Artificial Intelligence in Knowledge Management: A Topic Modeling Approach for Construction Specific Documents

Ezra Kassa

Ethiopian Institute of Architecture Building Construction and City Development, Addis Ababa University, Addis Ababa, Ethiopia

Email address:

ezrakassa16@outlook.com

Abstract: To make sure that all significant contractual obligations are documented and managed, it is essential to have a clear understanding of construction contract agreements. However, the text data that makes up the content in these papers frequently necessitates the use of text mining. The topic modelling method of text mining, which is based on the document's topic, is one possible strategy to address these. The objective of this research is to demonstrate whether meaningful knowledge relationships can be extracted from a sample construction specific document using topic models (i.e. LDA). The research used a contract administration manual for topic modelling which is prepared by the Ethiopian Roads Authority (ERA) for use by the Regional Roads Authorities (RRAs). A total of 3217 unique tokens were available for text analysis. Between 5 and 25 topics were specified for LDA training and the one with 5 topics had concise result. To enhance the interpretability of the topics; topic visualization, relevance metric and filtered noun-types approaches were used. The tuning parameters in LDA Gensim with 5 topics gave the highest coherence score of 0.5163. Topic 1 made up the biggest portion of topics constituting 27% of the tokens. In addition, topics were made more interpretable by adjusting their setting. A total of 24300 bigrams and trigrams were also filtered with noun structures to form a unique concept. Construction companies benefit much from knowing what is under construction documents. An automated domain-specific model is required that can precisely extract all the explicit and implicit criteria from the construction contracts since construction-specific contracts differ from those used in other industries. In order to ensure that all pertinent project requirements are recorded, the research aims to demonstrate how knowledge linkages may be derived from construction-specific documents using topic models (i.e. LDA).

Keywords: Artificial Intelligence, Knowledge Management, Topic Modelling, Latent Dirichlet Allocation (LDA), Natural Language Processing (NLP), Construction Document

1. Introduction

In recent years, it has become clear that an organization's performance is determined by its intellectual assets rather than the value of its physical assets [1]. And technologies like document management systems, which use information technology (IT) tools to implement the KM process, help to support these types of knowledge [2]. As a result, combining IT with the discipline of knowledge management has lately become a particularly active subject of research.

To promote knowledge acquisition in the construction industry, tools are required to extract information from documents. Construction papers generally contain a great quantity of text, which includes both trivial (i.e., instructions and supporting statements) and vital (i.e., instructions and supporting statements) contents (i.e., contractual requirements). This necessitates specialists reading the entire material attentively and identifying texts indicating the needs [3]. Furthermore, many requirements, such as those referencing standards or regulatory codes, are addressed implicitly in written documents. During the project scope understanding stage, such implicit needs are likely to be overlooked [4].

When dealing with knowledge contained in unstructured contract papers, current knowledge management systems have flaws [5]. There is a need for approaches for describing knowledge in a way that both people and robots can grasp. Knowledge organization, sharing, and reuse would be aided by the availability of such approaches and instruments.

Advances in various computing disciplines such as information retrieval, machine learning, natural language processing, and ontology have been made in the last two decades, laying the technological foundations for the future development of knowledge management systems [6, 7]. These approaches are significantly distinct from other commonly used technologies, and their impact on knowledge management deserves specific attention. This is especially important because machine learning is based on learning [8]. The nature of knowledge is said to be intimately tied to the capabilities and limits of AI and KM [9]. This study utilizes the natural language processing (NLP) and machine learning approach to develop a meaningful topics from a standard construction specific document.

1.1. Problem Statement

Construction projects are prone to conflicts, which are typically triggered by parties failing to meet their contractual duties [10], and if not handled, can lead to complicated legal challenges that can take years to settle [11]. To guarantee that all relevant contractual requirements of the project scope are captured and maintained, it is vital to have a thorough understanding of the contract documentation. As a result, examining text-rich documents like construction contracts aids project participants in determining their responsibilities and reduces the time and effort required for contract analysis. However, the information in these documents is frequently in the form of text data, necessitating the use of text mining [12]. Language is just beyond the pale for a computer, which can only grasp the most basic structures [13]. As a result, text is unstructured in the perspective of the computer, and unstructured data (in the computer sense) encompasses much more than text [14].

Extraction of explicit knowledge from unstructured text sources may be accomplished using a variety of ways, especially when the amount of data is significant. The employment of a text mining technique based on the document's topic, known as topic modeling, is one option [15]. Topic modeling is regarded a good solution to this problem in this study since it may discover groupings or related subjects that would enable the utilization of such information [16]. It presents a method for extracting knowledge concepts from construction-related documents.

1.2. Objective

The objective of the research is to demonstrate whether meaningful knowledge relationships can be extracted from a sample construction specific document using topic models (i.e. LDA).

1.3. Significance of the Study

Knowing what is under those documents and understanding opinions is highly valuable to construction businesses. And it's practically hard to manually read through such large volumes and compile the topics. Thus, it is required to have an automated algorithm that can read through the text documents and automatically output the topics discussed.

2. Literature Review

2.1. IT based Knowledge Management

Individuals and businesses are starting to recognize the importance of knowledge in today's competitive and complicated economy [17]. Simultaneously, the impact of information and communication technology on Knowledge Management is growing at a breakneck pace [18]. As a result, knowledge and understanding are becoming increasingly necessary in order to acquire a competitive advantage. Knowledge, on the other hand, is a messy and cryptic idea [19]. As a result, capturing it is a challenging undertaking. Knowledge, on the other hand, may be a driving factor behind any type of invention if it is recorded and placed into clear form.

Because, most work activities are carried out by/with/through IT-based equipment, it's important to look into the interactions between people and AI-based technology when it comes to KM-related duties. Because of the rapid pace of technical and commercial development, knowledge management is becoming more appealing and vital [20]. Data, facts, and information can now be gathered, transferred, organized, and shared in ways that were unimaginable just a few years ago thanks to information technology [21]. Over the last decade, computational research has gotten a lot of attention, and it now fits under the umbrella of Artificial Intelligence. Knowledge may be modelled to obtain the specific domain from specialists using Artificial Intelligence (AI) [22].

2.2. Integrating Machine Learning with KM

IT is at the heart of learning because of its ability to process data [23]. More importantly, ML is fundamentally different from traditional IT in that learning occurs both around and within the system [8]. Domain knowledge in the form of equations, logic rules, and prior distribution may be included into machine learning algorithms [24]. These technical advancements highlight the importance of information technology as a tool for knowledge production, distribution, and application [25]. Thus, Integrating human knowledge into machine learning can improve machine learning's dependability and robustness while also allowing for the creation of explainable machine learning systems.

2.3. The Problem with Unstructured Text

Humans are finding it more difficult to digest large amounts of information contained in text documents that are presented in an unstructured textual style. This issue emphasizes the significance of automatic reading and natural language comprehension [26]. It may be necessary to prevent manual reading of texts inside the initial corpus, depending on the domain. Hence, knowledge extraction must be done on a global scale, as a synthesis of the entire corpus [27].

Natural language's importance and promise in human-computer interaction should not be neglected, as it normally delivers effortless and successful communication in human-human interactions [28]. Because the data is generally unlabeled, and the meaning of words and phrases might offer vital clues, unsupervised and semantic approaches are used [29]. In light of these considerations, NLP's capacity to describe data as a collection of interconnected entities based on semantics and meaning makes it a valuable tool for knowledge representation, integration, and autonomous reasoning inference.

2.4. Knowledge Extraction in Topic Modelling

Unsupervised topic models are commonly employed for domain exploration in a variety of study domains where training data is either unavailable or prohibitively expensive to obtain [30-33]. Machine learning (ML) topic models are widely utilized for a variety of Natural Language Processing (NLP) activities.

Within the broader realm of artificial intelligence, topic modeling is one of numerous Natural Language Processing approaches. Topic models are statistics-based strategies for analyzing word cooccurrence probability, with better results as the data set grows larger [34]. The main goal of topic modeling is to develop a probabilistic model for a group of text texts. Documents are probability distributions over subjects in topic models, with each topic represented as a multinomial distribution over words [35]. As a result, topic models may be used as a powerful tool for uncovering hidden meanings in unstructured text data.

Probabilistic topic models, such as Latent Dirichlet Allocation (LDA), have recently received a lot of attention [36-38]. This approach has been successfully used to identify latent themes from text documents for a variety of text mining applications, including machine translation [37], word embedding [39], automated topic labeling [40] and many more.

2.5. Construction Specific Documents

Contract agreements are a vital legal component of every construction project since they detail the owner's intentions and expectations for the project's design, development, and handover [41]. Because these construction papers are complicated [42], it is critical to understand the exact client needs prior to participating. As a result, in order to effectively administer contracts, there must be an appropriate level of commonality in the interpretation of contract requirements [43].

Extraction of reporting requirements from long construction contracts is frequently conducted manually due to a lack of acceptable technologies [44]. As a result, the time and expenses involved with meeting reporting obligations are frequently estimated informally, leading to underestimate. When depending on a person to extract requirements, the procedure is time-consuming and error-prone [45], as project documentation are frequently voluminous and confusing, resulting in diverse interpretations.

Furthermore, due to numerous imbalanced contracting techniques such as lump-sum turn-key and low-bid selection, contractors responsible for the whole execution of engineering, procurement, and construction (EPC) projects are vulnerable to many hazards [3]. Furthermore, it requires a domain specialist who is familiar with the implicit criteria for translating regulatory codes into contractual obligations [46]. As a result, text mining aids in the extraction of keywords that should be regularly watched in correspondence.

3. Method

This study focuses on implementing Latent Dirichlet Allocation (LDA) method. The method is chosen to derive a set of classes based on underlying semantic data: a collection of automatically extracted propositions conveying general world knowledge. The eventual goal is the construction of corpus-specific topic that allow for multiple word senses, and whose structure reflects the distributional semantics found in text. Such outcome will allow for robust generalization of extracted knowledge. Implementation of topic modelling with the LDA method will use the Genism package. The method is implemented in Python programming language.

3.1. Data

This research used secondary data for text analysis. A contract administration manual which is prepared by the Ethiopian Roads Authority (ERA) for use by the Regional Roads Authorities (RRAs) in Ethiopia has been used for topic modelling. The primary reason to utilize the document was because of its knowledge content on proper planning, procurement, administering contracts, managing claims and resolving disputes as well as environmental and social safeguards in executing road works. They are primarily aimed at Road Engineers, Planners, Procurement Officers, and Managers of the Regional Road Authorities in Ethiopia. They provide users with standard reference material and a ready source of good practice in the efficient and timely delivery of road projects. It is also intended for the supervising consultants to make good use of the standardized formats to ensure uniformity of approach and compliance with internal control systems.

The manual and its guidelines focus on the physical implementation stage of the RRA projects which is the period between the acceptance of a contractor's tender and the end of the Defects Liability Period. The manual is presented in seven sections as follows:

1. Introduction;
2. Parties to a Contract;
3. Contract Documentation;
4. Contract Commencement;
5. Financial Control;
6. Project Management;
7. Project Conclusion.

As the intention of the manual is for daily use by the staff in the Construction Project Management Directorate, proper implementation of topic modelling can easily refer, various aspect of contract administration and the numerous concepts, principles, standards, rules, theories and practices commonly encountered in the implementation of civil works projects.

3.2. Data Processing

Data preprocessing begins with data preparation, which tries to transform text from human language into a machine-managed format, synthesize unstructured text, and keep keywords relevant for expressing subjects [47].

3.2.1. Text Normalization

NLP systems include common steps of text processing. Text normalization, which involves converting text to non-capital letters/case folding, eliminating special characters and punctuation, and removing white space, is the first stage of preprocessing. After that, the text will be tokenized, which involves deleting words that appear often (stop words), transforming words into important words (stemming), and removing words that appear frequently (stop words).

3.2.2. Bigrams and Trigrams

Through the Bag of Terms (BOW) technique, the text transformation step translates text data to the proper format by constructing bigram models and trigrams to group often simultaneous words [48], including the corpus and data dictionary. Until date, the most popular technique has been Bag of Words, which involves converting the corpus into a word-document matrix and disregarding the order of words [35, 49].

3.2.3. Term Frequency–Inverse Document Frequency (TF-IDF) Vectorizer

The TF-IDF score represent the titles as vectors, which means that the data must be represented numerically so that the model can handle it [50]. This method transforms a string of text into a word count matrix. Once the criteria are defined, the relevance score generated for each subject is assigned to each title as a column, and the dominant topic is determined by picking the topic with the greatest relevance score. This will eventually eliminate the terms with the lowest scores, making subsequent training more efficient.

3.3. Latent Dirichlet Allocation (LDA)

The word-topic and topic-document distributions in the LDA model are trained fully unsupervised [15]. After randomly assigning words to subjects, the algorithm iterates over all of the words in the training texts for a number of iterations [38]. The method achieves a stable state when the iterations are completed, and the word topic probability distributions may be approximated using word topic assignments.

Training parameters can be modified as per the result of the training [51]. Increasing the number of passes, iterations, and chunksize might help make the subjects more understandable.

The chunk size refers to the amount of documents that must be put into memory for each training session. The number of training iterations throughout the full corpus is called passes. The maximum number of iterations required to obtain convergence for each document is iterations. The Algorithm iterates each document's probability distribution assignments in these rounds, moving on to the next document if it has already attained convergence. This is essentially the major steps of the method, repeated for the amount of passes. When the themes still don't make sense, these criteria can be raised. Because topic distribution update is costly, increasing chunksize to the maximum memory capacity will improve performance. Iterations must also be long enough to guarantee that a sufficient number of documents have reached convergence before proceeding.

3.4. Model Performance

Following a succession of topic modeling procedures, the final topic models must be assessed to see how effective they are at grouping subjects by computing the coherence score. By differentiating semantically interpretable subjects from statistical inference artifacts, the coherence score may assess the degree of se-mantic similarity between high-scoring terms in the topic [52]. The CV algorithm [53] may be used to calculate the coherence score. In addition, the pyLDAvis module will be used to visualize the distribution of topics created by topic modeling using the LDA approach [54].

3.5. Interpretability

Making the output of topic models interpretable is an important aspect. A range of approaches are available to identify the importance of words both within topics and their relative frequency in the entire corpus. Three of the alternatives are discussed here.

3.5.1. n-Grams and Filtered Noun-Type

Its good practice to identify phrases so the topic model can recognize them. Bigrams are phrases containing 2 words e.g. "final_acceptance". Likewise, trigrams are phrases containing 3 words e.g. "road_sector_development". Filtering bigrams or trigrams with noun structures aids the LDA model in better clustering topics, since nouns are stronger indications of the topic being discussed.

3.5.2. Topic Visualization

pyLDAvis [54] is a general-purpose topic model visualization interface that may be used to gain a high-level overview of a model, examine topics in further detail, and examine terms linked with themes. The relevance meteric function, which allows the user to change the view of words in a topic for better understanding, stands out among the general-purpose interfaces. Each topic is represented by a circle. Topic relatedness is shown by the distance between the circles. These are translated into 2D space using dimensionality reduction on the distances between each topic's probability distributions. This indicates if the model produced separate themes.

3.5.3. Prioritize Terms More Exclusive to a Topic

The words for a topic are arranged using a relevance metric slider scale at the top of the right panel. Relevance combines two alternative methods of thinking about the degree to which a word is related with a subject. On the one hand, if a term appears frequently in a topic, it is said to be significantly linked with that issue. The lambda value (λ) on the slider is set to 1 by default, which arranges words according to their frequency in the topic. On the other hand, if a the ratio of a term's probability within a topic to its marginal probability across the corpus is high, it is as closely related with a topic as possible.

4. Method

The primary application of natural language processing was to automatically extract what topics are discussed from large volumes of a contract administration manual. In this study, a contract administration manual which supports road sector development program is used as a dataset and LDA method is applied to extract the naturally discussed topics. The core packages used in this research are re, gensim, spacy and pyLDAvis. Besides this, matplotlib, NumPy and pandas are applied for data handling and visualization. The Latent Dirichlet Allocation (LDA) from Gensim package is implemented for better topics segregation. The volume and percentage contribution of each topic is also extracted to get an idea of how important a topic is.

4.1. Preprocessing Results

The first prerequisite in any processing is to prepare the data for efficient training. The data (the corpus) was originally in a PDF format which needs to be converted in to a TXT for ease of processing. Each line in the file is considered a document. Before preprocessing though, texts are extracted from PDF file and are written into TXT file. TXT files are plain text documents with very little formatting. Text-based information is stored in them. Python has built-in file creation, writing, and reading capabilities.

The next step was to sanitize the data in order to place as much attention on the theme substance as possible. The use of python programming skills aided in this procedure. This is done by deleting stop words (words that are filtered out before or during natural language data processing) and lemmatizing (reducing a word to its root word) from the text. Python tools like NLTK and spacy model make this procedure easier. Following the removal of superfluous spaces and conditions, tokenization is used to break down the sentences into a list of terms, deleting all of the jumbled text in the process. Each phrase is tokenized into a list of words, obviating the need for punctuation and other characters.

4.2. Dictionary and Corpus

The dictionary and the corpus are the two basic inputs to the LDA topic model. Each word in the manuscript is assigned a unique id by Gensim. The resulting corpus is a

(word id, word frequency) mapping. For instance, (0, 1) indicates that word id 0 appears just once in the first document. If it is necessary to determine what word a particular id belongs to, the id can be supplied to the dictionary as a key, converting it to a human-readable format. The LDA model uses this as its input. A total of 3217 distinct tokens were discovered.

4.3. Optimize Interpretability Using TF-IDF

The dictionary and the corpus are the two basic inputs to the LDA topic model. Each word in the manuscript is assigned a unique id by Gensim. The resulting corpus is a (word id, word frequency) mapping. For instance, (0, 1) indicates that word id 0 appears just once in the first document. If it is necessary to determine what word a particular id belongs to, the id can be supplied to the dictionary as a key, converting it to a human-readable format. The LDA model uses this as its input. A total of 3217 distinct tokens were discovered.

To infer themes, LDA does not require TF-IDF [51]. The TF-IDF score, on the other hand, is particularly beneficial for LDA since it filters out low-value words and terms that aren't in the TF-IDF. Using the complete language is typically computationally costly. TFIDF picked the top 5 terms as an efficient strategy to reduce the lexicon. With the aid of TF-IDF, it is possible to get a preliminary look at word clusters (Figure 1).
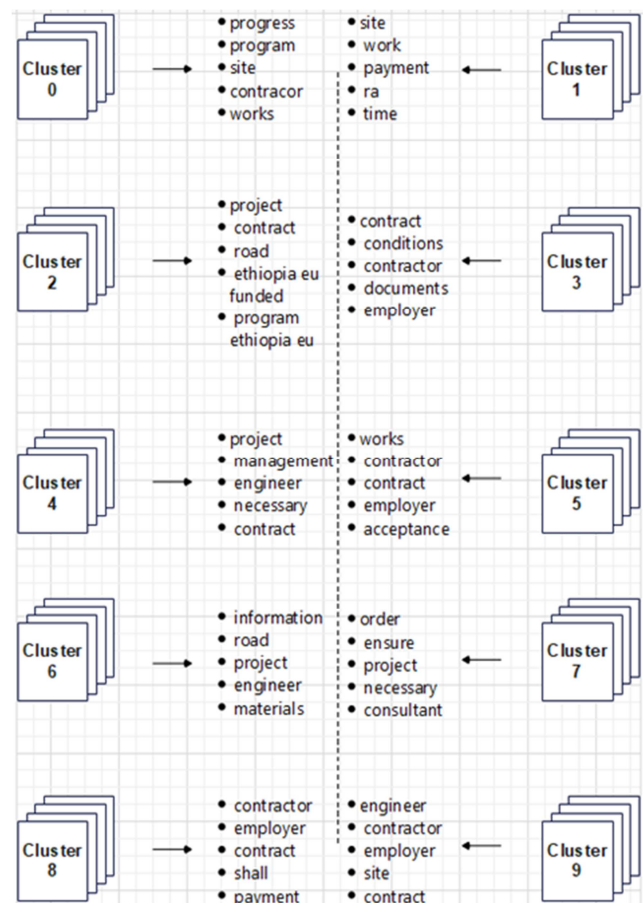

Figure 1. *Top 5 TF-IDF words in 10 clusters.*

Each cluster is represented by a different color (Figure 2). For example, the first cluster (i.e. cluster 0) is designated by red color and the second color (i.e. cluster 1) is designated by blue and so on and so forth. It can be used to visualize topics or to chose the vocabulary.
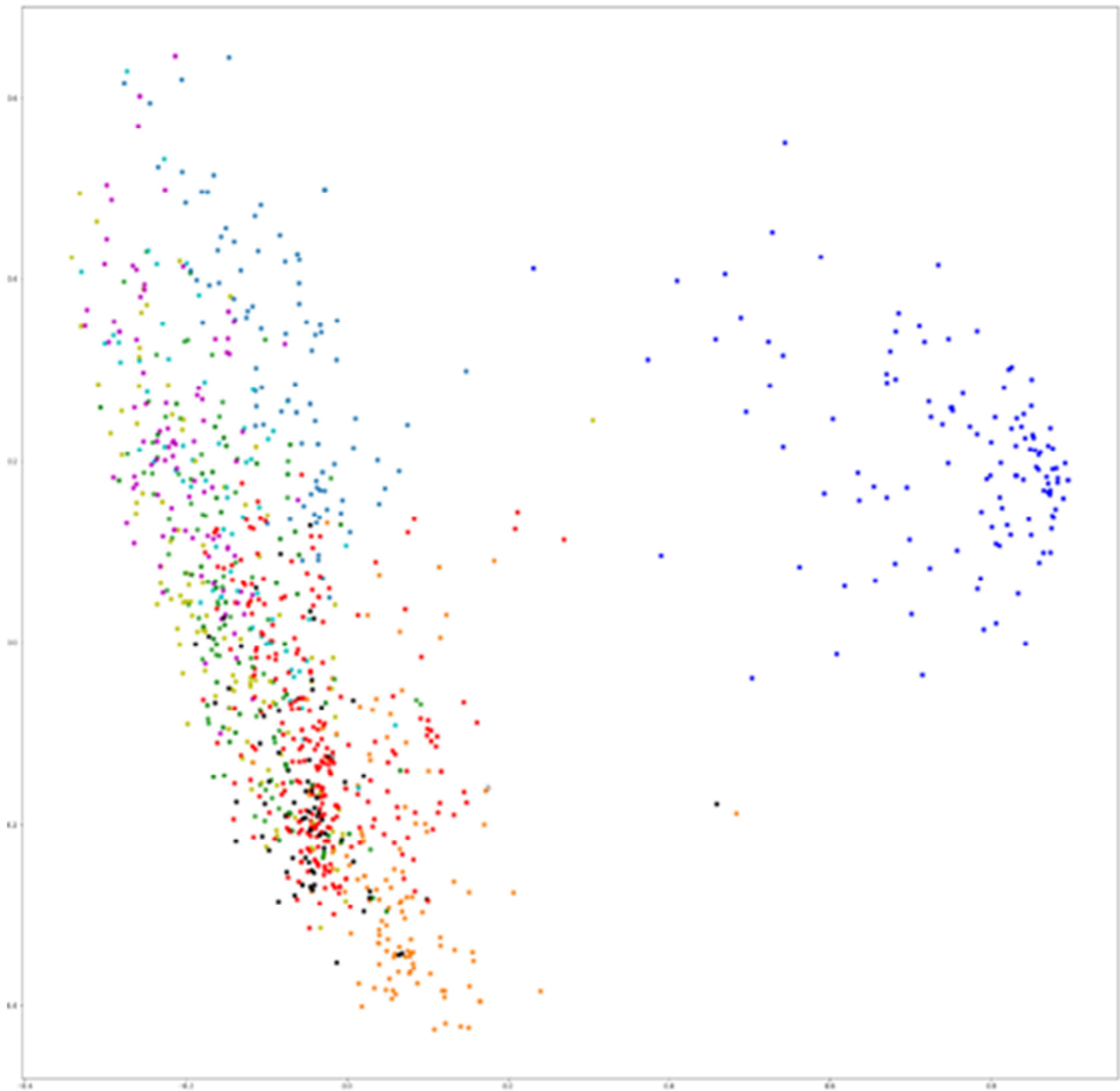

Figure 2. *TF-IDF cluster color designation.*

4.4. Coherence Measure

The number of subjects must be specified in LDA. This may be tweaked by adjusting factors like prediction probability, confusion, and coherence. The Cv measure was used to examine a variety of issues. Adding subjects can aid in the discovery of new subtopics. However, if the same terms appear across several subjects, the number of topics is too large. 5 to 25 subjects were examined, with 5 yielding clearer findings (Figure 3). After five subjects, the improvement ceases.
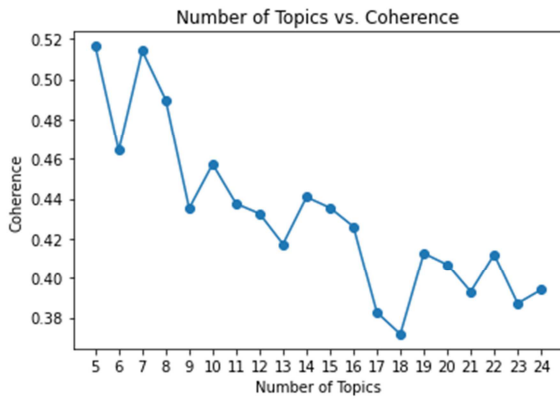
Figure 3. *Coherence measure.*

Apart from that, hyperparameters that affect sparsity of the topics were set according to Gensim documentation. The implementation of the LDA method in this study will be carried out using the Gensim pack-age. The topic distributions for entire corpus is updated after each chunksize, and after each passes. To ensure sufficient corpus topic distribution updates especially for small corpuses as this one, chunk size as well as passes are increased.

4.5. LDA Model

The LDA model is built with 5 different topics where each

topic is a combination of keywords, and each keyword contributes a certain weightage to the topic. The keywords for each topic and the weightage (importance) of each keyword can be shown using lda_model.print_topics() command. This means the top 10 keywords that contribute to each topic are given. The weights reflect how important a keyword is to that topic.

4.6. Interpreting Topics

Topic modelling outputs are often difficult to interpret for useful insights. However there are some techniques to enhance interpretability.

4.6.1. Visualize Topics-Keywords

After the LDA model has been developed, the following step is to review the generated themes and keywords. The pyLDAvis package is an interactive charting tool that is meant to operate with jupyter notebooks. Instead of being grouped in one quadrant, a strong topic model will have reasonably large, non-overlapping bubbles dispersed around the chart. The greater the bubble, the more popular the subject is. With 5 topics, the tuning settings in LDA Gensim produced the greatest coherence score of 0. 5163. Although some subjects cross, as seen in Figure 4, the distribution and separation of topics created is considerably better with a more diversified distribution.
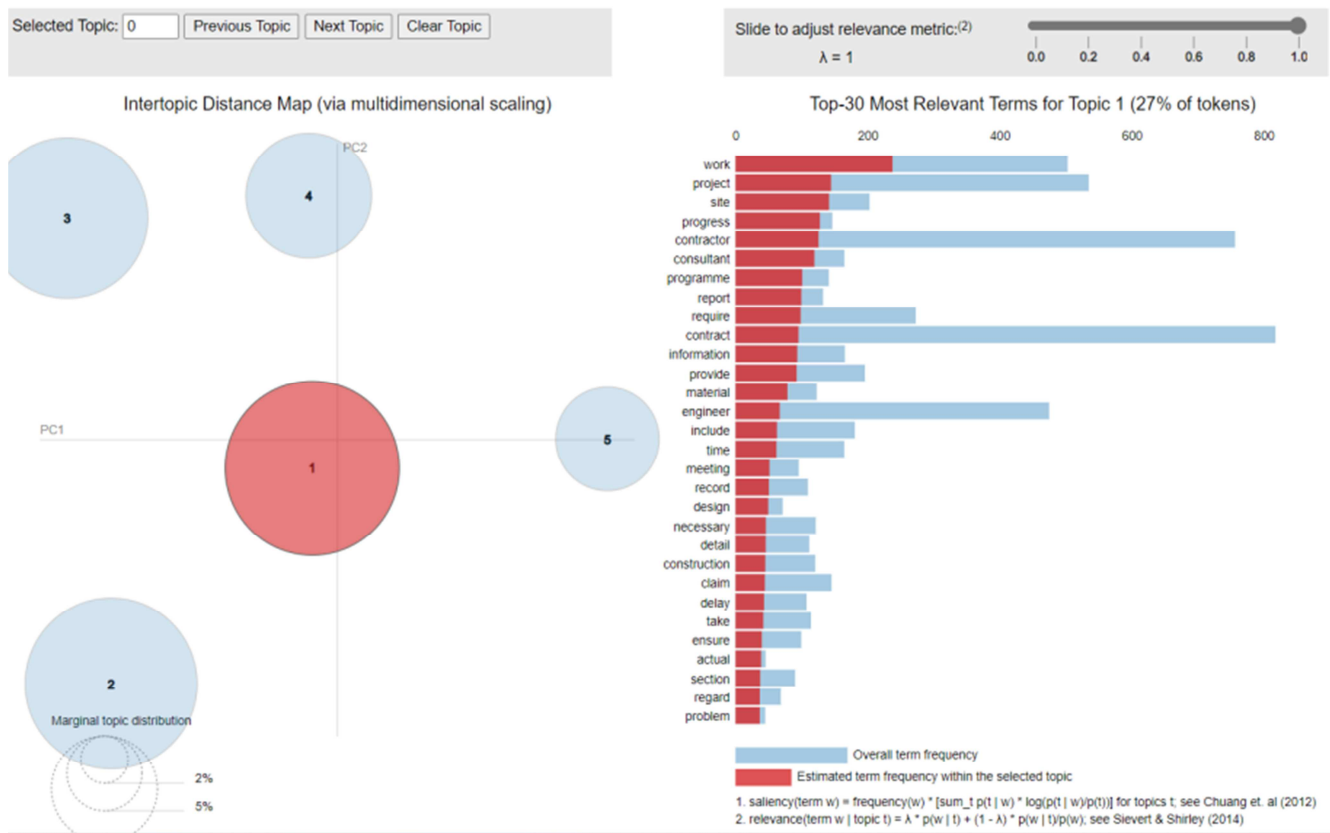

Figure 4. *Distribution of topic.*

The topics in a model are shown as circles in the

pyLDAvis interface's left panel, with circle size indicating

the relative statistical weight of topics. A certain topic is selected for inspection by clicking on a circle or entering a topic number in the search area at the top. The left panel is also a "intertopic distance map (multidimensional scaling)," as defined by pyLDAvis. This graphic depicts the statistical closeness or distance of subjects to one another. Topic distance also indicates how closely connected two topics are. The size of the circles denotes the frequency of a given topic. For example, as seen in Figure 4, topic 1 makes up the biggest portion of topics being talked about amongst documents, constituting 27% of the tokens.

The top words related with the specific subject selected in the left panel are displayed in the right panel of the pyLDAvis interface, along with bar graphs indicating their weight. The frequency of any word is represented by the blue bar in the overall topic model. The frequency of the term inside the specified topic is represented by the red bar. Table 5 shows the distribution of the top ten most often used terms in organizing the five subjects and their interpretations.

Extracted topics

Topic 1 cites most of the key concepts covered during a project management (Table 1). It states that monitoring progress and resources is a vital reference during construction and after. Along with that, it raises the issue of work progress measurement and representatives leading the contract administration.

Table 1. Extracted terms for topic 1.

Top 10 Key words	Weight
"work"	0.028
project"	0.017
"site"	0.016
"progress"	0.015
"contractor"	0.015
"consultant"	0.014
"programme"	0.012
"report"	0.012
"require"	0.011
"contract"	0.011

The second topic talks about the type of the document and its representatives (Table 2). It shows a clue as to how the document is a prepared through a technical Cooperation Program with funding from the European Union aimed for use by the Regional Roads in Ethiopia.

Table 2. Extracted terms for topic 2.

Top 10 Key words	Weight
"contract"	0.051
"project"	0.032
"volume"	0.018
"fund"	0.016
"support"	0.014
"road_sector"	0.014
"technical_cooperation"	0.014
"ethiopia_eu"	0.014
"europeaid_ih"	0.014
"contractor"	0.013

The third topic is about form of security and issues to be

addressed when negotiating contracts (Table 3). The topic deals with financial assurance given by one party to another to ensure the due and proper performance of its obligations under a contract issue.

Table 3. Extracted terms for topic 3.

Top 10 Key words	Weight
"contractor"	0.054
"contract"	0.028
"employer"	0.026
"work"	0.022
"payment"	0.019
"engineer"	0.018
"amount"	0.012
"require"	0.012
"security"	0.011
"issue"	0.01

Topic 4 shows the administrative inter-relationships between the three parties involved in the contract (Table 4). The construction team for a major project most commonly consists of three primary parties, the Employer (Public Body or Client) who initiates, pays for and is the ultimate owner of the project, the Contractor who carries out the actual construction and a Consultant, who may have designed the works and who supervises the works in the role of the Engineer.

Table 4. Extracted terms for topic 4.

Top 10 Key words	Weight
"engineer"	0.017
"contractor"	0.017
"contract"	0.015
"employer"	0.014
"task"	0.013
"require"	0.01
"delay"	0.009
"representative"	0.009
"action"	0.009
"engineer"	0.017

The last topic describes about all parties involved in a road project, whether they are the owner or contractor, protecting their interest via insurance and its arrangement framework (Table 5).

Table 5. Extracted terms for topic 5.

Top 10 Key words	Weight
"contractor"	0.016
"insurance"	0.012
"material"	0.011
"project"	0.01
road"	0.009
"value"	0.009
"information"	0.008
"pay"	0.007
"employer"	0.007
"type"	0.007

The left and right panels in pyLDAvis also interact in another way. Any terms that create only a few subject circles in the left panel (indicating that the word is significant in only a few of the model's themes) deserve

extra attention. Among the subset of words that produce only a few topic circles in the left panel, those that produce either a very large or very small circle for the topic whose word list you are looking at relative to the circles of other topics is particularly noteworthy (meaning that while the word is prominent in your currently selected topic, it is much more so or less so than in other related topics). As a demonstration (Figure 4), when topic 1 is selected, hovering

a mouse pointer over the word "contract" in the right panel changes the left panel to show only the other topics in which that word is prominent.

For example, the word "engineer" is prominent in topic 4, topic 3 and topic 2 respectively (Figure 5). And the word "volume" is solely distributed under topic 2 only. In another case, the word "contractor" is dominantly showed in all the topics (Table 6).
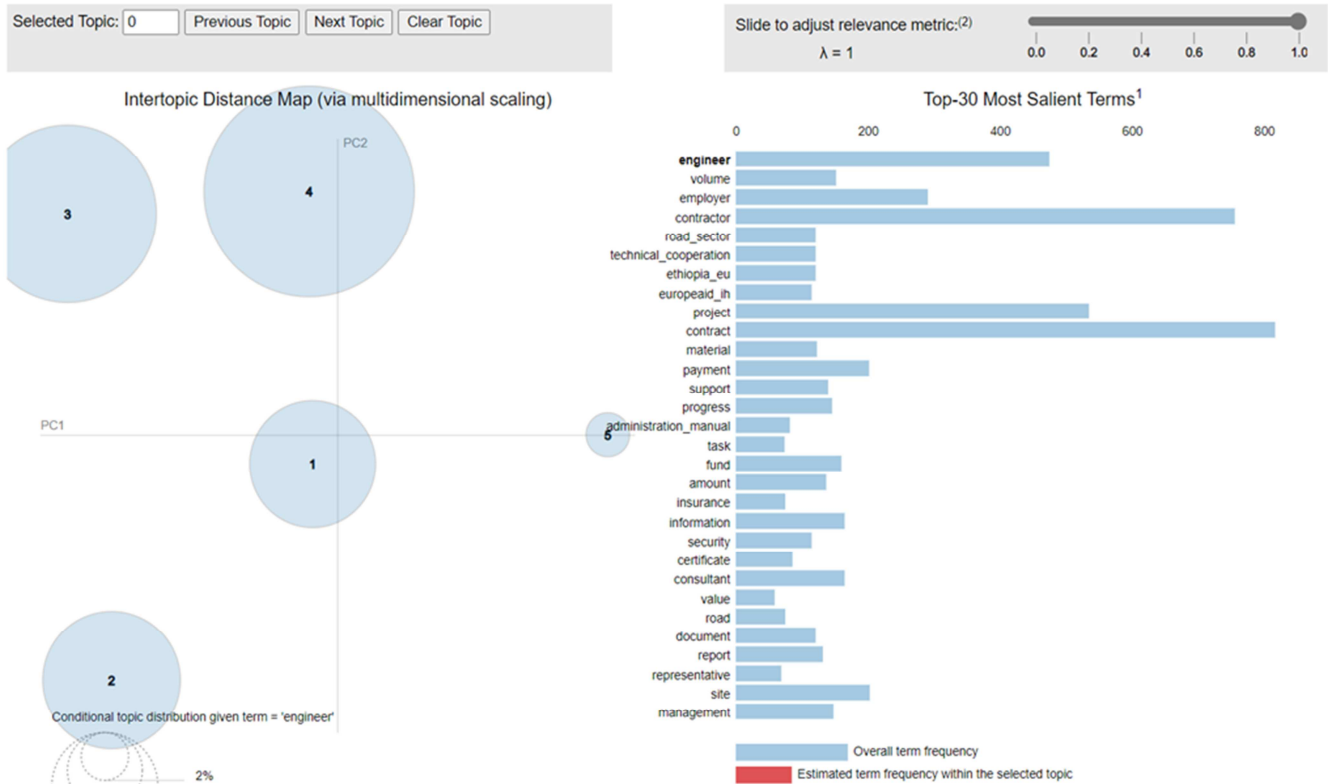

Figure 5. *Word distribution in each topic.*

Table 6. *Top 10 most salient terms.*

Top 10 Salient Terms	Conditional topic distribution
engineer	Topic 4
volume	Topic 2
employer	Topic 3
contractor	Topic 4
road_sector	Topic 2
technical_cooperation	Topic 2
ethiopia_eu	Topic 2
europeaid_ih	Topic 2
project	Topic 2
contract	Topic 2

4.6.2. Tune Relevancy Score

Words that indicate a topic are given a high ranking since they are used often across a corpus. The relevance score aids in the prioritization of phrases that are more specifically related to a certain topic, making the topic more evident. pyLDAvis is set to = 1 by default, which ranks words solely by their frequency inside a certain subject (by their red bars). Setting = 0 words, on the other hand, arranges words according to their "lift." This means that words with almost

as lengthy a red bar as their blue bar will be sorted first. It was attempted to provide a more consistent emphasis on what the issues are about using three lambda () relevance metric settings: 1, 0, and 0.6.

Table 7. *Word relevance metrics under different settings for topic 3.*

$\lambda = 0$	$\lambda = 0.6$	$\lambda = 1$
"backfill_mm"	"engineer"	"engineer"
"labourer"	"task"	"project"
"calendar"	"representative"	"contractor"
"week"	"delay"	"contract"
"lie"	"project"	"employer"
"lab"	"resource"	"task"
"lay_bedde"	"action"	"require"
"kiss"	"contractor"	"delay"
"forward"	"define"	"representative"
"initiate"	"duty"	"action"

In the example shown in table 4, where topic 3 is about administrative inter-relationships between the three parties involved in the contract when $\lambda = 1$. However, most of the words do not appear in the top words when $\lambda = 0$, where the top words describe the amount of work involved and

identification of the individual tasks to be undertaken. Dialing the lambda around 0 reveals that the availability of programme consultants and the contractor's effort into the preparation of an effective works programme to specify exactly what is actually necessary keeping in mind the "KISS" principle i.e. Keep it Short and Simple. The words ("lay_bedde", "lie","kiss", "backfill_mm") infer one should not include or insist on unnecessary or overly complicated details specifically for a pipe laying exercise.

4.6.3. Identify Phrases Through n-Grams and Filter Noun-Type Structures

Detecting n-grams measures how much more likely the words co-occur than if they were independent. Because the metric is sensitive to uncommon word combinations, it is used with an occurrence frequency filter to verify that phrases are relevant. A total of 24300 bigrams and trigrams are filtered with noun structures.

Table 8. n-grams with noun structures.

Words	Bigrams	Trigrams
1	"Defect liability"	"Defects liability period"
2	"Retention money"	"Interim payment certificate"
3	"Liability period"	"Final acceptance certificate"
4	"Cash flow"	"Sector development program"
5	"Payment certficate"	"Volume administration manual "

5. Conclusion

The impact of information and communication technology on Knowledge Management is growing at a breakneck pace. Because the majority of work activities are carried out by/with/through IT-based equipment, it's important to look into the interactions between people and AI-based technology when it comes to KM-related duties. More importantly, ML is fundamentally different from traditional IT in that learning occurs both around and within the system.

Within the subject of artificial intelligence, topic modeling is one of numerous Natural Language Processing approaches. It may be used to find underlying, hidden, or latent themes, patterns, or groups in enormous amounts of text, referred to as the "corpus." In light of these considerations, text mining in construction-related documents aids in the extraction of keywords that should be actively watched in correspondence.

At the early stage of data processing, a raw text was normalized by /case folding, removing special characters and punctuation, and removing white space. After that, the text went through a process of removing words that often appear (stop words), converting words into essential words (stemming), and tokenization (tokenization). The text was also transformed to the appropriate format by forming bigram models and trigrams to group frequently simultaneous words. Finally, lowest scoring words were removed using the TF-IDF vectorizer for more efficient training. Top 5 words with the highest TF-IDF for every document was calculated.

To make the most sense out of the topics, hyperparameters were modified as per the result of the LDA training. Between 5 and 25 topics were experimented and the improvement stopped after 5 topics using the Cv measure. The tuning parameters in LDA Gensim with 5 topics gave the highest coherence score of 0. 5163. In a respective manner, project management, document content, security issues, stakeholders of a project and project insurance were labeled as extracted topic from the corpus. Once irrelevant keywords were filtered out, the top-scoring keywords for each topic were visible. The top-10 most salient terms "engineer", "volume", "employer", "contractor", "road_sector", "technical_cooperation", "ethiopia_eu", "europeaid_ih", "project", "contract" were also listed along with their topic distribution.

Using three lambda (λ) relevance metric settings: 1, 0, and 0.6, it was experimented to bring a more coherent focus on what the topics are about. Furthermore, multi-word phrases that belong together as a phrase were identified as a sequences of tokens whose joint probability are high in comparison to the individual probabilities of their constituent tokens.

References

[1] E. Arafa, "THE IMPACT OF KNOWLEDGE MANAGEMENT ON PROJECT SUCCESS," UNIVERSITY OF PORTSMOUTH, 2015.

[2] M. Lindvall, I. Rus, and S. S. Sinha, "Technology Support for Knowledge Management," 2002. https://www.researchgate.net/publication/2535021_Technology_Support_for_Knowledge_Management (accessed May 30, 2022).

[3] S. J. Choi, S. W. Choi, J. H. Kim, and E. B. Lee, "AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects," *Energies,* vol. 14, no. 15, p. 4632, Jul. 2021, doi: 10.3390/EN14154632.

[4] A. K. Jallow, P. Demian, A. N. Baldwin, and C. Anumba, "An empirical study of the complexity of requirements management in construction projects," *Eng. Constr. Archit. Manag.*, vol. 21, no. 5, pp. 505–531, Sep. 2014, doi: 10.1108/ECAM-09-2013-0084.

[5] N. Bing Chong, L. Uden, and M. Naaranoja, "Knowledge management system for construction projects in Finland," *Int. J. Knowl. Manag. Stud.*, vol. 1, no. 3–4, pp. 240–260, 2007, doi: 10.1504/IJKMS.2007.012524.

[6] I. Ö. Arnarsson, O. Frost, E. Gustavsson, M. Jirstrand, and J. Malmqvist, "Natural language processing methods for knowledge management—Applying document clustering for fast search and grouping of engineering documents:," *SAGE Journals*, vol. 29, no. 2, pp. 142–152, Mar. 2021, doi: 10.1177/1063293X20982973.

[7] Q. Tang, "Knowledge management using machine learning, natural language processing and ontology," Cardiff University, 2006.

[8] Z. Zhou, Y. Liu, H. Yu, and L. Ren, "The influence of machine learning-based knowledge management model on enterprise organizational capability innovation and industrial development," *PLoS One*, vol. 15, no. 12 December, Dec. 2020, doi: 10.1371/JOURNAL.PONE.0242253.

[9]   M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith, "Artificial intelligence and knowledge management: A partnership between human and AI," *Bus. Horiz.*, Mar. 2022, doi: 10.1016/J.BUSHOR.2022.03.002.

[10]   A. K. Agrawal, M. Jagannathan, and V. S. K. Delhi, "Control Focus in Standard Forms: An Assessment through Text Mining and NLP," *J. Leg. Aff. Disput. Resolut. Eng. Constr.*, vol. 13, no. 1, p. 04520040, Oct. 2020, doi: 10.1061/(ASCE)LA.1943-4170.0000441.

[11]   K. M. J Harmon, "Resolution Of Construction Disputes: A Review of Current Methodologies," *Leadersh. Manag. Eng.*, vol. 3, no. 4, pp. 187–201, 2003.

[12]   N. Sajadfar, S. Abdollahnejad, U. Hermann, and Y. Mohamed, "Text detection and classification of construction documents," in *36th International Symposium on Automation and Robotics in Construction*, 2019, pp. 446–452.

[13]   Ray Jackendoff, *Why can't computers use English?* Linguistic Society of America, 2022.

[14]   B. Inmon, "Why Do We Call Text 'Unstructured'?," May 28, 2016. https://tdwi.org/articles/2016/06/28/text-unstructured.aspx (accessed May 30, 2022).

[15]   R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 1, 2015, doi: 10.14569/IJACSA.2015.060121.

[16]   P. Xie and E. P. Xing, "Integrating Document Clustering and Topic Modeling," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013, pp. 694–703.

[17]   H. K. Mohajan, "(2) (PDF) The Roles of Knowledge Management for the Development of Organizations," *J. Sci. Achiev.*, vol. 2, no. 2, pp. 1–27, 2017, Accessed: May 31, 2022. [Online]. Available: https://www.researchgate.net/publication/314063315_The_Roles_of_Knowledge_Management_for_the_Development_of_Organizations.

[18]   P. V Krishna and M. R. Babu, "The Role of ICTs in Knowledge Management (KM) for Organizational Effectiveness," *CCIS*, vol. II, pp. 542–549, 2011.

[19]   T. Maqsood, A. D. Finegan, and D. H. T. Walker, "Biases and Heuristics in Judgment and Decision Making: The Dark Side of Tacit Knowledge," *Issues Informing Sci. Inf. Technol.*, pp. 224–301, 2004.

[20]   R. M. Grant, "The Development of Knowledge Management in the Oil The Development of Knowledge management in the oil and Gas Industry," *Netw. Sci. Journals from Lat. Am.*, pp. 92–125, 2013, Accessed: May 31, 2022. [Online]. Available: http://www.redalyc.org/articulo.oa?id=43328679006.

[21]   S. Gupta, "Organizational Barriers to Digital Transformation," KTH ROYAL INSTITUTE OF TECHNOLOGY, STOCKHOLM, SWEDEN, 2018.

[22]   A. A. Kornienko, A. V. Kornienko, O. B. Fofanov, and M. P. Chubik, "Knowledge in artificial intelligence systems: searching the strategies for application," in *International Conference on Research Paradigms Transformation in Social Sciences 2014*, 2015, pp. 589–594, doi: 10.1016/j.sbspro.2014.12.578.

[23]   G. Sucharitha, A. Matta, K. Dwarakamai, and B. Tannmayee, "Theory and Implications of Information Processing," in *Emotion and Information Processing, A Practical approach*, Springer International Publishing, 2020, pp. 39–54.

[24]   C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating Machine Learning with Human Knowledge," *iScience*, no. 23, pp. 1–27, 2020, doi: 10.1016/j.isci.

[25]   M. Chugh, N. Chugh, D. K. Punia, and A. Agarwal, "THE ROLE OF INFORMATION TECHNOLOGY IN KNOWLEDGE MANAGEMENT MitaliChugh,*," in *Conference on Advances in Communication and Control Systems (CAC2S 2013)*, 2013, vol. 2013, no. Cac2s, pp. 688–693.

[26]   V. Rus, P. M. McCarthy, D. S. McNamara, and A. C. Graesser, "Natural Language Understanding and Assessment," in *Encyclopedia of Artificial Intelligence*, IGI Global, 2011.

[27]   G. Frisoni, G. Moro, and A. Carbonaro, "Unsupervised descriptive text mining for knowledge graph learning," *IC3K 2020 - Proc. 12th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, pp. 316–324, 2020, doi: 10.5220/0010153603160324.

[28]   B. Manaris, "Natural Language Processing: A Human-Computer Interaction Perspective," *Adv. Comput.*, vol. 47, no. C, pp. 1–66, 1998, doi: 10.1016/S0065-2458(08)60665-8.

[29]   M. Martinc, J. Stefan, and M. Robnik-˘ Sikonja, "Supervised and Unsupervised Neural Approaches to Text Readability," *Assoc. Comput. Linguist.*, vol. 47, no. 1, 2021, doi: 10.1162/COLI.

[30]   M. Jagannathan, D. Roy, V. Santosh, and K. Delhi, "Application of NLP-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies," *CSI Trans. ICT 2022*, pp. 1–10, May 2022, doi: 10.1007/S40012-022-00355-W.

[31]   R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front. Artif. Intell.*, vol. 3, p. 42, Jul. 2020, doi: 10.3389/FRAI.2020.00042/BIBTEX.

[32]   Tony Yiu, "Understanding NLP and Topic Modeling Part 1 - KDnuggets," *KD Nuggets*, 2019. https://www.kdnuggets.com/2019/11/understanding-nlp-topic-modeling-part-1.html (accessed May 31, 2022).

[33]   Skim AI, "Topic Modeling for Product Managers - A Beginner's Guide," 2016. https://skimai.com/topic-modeling-for-product-managers/ (accessed May 31, 2022).

[34]   R. M. Snyder and R. Com, "An Introduction to Topic Modeling as an Unsupervised Machine Learning Way to Organize Text Information," in *ASCUE Proceedings*, 2015, vol. 86, Accessed: May 31, 2022. [Online]. Available: http://www.robinsnyder.com.

[35]   L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *Springerplus*, vol. 5, no. 1, pp. 1–22, Dec. 2016, doi: 10.1186/S40064-016-3252-8/TABLES/4.

[36]   D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[37] H. Jelodar *et al.*, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2017, Accessed: May 31, 2022. [Online]. Available: https://www.researchgate.net/publication/321069759.

[38] I. R. Putri and R. Kusumaningrum, "Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia," *J. Phys. Conf. Ser.*, vol. 801, no. 1, Mar. 2017, doi: 10.1088/1742-6596/801/1/012073.

[39] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for Topic Models with Word Embeddings," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Jul. 2015, pp. 795–804, Accessed: Jun. 01, 2022. [Online]. Available: https://code.google.com/p/word2vec/.

[40] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models: A survey," *Front. Comput. Sci. China*, vol. 4, no. 2, pp. 280–301, 2010, doi: 10.1007/S11704-009-0062-Y.

[41] I. Saad, "Construction Contracts: From Zero-Sum to Win-Win," *University of Cincinnati*, 2020. https://www.researchgate.net/publication/343107480_Construction_Contracts (accessed Jun. 01, 2022).

[42] H. Jelodar *et al.*, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2018, Accessed: Jun. 01, 2022. [Online]. Available: https://www.researchgate.net/publication/321069759.

[43] K. Koc and A. P. Gurgun, "Ambiguity factors in construction contracts entailing conflicts," *Eng. Constr. Archit. Manag.*, 2021, doi: 10.1108/ECAM-04-2020-0254.

[44] P. Jafari, M. Al Hattab, E. Mohamed, and S. Abourizk, "Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation," *Appl. Sci.*, vol. 11, no. 13, Jul. 2021, doi: 10.3390/app11136188.

[45] V. Ivanov, A. Sadovykh, A. Naumchev, A. Bagnato, and K. Yakovlev, "Extracting Software Requirements from Unstructured Documents," Feb. 2022, Accessed: Jun. 01, 2022. [Online]. Available: http://arxiv.org/abs/2202.02135.

[46] F. ul Hassan and T. Le, "Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing," *J. Leg. Aff. Disput. Resolut. Eng. Constr.*, vol. 12, no. 2, p. 09, May 2020, doi: 10.1061/(ASCE)LA.1943-4170.0000379.

[47] T. Aksoy, S. Celik, and S. Gulsecen, "DATA PRE-PROCESSING IN TEXT MINING," in *Who Runs the World*, Istanbul University Press, 2020, p. 125.

[48] J. Daniel and J. H. Martin, "N-gram Language Models," in *Speech and Language Processing.*, 2021.

[49] J. Brownlee, "A Gentle Introduction to the Bag-of-Words Model," *Deep Learning for Natural Language Processing*, 2019. https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (accessed Jun. 01, 2022).

[50] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/IJCA2018917395.

[51] R. Řehůřek, "Corpora and Vector Spaces — gensim," May 06, 2022. https://radimrehurek.com/gensim/auto_examples/core/run_corpora_and_vector_spaces.html (accessed Jun. 01, 2022).

[52] J. Wira, G. Putra, and T. Tokunaga, "Evaluating text coherence based on semantic similarity graph," in *the Workshop on Graph-based Methods for Natural Language Processing*, Aug. 2017, pp. 76–85.

[53] E. Zvornicanin, "When Coherence Score is Good or Bad in Topic Modeling? | Baeldung on Computer Science," Dec. 07, 2021. https://www.baeldung.com/cs/topic-modeling-coherence-score (accessed Jun. 01, 2022).

[54] S. Kleinman and L. Thomas, "WE1S 'pyldavis' module," *WE1S Tools and software*, Sep. 18, 2020. https://we1s.ucsb.edu/wp-content/uploads/S-22.pdf (accessed Jun. 01, 2022).