. Among all these research, one of the most popular methods used for topic modeling is Latent Dirichlet Allocation (LDA), which also shows its power in classifying BIM cases [23]. In addition to unsupervised methods like LDA, supervised methods including Bayes classifier were also used for topic modeling with the assumption that one sentence has only one topic [41].

With the above-mentioned literature, it is concluded that the current research mainly aims to:

1) Retrieve similar documents to aid decision-making process, which has applied in risk management, green buildings, etc.;

2) Cluster and classify documents into different groups for easy access and management of documents;

3) Extract useful data from text to predict, evaluate the performance of the project;

4) Identify key concerns or thematic structure from literature, social-media posts, etc. to help understand overall trends and characteristics of a special field.

It is concluded that text mining techniques could extract valuable insights from unstructured textual data, such as literature, CAD documents, social-media posts. With extracted data, it is possible to analyze performance of a construction project, identify patterns of public attitude and research trends.

Given that the construction projects are highly dynamic, plenty of texts are generated during on-site inspection and collected through various applications like BIM-based inspection tools. Thus, it is important to investigate how to identify hidden patterns based on text mining and analyze their changes with time, providing insights on how to resolve on-site issues more efficiently and make better decisions in construction management.

C. SUMMARY

As mentioned before, though various methods are proposed to achieve timely on-site inspection and diverse data formats including images, videos, etc. are generated, text data is still a majority data type and is widely used in on-site inspection. Although techniques and tools adopted for text analytics have been adopted for document retrieval, information extraction and identification of concerns, their application in mining unstructured texts generated during on-site inspection still needs further investigation to find key concerns and their dynamics for decision-making purpose.

III. METHODOLOGY

To address the above-mentioned problems and improve the efficiency for on-site inspection and decision-making, an integrated approach based on various text mining techniques is proposed (Figure 1). First of all, issue records are collected with a previously developed mini-program based on WeChat [42]. After data preprocessing, text mining technologies including segmentation and tagging, keyword extraction, and topic modeling are adopted to discover key concerns of the managers and engineers implicitly embedded in the issue records, and their changes with time are analyzed. Finally, based on extracted key concerns, detected changes with time, construction managers could make better decisions and eliminate on-site issues. Details of each part are as follows.

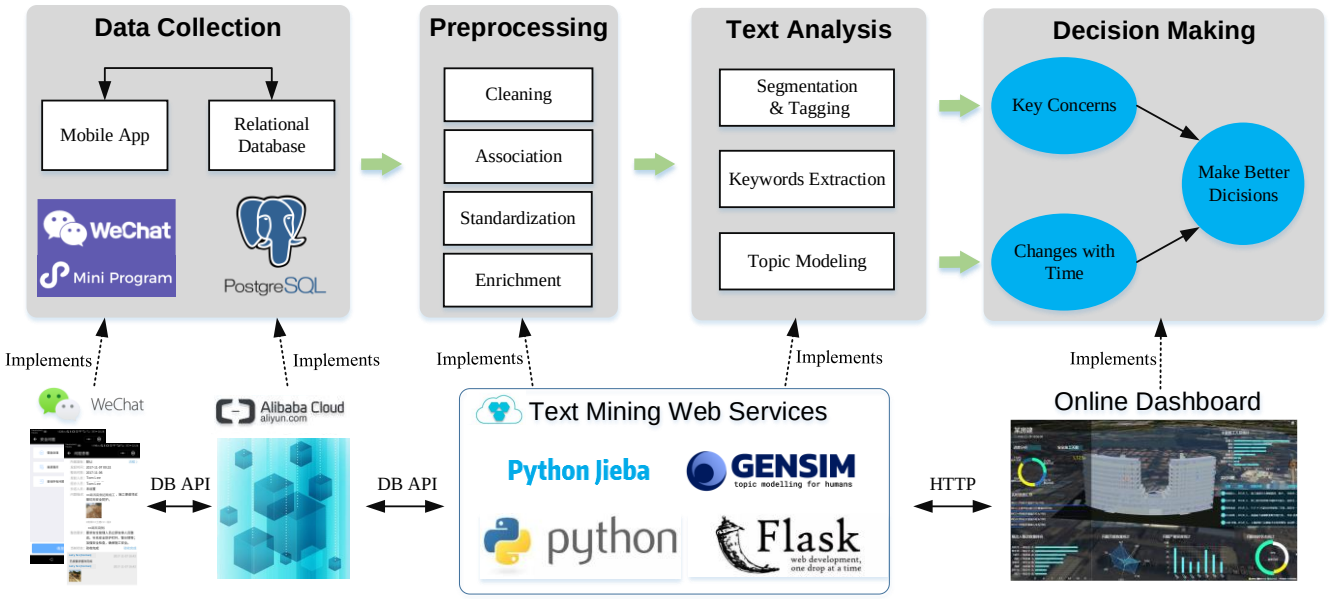

FIGURE 1. Proposed Methodology for Automatic Mining of On-Site Inspection Texts.

A. DATA PRE-PROCESSING

The presented study utilized data collected by an App developed before for on-site inspection issue collection, tracking, and resolving [42]. With this App, description, creator, type, level of importance of a specific on-site inspection issue are captured [42]. And all the collected data is persisted in a relational database called PostgreSQL, and the data could be easily accessed based on structured query language(SQL) queries. For easy deployment, the database is hosted on Alibaba Cloud, a widely used public cloud platform in China.

Since the quality of collected texts will highly impact the text mining process. For example, descriptions of some issue records are too short or empty, and words or abbreviations with the same meaning are mix-used sometimes. And considering that all collected data, including issue records and user information, are all kept in different data tables, a 4-step data pre-processing process is adopted:

1) Data cleaning: issue record whose description is empty or too short (less than 5 words in this research), or issue

records without creation time are omitted. This is done by specifying certain constraints when querying the database. For example, the following SQL query skipped all issue records without short or empty descriptions:

```
SELECT * FROM issue_record WHERE Description
IS NOT NULL AND LENGTH(Description)>5
```

2) Data association: users involved in the project have different roles and belong to different organizations, since these data are kept in different tables, a data association process is needed to connect or join different tables. In this research, data kept in different tables is first associated based on JOIN clause of SQL query in accordance with the defined database schema before [42]. Then, the other part is performed using python scripts.

3) Form standardization: generally, people may use different words and forms to express the same meaning. That is, synonyms, phrases and their abbreviations, are usually mixed used in describing on-site inspection issues. For example, "混凝土"(hùnníngtǔ in Chinese pinyin) and "砼"(tóng) are both utilized to represent concrete in descriptions of issue records. Moreover, one engineer may use English abbreviations to represent zones or spaces on a construction site, while another engineer uses Chinese phrases to represent the same zone or space. Due to limited time for on-site inspection, upper cases and lower cases are often used equally from an engineer's point of view. Thus, English abbreviations and Chinese phrases that express the same meaning should be converted to the same form. To give an example, "L2", "l2", and "二层" (Level 2, èr céng in pinyin) are all transformed to "L2" for standardization purpose. In this research, a term dictionary was built and introduced to map different forms of a term to its standardized form.

4) Data enrichment: on one hand, raw data extracted from the database lacks semantic information for analytics purpose. In this research, meanings of TypeID, LevelID, RoleID, and StatusID, etc. are added by discussing with software developers and product managers. On the other hand, we may look at the data at different granularities to the support decision-making process. Therefore, date, daytime, and even hour of a day are extracted from time fields of different data tables for the coming data mining process.

B. WORD SEGMENTATION AND TAGGING

Since all the descriptions, comments of inspection issue records are in Chinese, and there are no explicit characters like white space to separate different words, so word segmentation is usually the first step to mining Chinese texts. As a well-known problem since NLP emerged, there already exist many methods for word segmentation, for example, n-gram model, hidden markov model. And ready-to-use tools including jieba, Tsinghua university lexical analyzer for Chinese(THULAC), Peking university word segmentation(PKUSeg), have been developed and used for years [43]. Thus, considering the popularity and ease of use, this research just takes jieba as the tool for word segmentation.

However, the since pre-trained model of above-mentioned NLP tools is built on common texts without specific consideration of domain related concepts, they could not properly detect and handle construction related terms like "承台"(pile cap, chéngtái in pin yin) or "基坑"(foundation pit, jīkēng in pinyin). For example, jieba tokenizes "承台基坑无防护措施"(there is no protection around the foundation pit of the pile cap, chéngtáijīkēngwúfánghùcuòshī in pinyin) as "承/台基/坑/无/防护/措施" in default, while what we want is "承台/基坑/无/防护/措施". To address this issue, a term list is established together with on-site managers and engineers, and jieba could directly load a text-based files custom word dictionary. After word segmentation, a tagging process that classifies words into nouns, verbs, etc., is conducted with jieba. With tagging results, we could extract keywords in certain categories in the future. For example, adjectives, conjunctions, and prepositions are omitted when conducting keyword extraction.

C. KEYWORD EXTRACTION

As mentioned above, lots of text data is generated everyday on a construction site. It is difficult for the managers to review all of them. While keyword extraction allows them to lift the most important words from huge sets of texts in just seconds and obtain insights about the topics the engineers are talking about. Generally, each issue record contains short text describing inspected on-site issues submitted by inspectors, as well as potential requirements from construction managers. Keyword extraction could help us get a big picture of a project and obtain insights on how the dominant topic changes as the project goes on? With which, the managers could make better decisions to control and eliminate on-site issues.

Statistical approaches are usually used for identifying main keywords and key phrases within texts. Relevant methods include word frequency, word co-occurrences, term frequency-inverse document frequency (TF-IDF), TextRank algorithms [44] and others. As implied in the name of TF-IDF, it is calculated by multiplying two metrics: term frequency of a word in a document, and inverse document frequency of the word across a set of documents, which means how common or rare a word is in the entire document set. Formally, the TF-IDF calculation process of $word_{ij}$ in document d_i within document set D can be noted as:

$$tfidf(word_{ij}, d_i, D) = tf(word_{ij}, d_i) \cdot idf(word_{ij}, D) \quad (1)$$

Where

$$tf(word_{ij}, d_i) = log\left(1 + tf(word_{ij}, d_i)\right) \quad (2)$$

$$idf(word_{ij}, D) = log\left(N / count(d_i \in D : word_{ij} \in d_i)\right) \quad (3)$$

Fortunately, jieba has built-in support with all these algorithms and it's handy to compare them and find the best one suitable for a certain problem. In this research, after

comparing word frequency, TF-IDF, and TextRank methods, we found that they provide similar results when processing on-site inspection records, so word frequency is selected for keyword extraction as it is easy and fast. Usually, text-based descriptions of the on-site issues are 50-70 characters, or about 25 Chinese words considering the fact that each Chinese word has 2 or 3 characters. After removing a few stop words during the data cleaning process, 16 keywords could cover most of the information contained in a description. Therefore, the top 16 words with the highest frequency are selected for further keyword analysis. Meanwhile, word cloud, a popular visualization method for word frequency, is utilized to help people understand the results easily.

Finally, keyword extraction is used in two scenarios in this research: 1) extract keywords from combined descriptions of all issue records, and the managers can get a big picture of a project, including most frequent issues, and relevant locations or areas of on-site issues; 2) divide collected issues records into a few sets by date of creation, and extract keywords of each set, then changes of dominant topics with time could be identified.

D. TOPIC MODELING

Generally, each issue record should contain one topic or one problem. However, on-site inspectors may add more than one issue to a single record due to lack of enough time. It is difficult for text clustering or classification methods to handle this problem. Thus, topic modeling is adopted for more accurate processing of on-site issues.

With word segmentation, each text or document d_i can be taken as a set of words $\{word_{i1}, word_{i2}, \cdots\}$ regardless of their orders. This is also called bag-of-word (BOW) model [44], which is widely used in the NLP area. Put all words from different documents together, we can get a corpus noted as:

$$C = \{word_1, word_2, \cdots, word_n\} \tag{4}$$

Where $word_j$ is word j in corpus, and n is the total amount of words occurred in all the documents. Then, document d_i can be represented as:

$$d_i = [w_{i1}, w_{i2}, \cdots, w_{in}] \tag{5}$$

Where w_{ij} is the contribution of word j to document d_i. Word frequency, TF-IDF score are two usually used measures for w_{ij}. Thus, all the documents could be represented as a matrix D, which is also known as document-term matrix.

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix} \tag{6}$$

Similarly, suppose k topics are discovered in the documents and each of them can be taken as:

$$t_l = [w_{t1}, w_{t2}, \cdots, w_{tn}] \tag{7}$$

While all the topics could be noted as a topic-term matrix T as follows:

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{bmatrix} \tag{8}$$

Then, assume that each document is a mixture of various topics with different coefficients, we can get the following statement by putting all the documents and topics together. Here the matrix Ω represents the coefficients of topics for all the documents and is called document-topic matrix.

$$D = \Omega \cdot T = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1k} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{m1} & \omega_{m2} & \cdots & \omega_{mk} \end{bmatrix} \cdot T \tag{9}$$

In this way, topic modeling is to find the best combination of Ω and T. One way to solve this problem is to decompose D using singular value decomposition (SVD) and check the largest k singular values to find potential topics. This is an early topic modeling method called probabilistic latent semantic analysis (pLSA). In 2002, pLSA was generalized as LDA [45], which is widely adopted until now [46]. Nowadays, there already are a few tools. Among them a python tool called genism was selected for this research, so that word segmentation results could be easily passed to the topic modeling phase.

To evaluate the learned topic model and help us find the best topic number k, topic coherence model [47] and perplexity [45] are two commonly used methods. The former calculates topic coherence with different measures to evaluate assess the topic model, while the latter measures how well the learned topic model predicts a new sample. Generally, the higher coherence or lower perplexity imply a topic model is better. Formally, perplexity is calculated as follows:

$$per(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} log\, p(word_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{10}$$

Where M is the number of documents, $word_d$ represents the words in document d, and N_d is the number of words in document d.

In this research, perplexity is adopted together with a grid search strategy to find the best topic number *k*. That is, with the preset minimum and maximum value of topic number *k*, the algorithm will iterate possible topic numbers and automatically calculate perplexity of each model, then choose the model and topic number with lowest perplexity as the best one. Moreover, a few data visualization strategies are also used to help users interpret the topic modeling results.

E. IMPLEMENTATION

Considering the dynamic nature of a construction project, timely processing and analysis of on-site issues are important for decision-making and construction management. In other words, the process of data collection, preprocessing, text

analysis should be integrated and automated. To this end, a text mining pipeline is established to integrate modules implementing different steps of the proposed method (bottom of Figure 1).

First of all, the previously developed mini-program based on WeChat is deployed and adopted as a tool for everyday on-site issue collection and management. Collected data is then persisted in a relational database called PostgreSQL on Alibaba Cloud. PostgreSQL provides a standard database access protocol and application programming interface (DB API), which are used to save data sent from WeChat mini-program. Then, the steps of data preprocessing and text analysis are embedded in a python back-ended web service. Following the same way, collected data in Alibaba Cloud is filtered and retrieved from PostgreSQL database based on DB API, and sent to data association, standardization, and enrichment module based on python. Meanwhile, a python package called jieba is used for word segmentation and keyword extraction. After that, genism is adopted for topic modeling. Putting these packages scripts together, data preprocessing and text analysis could be conducted automatically. In this way, data preprocessing and text analysis are conducted periodically, and the results are cached at the service. Finally, Flask has adopted to integrate all the python scripts as a web service, through which front-end web pages such as online dashboard could access the results for visualization and decision-making purposes.

IV. Results and Discussion

This research utilizes the data of a real world project from one of the co-authors of this paper, making it easier to access the data and test the proposed method. All the data was collected from March 20, 2019 to October 28, 2019. Except for March, which is the beginning of data collection, there are about 800-1200 on-site issues collected each month (Figure 2). In about 7 months, 7821 issue records are created in total. Using previously mentioned strategies for data cleaning, 571 issue records are removed and 7250 records are kept for further data mining purposes.

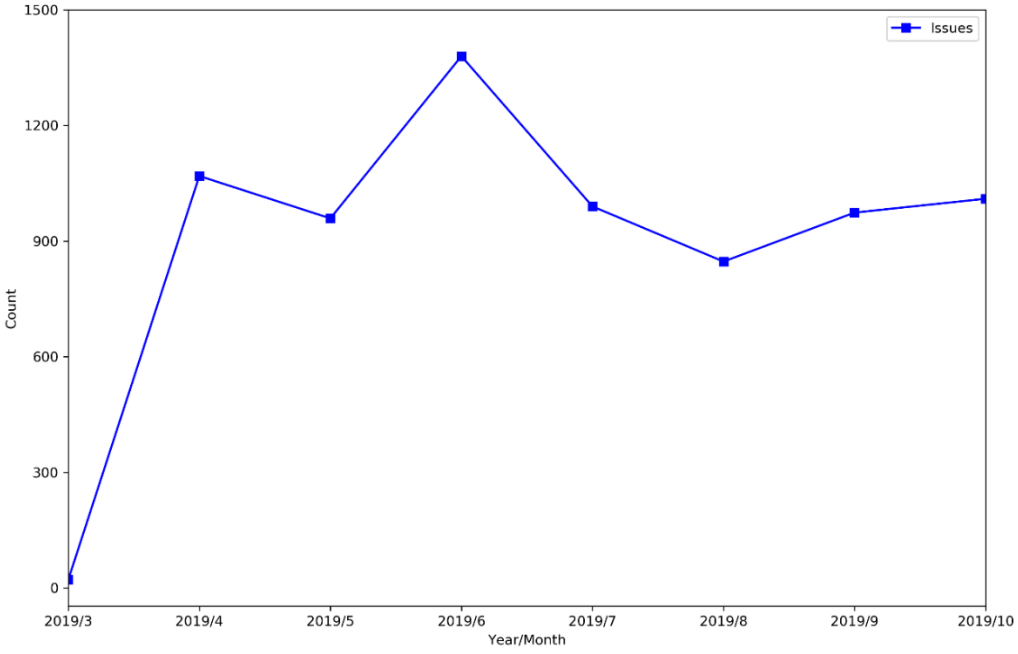

FIGURE 2. Number of Issues Generated Every Month.

A. KEYWORD EXTRACTION AND ANALYSIS

Firstly, all descriptions of cleaned issue records were put together, then word segmentation and tagging were conducted. After that, keywords are extracted and word cloud as well as word frequencies are visualized. In Figure 3, it could be concluded that the leading on-site issues are work violations (keyword "main building" and "violation"), and the workers forgot to wear safety belts quite frequently (keyword "worker", "safety belt" and "wear"). Meanwhile, it is also shown that these issues usually occur at the location called "main building". Besides, material-related hazards (keyword "material"), fire extinguishers, as well as fall protection (keyword "protection", "edge") are also common safety issues in this project.

In this way, managers could get a big picture of potential issues of a construction project in seconds without reviewing any of the submitted records.

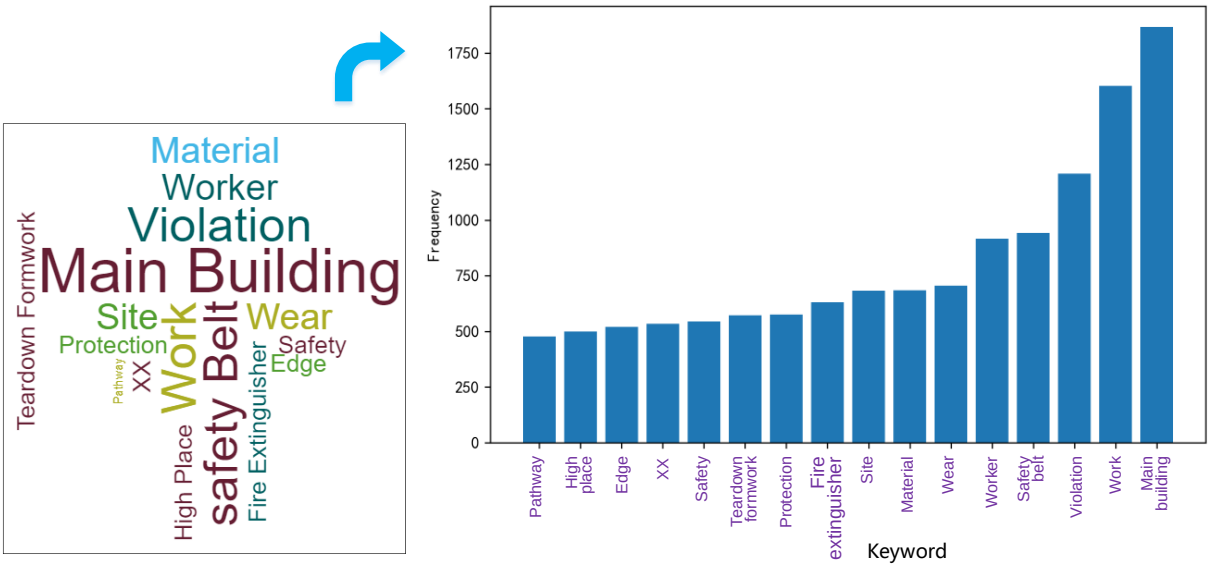

FIGURE 3. **Keywords and Frequency of Top 16 Keywords.**

Similarly, divided collected issue records into different months, word cloud and frequency of extracted keywords of each month are obtained. Combining the frequencies of the top 12 keywords of each month, it could be found that though the majority issues of each month are the same as the whole project, there are significant differences between different months. Take the top 12 keywords of 2019/4, 2019/8, 2019/10 as an example (Figure 4), one could conclude that there are more issues (word frequency is larger) related to work violations at the location of main building when the project starts in 2019/4, while things get better later. Moreover, in 2019/4, many issues are related to location at the main building, zone A, the foundation pit as well as subcontractor called "XX" (real name of the company is masked by XX for privacy consideration). This is, more attention should be paid to the above-mentioned areas and subcontractor to ensure the safety and quality of the construction. As the construction project goes, more efforts should shift to issues related to L40 storey, materials and formwork in 2019/8. While in 2019/10, in addition to common issues like work violations, construction cleanup and fall protection are also important for the safety of the project. Note that 2019/3 is not considered in comparison since it is the beginning of the application of the developed mobile App and there are only a few data records collected.

With comparison of keywords of each month, the managers can get a deep understanding on how the on-site issues change with time, and make better decisions to adapt to the change of issues. It is also worth noting that relationship between the identified changes of on-site issues and ongoing tasks could be established if construction progress or schedule of the project are considered in the analysis.

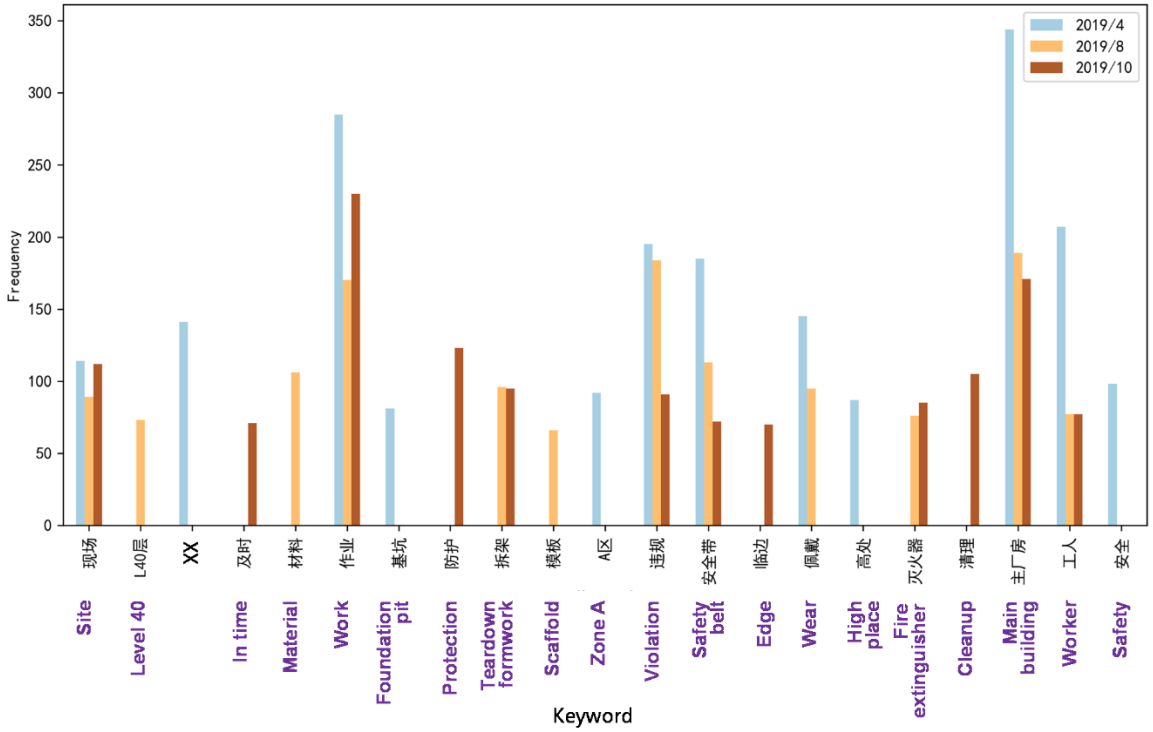

FIGURE 4. Frequency of Top 12 Keywords of 2019/4, 2019/8, and 2019/10.

B. TOPIC MODELING AND ANALYSIS

Keyword extraction provides managers a big picture of the dominant issues and how they change with time. However, types of issues and the amount of each type are still unclear to users. This is why topic modeling is introduced in this research.

First of all, the best number of topics should be determined. Since issue records are usually submitted by different engineers or foremen that belong to different divisions or groups of a construction team, therefore, this study assumes that each group is responsible for a limited kinds of on-site issues, so reasonable value of topic numbers should be between 2 and 15. A topic number larger than that would lead to topics with a small set of issues or cause problems for the interpretation of learned topic model. If it is the case, it is suggested to first divide the collected data records into different groups based on the division a creator belongs to. Thus, by enumerating all possible topic numbers between 2 and 15, and calculating the perplexity of each topic model, Figure 5 could be generated. As mentioned above, the lower the perplexity is, the better the topic model is, which means the best topic number is 5.

Fine tuning the model by gradually changing parameters of LDA and its learning process with the determined best topic number, 5 topics and their keywords are discovered from the collected issue records (Table I and Figure 6). With generated word clouds from learned topics in Figure 6, it is easy to conclude that: 1) topic 0 covers issues related to fire and electricity caused issues based on keywords “fire extinguisher”, “fire work”, and “socket” respectively; 2) topic 1 is mainly about issues related to site cleanup and material stacking since the top keywords are “cleanup”, “material”, and “stacking” respectively; 3) topic 2 focuses on work violations and safety belt based on keywords “work”, “worker”, “safety belt”, and “violation”; 4) topic 3 describes issues related to object falling with keywords “safety”, “tear down”, “below”, “warning sign”; 5) and finally, topic 4 is about edge protection since the keywords are “protection”, “edge”, and “hole”. Moreover, weights of each keyword that composes the topic are also provided with Table I.

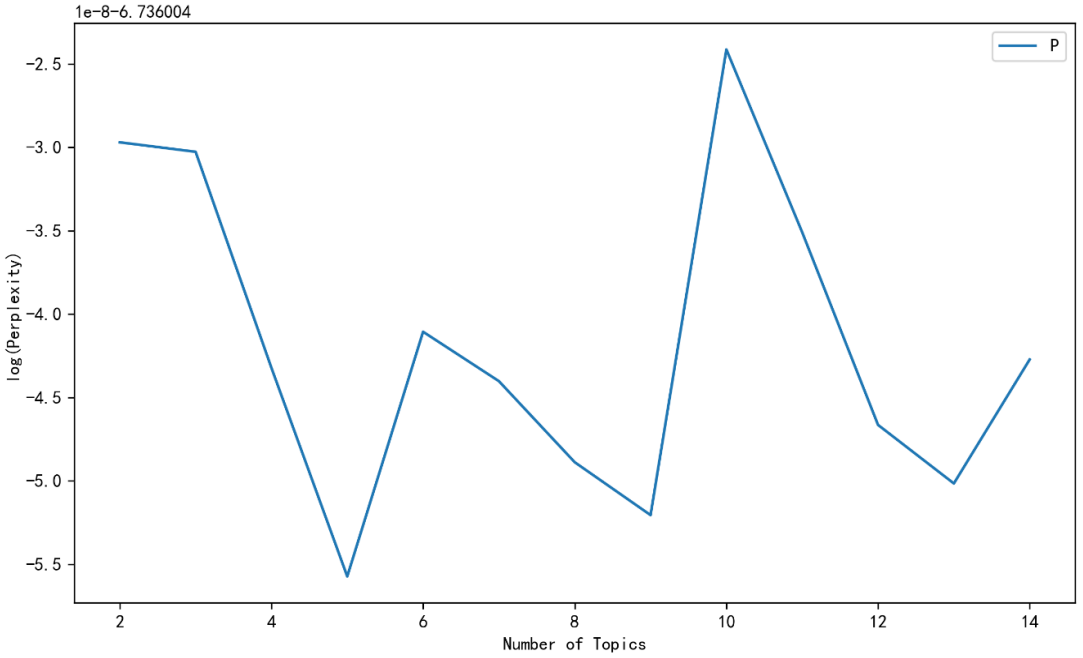

FIGURE 5. **Relationship between number of topics and perplexity of the topic model.**

FIGURE 6. **Word Clouds of Detected Topics.**

TABLE I
LEARNED TOPICS AND THEIR KEYWORDS

Topic #	Summary of Topic	Keywords
0	Issues related to usage of fire and electricity	0.084*“fire extinguisher” + 0.058*“usage” + 0.037* “socket” + 0.035*“box factory” + 0.032*“fire utilization” + 0.028*“roof” + 0.027*“inadequate” + 0.025*“north” + 0.023*“sundries” + 0.022*“steel structure”
1	Issues related to site cleanup and material stacking	0.061*“cleanup” + 0.055*“zone” + 0.048*“in time” + 0.046*“material” + 0.036*“scaffold” + 0.029*CUB + 0.022*“stack” + 0.021*“east” + 0.021*“lots of” + 0.021*“glass factory”

2	Issues related to work violations and safety belt	0.106*"work" + 0.075*"main building" + 0.046*"level 40" + 0.046*"tear down formwork" + 0.042*"site" + 0.036*"violation" + 0.034*"worker" + 0.028* "safety belt" + 0.028*"level 30" + 0.026*"people"
3	Issues related to work violations and safety belt	0.044*"safety" + 0.040*"below" + 0.038*"exist" + 0.035*"tear down" + 0.027*"steel structure" + 0.027*"fence" + 0.024*"store" + 0.022*"install" + 0.021*"lift" + 0.021*"warning"
4	Issues related to edge protection	0.122*"protection" + 0.055*"missing" + 0.042*"pathway" + 0.040*"edge" + 0.039*"method" + 0.024*"part" + 0.020* "waterproof" + 0.019*"hole" + 0.018*"expose" + 0.018*"stair"

Given that a document may have more than one topic, here we define the dominant topic of a document as the topic that takes the biggest portion (or has the biggest weight) of the document, and correspondingly define the document as a representative document of its dominant topic. In this way, the top 2 representative issue records of each topic are extracted and listed in Table II. Comparing the keywords of a topic and its representative records, it is clear that they share several words which reveals a strong connection between them (representative records of topic 0, 1, 2, 4). What's more, though a topic and its representative records share less or even no keywords, there still exists a reasonable connection between them (representative records or topic 3), this is because the introduced topic modeling approach also learns hidden relations between words and applies learned word relations in topic detection.

According to Table II, we can summarize different topics based on their keywords and also connect collected issue records to a topic by comparing the extracted keywords. The results are easy to explain and understand. Thus, one could say that topic modeling could extract hidden topics and learn their connections to related issue records, and shows good interpretability in solving real world problems. And, it could be concluded that topic modeling is quite appropriate for mining texts collected during on-site inspection, and the learned topics extract knowledge and information implicitly hidden in a large amount of texts. In this manner, topic modeling provides a new way for managers and engineers to get insights on what kinds of issues are there and figure out how to deal with the issues.

TABLE II
REPRESENTATIVE ISSUE RECORDS OF DIFFERENT TOPICS

Topic #	Keywords	Weight	Representative Record
0	fire extinguisher, usage, socket, box factory, fire utilization, roof, inadequate, north, sundries, steel structure	0.71191	…space for **fire utilization** in zone A is **inadequate**…
		0.69424	…**usage** of **sockets** is not compliant with…
1	cleanup, zone, in time, material, formwork, CUB, stack, east, lots of, glass factory	0.78632	…lots of **materials** are **stacked** beside Road 2# and site **cleanup** is needed
		0.77098	…**cleanup** of **materials** and wastes was not conducted **in time**
2	work, main building, level 40, tear down formwork, site, violation, worker, safety belt, level 30, people	0.85229	**work violation** happened when **tearing down formwork** of the **main building**, quite a few **workers** forgot to wear **safety belt**…
		0.82647	a **worker** was found not wearing **safety belt** when working at a high place of the **main building**…
3	safety, below, exist, tear down, steel structure, fence, store, install, lift, warning	0.71540	…**fences** were not **installed** around the **steel structure** and…
		0.71325	…a worker is found working **below** a **lift** that is **installing** the **steel structure**…
4	protection, missing, pathway, edge, method, part, waterproof, hole, expose, stair	0.68238	There is no **edge protection** for the **hole** of elevator
		0.68121	There lacks **edge protections** at the left side of the main building, and workers are **exposed** to falling object hazard

With topics discovered, the question of what are the most discussed topics and the proportion they take is easy to answer. For example, the left part of Figure 7 counts the number of documents by their dominant topics, and there are 2983 documents whose dominant topic is topic 2 and 1744 documents with their dominant topic as topic 1. Similarly, if we accumulate contributions of each topic to the documents, the right part of Figure 7 is obtained. One could get the same results that topic 2 and 1 are the two leading topics. However, there are differences between the left and right parts of Figure 7 too. Counting the number of documents by their dominant topics (NumDT) would make strong topics even stronger and weak topics even weaker since it omits the contributions of non-dominant topics. Thus, utilizing the number of documents

by topic weightage (NumTW) could provide people with more balanced results. That is, in a certain scenario, weak topic matters, i.e. , to calculate the importance of a specific category of on-site issues, NumTW would be better, while if we only care about strong topics, i.e., to answer the question of what is the dominant category (dominant topic) of on-site issues, then NumDT is recommended.

In this manner, construction managers and decision makers could understand the on-site inspected issues much better, and these numbers could be used as quantitative measures to help them optimize the issue resolving process and the employees involved in this process.

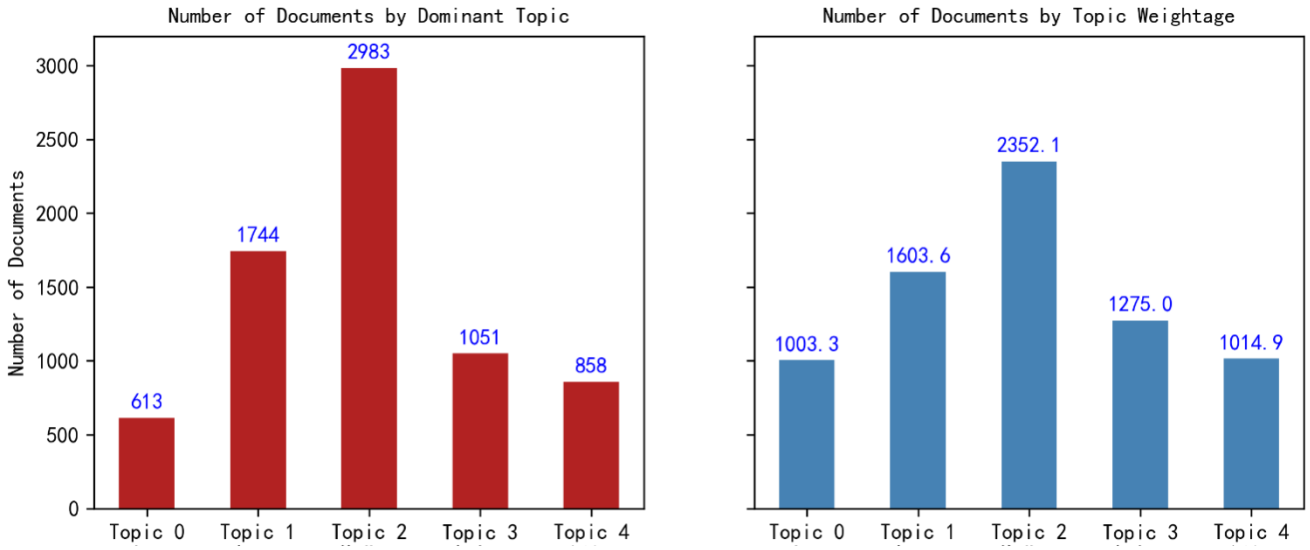

FIGURE 7. Most Discussed Topics.

In addition, Figure 8 is created by counting documents based on their dominant topics week by week. As illustrated in Figure 8, topic 2 and 1 are the leading topics, which means the majority of the efforts should be paid to them. The result is the same as mentioned before. Even so, topic 1 and 2 have different trends with time, the number of documents related to topic 2 increases dramatically at week 14, 23, 38, and 42 while the number of documents related to topic 1 keeps increasing during week 35-39. Thus, proper actions should be taken to prevent the increase of issues related to topic 1 and 2. Analysis of the other 3 topics could also be helpful in keeping the construction site safe and efficient.

As discussed above, Figure 8 shows how the number of documents of each topic changes with time, and the managers can utilize this method to identify the trends of different types of on-site issues, and decide how to eliminate them in the near future.

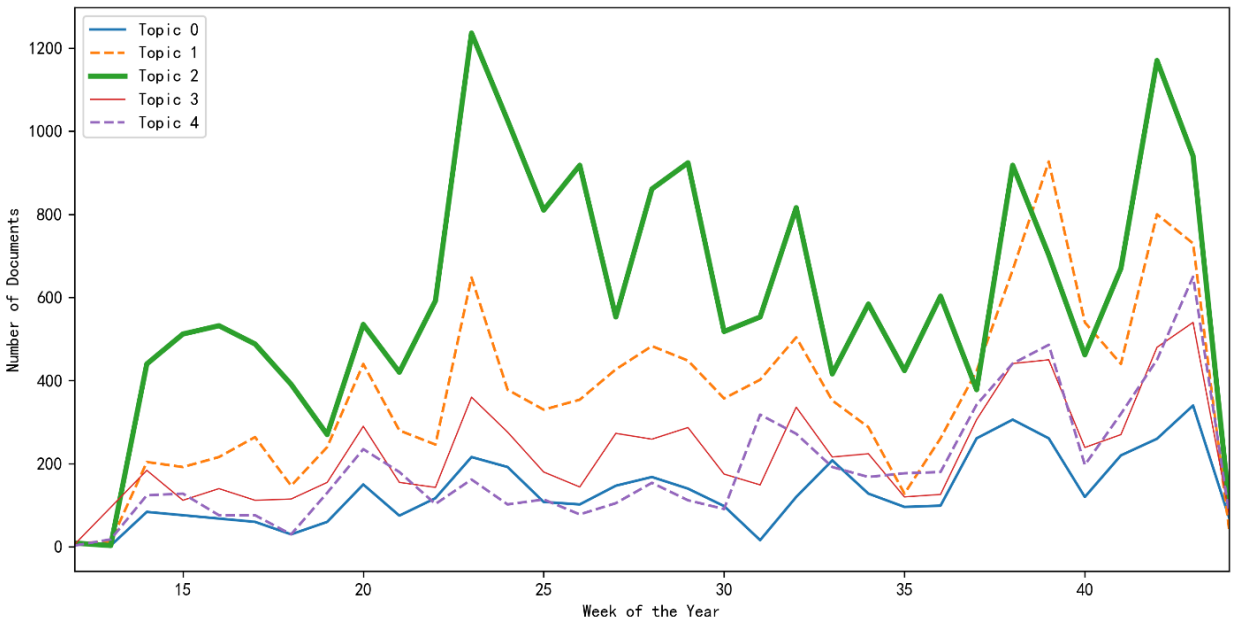

FIGURE 8. **Topic Change with Time.**

C. DISCUSSION

Though large dataset of images and video are collected and analyzed for on-site inspection and management, text data including inspection documents are still important since they capture ideas and knowledge directly from experienced engineers and workers. Lack of time and proper method to analyze on-site inspection documents could cause overlooking of on-site issues and lead to inefficient decision-making process.

To this end, this research introduces a new approach for mining on-site inspection texts based on keyword extraction and topic modeling, and tested the proposed method in a dataset with 7250 records which are collected in a real world project. Test and feedback from managers and engineers of the projects demonstrate the following results:

(1) Keyword extraction provides users a big picture of what's going on a construction site, while a fine-grained view with different topics is also available for the users based on topic modeling. With key concerns extracted from texts, managers could make decisions much easier and keep the construction site safer.

(2) Due to its dynamic nature, construction site and its related issues always change with time. Identifying changes of keywords and topics with time helps the managers adapt to changes of on-site issues in an agile way. Meanwhile, it could be possible to avoid more on-site issues once a rising trend of certain type of issues is detected. As a construction project usually involves various stakeholders, it is interesting to summarize on-site issues from different perspectives. However, the developed mini-program is mainly for site engineers and foremen, which means the collected data is limited, and it is hard for us to look at the problem from the perspective of a designer, or managing director. In the future, if more data is collected, it is valuable to further explore this problem.

(3) In this research, unsupervised learning method, namely, LDA, is utilized to discover hidden topics from on-site inspection records. To evaluate the performance of the model, perplexity is usually used. Lower perplexity means less overlap between different topics, and leads to a better model. Thus, a grid search strategy is adopted to find topic model with the lowest perplexity. Meanwhile, an engineer with rich experience in on-site inspection is also invited to further check the feasibility of the learned topic model. First, the issue records are divided into five groups based on their dominant topic number predicted by the learned model; then, 20 issue records are sampled from each group; finally, keywords of each topic and the sampled issue records are provided to the engineer, and he is asked to find the most relevant topic for each issue record. By comparing the predicted dominant topic and the most relevant topic selected by the engineer, Table III is obtained. It is concluded that the learned topic model could predict the dominant topic of on-site issues with an accuracy larger than 85%, and overall accuracy of the model is about 93%, which implies that the model is effective and the proposed approach is feasible.

TABLE III
ACCURACY OF TOPIC MODELING IN PREDICTING DOMINANT TOPICS

Dominant Topic #	Correct	Wrong	Accuracy
0	20	0	100%
1	18	2	90%

2	19	1	95%
3	17	3	85%
4	19	1	95%

(4) As mentioned before, accumulating topic contributions to each document is better than counting documents by dominant topics, and could provide more balanced results than the latter. Moreover, similar to the relationship between topics and documents, using weights of words is better than frequency of words for measuring the importance of keywords for different topics (Figure 9).

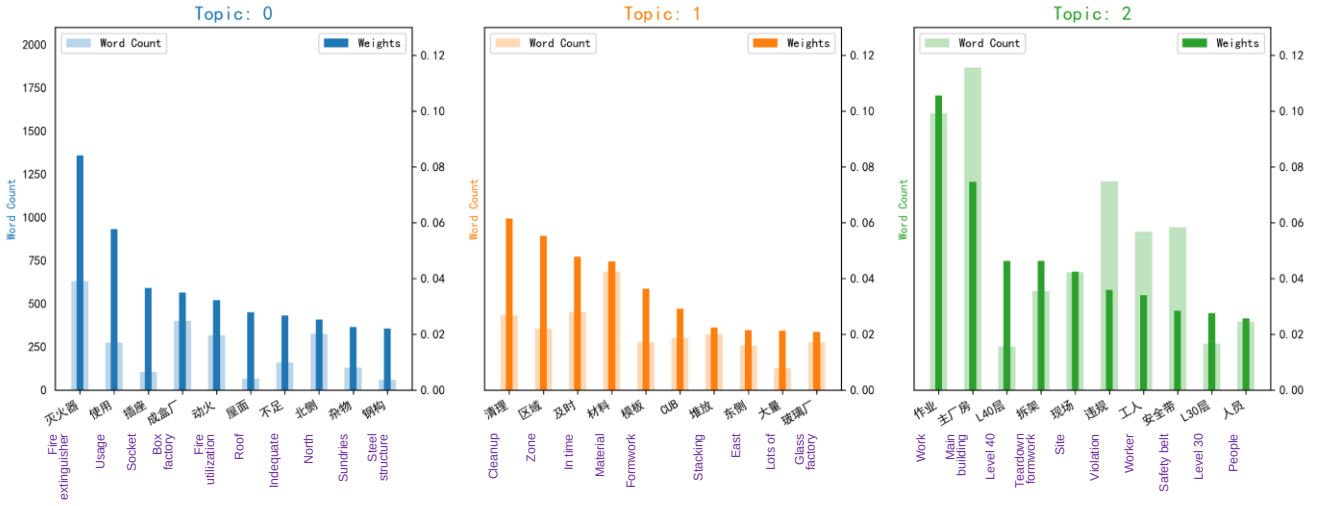

FIGURE 9. Importance of Keywords for Different Topics.

(5) Keyword extraction and even topic modeling identified a few words represent the name of subcontractors and workers (which are replaced as "XX" in this paper). Though it is useful to avoid construction issues, this is also a sign that privacy is more and more important in a big data era. And new methods are needed to explode the value of big data while protecting the privacy of people.

Moreover, on-site inspection records are generated and processed by engineers. It is also interesting to identify patterns between certain types of issues and the responsibilities of users or engineers, analyze the information flow and create new ideas to efficiently deal with on-site issues. In the future, as data of more projects is collected, it could be possible to horizontally compare different projects and find relationships between types of projects and issues.

Though Chinese texts are used in this research for validation purpose, the proposed method could be applied in mining textual data in other languages by utilizing corresponding tools for word segmentation. Since textual in English or other similar languages use whitespace and punctuations as explicit symbols to separate different words, word segmentation for these languages is quite simple and straightforward. While for Asian languages such as Korean and Japanese, due to lack of explicit separators, statistical models adopted in this research could be used, and then the proposed method in this research can be used to mine textual data in these languages.

V. CONCLUSION AND FUTURE WORK

Many new technologies are incorporated in collecting on-site inspection data nowadays, and a huge data set with images, videos, etc. is generated. Among different data formats, text data still plays an important role in construction. However, text mining techniques are not used for mining on-site inspection data based on the knowledge of the author. Therefore, to understand text-based on-site inspection data and improve the decision-making process, this research introduces a novel text mining approach for understanding on-site inspection issues and their dynamics based on keyword extraction and topic modeling. Validation and results showed that the proposed method could: 1) extract both course-grained (keywords) and fine-grained (topics) key concerns from a large text dataset, and 2) reveal their changes with time. Thus, this research provides construction managers with a deep understanding of what are the top issues and how many categories of issues are there as well as how these issues change with time, which are quite important in handling on-site issues in an efficient way and making better decisions. In a word, this research contributes to 1) the body of knowledge a novel approach to mining text-based construction data, and 2) to the state of practice a new way in applying text mining for understanding on-site inspection issues and their dynamics to make better decisions.

However, mining unstructured texts in the construction domain is still at the beginning, and future investigations are needed. Integration with building information modeling-based construction management [48], construction safety simulation and management [49], semantic analysis and reasoning

[50,51], as well as other data mining techniques [52] are encouraged. Moreover, looking at on-site issues from the perspective of different stakeholders is also valuable for both researchers and practitioners. Finally, as discussed in this paper, data privacy is more and more important in a big data era, more attentions such as sharing data in a cloud environment with consideration of data privacy [53] are also interesting to research and applications.

ACKNOWLEDGMENTS

This research is supported by the Natural Science Foundation of China (No. 51908323), the Beijing Municipal Science and Technology Project (No. Z181100005918006) and the Tsinghua University Initiative Scientific Research Program (No. 2019Z02UOT).

REFERENCES AND FOOTNOTES

REFERENCES

[1] M. Kopsida, I. Brilakis, and P. A. Vela, "A review of automated construction progress monitoring and inspection methods," in Proc. of the 32nd CIB W78 Conference 2015, 2015, pp. 421–431.
[2] M. Marzouk and M. Enaba, "Text analytics to analyze and monitor construction project contract and correspondence," Automation in Construction, vol. 98, pp. 265–274, 2019.
[3] K. S. Saidi, A. M. Lytle, and W. C. Stone, "Report of the NIST workshop on data exchange standards at the construction job site," in Proc. of 20th International Symposium on Automation and Robotics in Construction (ISARC), 2003, pp. 617–622.
[4] S. El-Omari and O. Moselhi, "Integrating automated data acquisition technologies for progress reporting of construction projects," Automation in Construction, vol. 20, no. 6, pp. 699–705, 2011.
[5] J. C. Garcia Garcia, "Construction Progress Control (CPC) application for smartphones," Journal of Information Technology in Construction, vol. 19, pp. 92-103, June 2014.
[6] B. McCulloch, "Automating field data collection in construction organizations," in Construction Congress V: Managing Engineered Construction in Expanding Global Markets, 1997, pp. 957–963.
[7] R. Navon and R. Sacks, "Assessing research issues in automated project performance control (APPC)," Automation in Construction, vol. 16, no. 4, pp. 474–484, 2007.
[8] M. Golparvar-Fard, J. Bohn, J. Teizer, S. Savarese, and F. Peña-Mora, "Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques," Automation in Construction, vol. 20, no. 8, pp. 1143–1155, 2011.
[9] K. Horak, S. M. DeLand, and D. S. Blair, "The feasibility of mobile computing for on-site inspection," SAND, vol. 2014, p. 18291, 2014.
[10] R. Y. M. Li, "An Institutional Economic Analysis on Construction Safety Knowledge Sharing and E-Learning via Mobile Apps," in Construction Safety and Waste Management, Springer Nature Switzerland, 2015, pp. 75–91.
[11] H. Zhang, S. Chi, J. Yang, M. Nepal, and S. Moon, "Development of a safety inspection framework on construction sites using mobile computing," Journal of Management in Engineering, vol. 23, p. Article–Number, 2016.
[12] Z. Zhu, S. German, and I. Brilakis, "Detection of large-scale concrete columns for automated bridge inspection," Automation in construction, vol. 19, no. 8, pp. 1047–1055, 2010.
[13] S. Han, S. Lee, and F. Peña-Mora, "Vision-based detection of unsafe actions of a construction worker: Case study of ladder climbing," Journal of Computing in Civil Engineering, vol. 27, no. 6, pp. 635–644, 2013.
[14] Z. Ma, S. Cai, N. Mao, Q. Yang, J. Feng, and P. Wang, "Construction quality management based on a collaborative system using BIM and indoor positioning," Automation in Construction, vol. 92, pp. 35-45, 2018.
[15] Y. Zou, A. Kiviniemi, and S. W. Jones, "Retrieving similar cases for construction project risk management using Natural Language Processing techniques," Automation in construction, vol. 80, pp. 66–76, 2017.
[16] J. Hsu, "Content-based text mining technique for retrieval of CAD documents," Automation in construction, vol. 31, pp. 65–74, 2013.
[17] H. Fan, F. Xue, and H. Li, "Project-based as-needed information retrieval from unstructured AEC documents," Journal of Management in Engineering, vol. 31, no. 1, p. A4014012, 2015.
[18] M. Al Qady and A. Kandil, "Automatic clustering of construction project documents based on textual similarity," Automation in construction, vol. 42, pp. 36–49, 2014.
[19] C. H. Caldas, L. Soibelman, and J. Han, "Automated classification of construction project documents," Journal of Computing in Civil Engineering, vol. 16, no. 4, pp. 234–243, 2002.
[20] C. H. Caldas and L. Soibelman, "Automating hierarchical document classification for construction management information systems," Automation in Construction, vol. 12, no. 4, pp. 395–406, 2003.
[21] T. P. Williams and J. Gong, "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers," Automation in Construction, vol. 43, pp. 23–29, 2014.
[22] S. Yarmohammadi, R. Pourabolghasem, and D. Castro-Lacouture, "Mining implicit 3D modeling patterns from unstructured temporal BIM log text data," Automation in Construction, vol. 81, pp. 17–24, 2017.
[23] N. Jung, and G. Lee, "Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning," Advanced Engineering Informatics, vol. 41, pp. 100917, 2019.
[24] S. Choo, H. Park, T. Kim, and J. Seo, "Analysis of trends in Korean BIM research and technologies using text mining," Applied Sciences, vol. 9, no. 20, pp. 4424, 2019.
[25] Y. Lai, and C. E. Kontokosta, "Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities," Computers, Environment and Urban Systems, vol. 78, pp. 101383, 2019.
[26] N. Arora and A. Ogra, "Mobile GIS for construction quality managers and surveyors," in Proc. of GISSA Ukubuzana 2012, 2012. Johannesburg, South Africa. October 15th, 2012. ISBN 978-0-620-52913-6
[27] Y.-H. Tsai, S.-H. Hsieh, and S.-C. Kang, "A BIM-enabled approach for construction inspection," in Computing in Civil and Building Engineering (2014), 2014, pp. 721–728.
[28] Y. Zhou, H. Luo, and Y. Yang, "Implementation of augmented reality for segment displacement inspection during tunneling construction," Automation in Construction, vol. 82, pp. 112–121, 2017.
[29] M. Kopsida and I. Brilakis, "BIM registration methods for mobile augmented reality-based inspection," in eWork and eBusiness in Architecture, Engineering and Construction: ECPPM 2016: Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM 2016), Limassol, Cyprus, 7-9 September 2016, 2017, p. 201.
[30] R. E. Asbahan and P. DiGirolamo, "Value of Tablet Computers in Transportation Construction Inspection: Ongoing Case Study of Projects in Pennsylvania," Transportation research record, vol. 2268, no. 1, pp. 12–17, 2012.
[31] J. Yamaura and S. T. Muench, "Assessing the Impacts of Mobile Technology on Project Inspection," in Proc. of Transportation Research Board 95th Annual Meeting, no. 16-6009, 2016.
[32] T. Omar and M. L. Nehdi, "Data acquisition technologies for construction progress tracking," Automation in Construction, vol. 70, pp. 143–155, 2016.
[33] N. Ur-Rahman and J. A. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," Expert Systems with Applications, vol. 39, no. 5, pp. 4729–4739, 2012.
[34] M. Martínez-Rojas, N. Marín, and M. A. Vila, "The role of information technologies to address data handling in construction

project management," Journal of Computing in Civil Engineering, vol. 30, no. 4, p. 04015064, 2016.
[35] J. Zhang and N. M. El-Gohary, "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking," Journal of Computing in Civil Engineering, vol. 30, no. 2, p. 04015014, 2016.
[36] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports," Automation in Construction, vol. 62, pp. 45–56, 2016.
[37] J. Zhang and N. M. El-Gohary, "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking," Journal of Computing in Civil Engineering, vol. 30, no. 2, p. 04015014, 2016.
[38] J.-R. Lin, Z.-Z. Hu, J.-P. Zhang, and F.-Q. Yu, "A natural-language-based approach to intelligent data retrieval and representation for cloud BIM," Computer-Aided Civil and Infrastructure Engineering, vol. 31, no. 1, pp. 18–33, 2016.
[39] Y.Wang, H. Li, and Z. Wu, "Attitude of the Chinese public toward off-site construction: A text mining study," Journal of Cleaner Production, vol. 238, pp. 117926, 2019.
[40] Y. Jallan, E. Brogan, B. Ashuri, and C. M. Clevenger, "Application of Natural Language Processing and Text Mining to Identify Patterns in Construction-Defect Litigation Cases," Journal of Legal Affairs and Dispute Resolution in Engineering and Construction vol. 11, no. 4, pp. 04519024, 2019.
[41] S. Zhou, S. T. Ng, S. H. Lee, F. J. Xu, and Y. Yang, "A domain knowledge incorporated text mining approach for capturing user needs on BIM applications," Engineering, Construction and Architectural Management, vol. 27, no. 2, pp. 458-482, 2019.
[42] J.-R. Lin, J. Zhang, D. Wu, B. Li, and H. Gao, "Application Framework for On-Site Quality and Safety Inspection based on WeChat," in Proc. of the 17th International Conference on Computing in Civil and Building Engineering, 2018.
[43] X. Sun, H. Wang, and W. Li, "Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2012, pp. 253–262.
[44] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," Expert Systems with Applications, vol. 57, pp. 232–247, 2016.
[45] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.
[46] M. Hajjem and C. Latiri, "Combining IR and LDA topic modeling for filtering microblogs," Procedia Computer Science, vol. 112, pp. 761–770, 2017.
[47] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," Expert Systems with Applications, vol. 42, no. 13, pp. 5645–5657, 2015.
[48] M. Kassem, N. Dawood, and R. Chavada, "Construction workspace management within an Industry Foundation Class-Compliant 4D tool," Automation in Construction, vol. 52, pp. 42–58, 2015.
[49] J.-R. Lin, J.-P. Zhang, X.-Y. Zhang, and Z.-Z. Hu, "Automating closed-loop structural safety management for bridge construction through multisource data integration," Advances in Engineering Software, vol. 128, pp. 152–168, 2019.
[50] Y.-W. Zhou, Z.-Z. Hu, J.-R. Lin, and J.-P. Zhang, "A review on 3D spatial data analytics for building information models," Archives of Computational Methods in Engineering, vol. 27, no. 5, pp. 1449–1463, 2020.
[51] T. He, J. Zhang, J. Lin, and Y. Li, "Multiaspect similarity evaluation of bim-based standard dwelling units for residential design," Journal of Computing in Civil Engineering, vol. 32, no. 5, p. 04018032, 2018.
[52] Y. Peng, J.-R. Lin, J.-P. Zhang, and Z.-Z. Hu, "A hybrid data mining approach on BIM-based building operation and maintenance," Building and environment, vol. 126, pp. 483–495, 2017.
[53] J. Zhang, Q. Liu, Z. Hu, J. Lin, and F. Yu, "A multi-server information-sharing environment for cross-party collaboration on a private cloud," Automation in Construction, vol. 81, pp. 180–195, 2017.

JIA-RUI LIN received the B.S. and Ph.D. degrees from the Department of Civil Engineering, Tsinghua University, China, in 2011 and 2016, respectively. He is currently a Research Assistant Professor with the Department of Civil Engineering, Tsinghua University.

His research interests are information technology for building and civil engineering, including building information model (BIM), augmented reality (AR), machine learning and internet of things (IoT).

ZHEN-ZHONG HU was born in Guangdong, China. He received the B.S. and Ph.D. degrees from the Department of Civil Engineering, Tsinghua University, China, in 2005 and 2009, respectively. He is currently an Associate Professor with the Department of Civil Engineering, Tsinghua University.

His research interests include information technology in civil engineering, building information model, and digital disaster prevention and mitigation.

JIU-LIN LI was born in HeBei, China. He received the master's degree from China University of Geosciences, China, in 1991 and 1992. He is currently the deputy chief engineer of Beijing Urban Construction Group Co., Ltd.

His research interests include modern construction technology of Olympic venues, construction technology of extra-large span and complicated steel structure, and green construction and smart construction technology.

LI-MIN CHEN was born in Shanxi, China. He received the Ph.D. degrees from the Department of Geological Engineering, China University of Mining and Technology-Beijing, in 2015. He is currently a Senior Engineer with the Beijing National Speed Staking Oval Operation CO., Ltd.

His research interests include building information model and smart construction.