ORIGINALE RESEARCH

Application of NLP-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies

Murali Jagannathan1 • Debopam Roy1 • Venkata Santosh Kumar Delhi2

Received: 11 March 2022 / Accepted: 5 May 2022 / Published online: 13 May 2022
© CSI Publications 2022

Abstract The construction industry is the backbone of a nation’s economy. It is a matter of great concern that such an industry suffers from time and cost overruns, especially in these challenging times. Coupled with the overrun issues, the sector is often criticized for lacking adequate quality and quantity of structured secondary data. The emerging technologies in data science and machine intelligence present a unique opportunity to understand the sector better and aid in effective decision-making. To better understand the utility of such technologies, the Management Discussion and Analysis sections of the annual reports of publicly listed top Indian construction contracting firms are analyzed to identify the presence of ‘strategy themes’ and further map them to the organizations considered. Natural Language Processing (NLP)-based topic modeling algorithms, namely Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), are used in this study to perform a qualitative content analysis to identify the latent themes. From a methodological standpoint, considering the context of this study, the NMF results are better in accuracy, precision, and recall compared with the LDA. The results show that while most construction contracting firms prioritized a ‘revenue-focused’ strategy to expand their order books, a smaller set of large-sized firms seem to prioritize process improvement to improve their execution productivity and therefore are ‘profit margin improvement focused’ or ‘lean-focussed’ in their approach. Although a proof-of-concept, this study unlocks the immense potential of unsupervised NLP-based topic-modeling tools to understand and infer from unstructured and freely available text data in the public domain to aid sectoral analysis and policymaking.

Keywords Strategy • Construction • NLP • Topic modelling • LDA • NMF

1 Introduction

The construction sector is operating in unprecedented times. The industry is now facing the onslaught of the Coronavirus pandemic (and its aftereffects), which critically affects the sector’s value generating potential. Top management decision-making is crucial to successfully steer organizations in these trying times [1]. ‘Top management,’ in this study, refers to the group of individuals in a construction contracting firm who are responsible for strategic decision-making. Construction management researchers and industry professionals can collaborate, and collective efforts can be crucial in informed decision-making. However, researchers require data to analyze the trends, patterns, and scenarios for informed decision-making to propose any solution irrespective of the domain. In addition to the quality and quantity requirements, if the available data is structured and quantifiable, it becomes easier for researchers to analyze them and report findings quickly. Unfortunately, the construction sector, at least in India, is often criticized for being poor in maintaining a research-ready database that can serve as a publicly available and dependable secondary data source for researchers [1–4]. Private-sector data is generally

---

1 National Institute of Construction Management and Research (School of Construction Management), Pune, Maharashtra, India
2 Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India
considered confidential and is out of bounds for researchers. Therefore, researchers often bank upon primary data collection techniques like questionnaire surveys and interviews to equip themselves with data [5].

While the structured-data inadequacy issue is a severe concern in the construction sector, unstructured text data related to the construction sector is abundantly available in the public domain. In the private sector context, companies’ websites contain information on vision and mission statements and promotional videos. Specifically, in the case of publicly listed private firms, annual reports and financial statements are mandatory disclosures in the public domain. Knowledge extraction from such unstructured data is now possible with the recent developments in computer-aided text mining and Natural Language Processing (NLP) [6–8]. In this research, the authors explore the efficiency of NLP-based topic modeling algorithms to extract keywords and topics from the publicly available annual reports of construction contracting firms and use the information obtained to analyze the strategies such firms adopt in dealing with emerging sectoral challenges explained in the next section.

2 Opportunities and sectoral challenges

Considering the stakeholder-intensive and labor-oriented nature of the industry, the pandemic has made construction execution challenging due to its impact on the logistics and supply chain efficiency and the working style of people involved in the project delivery. Notwithstanding the pandemic impact, the construction sector cannot pause even momentarily, as it plays a central role in shaping a nation’s economy. With the pandemic affecting virtually all sectors of the economy, a renewed growth of the construction sector can infuse a much-needed boost to propel economic growth. Realizing this, the Government of India has announced massive investments in the infrastructure sector that can revive the construction industry. Schemes like the National Infrastructure Pipeline (NIP) envisage providing world-class infrastructure to citizens and improving their quality of life [9]. NIP is an ambitious multi-billion dollar investment plan spread over infrastructure sectors like energy, roads, urban development, and railways, with 2465 projects currently under development in active partnership with the private sector [9].

While such massive investments open up much-needed avenues for the private sector to bounce back, there are challenges in converting the opportunity to benefits. The tangible positive impact of the opportunities rolled out by the Government on the construction sector can be seen only if there are structural changes in the execution style of construction contracting companies involved in infrastructure project deliveries. The observation above is considering the fact that the industry traditionally suffers from significant cost and time overruns, ultimately leading to disputes and litigation [10, 11]. Therefore, with such a high probability of overruns and a poor track record in claims management and dispute resolution [12], there is a risk of massive investments in infrastructure (like the NIP) turning into breeding grounds for claims and disputes. Overall, there seems to be a two-pronged attack on the contracting construction sector. Firstly, the burden of unresolved pre-pandemic claims and disputes stresses a construction contracting company’s balance sheet. Secondly, the onslaught of the pandemic is forcing organizations to secure their strained profit margins from deteriorating further. Considering the two-pronged attack (as described above) a challenge faced by construction contracting firms amidst opportunities, this article reviews the strategies Indian construction contracting firms adopted in response to the challenges.

Considering the evolving critical financial situation and the impending opportunities the Government is creating, construction contracting firms can strategize their future path in several ways. While embarking on a revenue-focused strategy [13] can be one option, profit margin improvement [14, 15] through cost reduction exercises, liquidating long-pending claims, mergers, and acquisitions, among others, can be the other option. While it is not intended to infer that the strategies above are mutually exclusive, the study explores the predominant strategy adopted by construction contracting firms in response to the evolving scenario. The study uses an NLP-based topic modeling algorithm to qualitatively analyze the management viewpoints from the annual reports of publicly-listed top construction contracting firms in India to understand the strategies adopted.

3 Literature review

In the face of calamities, the role of top management takes center stage [16, 17]. Therefore, understanding the top management view becomes crucial to decipher an organization’s strategy to manage an emerging scenario unfolding risks (due to the pandemic) and opportunities (Government’s infrastructure push through capital investment). Considering the nature of the study, it is essential to refer to those documents and materials that can help explain an organization’s strategic intent. Such reference documents should be authentic and also available in the public domain. Considering the requirement, the annual reports of publicly listed construction contracting firms are analyzed to understand top management’s commitment to
formulating and communicating strategies in response to the risks and opportunities evolving in the market.

Publicly listed companies have a regulatory requirement to publish annual reports containing the company’s financial information and the management opinion about the performance in the previous year and prospects [18]. Since an annual report is a public document read by a sizeable heterogeneous audience, companies may also treat them as an instrument of mass communication of their achievements and innovations [19]. In academics, annual reports have often been used as grey literature, especially for research on corporate strategy [20, 21]. Specifically, content analysis of annual reports of multiple companies is an effective way to gauge the industry’s strategy [22]. In particular, annual reports, among others, are used to understand leadership phenomena [23].

As per the statutory requirements laid down by the Ministry of Corporate Affairs (MCA), Government of India, and the relevant governing laws of the land, an annual report is typically divided into three parts, namely corporate overview, statutory reporting, and financial statements. Among the three parts, the second part, statutory reporting, consists of a section titled ‘management discussion and analysis’ (MDA). According to the Securities and Exchange Board of India (SEBI), MDA should include discussion on matters about a company’s competitive position, such as (1) overall business scenario (global, national, and sectoral), (2) opportunities and threats, (3) highlights of the performance, (4) business outlook, (5) risks, concerns and mitigation plans, (6) internal control systems and their adequacy, and (7) discussion on financial performance concerning operational performance and human resources/industrial relations initiatives [24]. From the very structure of this section, it is clear that the report captures the management’s view on the impending threats and opportunities and the organization’s response plan. Therefore, the contents of MDA sections of multiple construction contracting firms are analyzed to understand the direction of top management thinking in response to the emerging scenario.

The qualitative analysis technique is employed to analyze the text data in the MDA section of the annual reports. Earlier studies have attempted to understand such reports using a keyword-based search of specific pre-decided themes [25]. In this study, however, the themes will be an output of the study, and therefore the keyword-based search is not suitable. Since the study aims to identify various topics/themes evolving out of the annual reports’ MDA section, authors use tools that help cluster topics from raw text. Topic modeling algorithms like Natural Language Processing (NLP)-based Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) can be employed to meet the objectives of this study. Such NLP-based tools have been used widely in the construction management research domain [26]. LDA, in particular, is widely used in several domains as a tool for topic modeling [27]. Table 1 below presents a non-exhaustive summary of construction management domain research employing topic modeling, particularly LDA models. The articles are obtained by employing a keyword search in the Google Scholar Database. The keywords used are ‘topic modeling,’ ‘NLP,’ ‘Construction,’ and ‘Project.’

While the keyword-based search (in the Google Scholar database) highlighted many recent LDA-based articles in construction management, the search is further narrowed by adding the phrase ‘annual report’ to the keywords to identify construction management domain-specific topic modeling studies employing LDA on annual reports. The result indicated that topic modeling on annual report documents is quite common in finance [36]. Additionally, it is also seen that the search (with the key phrase – ‘annual report’) did not yield any similar articles in the construction domain. The study attempts to address the gap in the literature using NLP-based topic modeling techniques.

While NLP-based techniques are commonly used for detailed analysis of text from standard contract forms to identify problematic provisions and control-focus, among other things [7, 37], topic modeling algorithms are rarely used to extract keywords and topics from standard contract forms as the text from those documents are already classified into topics in the form of clauses and sub-clauses with clear headings and sub-headings. However, extraction of strategies (represented through keywords) from the annual reports in response to market challenges is relatively more challenging because companies do not explicitly lay down their action plan. The same can be gathered only after reading through all the relevant pages of the annual report. Therefore, topic modeling algorithms can potentially help in a quick and reliable extraction of strategies from annual report documents. The MDA section of the annual reports falls under relatively (relative to contract documents) more unstructured data categories such as news articles, social media data, and court cases where topics are not identified as explicitly as that of a standard form contract clauses. The following sections describe the methodology adopted for the study, followed by results, validation, discussion, research implications, future scope, and conclusion.

4 Methodology

As explained in the earlier sections, the study employs qualitative topic modeling techniques to thematically analyze the MDA section of annual reports of publicly listed construction contracting companies in India. The
The construction sector in India consists of more than 31,000 registered firms as well as a much larger number of contractors in the unorganized sector, but only around 100 companies are listed in the National Stock Exchange (NSE) and Bombay Stock Exchange (BSE) [38]. India’s leading 18 publicly-listed construction contracting companies are selected for the study considering their market capitalization, with representations from real estate, infrastructure, and commercial development sectors. Next, the NLP-based LDA and NMF approach models latent topics in the input text corpus (MDA section of annual reports). Anaconda 3, a pre-packaged distribution of an open-source Python programming interface, is used in the Jupyter Notebook web application to perform topic modeling.

### 4.1 Topic modeling process

Two topic modeling algorithms, namely LDA and NMF, are employed in this study. LDA is a Bayesian-based statistical model. It assumes that each document is a combination of a given number of topics, and each word in the corpus is associated with each of the given topics with some probability. The model builds a probability distribution according to which topics are modeled by allocating words under them [39]. NMF, on the other hand, is a linear algebraic model. While NMF attempts to achieve the same objective, topic modeling, NMF is a matrix factorization and multivariate analysis technique that generates coefficients (instead of probability) for each word while mapping them to a given topic. LDA and NMF models have been employed for topic modeling, and earlier studies in non-construction contexts (large text stream data analysis and review data analysis) report superior performance of one algorithm over the other [40, 41]. As there was no data on the robustness of the topic modeling algorithms in the construction context, both LDA and NMF algorithms are used in this study. The coding process of LDA and NMF is almost similar except for some changes in the commands and tools called by the program. Therefore, LDA coding is first described, followed by a few lines highlighting key differences in the NMF coding process.

Topic modeling is carried out in a sequential process. The process begins with preparing input corpus in a format recognizable by the programming language. The data from the MDA section of the annual reports (which are primarily published in the Portable Document Format or PDF) is copied and pasted into a Comma-Separated Values (CSV) file with Unicode Transformation Format, 8 Bit (UTF-8) encoding. This enables the programming interface to recognize and skim the data in the files. The coding process is then initiated in the Jupyter Notebook interface using open-source software libraries designed for data analysis and manipulation through Python programming language [42]. The input corpus (a CSV file) is then read into the program by invoking necessary programming commands. After importing the necessary libraries, the text data in the CSV file (input corpus) is pre-processed by instructing the program to ignore those words that are commonly found in more than 95% of the documents (95% of 18 companies) and also ignore the words that are found in less than or equal to 2 documents (2 out of 18 companies). Secondly, the common English language stop words such as “a,” “the,” “is,” and “are” are eliminated in the process. Overall, pre-processing eliminates commonly used words across documents that may not contribute in differentiating the companies. Similarly, words repeated in only two or fewer company MDA data indicate that the words are unique to a particular company and may not be relevant while modeling for latent topics hidden across the text corpus. The input data is now ready for the entire text data’s ‘fitting and transformation,’ called ‘unsupervised learning.’ Finally, the algorithm extracts four latent topics from the input corpus. While the number of topics is initially set to ‘4,’ the input parameters are subsequently modified to generate more models for comparison.

The output obtained is presented in the results section. While the above steps are for LDA, similar steps are followed for NMF-based topic modeling. However, the difference is that a ‘Term Frequency and Inverse

| Type of text data                  | Topic modeling technique                                      | Authors |
|-----------------------------------|----------------------------------------------------------------|---------|
| Permit descriptions and other urban data | LDA                                                            | [28]    |
| Site issue records                | LDA                                                            | [29]    |
| Defect litigation cases           | LDA                                                            | [30]    |
| Building information modeling (BIM) case studies | LDA, Latent Semantic Analysis (LSA), and Support Vector Machine | [31]    |
| Construction market news          | LDA                                                            | [32]    |
| Construction schedules            | LDA, LSA, word2vec and fastText                                 | [33]    |
| Construction specifications       | Named Entity Recognition (NER) models                         | [34]    |
| Building regulations (rule-checking) | NLP and Deep Learning-based semantic analysis processes       | [35]    |
Document Frequency (or ‘tfidf’) vectorization is used instead of the ‘CountVectorizer,’ and the ‘NMF’ tool is called instead of the ‘LatentDirichletAllocation’ tool. Finally, in both LDA and NMF-based models, topics are assigned to the input corpus text to display and interpret the results.

5 Results

The keywords assigned to the four topics are displayed by calling the LDA component function (NMF components in the NMF model). However, the output is in the form of vectorized values, as shown in Fig. 1. The output values have to be sorted in their probability of occurrence under a given topic and converted into actual words so that the topic names can be deciphered. The output in the form of keywords is shown in Fig. 2.

6 Naming the topics and validation of results

To begin with, as discussed in the methodology section, the number of topics sought from the model is four. On observing the components, there are broadly two themes of strategy spread across the four topics. One set of words referred to an inward-looking strategy to improve the execution processes (like ‘digital,’ ‘technology,’ ‘productivity,’ ‘employee,’ and ‘value’), and the second set indicated an outward-looking strategy that involves interaction with external stakeholders (words like ‘government,’ ‘clients’ and ‘business’). Considering the presence of two latent themes, the algorithm is again executed with an instruction to generate two topics. The LDA and the NMF topic modeling program results for the 2-topic model are shown in Figs. 3 and 4. The results show that except for the companies with index numbers 1, 6, and 15, the LDA and NMF models show similar results. Since the LDA and the NMF tools model the topics using different algorithms, there can be slight differences in the set of words under topics (topic 0 and topic 1).

The two topics modeled and allocated by the LDA and NMF tool (shown in Table 2) indicate ‘latent’ themes in the corpus text. On a keen observation, it is seen that topic-1 contains words such as ‘team,’ ‘digital,’ ‘process,’ ‘management,’ ‘productivity,’ ‘technology,’ ‘learning,’ ‘safety,’ ‘innovation,’ and ‘quality.’ Here, the term ‘digital’ refers to the implementation of technologies that help in increased visualization through tools and techniques such as Building Information Modeling (BIM). Similarly, the term ‘team’ can be equated to a ‘collaborative’ mindset amongst stakeholders. According to [25], the above terms correspond to ‘lean signs,’ and such signs indicate a construction contracting organization’s keenness to implement lean construction techniques. Lean construction techniques help reduce waste, thereby increasing the productivity of the construction execution operations [43] and improve the profit margin. Therefore, the term profit margin improvement-focussed strategy is renamed lean-focused-strategy considering the positive impact of lean on profit-margin improvement (through waste reduction and productivity improvement). Rightly, even the word ‘productivity’ has popped up under the ‘topic-1’. Therefore, the ‘topic-1’ will be named the ‘lean-focussed’ strategy. On the other hand, the words grouped under ‘topic 0’ stress ‘growth,’ ‘government,’ ‘capital,’ ‘order,’ and ‘industry’ indicating that the strategy focuses on external opportunities to propel revenues. This is a seemingly different strategy compared to the inward-focused lean strategy of certain other organizations. Therefore, the name ‘revenue-focused strategy’ is retained.

Choosing the keywords “collaboration” and “team”, a phrase extraction algorithm is applied to extract 30 words on either side of the keywords from one company to check if the extracted keywords are relevant considering the lean context. The result in one case is shown in the Fig. 5. The result shows that the keywords are used in the lean context, aimed at improving the work productivity.
THE TOP TWENTY WORDS FOR TOPIC #0
['quality', 'lead', 'sector', 'risks', 'equipment', 'government', 'internal', 'execution', 'time', 'cost', 'growth', 'profitability', 'business', 'raw', 'management', 'construction', 'project', 'infrastructure', 'risk', 'company']

THE TOP TWENTY WORDS FOR TOPIC #1
['safety', 'technologies', 'impact', '19', 'covid', 'solutions', 'based', 'order', 'infrastructure', 'process', 'quality', 'digital', 'new', 'risk', 'construction', 'management', 'risks', 'project', 'business', 'company']

THE TOP TWENTY WORDS FOR TOPIC #2
['operations', 'process', 'technology', 'employee', 'new', 'processes', 'execution', 'employees', 'time', 'team', 'work', 'quality', 'construction', 'various', 'digital', 'business', 'safety', 'management', 'project', 'company']

THE TOP TWENTY WORDS FOR TOPIC #3
['clients', 'india', 'book', 'control', 'rs', 'working', 'equipment', 'mitigation', 'capital', 'management', 'cost', 'construction', 'risks', 'time', 'order', 'infrastructure', 'execution', 'business', 'project', 'company']

**Fig. 2** Four topics with the top twenty keywords mapped under them

THE TOP TWENTY WORDS FOR TOPIC #0
['operations', 'industry', 'capital', 'growth', 'government', 'sector', 'cost', 'quality', 'india', 'order', 'execution', 'time', 'risks', 'business', 'management', 'construction', 'project', 'risk', 'infrastructure', 'company']

THE TOP TWENTY WORDS FOR TOPIC #1
['technologies', 'execution', 'employees', 'processes', 'time', 'risk', 'team', 'process', 'various', 'quality', 'work', 'new', 'risks', 'safety', 'construction', 'digital', 'company', 'management', 'business', 'project']

**Fig. 3** Two topics with the top twenty keywords mapped under them (LDA)

THE TOP TWENTY WORDS FOR TOPIC #0
['iso', 'crores', 'growth', 'equipment', 'government', 'industry', 'internal', 'operations', 'quality', 'cost', 'order', 'india', 'time', 'sector', 'rs', 'risks', 'project', 'risk', 'infrastructure', 'company']

THE TOP TWENTY WORDS FOR TOPIC #1
['innovation', 'design', 'quality', 'employees', 'risks', 'learning', 'new', 'time', 'process', 'work', 'solutions', 'technology', 'processes', 'productivity', 'team', 'technologies', 'solution', 'safety', 'project', 'digital']

**Fig. 4** Two topics with the top twenty keywords mapped under them (NMF)

Considering the unsupervised nature of the modeling process, the results need validation. To validate the results, the authors consider the companies under the lean-focussed strategy (allotted to topic-1) and compare the results with those of the earlier keyword-based non-NLP work [25]. This study considers the 49 ‘lean signs’ identified by [25] (in which lean signs are based on the keywords from lean-related research articles) and performs a simple keyword search in the annual reports of the 18 companies considered in the analysis. It is found that the four out of the top five companies with the highest count of the ‘lean-signs’ are identified correctly by both LDA and the NMF-based models (see column 4 of Table 2 for lean signs frequency). However, the LDA-based model identified four additional companies under the ‘lean-focussed’ category while they are not (false positives) and one under the ‘revenue-focussed’ category, even when they have a high frequency of ‘lean-signs’ (false negative). In the NMF model, one case of false positive and false negative is observed. A confusion matrix and the associated validity parameters are shown in Tables 3 and 4. It is evident that in the context and the data considered for the study, NMF performs better than LDA.

### 7 Discussion

The results confirm the presence of two strategies to counter the emerging construction scenario. The four companies with a lean-focussed strategy seem to be in the minority compared with the revenue-oriented companies (opportunity-focussed). The results align with extant...
literature that finds a shallow extent of lean penetration in construction contracting organizations, especially in India [44, 45]. It is observed from the results that the few companies categorized as ‘lean-focussed’ are strategizing collaboration, increased visualization, and productivity improvement so that the work efficiency is improved, resulting in a more significant contribution to the profit margin. Extant studies elaborate on the low penetration of lean philosophy in construction, and the absence of a ‘lean mindset’ is considered a cultural barrier to implementing lean [46].

At the heart of all construction operations lies the concept of productivity. Construction productivity is the efficiency with which a firm converts inputs (resources like workforce, material, machinery, and money) to saleable outputs (work done or revenue earned). When productivity, which represents the execution efficiency of a construction contracting organization, is high, then the balance sheet tends to be healthy [47]. Three out of the four companies highlighted as lean-focussed are the top three companies in market capitalization. Therefore, the lean elements in the MDA section of the annual reports of large companies are a ray of hope that large companies are setting benchmarks that the companies in the middle and the lower rung can follow. As a testimony to this fact, in the October 2021 edition of the ‘Construction World’ magazine, construction industry leaders have reported the positive impact of implementing lean in their organizations [48]. Another interesting observation from the study is that words like ‘learning’ and ‘employees’ appear in the top twenty words mapped to topic-1 (or the lean-focussed organization). Employee-focussed and learning-oriented strategy is a hallmark of lean organizations [49].

| Company ID | Topic allocation as per LDA | Topic allocation as per NMF | Companies with a high frequency (> = 10) of ‘lean signs’ (keyword search) |
|------------|-----------------------------|-----------------------------|-------------------------------------------------------------------------|
| 0          | 0                           | 0                           | Low                                                                     |
| 1          | 1                           | 0                           | Low                                                                     |
| 2          | 0                           | 0                           | Low                                                                     |
| 3          | 0                           | 0                           | Low                                                                     |
| 4          | 0                           | 0                           | Low                                                                     |
| 5          | 1                           | 1                           | Low                                                                     |
| 6          | 1                           | 0                           | Low                                                                     |
| 7          | 0                           | 0                           | Low                                                                     |
| 8          | 0                           | 0                           | Low                                                                     |
| 9          | 0                           | 0                           | Low                                                                     |
| 10         | 0                           | 0                           | Low                                                                     |
| 11         | 1                           | 1                           | High                                                                    |
| 12         | 0                           | 0                           | Low                                                                     |
| 13         | 0                           | 0                           | Low                                                                     |
| 14         | 1                           | 1                           | High                                                                    |
| 15         | 1                           | 0                           | Low                                                                     |
| 16         | 1                           | 1                           | High                                                                    |
| 17         | 1                           | 1                           | High                                                                    |

Bold indicates the ‘True-Positive’ results in the case of companies with a lean-focussed strategy.

---

6784319545677477701 Team 1143 1144, L&T also achieved an exciting milestone by implementing Aerial LiDAR survey and sub-surface investigation for the first-ever bullet train project in India. The geospatial team won the FICCI award for Geospatial Excellence for Business Applications. Further work is happening on automated drone operations for remote surveillance and inspection purposes. The Path Forward Digital

10122934361582390631 Collaboration 1400 1401 workplace from anywhere and anytime. Adequate training to employees was provided to deal with this new situation. The result was a very high level of employee communication and collaboration through the digital mode. This culture shift has had the positive impact of improved productivity. Today, technologies are on the cusp of a Digital Perfect Storm™

6784319545677477701 Team 2751 2752 and claims and close follow-up with the client for pending payments. Significant Initiatives As the electricity system gets increasingly decentralised, decarbonised and digitalised, a specialised team of top-notch experts has been created to develop and deliver smart digital solutions worldwide, leveraging the engineering know-how and integration abilities. The first international battery energy

6784319545677477701 Team 2831 2832 grid stability in a region with a large, inherently intermittent renewable generation. A special thrust was given on Stakeholder Connect, a drive through which the supply chain team reached out to different levels of vendors to deliberate on challenges and mitigation measures. Digitalisation A digitally integrated project management eco system has been created to enable timely completion of

6784319545677477701 Team 2898 2899 and enables sophisticated analytics of all the processes in the project lifecycle, right from prospecting to post-completion obligations. The predictive abilities of the platform equip the team with the requisite tools and analytics to create optimised plans initially and during the course of the project. Significant process changes and digital initiatives were adopted in the supply

10122934361582390631 Collaboration 3006 3007 techniques, usage of drones, user-friendly incident-reporting systems and deployment of smart glasses, mobility devices, etc. have yielded tangible benefits. Virtual collaboration tools came in handy during remote design review, inspection, installation supervision and testing processes and have now become the norm. Risks and Concerns Mature stage gate and

6784319545677477701 Team 3527 3528 of the business are linear in nature and they involve extensive approval for ROW/ROU as well as land acquisition. The business engages a strong IR/Admin team to ensure all approvals are taken well in advance. Any delays on account of external factors are also properly documented and appropriate EOTs/cost claims are filed with

---

Fig. 5 Paragraph extraction results
Regarding the ‘revenue-focused’ organizations, the firms’ primary focus seems to expand their order books by tapping into the infrastructure push by the Government. While it is not argued that the two strategies are mutually exclusive, it is only highlighted that there is a visible difference in the top priorities of the top management when it comes to responding to the uncertain scenario that the world is currently facing. With the inclusion of words such as ‘time,’ ‘cost,’ ‘quality,’ and ‘operations,’ the absence of evident mutual exclusivity can be established by noting that some aspects of lean-focussed thought process in the top management can be seen in the topic-0 (or revenue-focussed) organizations as well. However, such words are few in the topic-0 firms compared with the lean-focussed topic-1 firms.

The results demonstrate that the NLP-based topic modeling algorithms can be helpful in clustering and understanding the themes in the large text corpus sourced from the MDA section of the annual reports of Indian construction contracting firms. While this study focuses on the contracting firms, a similar analysis of the MDA sections of the employer firms will indicate the extent to which employers find lean implementation necessary compared to the contracting firms’ interest in lean.

### 8 Research contributions and limitations

Even though the study is only a proof-of-concept, the contribution is the findings opening up many new research directions. The observations from the study pave the way for expanding the use of such algorithms also to analyse other sections of annual reports and vision and mission statements of construction organizations. Policymakers and researchers can understand the overall strategies adopted by the construction sector and juxtapose them with the topics extracted from Government’s industrial policy documents so that the extent of the alignment of industry and the Government can be assessed. In addition, a comparison of results with the keywords obtained from similar strategies adopted earlier in the timeline or across a cross-section of various countries or in both dimensions can reveal crucial information on the effectiveness of the policy implementation, and necessary course correction can be adopted well in advance. In terms of contribution to the body of practice, the study can help organizations assess their strategy compared to their competitors and embrace newer and better strategies to set their organizations on the growth trajectory.

The accuracy and the precision levels of the model need further improvement. Improvement in model validity can be achieved by incorporating the MDA section of more companies across various stakeholder types such as consultants, owners, and developer organizations. Further, data from international organizations can also be included as a part of input data to understand its effect on the topic modeling results and its validity. Additionally, it will be interesting to identify keywords representing strategies that combine both revenue and lean focus.

### 9 Conclusion

The study was initiated with a primary intention to develop a proof-of-concept to test the applicability of NLP-based unsupervised topic modeling algorithms in deciphering
latent topics prevalent in unstructured publicly available construction sector data. Considering the need to understand a construction contracting organization’s strategy to respond to both risks and opportunities evident in the evolving scenario in the Indian construction sector, the MDA section of the annual reports of construction contracting firms is used as a text data corpus to fulfill the primary objective. The results show that the top management of construction contracting firms laid down their strategy in two broad ways. A majority of the organizations are revenue-focused to propel growth by tapping into the massive infrastructure investment unveiled by the Government. On the other hand, a small number of organizations, typically having large market capitalization, focus on improving their work productivity by embracing lean construction concepts dominated by digitalization, team building, and process improvement. It is important to note that the two groups indicated above are not strictly mutually exclusive but only indicate an organization’s priority in trying times. While the finding above confirms the observation of lean experts and researchers, the uniqueness of the study is in the demonstration of the potential of NLP-based topic modeling algorithms to analyse and cluster keywords in a manner that can help researchers to visualise the strategies of construction contracting firms, which otherwise would require a significant amount of time if manually analysed. The demonstrated potential of the NMF algorithm paves the way to greater use of NLP-based topic modeling techniques for the systematic content analysis of text documents. Notwithstanding that the work is just a ‘proof-of-concept,’ it is an essential step in developing a robust text mining algorithm that can simplify the process of reading, understanding, and summarizing unstructured text data.

References

1. Jha KN (2013) Research method. in determinants of construction project success in India. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6256-5_2
2. Ram VG, Kishore KC, Kalidindi SN (2020) Environmental benefits of construction and demolition debris recycling: evidence from an Indian case study using life cycle assessment. J Clean Prod 255:120258. https://doi.org/10.1016/j.jclepro.2020.120258
3. Dixit S (2021) Impact of management practices on construction productivity in Indian building construction projects: an empirical study. Organ Technol Manag Constr 13(1):2383–2390. https://doi.org/10.2478/otmcj-2021-0007
4. Erumban AA, Das DK (2016) Information and communication technology and economic growth in India. Telecommun Policy 40(5):412–431. https://doi.org/10.1016/j.telpol.2015.08.006
5. Tabish SZS, Jha KN (2011) Identification and evaluation of success factors for public construction projects. Constr Manag Econ 29(8):809–823. https://doi.org/10.1080/01446193.2011.611152
6. Baviskar D, Ahirrao S, Potdar V, Kotecha K (2021) Efficient automated processing of the unstructured documents using artificial intelligence: a systematic literature review and future directions. IEEE Access 9:72894–72936. https://doi.org/10.1109/ACCESS.2021.3072900
7. Agarwal AK, Jagannathan M, Delhi VSK (2020) How control-focused are the standard forms? an assessment through text mining. J Leg Aff Disput Resolut Eng Constr 13(1):04520040. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000441
8. Marzouk M, Enaba M (2019) Text analytics to analyze and monitor construction project contract and correspondence. Autom Constr 96:265–274. https://doi.org/10.1016/j.autcon.2018.11.018
9. Department of economic affairs. (2022) National infrastructure pipeline. Government of India. Retrieved February 28, 2022, from https://indiainvestmentgrid.gov.in/national-infrastructure-pipeline
10. Sun M, Meng X (2009) Taxonomy for change causes and effects in construction projects. Int J Project Manage 27(6):560–572. https://doi.org/10.1016/j.ijproman.2008.10.005
11. Jagannathan M, Delhi VSK (2020) Litigation in construction contracts: literature review. J Leg Aff Disput Resolut Eng Constr 12(1):1–9. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000342
12. The World Bank. (2021) Ease of doing business in India. world bank group. Retrieved November 14, 2021, from https://www.doingbusiness.org/en/data/exploreeconomies/india#DB_ec
13. Baumol WJ (1959) Business behavior, value and growth. Princeton University, Princeton, NJ
14. Eisenberg T, Farber HS (1997) The litigious蚍ntroduction hypothesis: case selection and resolution. RAND J Econ 28:S92–S112. https://doi.org/10.2307/3087457
15. Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. Calif Law Rev 88(4):1051–1144. https://doi.org/10.2307/3481255
16. Camelo C, Fernández-Allés M, Hernández AB (2010) Strategic consensus, top management teams, and innovation performance. Int J Manpow 31(6):322–333
17. Singh B, Garg SK, Sharma SK (2010) Scope for lean implementation: a survey of 127 Indian industries. Intl J Rapid Manuf 1(3):322–333
18. Penrose JM (2008) Annual report graphic use: a review of the literature. J Bus Commun 45(2):158–180. https://doi.org/10.1177/021943607313390
19. Parker LD (1982) Corporate annual reporting: a mass communication perspective. Account Business Res 12(48):279–286
20. Bowman EH (1984) Content analysis of annual reports for corporate strategy and risk. Interfaces (Providence, Rhode Island) 14(1):61–71. https://doi.org/10.1287/inte.14.1.61
21. Santema S, van de Rijt J (2001) Strategy disclosure in Dutch annual reports. Eur Manag J 19(1):101–108. https://doi.org/10.1016/S0263-2373(00)000075-X
22. Azis Y, Osada H (2010) Innovation in management system by six sigma: an empirical study of world-class companies. Intl J Lean Six Sigma 1(3):172–190. https://doi.org/10.1108/2041461011074991
23. Tomidandel S Summerville KM Gentry WA and Young SF (2021) Using structural topic modeling to gain insight into challenges faced by leaders. Leadership Quarterly, In Press. https://doi.org/10.1016/j.leaqua.2021.101576
24. SEBI. (2021). Clause 49 - Corporate governance. securities and exchange board of India. Retrieved March 1, 2022, from https://www.sebi.gov.in/sebi_data/commanddocs/cir2803an1.pdf
25. Roy D & Jagannathan M (2021) Exploring the reach of lean philosophy in Indian construction industry. In Proceedings of the fourth biennial conference of the Indian Lean Community - Indian Lean Construction Conference (pp. 203–212). Ahmedabad: CEPT University Press.

26. Hassan FU, Le T, Lv X (2021) Addressing legal and contractual matters in construction using natural language processing: a critical review. J Constr Eng Manag 147(9):03121004. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002122

27. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211. https://doi.org/10.1007/s11042-018-8994-4

28. Lai Y, Kontokosta CE (2019) Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. Computers, Environ Urban Syst 78:101383. https://doi.org/10.1016/j.compenvurbsys.2019.101383

29. Lin JR, Hu ZZ, Li JL, Chen LM (2020) Understanding on-site inspection of construction projects based on keyword extraction and topic modeling. IEEE Access 8:198503–198517. https://doi.org/10.1109/ACCESS.2020.3035214

30. Jallan Y, Brogan E, Ashuri B, Clevenger CM (2019) Application of natural language processing and text mining to identify patterns in construction-defect litigation cases. J Leg Aff Disput Resolut Eng Constr 11(4):04519024. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000308

31. Jung N, Lee G (2019) Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. Adv Eng Inform 41:100917. https://doi.org/10.1016/j.aei.2019.04.007

32. Moon S, Chung S, Chi S (2018) Topic modeling of news article about international construction market using latent dirichlet allocation. J Korean Soc Civil Eng 38(4):595–599

33. Hong Y, Xie H, Bhumbrta G, Brilakis I (2021) Comparing natural language processing methods to cluster construction schedules. J Constr Eng Manag 147(10):1–11. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002165

34. Moon S, Lee G, Chi S, Oh H (2021) Automated construction specification review with named entity recognition using natural language processing. J Constr Eng Manag 147(1):1–12. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001953

35. Song J, Kim J & Lee JK (2018) NLP and deep learning-based analysis of building regulations to support automated rule checking system. In ISARC 2018 - 35th International Symposium on Automation and Robotics in Construction and International AEC/FM Hackathon: The Future of Building Things. https://doi.org/10.22260/isarc2018/0080

36. Zhang B (2020) Financial Risk Disclosure Return Premium: A Topic Modeling Approach. Stevens Institute of Technology. Retrieved from http://repository.unan.edu.ni/2986/1/5624.pdf

37. Padhy J, Jagannathan M, Delhi VSK (2021) Application of natural language processing to automatically identify exculpatory clauses in construction contracts. J Leg Aff Disput Resolut Eng Constr 13(4):1–9. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000505

38. Samanta PK, Singla HK (2019) Factors affecting the success of joint ventures in Indian construction firms. IUP J Manag Res 18(3):39–50

39. Ding Y, Jie M, Luo X (2022) Applications of natural language processing in construction. Autom Constr 136(2022):1–19. https://doi.org/10.1016/j.autcon.2010.09.005

40. George S, Vasudevan S (2020) Comparison of LDA and NMF topic modeling techniques for restaurant reviews. Indian J Nat Sci 10(62):28210–28216

41. Suri P & Roy NR (2017) Comparison between LDA & NMF for event-detection from large text stream data. In 3rd IEEE International Conference on “Computational Intelligence and Communication Technology” (IEEE-CICT 2017) (pp. 1–5). IEEE. https://doi.org/10.1109/CICT.2017.7977281

42. McKinney W (2008) Pandas. The pandas development team. Retrieved from https://pandas.pydata.org/docs/getting_started/overview.html

43. L Koskela T, Bølviken J, Rooke 2013 Which are the wastes of construction? 21st Annual Conference of the International Group for Lean Construction 2013 IGLC 2013 905 914

44. Malla V, Jagannathan M, Delhi VSK, Nair BS (2022) BIM-specific prequalification criteria in construction projects: exploring the nature and timeline of their inclusion. J Leg Aff Disput Resolut Eng Constr 14(2):1–12. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000540

45. Raghavanan N, Kalidindi S, Mahalingam A, Varghese K & Ayesh A (2014) Implementing lean concepts on Indian construction sites: organisational aspects and lessons learned. In 22nd Annual Conference of the International Group for Lean Construction: Understanding and Improving Project Based Production, IGLC 2014 (pp. 1181–1190). Oslo, Norway: International Group for Lean Construction.

46. Aslesen AR, Nordheim R, Varegg B & Lædre O (2018). IPD in Norway. In IGLC 2018 - Proceedings of the 26th Annual Conference of the International Group for Lean Construction: Evolving Lean Construction Towards Mature Production Management Across Cultures and Frontiers (pp. 326–336). Chennai, India: International Group for Lean Construction. https://doi.org/10.24928/2018/0284

47. Cyril EJ, Singla HK (2021) The mediating effect of productivity on profitability in Indian construction firms. J Adv Manag Res 18(1):152–169. https://doi.org/10.1108/JAMR-05-2020-0092

48. Narayanan SR (2021) Adopting lean, cycle time got reduced from 15 days to 7 days. Construction World, pp 48–48

49. Marin-Garcia JA, Bonavia T (2015) Relationship between employee involvement and lean manufacturing and its effect on performance in a rigid continuous process industry. Int J Prod Res 53(11):3260–3275. https://doi.org/10.1080/00207543.2014.975852