Smart Data and Business Analytics: Propagation of a Performance Framework to Mitigate Rework in Mega-Projects

Matthews, J. and Love, P.E.D. and Porter, S.R. and Fang, W.

***Abstract*:** Within construction, we have become increasingly accustomed to relying on the benefits of digital technologies, such as Building Information Modelling, to improve the performance and productivity of projects. We have, however, overlooked the problems that technology is unable to redress. One such problem is rework, which has become so embedded in practice that technology adoption alone can not resolve the issue without fundamental changes in *how* information is managed for decision-making. Hence, the motivation of this paper is to bring to the fore the challenges of classifying and creating an ontology for rework that can be used to understand its patterns of occurrence and risks and provide a much-needed structure for decision-making in transport mega-projects. Using an exploratory case study approach, we examine 'how' rework information is currently being managed by an alliance that contributes significantly to delivering a multi-billion dollar mega-transport project. We reveal the challenges around location, format, structure, granularity and redundancy hindering the alliance's ability to classify and manage rework data. We use the generative machine learning technique of Correlation Explanation to illustrate how we can make headway toward classifying and then creating an ontology for rework. We suggest that when an ontology is created, it will enable a smart data and business analytics approach to be enatced, providing construction organisations with the foundations to formulate descriptive, predictive and prescriptive indicators to support their continuous improvement strategy.

Keywords: Business analytics, machine learning, rework, risk, smart data, topic modelling.

1.0 INTRODUCTION

"Information is a source of learning. But unless it is organised, processed, and available to the right people in a format for decision-making, it is a burden, not a benefit". William G. Pollard (1911 – 1989),

The context for this paper is the delivery of a mega transport[1] project. Such infrastructure projects are drivers of economic and social progress in countries worldwide. Transport projects are typically complex, and managing information throughout the life cycle of a mega-project poses many challenges (Jiang *et al*., 2018; Love *et al*., 2018a; Tian *et al*., 2021). Information needed for decision-making to deliver projects is created by many organisations in differing formats (e.g., structured and unstructured data) and is exchanged using various mediums (e.g., digital and paper). The result is that the *right* information (i.e., accurate and reliable) is often inaccessible, hindering the implementation of continuous improvement initiatives. The ability to provide accurate and reliable has been thwarted by the absence of an agreed industry-wide standardised and structured approach for managing information in projects (Tolam, 1999; Laasko and Kiviniem, 2012; Love *et al*., 2020).

The digital transformation process in construction, enabled by Building Information Modelling (BIM) and technologies such as Augmented Reality, Blockchain, Computer Vision, and the Internet of Things, has led to an abundance in the availability and accessibility of data and information. Issues, however, surrounding the collection, storage, retrieval and dissemination of information remain a pervasive problem, which can hinder the efficacy of decision-making and the ability to enact a robust continuous improvement strategy (Love *et al*., 2018a; Love *et al*., 2020; Tian *et al*., 2021). Contrastingly, in industrial sectors such as manufacturing, retail, and services (e.g., banking), that possess more mature digital transformation strategies,

[1] Typically defined by the monetary budget (e.g., >$1 billion), physical size, labour force requirements, complexity, and impact (e.g., economic, social, and political) (Hamdy, 2010).

emphasis has been placed on managing activities that change customers, people and organisations behaviour and the way they use their information to create knowledge (Kar and Dwivedi, 2020).

Within the information management literature, we seldom see research describing the 'wicked problems' that permeate and negatively impact the performance and productivity of organisations and, at the same time, remain unresolved issues. Instead, we are more likely to see new methodologies, technologies, theories, systems and success stories presented that provide learning opportunities for organisations (e.g., Duan *et al*., 2019; Carvalho *et al*., 2021; Ranjan and Foropon, 2021). Cynically, it can be argued that many methodological and technological solutions developed do not address fundamental problems that confront practice, as little consideration is given to the way data needs to be structured, documented and transferred for decision-making (Love *et al*., 2020).

We have become increasingly accustomed to celebrating advances in technology success and, in doing so, have overlooked the problems that it does not solve (Love and Matthews, 2019). As Methta (2019) suggests, "the collective bias towards finding positive results in the face of failure is a dangerous motivation". Hence, the motivation of this paper is to bring to light the challenges of addressing a problem that is so embedded in practice that the adoption of technology (e.g., BIM) will not resolve the issue unless fundamental changes in the way information are managed is undertaken (i.e., its collection, storage, management and maintenance).

When the flow of information is impeded in an organisation, it can adversely affect its ability to function effectively (Westrum, 2014), and problems may become masked. A case in point

is rework[2] performed during the construction of mega-transport projects where, on average, it has been shown to increase construction costs by 12% (Li and Taylor, 2014). Rework often remains a hidden issue as many construction organisations do not have specific information systems to collect, process, store, and disseminate information about its costs and causation (Robinson-Fayek *et al*., 2004). The upshot is that organisations cannot analyse and visualise the patterns of rework occurrence, and therefore struggle to assess and manage its risks (Love *et al*., 2021a;b). Germane examples of projects experiencing high rework costs in Australia include the Sydney Skytrain Gold Coast Light Rail Transit (LRT) and Sydney's LRT (Webb, 2017; Coultan, 2016; Bungard and Rabe, 2020).

Having to re-do a task or process is often frowned upon by managers and supervisors as it may result in additional time and cost on a project and organisation. However, when there is a need for rework, it is often rationalised as a one-off event and is considered to be 'uncomfortable knowledge[3]'. While organisations do record non-conformances that may require rework through their quality management systems, such rework only represents a small portion of the total amount that may occur in a project (Love *et al*., 2021b). More often than not, low levels of psychological safety prevail in construction projects, contributing to a reluctance to formally report quality issues requiring rework (Love *et al*., 2018b). Thus, it is rare for a construction organisation to know to what extent rework impacts their bottom-line, safety, or a project's environmental performance (Love and Matthews, 2020).

It is outside the scope of this paper to explain rework causation[4] and why people do not report such events (also the errors leading to its manifestation) in projects. However, it is important

[2] Defined as the "unnecessary effort of having to redo a process or activity that was incorrectly implemented the first time (Love, 2002: p.19)
[3] The presence of rework may be therefore: (1) denied; (2) dismissed, (3) diverted or (4) displaced (Rayner, 2012).
[4] An in-depth exposition of rework causation can be found in Love *et al.* (2018b).

to note that rework can be classified as *change* or *quality-based* (Love and Matthews, 2020). While quality-based rework forms only a fraction of the total cost a project may experience, it can have a far more adverse impact on the performance of projects as organisations are seldom reimbursed for its occurrence. It generally arises due to errors and violations (Love *et al*., 2018b).

Rework is not a 'tame' problem (i.e., well-defined, with a single goal and a set of well-defined rules) but is instead 'wicked' (i.e., loosely formulated) (Newell and Simon, 1972; Coyne, 2005). By their very nature, 'wicked 'problems are difficult or even impossible to solve due to their incomplete or contradictory knowledge, interconnectedness with other issues, and, as in the context of this research, the number of organisations and people involved (Love and Smith, 2019).

In making strides to address the 'wickedness' of rework and drawing on our observations from transport mega-projects practices, we report, as part of an ongoing study, the information management challenges that impede an organisation's ability to capture, monitor, and analyse its rework and associated wastes. As rework data is documented in a large number of records and reports, which contain considerable noise, we need to extract only that information required for decision-making (Love *et al*., 2021a).

In the next section of our paper, we examine the issues associated with managing rework information, emphasising the role of topic modelling in creating an ontology to determine the patterns of occurrence and risks and provide much-needed structure for decision-making (Section 2). We then describe our case study approach to examining the information management challenges associated with containing (i.e., measures designed to enhance the

detection and recovery from errors, as well as seeking to minimise adverse consequences) and reducing (i.e., measures designed to limit its occurrence) errors resulting in the need for rework (Section 3). Next, we use the generative machine learning technique of Correlation Explanation (CorEx) to illustrate how we can make headway toward creating an ontology for rework (Section 4). We suggest that an ontology will enable a smart data approach that can be used to develop a suite of business analytics to mitigate rework (Section 5). Finally, we conclude our paper by identifying its limitations, future research directions, and contributions (Section 6).

2.0 MANAGING REWORK DATA AND INFORMATION

In response to the prevalence of rework in construction, particularly in mega transport projects, many studies that have sought to understand its causes, costs and impacts (e.g., Barber *et al*., 2000; Rogge *et al*., 2001; Li and Taylor, 2014; Love *et al*., 2021b). This has stimulated a dialogue within the construction industry, which has led to many organisations embracing lean tools to reduce work (e.g., Last Planner® and Value Stream Mapping) through improving workflows, visualising activities and engendering a social network among contractors to enhance coordination (Freire and Alarcón, 2002; Priven and Sacks, 2015; Michaud *et al*., 2019). Despite the widespread use of digital technologies and lean tools to combat rework, it still prevails in construction (Love *et al*., 2020). Indeed, digital technologies and lean practices have somewhat contributed to addressing rework. Still, their impact has been minimal as practitioners have had limited access to data in a format that can be used for decision-making that can be used to assess its risks and identify strategies to mitigate its occurrence.

Identifying the properties of rework and their relatedness enables us to retrieve and organise data into information and knowledge about rework. In making strides towards this, an ontology of rework is required. An ontological knowledge representation enables the “physical and

abstract objects, relations between these objects, and events influencing these objects" for rework to be formally depicted and can form a basis for computational development and improved problem solving (Hartman and Trappey, 2020: p.6).

Within construction, several information systems have been designed and developed to manage rework in projects (e.g., Farrington, 1987; Willis and Willis, 1996; Low and Yeo, 1998; Love and Irani, 2003; Robinson-Fayek *et al*., 2004). However, the ontologies underpinning these information systems are artificially created and eschew relations, rendering them impractical for decision-making. This situation arises as rework information is shoehorned into an ontology to match the purpose of their analysis. Thus, if we are to accommodate the actuality of rework, a "bottom-up approach", whereby data is harvested from existing systems and peoples experience, is required to develop its ontology (Hartman and Trappey, 2020: p.6). Our intention here is not to be critical of the studies mentioned earlier, quite the contrary. They have laid essential building blocks to be considered when creating a robust ontology representing all aspects of rework and its relations in practice.

2.1 Topic Analysis Modelling

In line with previous text classification studies in construction, we have adopted topic analysis modelling[5], a machine learning technique and basic activity of Natural Language Processing (NLP). Topic analysis modelling can discover latent topics in documents and determine relationships between words, topics and documents and contribute to the creation of an ontology (Caldas *et al*.,2002; Chi *et al*., 2014; Zhou and El-Gohary, 2016; Zhang *et al*., 2019; Fang *et al*., 2020; Zhong *et al*., 2020a). We intend to make headway toward developing a rework ontology to describe the relationships and interconnectedness between rework

[5] Topic analysis machine learning comprises of: (1) topic modelling, which is an 'unsupervised' and does not require training of data; and (2) topic classification, which is 'supervised' requiring data to be trainined to be able to automatically analyze texts.

activities and processes. The creation of an ontology will form the basis for modeling high-quality, linked and coherent smart data that can be used to determine patterns and risk of rework occurrence. However, this is not a straightforward task given the number of different document types rework data is typically recorded and stored in, both formally and informally, especially in mega-projects (Love *et al*., 2018a; Fang *et al*., 2020; Tian *et al*., 2021).

Traditional or shallow[6] machine learning classifiers such as Support Vector Machine, Hidden Markov model or Conditional Random Fields have been widely used to classify text from accident and near-miss reports (Goh and Ubeynarayana, 2017; Zhang *et al*., 2018; Zhong *et al*., 2020a) and building regulations in construction (Zhou and El-Gohary, 2017; Zhong *et al*., 2020b). Notably, topic modelling and classification of rework issues have not been previously examined in the literature, as data has not generally been made available to researchers. While traditional machine learning approaches have become popular for classifying text from documents in construction and can yield good classification performance, they are "time-consuming and inefficient to use due to their reliance on manual-handcrafted features" (Zhong *et al*., 2020: p.2).

The shortcoming of traditional machine learning has resulted in the increasing use of a generative (i.e. unsupervised learning) statistical model for topic modelling, the Latent Dirichlet Allocation (LDA) algorithm (Zhong *et al*., 2020a; Tian *et al*., 2021). The LDA algorithm can automatically extract features from text descriptions and identify the topics within documents. While LDA has proven successful in topic modelling, it requires "detailed assumptions and careful specification of hyperparameters" to ensure valuable results

[6] Shallow is a types of machine learning where we learn from data described by pre-defined features

(Gallagher *et al*., 2017: p.529). Additionally, LDA cannot model relationships between topics and performs poorly with short sentences (Lin *et al*., 2017).

Rework data is often embedded in multiple documents that contain a lot of noise (i.e., unconnected data or data intended for a different purpose). Thus, there is a need to reduce the complexity and *learn* topics without imposing pre-existing notions if progress is to be made towards developing an ontology capable of supporting a knowledge engineering system and a continuous improvement strategy.

We can bypass the limitations of implementing LDA by applying CorEx to topic modelling and achieve similar performance and results with minimal human intervention (Gallagher *et al*., 2017). Unlike LDA, CorEx does not assume a particular data generating model but "instead searches for topics that are maximally informative about a set of documents" (Gallager *et al*., 2017: p.29). Thus, by learning from informative rather than generated topics in the context of rework, we can avoid relying solely on theoretical models and learn from practice (Gallager *et al*., 2017). CorEx allows the introduction of words as domain-knowledge anchors (Tisby *et al*., 1999). It combines *non-negative matrix factorisation* (Dhillion and Sra, 2006) to improve the coherence of the overall document classification and the topic model's predictability (Namekawa and Tezuka, 2021). As we stated in the introduction, this paper will use CorEX for topic analysis to build representations and relationships for quality-based rework. The open-source code for implementing CorEx is available at Github[7].

[7] The open-source code for CorEX at; https://github.com/gregversteeg/CorEx originally presented in Ver Steeg and Galstyan (2014). A further vision that features continuous variables, missing values, and Baynesian Smoothing is available at: https://github.com/gregversteeg/bio_corex/ with details being provided in Ver Steeg and Galstyan (2015).

3.0 CASE STUDY

A case study is a research approach used to generate an in-depth, multi-faceted understanding of a complex issue in its real-life context (Crowe *et al*., 2011). It is a well-established research design used extensively within various disciplines, particularly information systems (Tsang, 2014). Considering the limited research that has examined 'how' rework information is managed in projects, we adopt an exploratory case study to explain and classify its context within a project setting (Yin, 2018).

Our exploratory approach akin to a discovery process will provide organisations with a better understanding of the nature of information needed to adapt, embrace and respond to rework in mega transport projects. Our exploratory research aims to address the following research questions: How is rework information collected, and in what format by constructions organisations? Where is the information stored? And how can we use this information to manage the risks of rework? Having access to information to address these questions has been a challenge as it is typically commercially sensitive, and organisations are reluctant to admit that they have an issue with rework.

3.1 Case Selection

The backdrop for our research is a transport mega-project, initiated by an Australian State Government, comprising a program of works to remove existing and construct new road and rail infrastructure across the metropolitan area of a major Australian city. This multi-billion dollar project is being delivered using a series of program alliances. Due to issues of political sensitivity, we are unable to name the project and provide specific details about its characteristics, though it predominately comprises civil engineering works.

It has been acknowledged within the program alliance that rework is an issue. In late 2018, the project client (i.e., owner) held a one-day technical symposium, with representatives from each of the alliances involved in delivering the program of works, to examine how rework could be mitigated. After the symposium, we had several discussions with one of the alliances about 'how' to determine the causes, costs and impacts of rework, as no dedicated systems were in place to measure its occurrence. Consequently, we were invited to work with the alliance and examine how they managed their rework information, to understand by what means they could effectively use it as part of their continuous improvement strategy. The alliance had completed four projects between 2015 and 2021. Another four are currently under construction.

In striving to stimulate learning, innovation and continuous improvement, the alliance has developed and implemented an initiative focusing on containing and reducing errors resulting in rework and associated wastes. The researchers' involvement in the initiative is to design and develop a knowledge-based engineering system (KBES) to capture rework events to determine and manage future risks. However, as we show in this paper, achieving this goal requires analysis across multiple existing data structures to generate views suited for decision-making. Developing a KBES with practical relevance depends on having access to real-life projects and documentation, which, as we mentioned above, construction organisations seldom share with researchers.

3.1 Data Collection

A working and steering group were created, comprising representatives from three of the four organisations that formed the alliance. In addition, representatives from various contractors also attended the working group so their expertise and experience could be drawn upon to ensure the outcomes have relevance and add value to their work. The researchers acted as

participant observers during the working group's meetings, which occurred every other week. Steering group meetings were undertaken every month. As participant observers, the researchers participated in the s working and steering group meetings and actively listened to the rework issues raised over six months. The researchers familiarised themselves with the alliance's existing approaches used to record rework in various guises, with input from design managers, quality managers, projects engineers and site supervisors. Access was provided to the completed and ongoing project documentation sources, where rework data may have been recorded, such as site dairies, design change requests/notifications, non-conformance reports (NCR), (internal) requests for information (i/RFI), site instructions, punch lists and Touchplan® data (Figure 1). The discovery process served to clarify and document the current situation and provide us with a frame of reference to understand the nuances and context of the rework data available.

3.2 Dataset

Our NLP approach takes natural language text from the alliance's documents, identifies common rework topics, and then uses this knowledge to identify recurring themes. The procedure to classify text and identify topics from the alliance's documents is presented in Figure 2. As shown in Figure 1, rework data in the alliance is stored in a vast array of documents, but the most common sources, particularly for quality-based rework, are NCRs, iRFIs and RFIs. Our sample data is extracted from the completed projects and comprises iRFI (n=1987), RFIs (n=532) and NCRs (n=274).

To train our CorEX topic model, we used contract documents from the State's Road Standards[8] (SRStds)(n=109) to inform the quality standards that need to be adhered to in the project. With

the NCRs, RFIs and SRStds, our training set comprised 915 documents. The iRFIs were excluded at this stage as the text within them contained too much noise, with many records deemed too short (i.e., containing less than ten words) for analysis.

3.3 Procedure

The research was performed on a Jupyter[9] Notebook environment, using Python (3.7) as the programming language. As noted in Figure 3, data needs to be pre-processed before applying the CorEX topic model, which involves the following key steps (Rani and Kumar, 2021:p.5592):

- *Tokenisation*: Text is split into sentences and the sentences into words. Then, the words are converted into lowercase and punctuation is removed.
- *Stopword removal*: Stopwords are removed to give importance to significant words in a document, which provides the meaning of the text.
- *Stemming:* Attempts to reduce words to their most basic form or 'stem', typically removing the ends of words ([Culling, Culled] -> Cull) and, in some instances, the removal of derivational affixes. We have implemented the Snowball Stemmer provided by the Natural Language Toolkit (NLTK).
- *Lemmatisation:* The word's context is considered and returns the dictionary form, referred to as the *lemma*. This step is undertaken to reduce the data size to process, which helps improve the model's performance. This process enables the identification of 'nouns' to be used. Our study used the SpaCy[10] system to lemmatise inputs as it has a greater level of grammatical awareness (Lagus and Klami, 2021).

[9] Details can be found at: https://jupyter.org/
[10] Available at: https://spacy.io/

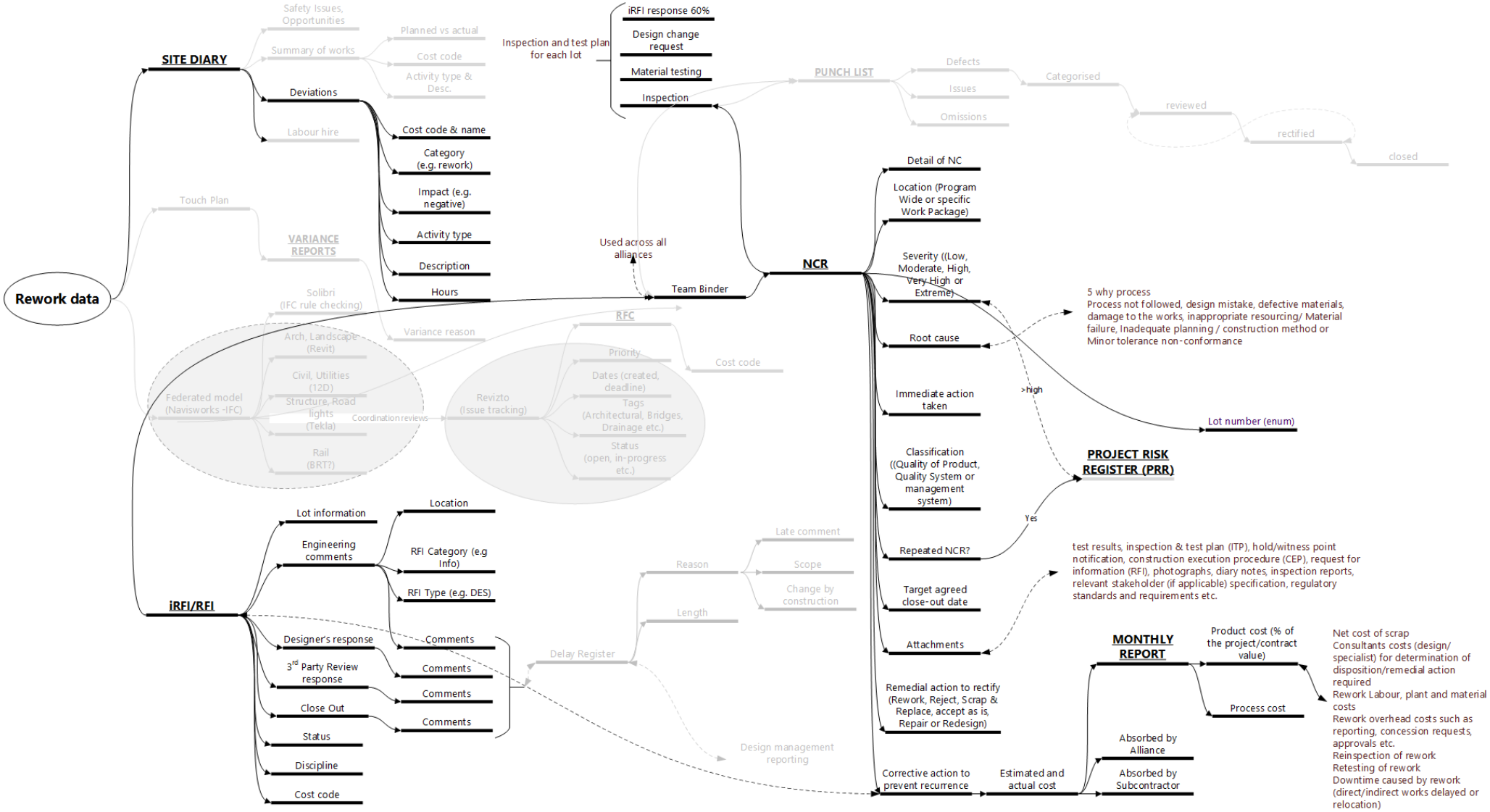


Figure 1. Mapping potential data sources for rework (greyed out sources not yet analysed in detail)

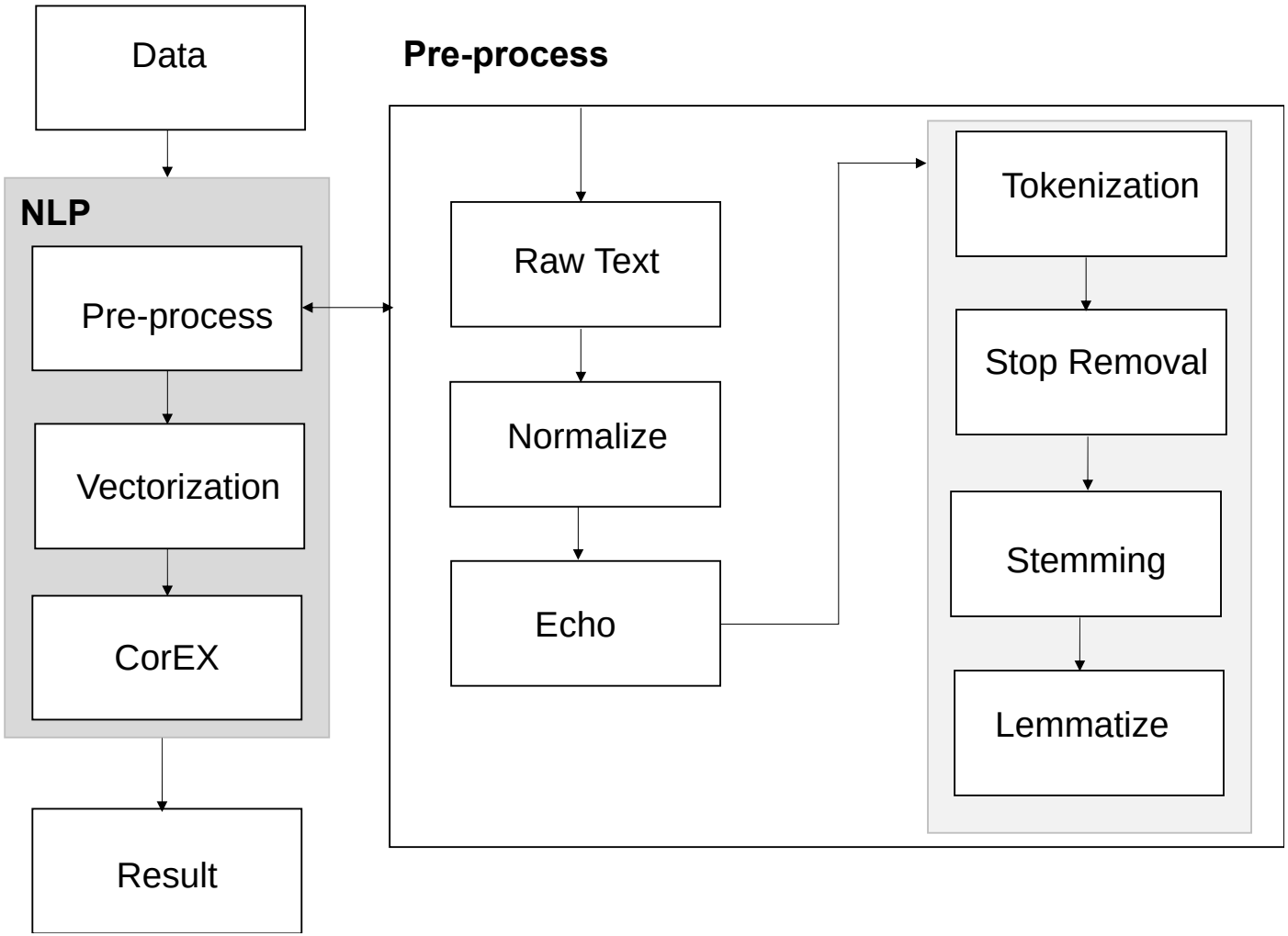
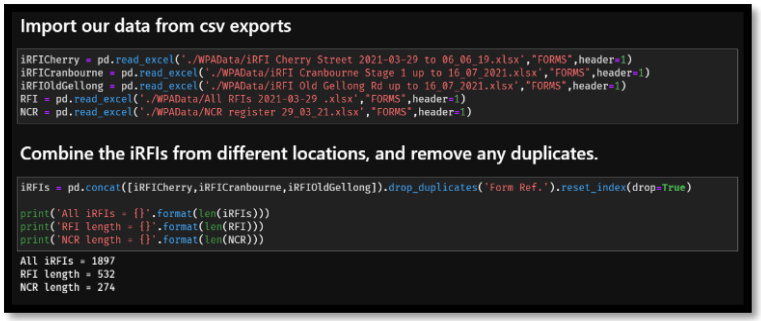
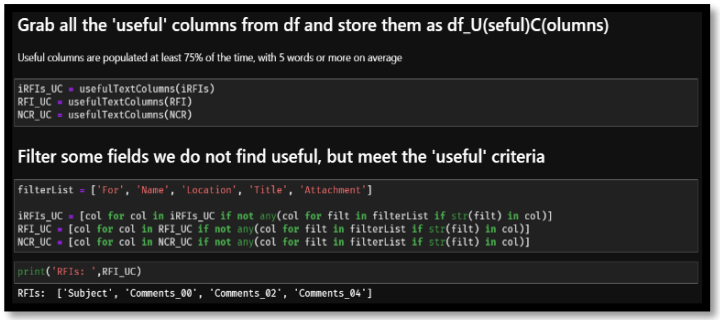


Figure 2. Research procedure

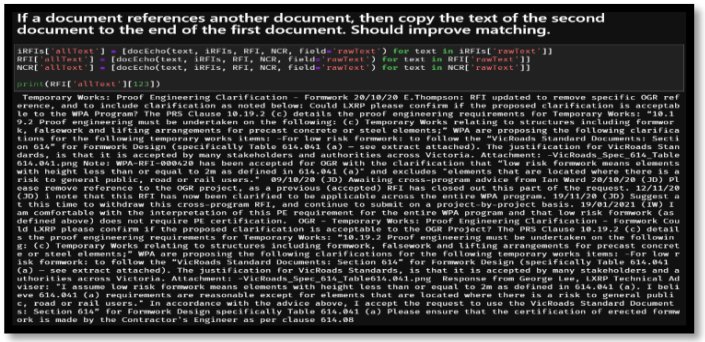


(a) Import data from TeamBinder

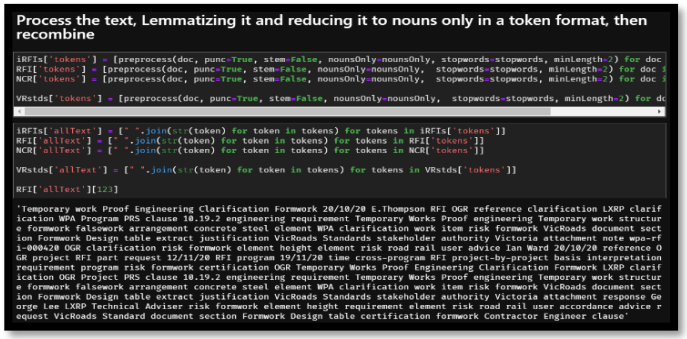


(b) Select and remove fields

(c) Combining of fields



(d) Document Echo



(e) Lemmatizing, stemming and nouns

Figure 3. Examples for a screenshot from pre-processing of data

The NCR and i/RFI data were stored in a cloud-based centralised document management system called *TeamBinder*[11], and retrieved in a comma-separated value (CSV.) file format. Each record comprises up to 74 columns, standardised to capture specific data (e.g., NCR or RFI/iRF) at various stages of a project's life cycle. Back end meta-data, such as when a particular record was first created or a response received, was also available. A purposeful selection of fields for each record was extracted and combined into a single 'Pseudo-Document' for that record (Figure 3c).

A 'document echo' function is searched within the pseudo-document for references to other documents. If one is found, the second record pseudo-document is appended to the first, as noted in Table 1. This function assists the NLP process with learning relations between words such as concrete, MPA, strength and requirements when the CorEX is initiated. The stopword removal, stemming (e.g., concrete and concreting results in concret), and lemmatisation is then enacted. Subsequently, the dataset is reduced to only 'nouns' to improve data density.

After pre-processing, the data is vectorised using *CountVectorizer*[12], provided by the scikit-learn library in Python. (Figure 4). Here words are converted into a numerical representation, which the CorEX algorithm can use to determine the similarity between words and their rate of occurrence across all documents. The results of the modeling process are presented in the next section.

[11] Details can be found at: https://www.teambinder.com/teambinder5/Home/
[12] Details can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

Table 1. Document 'Echo' function

Before the initiating Echo	**After initiating Echo**
RecordA: "Record B refers to concrete strength requirements." *RecordB*: "Concrete needs a MPA of at least xx to be useful."	*RecordA*: "Record B refers to concrete strength requirements. Concrete needs a MPA of at least xx to be useful." *RecordB*: "Concrete needs a MPA of at least xx to be useful."

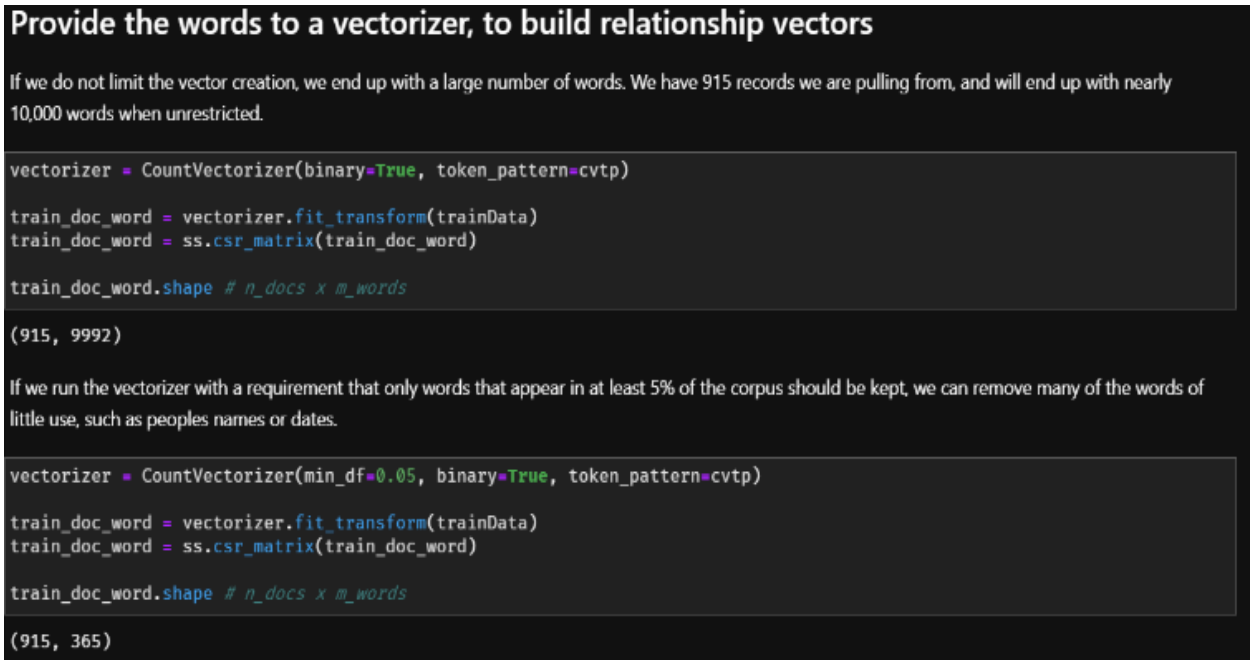


Figure 4. Vectorisation of data using *CountVectorizer*

4.0 EMERGENT OBSERVATIONS OF PRACTICE

Rework data were documented and stored in several locations within the alliance's overall information architecture (Figure 1), and the quality of data (e.g., granularity, consistency) captured and recorded was variable. This may be due to people's understanding of the information that needs to be documented (e.g., level of detail and format) and the time it takes to enter into a system (e.g., a non-conformance). In addition, the systems in place are primarily

intended for purposes other than capturing rework data. Site diaries, for example, provide an extensive record of events that occur daily during construction. The data entered is brief and often lacks a context, rendering them difficult, at times, to fully understand. While a reference can be made to another i/RFI, NCR, DCR, and the like, which may contain detailed information, searching for this information is a time-consuming process and a non-value adding activity. Handling the mismatch of information within the prevailing designed system requires developing and generating significant domain-specific and potentially project-specific rules.

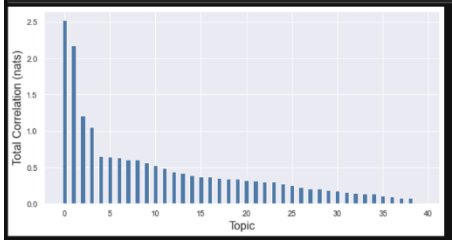
Figure 5 presents the results after training the CorEX algorithm with our data. Figure 5a identifies the Total Correlations graph, which displays the hierarchy of informative latent factors for quality-based rework. The top ten topics generated by a single run of the CorEx algorithm and associated with rework events occurring in the civil engineering works can also be seen. A cursory examination of the data reveals that reinforced concrete (e.g., concrete, reinforcement, formwork) is a source of quality-based rework and that issues around its compaction, fill, tolerance, and inspection prevail. These findings align with the research presented in Love and Matthews (2020)

We can now run out chosen NLP algorithm, CorEx, on our preprocessed text

```
#Sets the number of topics CorEx will train for
numTopics = 40

# Train the CorEx topic model with numTopics topics
topic_model = ct.Corex(n_hidden=numTopics, max_iter=1000, verbose=False, seed=1)
topic_model.fit(train_doc_word, words=words)

plt.figure(figsize=(10,5))
plt.bar(range(topic_model.tcs.shape[0]), topic_model.tcs, color='#4e79a7', width=0.5)
plt.xlabel('Topic', fontsize=16)
plt.ylabel('Total Correlation (nats)', fontsize=16);
```



(a) Total correlation graph

The top 10 topics generated by this CorEx run

```
topics = topic_model.get_topics()
topics = topics[:10]
for n,topic in enumerate(topics):
    topic_words,_,_ = zip(*topic)
    print('{}: '.format(n) + ', '.join(map(str, topic_words)))
```

```
0: superintendent, general, text, sections, method, contract, reference, section, definitions, minimum
1: action, ncr, corrective, remedial, designer, conformance, irfi, prs, wpa, tolerance
2: commencement, completion, interval, manner, ground, personnel, end, operation, hole, care
3: test, testing, result, sample, day, vicroads, frequency, compaction, quality, fill
4: limit, size, moisture, content, table, value, source, state, future
5: reinforcement, inspection, repair, bar, steel, measurement, position, rectification, plate, crack
6: type, time, property, rate, plant, week, requirements, amount
7: edge, drawing, cover, level, grade, top, face, height, alignment, dimension
8: damage, manufacturer, identification, name, product, coating, defect, form, way
9: contractor, specification, period, rock, class, transport, production, representative
```

(b) Top 10 topics

Figure 5. CorEX modeling results

Like most NLP algorithms, CorEx has several variables that influence its performance. Additionally, the quality and quantity of training data impact the ability of CorEx to assign topics to documents accurately. For example, distinguishing between topics that have a causal relationship or correlation with quality-based rework. Correctly processing the data to select what is provided to the NLP algorithm is itself a 'wicked problem'. Tian *et al*. (2021) reiterate this point, stating that "simplification of sentences may lead to the loss of construction text knowledge. This unsurprising finding explains the exact challenge of natural language processing".

To develop and refine our NLP system, we use a combination of metrics to evaluate its performance. Firstly, CorEx optimises for and provides a Total Correlation (TC) value, representing the overall correlation of the topics. A higher TC value indicates the overall group of topics better matches the provided corpus of documents. The TC value is graphed in Figure 5a, enabling a judgement on the number of topics included in a given data set. This process contrasts with methods such as LDA, where the topic number must be arrived at through trial and error.

An in-group versus out-group analysis was also performed, with the results displayed in Figure 6. After CorEx has generated its topics, this provides a binary vector for each document, indicating if that topic exists within it or not. Using the *Jensen Shannon Divergence* (JSD) algorithm (Schütze and Manning, 1999), we can then compare the vector representation of any two documents to judge how closely related they are, based on the topics generated.

Before running the algorithm, we manually identified four pairs of documents that we believed to be strongly related to their paired partner but not related to the other six documents. The JSD

algorithm is used to measure the distance between the associated pairs, or in-group, and the distance between them and the non-related documents, or out-group, which we then average. This provides us with four reference points to determine how well a given iteration of training data and the CorEx settings have generated topics that distinguish between related and unrelated documents. The values generated are averaged to produce the ALL AVG identified in Figure 6. The standard deviation (ALL STD) is also specified, assessing the solution's stability.

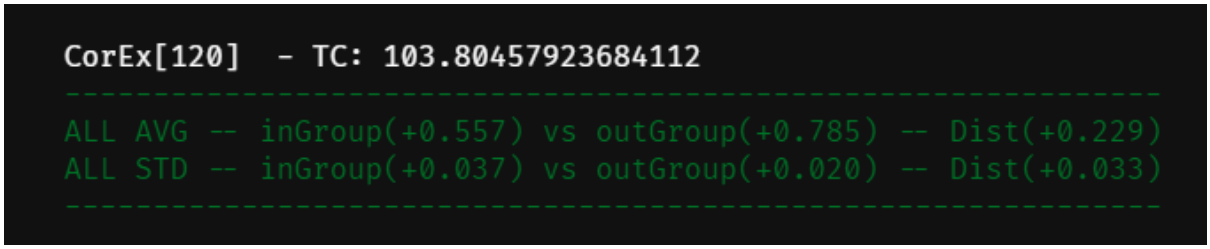


Figure 6. Sample in-group versus outgroup analysis

Finally, manual analysis is performed for runs that show promising numerical results to examine the generated topics and generate a list of most similar documents to a known document. The list is checked for documents known to be related, and those unknown are manually classified. While we continue to evaluate all elements of our NLP process to improve its performance (e.g., training as data becomes available), the use of CorEx demonstrates promising results. With additional data and further development, the identification and labelling of topics will improve, enabling us to create a rework ontology that can be used as a basis for performance management and risk analysis.

5.0 DISCUSSION

A significant amount of data is being collected within the exploratory case examined. Still, the data we analysed was unstructured and contained noise, making it difficult to determine the

right information needed for the training of algorithms and ensure the results are helpful for rework risk analysis. For example, the data collated from quality-based rework incidents under the auspices of NCRs tended to be unstructured, scant in detail, often duplicated and not cross-referenced with other events that may have coincided as a consequence of its occurrence. We should not be surprised by this observation as it is the norm in construction (Love *et al*., 2018b). However, it's important to note that when rework is required, the likelihood of an adverse safety event increases (Love *et al*., 2018).

Similarly, the potential for environmental impact (e.g., contamination and pollution) also amplifies. Notably, "most construction organisations mimic each other's practices so that no competitor has an overwhelming strategic advantage in their respective marketplaces. The downside here is that established rules and norms of the organisation-project dyad are, rarely if at all questioned and changed" (Love *et al*., 2018b: p.1114). As a matter of course, we have provided a detailed description and some analysis to better understand how rework information is managed in practice. We can now expand constructs and interrelationships based on our distinct settings as a means to develop a theoretical platform to examine the data (e.g., patterns of behaviour) required to improve decision-making is established.

Big data analytics provides the basis to identify patterns and derive insights into construction organisations and mega-projects business performance. However, many construction organisations remain unprepared to effectively utilise big data analytics, especially as there has been a slow uptake of digital technologies and an under-appreciation of the importance of data to their business operations (Gandomi and Haider, 2015; Bilal *et al*., 2016; Bilal *et al*., 2019; Ngo *et al*., 2020; Love *et al*., 2020).

5.1 Theoretical Contributions and Implications

Indeed, a commodity that construction organisations can draw upon to differentiate themselves from their competitors is *data*. The insights that can be mined from data are an invaluable resource that can provide the currency and means to drive an organisation's organic growth. Thus, it is essential for construction organisations to collect data to facilitate meaningful analysis and obtain insights into patterns of behaviour that can be used for decision-making and, therefore, formulate a competitive advantage.

Our exploratory analysis of the top ten topics revealed that reinforced concrete and associated activities are a major source of quality-based rework. This finding is not surprising considering the civil engineering works are highly reliant on reinforced concrete. However, it bolsters the need to examine where and why quality problems materialise in greater detail. By deriving topics from the documentation and with the help of experts, we will be able to create a taxonomy (i.e., "define the classes and the class hierarchy") and domain ontology, enabling knowledge about rework to be shared and re-used (including information retrieval) for decision-making and engendering performance management (Niu and Issa, 2015:p.473). While there this some way to go to develop a rework ontology, our paper provides a theoretical foundation to commence this journey. As a consequence of the work presented, we are better positioned to understand the rework problem, its complexities and nuances. Moreover, it contributes to the process of theorising and building an "epistemology of questioning", ensuring relevance to practice (Turnbull, 2017:p.3)

With advancements in Artificial Intelligence (AI), construction organisations are well-positioned to harness the benefits[13] of big data, if priority is given to managing information, to develop their competitive advantage, which can be transferred into the projects they deliver (Bilal *et al*., 2019). Strang and Sun (2020) cogently point out that "each discipline and industry has unique big data analytics issues" (p.982). Within the context of construction, big data has been identified as having three[14] defining characteristics, which mega-projects can utilise to ensure their successful delivery (Laney, 2001; Bilal *et al*., 2016; Bilal *et al*., 2019; Ngo *et al*., 2020):

1. Volume (i.e., terabytes, petabytes of data and beyond that may be derived from design and cost data and the like).
2. Variety (i.e., heterogeneous formats such as drawings, text, sensors, audio, video, and graphs).
3. Velocity (i.e., continuous streams of the data from dynamics sources such as sensors and Radio Frequency Identification).

These characteristics, referred to as the 3Vs, may enhance value generation by enabling organisations within a mega-project, particularly during its operations, "to automate decisions that were previously dependent on human judgment and intuition" (Ghasemaghaei, 2019: p.2). With a large amount of available data and AI technologies to process them, mega-projects can position themselves to quickly exploit new information to create and implement new ideas (Sivarajah *et al*., 2017).

[13] The espoused benefits of big data are widespread and far reaching including (Bilal *et al*., 2019): (1) increases in efficiency (e.g., use of google maps, geographical information systems, and social media); (2) accurate budget estimates and margins; (3) innovation; (4) effective decision-making; (5) safety management; (6) risk reduction; (7) improved connectivity

[14] Other scholars such as Jovanovi *et al*. (2015) have suggested 5V's: (1) high volume; (2) complex variety; (3) large velocity; (4) strategic value; and (5) veracity. However,

Despite the claimed benefits of big data and the increasing drive for construction organisations to embrace and apply its dimensions in their respective projects (Construction Industry Council, 2014; Han and Golparvar, 2016; Ngo *et al*., 2020), its adoption should be treated with a degree of scepticism "as big data is not always better data" (Ghasemaghaei and Calic, 108: p.147). Studies have shown its adoption does not necessarily enhance innovation performance or provide the positive business outcomes initially expected (Kwon *et al*., 2014; Johnson *et al*., 2017; Ghasemaghaei *et al*., 2017a;b). Additionally, organisations need to be cognisant that the impediments to successfully adopting big data analytics are "insufficient organisational alignment, lack of middle management adoption and understanding and business resistance (McShea *et al*., 2016). Affirming this resistance and inability to adapt and respond to change, construction organisations have struggled to realise the benefits of digital technologies as information management is deemed to be a secondary function of their operations (Matthews *et al*., 2018). A case in point, as we show, in this paper pertains to the documentation, structure and format of rework data.

Considering the above discussion and the unearthed observations during the discovery process with the alliance, regardless of the purported benefits[15] of big data analytics, we believe it would be unfeasible for a construction organisation in the short-to-medium term to predict the likelihood of rework considering their prevailing information management constraints. So, if we were to focus on collecting large amounts of rework data and using sophisticated machine learning algorithms then, there is a likelihood we will end up with sub-optimal outcomes for assessing rework risks. In this instance, progress toward improving project performance will be stymied (Ghasemaghaei and Calic, 2020). Accordingly, Fenton and Neil (2018) maintain that "big data, even when carefully collected, is typically unstructured and noisy; even the

[15] Benefits include increasing productivity and efficiency (e.g., increased processing of data), reduced costs (e.g. streamlining operations), improveing customer service and experience (e.g., technical support), and improved decision-making (e.g, data-driven sights).

‘biggest data’ typically lack crucial, often hidden, information about key causal or explanatory variables that generate or influence the data we observe”. (p. 1). Therefore, we suggest that emphasis be placed on acquiring ‘smart data’ rather than big data.

Computing algorithms have varying levels of complexity and efficiency, which can be described using ‘Big O’ notation. Machine learning algorithms can generally be expected to have a minimum complexity of O(m*n), where *n* is the number of records and *m* is the number of fields on each record that need to be compared. Even if it was possible to input all an organisation’s big data into a machine learning algorithm and obtain results, the more data loaded into the algorithm, the more computational resources and storage is needed, both of which are finite and costly.

A smart data approach is driven by what data are required for prediction rather than what is available (Constantinou and Fenton, 2017; Fenton and Neil, 2018). Such data is processed and turned into actionable information, empowering an organisation. Therefore, it makes big data manageable and actionable, and in doing so, trades ‘volume’ for ‘value’ and ‘veracity’. However, the challenge is determining what data is required to predict and mitigate rework. It is this problem that the alliance we have examined in this paper is endeavouring to address through its continuous improvement strategy.

We cannot rely on machine learning techniques alone to determine rework relationships no matter how large the dataset. As mentioned above, we also need to draw on practitioners’ knowledge and insights to help understand and develop relationships to structure the rework data and contribute to the development of an ontology. Ontologies “help humans and computers understand and fully utilise domain knowledge” and, therefore, are the first step

towards the development of a KBES that can be used for managing rework risks (Hartman and Trappy, 2020: p.5).

5.2 Implications for Practice

By leapfrogging over the discourse that surrounds big data, we can build a "coherent understanding and a nomenclature" for a smart data approach, which can be used to formulate our intended ontology (Gandomi and Haider, 2015:p.137), which we have yet to develop. Once developed based on our topic modelling approach, it will provide the basis for data-driven decision-making, forming the heart of business analytics.

The success of deploying business analytics is dependent on data quality (e.g. smart data), an organisational commitment to using data to garner insights that inform decisions and analysts who understand the whys and wherefores of technology and business performance. Armed with an ontology based on smart data, we suggest that business analytics can facilitate the utilisation of three performance measures, which can be part of a continuous improvement strategy and enable benchmarking and learning in projects (Table 2). By recognising trends in rework, testing hypotheses about its causes and effects, and drawing conclusions from the analysis, an analytical framework such as the KBES, which we aim to develop, can be used in everyday decision-making. In particular, the smart data and ontologies need to be aligned to the specific questions that we propose in Table 2.

The rework data in its current format within the alliance can only be used to develop lagging indicators for rework. Lagging indicators occur after the event and are useful for determining outcomes but are not helpful when an organisation needs to be proactive (instead of being reactive) and adjust to change to create positive outcomes. In some organisations, the cost of

non-conformances requiring rework has been used as a lag indicator of quality. However, this measure only provides a snapshot of the total amount of rework that can arise in construction and does not reflect the actual levels of quality in a project (Love and Matthews, 2020). As already noted, most construction organisations do not formally measure their rework.

Well aware of the negative impact that rework can have on a project's performance, the alliance is committed to undertaking a more in-depth exploration of its data, not only to create lag indicators but also as a suite of lead indicators, based on predictive (insight) and prescriptive (foresight) analytics. Noteworthy, the indicators identified in Table 2 can also be used to manage other issues such as safety and environmental performance.

5.2.1 Recommendation for Project Stakeholders

The mega-project transport project we have examined comprises several alliances that deliver a program of works. The alliance discussed in this paper consists of four organisations: (1) two engineering design-houses; (2) a contractor; and (3) an operator. A 'gain-share, pain-share' regime forms an integral part of the alliance's contract. An alliance is underpinned by a risk and reward compensation regime. Thus, parties equitably share in the financial 'gain' of a project's success or the financial 'pain' of its underachievement. The 'gain-share, pain-share' payment model fosters a 'win, win', lose, lose' mindset. Thus, as rework has can adversely impact a project's performance, it is beneficial to all parties to develop a KBES, which can be used to support the implementation of the business performance management framework for managing the risks of rework we have proposed in Table 2. In addition, the KBES can be applied and used by the mega-project's other alliances and used for benchmarking and comparing their performance and therefore can stimulate learning and innovation. Such benefits would not only be felt by the alliances but the wider community.

Table 2. Performance management framework to manage rework

PERFORMANCE MANAGEMENT			
Indicator Type	Lagging (*Hindsight*)	Lagging/Leading (*Insight*)	Leading (*Foresight*)
Analytic Type	Descriptive (Insight into the past)	Predictive (Understanding the future)	Prescriptive (Handling similar situations in the future)
Questions	• What has happened? • Why did it happen? • What is happening now?	• What will happen? • Why will it happen?	• What should be done? • Why should it be done?
Techniques	• Statistical Analysis (e.g., Descriptive Statistics, Data Mining, and Data Aggregation)	• Probabilistic Models (e.g., Bayesian Networks and Markov Chains), • Machine Learning (e.g., Pattern Recognition, Support Vector Machine, Random Search, Artificial Neural Networks) • Statistical Analysis (e.g., Linear Regression, Multiple Regression, Logistic Regression)	• Probabilistic Models (e.g., Markov Decision Process and Hidden Markov Model) • Machine Learning/Data Mining (e.g., K-means clustering, Convolutional Neural Networks and Reinforcement Learning) • Mathematical Programming (e.g., Mixed Integer Program and Linear Program) • Evolutionary Computation (e.g., Genetic Algorithm and Particle Swarm) • Simulation (e.g., System Dynamics and Monte-Carlo Simulation) • Logic-based Models (e.g., Association Rules, Decision Rules and Criteria-based Rules)

Adapted from Lepenioti *et al*. (2020)

5.3 Limitations and Future Research

While our exploratory research has highlighted the absence and issues with developing an information management system to manage the risks of rework in a mega-project setting, it has limitations. The research is ongoing, and the dataset used for the topic modelling is small and contains noise. But, as more data becomes available, we will be better positioned to improve the accuracy of our topic analysis and develop our proposed ontology and KBES. Needless to say, in a world's first, we have been able to map the data sources of rework in a mega-project. The upshot is that we are now well-positioned to apply a smart-data approach based on the topic modelling to develop an ontology of rework that has relevance to practice in the future.

6.0 CONCLUSION

The issue of rework can be a problem during the construction of mega-transport transport projects. An absence of psychological safety in projects results in errors requiring rework often going unreported. However, the documented rework information is usually scant, fragmented and ambiguous, making it difficult for construction organisations to implement mechanisms to mitigate its occurrence. As we pointed out at the commencement of our paper, information is a source of learning. It must be structured, processed, and available in the right format for decision-making; this has been a challenge for the alliance we have examined in this paper, and the construction industry worldwide, within the context of rework.

As part of a continuous improvement strategy and drive to instigate learning, the alliance has embarked on an initiative to re-design how its rework data is managed to unlock productivity and performance improvements. Ensuring decision-makers are provided with the right rework information is a proverbial quest, but unfortunately, roadblocks tend to prevail. Indeed, the

roadblocks are numerous (e.g., behavioural and cultural) but improving the flow of information between people and systems will provide the basis for improved decision-making.

Initially, we mapped the sources of rework data to understand its location, format and relevance. Then, we examined NCRs and documents referencing them from four projects completed by the alliance over six years. The generative machine learning technique of CorEx was used to discover the top ten abstract topics related to rework that occurred in the documents. Such modelling enabled us to create a theoretical setting to develop an ontology. A limitation of our work is that it is exploratory and only takes account of quality-based rework (i.e., NCRs). Research needs to also focus on change-based rework occurring in construction. We proffer that an ontology will provide construction organisations with the foundation to institutionalise a smart data and business analytics approach. As a result, we propose a performance management framework that comprises descriptive (hindsight), predictive (insight) and prescriptive (foresight) analytics that can be used to support such a strategy. Nonetheless, determining the information needed to enact the proposed performance management framework poses a challenge, but we strive to address this issue with our continued research with the alliance. Thus, our future research will develop a rework ontology to support our smart data and business analytics approach.

To this end, the contributions of our paper are twofold as we: (1) propose the use of a smart data approach (i.e., driven by the data required for its prediction, rather than all the data available) to support the development of new insights and understand the patterns and risks of rework; and (2) recommend a series of novel business analytic indicators that can be incorporated into an organisation's continuous improvement strategy to measure and predict rework, which will emanate from the creation of a rework ontology.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, Professor Yogesh Dwivedi and anonymous reviewers for their insightful and constructive comments, which have helped improve the quality of this manuscript. The authors would also like to thank the *Australian Research Council* for providing the funding to enable the work presented in this paper to be undertaken (DP210101281). Finally, we are indebted to the alliance and its stakeholders for their ongoing support of the study presented in this paper.

REFERENCES

Barber, P., Graves, A., Hall, M., Sheath, D. and Tomkins, C. (2000). Quality failure costs in civil engineering projects. *International Journal of Quality and Reliability Management*, **17** (4/5), pp. 479-492, doi.org/10.1108/02656710010298544

Bilal, M. Oyedele, L.O., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A., Pasha, M. (2016). Document details – big data in construction the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, **30**(3), pp.500-521, doi.org/10.1016/j.aei.2016.07.001

Bilal, M., Oyedele, L.O., Kusimo, H.O., Owolabi, H.A., Anuoluwapo, A.O., Akinade, O.O., and Delgado, J.M.D. (2019). Investigating profitability of construction projects using big data: A project analytics approach. *Journal of Building Engineering*, **26**, 100850, doi.org/10.1016/j.jobe.2019.100850

Bungard, M., and Rabe, T. (2020). Light rail breaks down in CBD. *The Sydney Morning Herald*. January 21st, Available at: https://www.smh.com.au/national/nsw/light-rail-breaks-down-in-cbd-20200121-p53tgs.html, Accessed 25th February 2020.

Caldas, C., Soibelman, L., and Han, J. (2002). Automated classification of construction project documents. *ASCE Journal of Computing in Civil Engineering*, **16** (4) pp. 234-243, doi.org/10.1061/(ASCE)0887-3801(2002) 16:4(234)

Carvalho, A., Merhout, J.W., Kadiyala, Y., and Bentley II, J. (2021). When good blocks go bad: Managing unwanted blockchain data. *International Journal of Information Management*, 57, 102263, doi.org/10.1016/j.ijinfomgt.2020.102263.

Chi, N.W., Lin, K.Y., and Hsieh, S.H. (2014). Using ontology-based text classification to assist Job hazard analysis. *Advanced Engineering Informatics*, 28 (4) pp. 381-394, doi.org/10.1016/j.aei.2014.05.001

Construction Industry Council (2014). *Built Environment 2050: A Report on our Digital Future. BIM 2050*, Available at: be2050-cicbim2050-2014-1.pdf, Accessed 1st December 2020

Constantinou, A. C. and Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, 124, pp.93-104, doi.org/10.1016/j.knosys.2017.03.005

Coultan, M. (2016). Cost overruns 'new normal' in transport projects. *The Australian*, 2nd December, Available at: https://www.theaustralian.com.au/national-affairs/state-politics/cost-overruns-the-new-normal-in-transport-projects/news-story/b73a56a972e8b052d0d2ed2b72b50280, Accessed 13th July 2018.

Coyne, R. (2005). Wicked problems revisited. *Design Studies*, **26**(1), pp. 5-17, doi.org/10.1016/j.destud.2004.06.005

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., and Sheikh, A. (2011). The case study approach. *BMC Medical Research Methodology*, 11, Article number 100, https://doi.org/10.1186/1471-2288-11-100

Dhillion, I.S., and Sra, S. (2005). *Generalised non-negative matrix approximations with Bregman divergences*. NIPS'05: Proceedings of the 18th International Conference on Neural Information Processing Systems pp. 283–290, Available at: https://papers.nips.cc/paper/2005/file/d58e2f077670f4de9cd7963c857f2534-Paper.pdf, Accessed 21st August 2021

Duan, Y., Edwards, J.S., and Dwivedi, Y.K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, **48**, pp.63-71, doi.org/10.1016/j.ijinfomgt.2019.01.021

Gallagher, R.J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, [S.l.], v. 5, p. 529-542, dec. 2017. ISSN 2307-387X. Available at: <https://transacl.org/ojs/index.php/tacl/article/view/1244>. Date Accessed: 21st August. 2021

Gandomi, A., and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, **35**(2), pp.137-144, doi.org/10.1016/j.ijinfomgt.2014.10.007

Ghasemaghaei, M. (2019). Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics. *International Journal of Information Management*, 102055, doi.org/10.1016/j.ijinfomgt.2019.102055

Ghasemaghaei, M. and Celic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, **108**, pp.147-162, doi.org/10.1016/j.jbusres.2019.09.062

Ghasemaghaei, M., Ebrahimi, S., and Hassanein, K. (2017a). Data analytics competency for improving firm decision making performance. *The Journal of Strategic Information Systems*, **27**(1), pp.101–113, doi.org/10.1016/j.jsis.2017.10.001

Ghasemaghaei, M., Hassanein, K., and Turel, O. (2017b). Increasing firm agility through data analytics: The role of fit. *Decision Support Systems*, **101**, pp.95–105, doi.org/ doi.org/10.1016/j.dss.2017.06.004

Fang, W., Luo, H., Xu, S., Love, P.E.D., Lu, Z., and Ye, c. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*, 44, 101060, doi.org/10.1016/j.aei.2020.101060

Farrington, A. (1987). *Methodology to Identify and Categorise Costs of Quality Deviations in Design and Construction,* Ph.D. Dissertation, Graduate School of Clemson University, Clemson, South Carolina, USA,

Fenton, N., and Neil, M. (2018). How Bayesian Networks are pioneering the 'smart data' revolution. *Open Access Government*, 4th June, Available at: https://www.openaccessgovernment.org/how-bayesian-networks-are-pioneering-the-smart-data-revolution/46329/, Accessed 20th December 2020.

Freire, J., and Alarcón, L.F., (2002). Achieving lean design process: Improvement methodology. *ASCE Journal of Construction Engineering and Management* **128** (3): 248-256, doi.org/10.1061/(ASCE)0733-9364(2002)128:3(248)

Goh, Y.M., and Ubeynarayana, C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis and Prevention*, **108**, pp.122-130, doi.org/10.1016/j.aap.2017.08.026

Han, K.K., and Golparvar, M. (2016). Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. *Automation in Construction*, **73**, pp.184.198, doi.org/10.1016/j.autcon.2016.11.004

Hamdy, K. (2010). The essential role of leadership in managing mega projects. Paper presented at PMI® Global Congress 2010—North America, Washington, DC. Newtown Square, PA: Project Management Institute.

Hartman, T. and Trappy, A. (2020). Advance engineering informatics: Philosophical and methodological foundations with example from civil and construction engineering. *Developments in Built Environment*, doi.org/10.1016/j.dibe.2020.100020

Jiang, H., Qiang, M., Fan, Q., and Zhang, M. (2018). Scientific research drive-by large-scale projects: A case study of the Three Gorges project in China. *Technological Forecasting and Social Change*, **134**, pp.61-71, doi.org/10.1016/j.techfore.2018.05.012

Johnson, J. S., Friend, S. B., and Lee, H. S. (2017). Big Data facilitation, utilisation, and monetisation: Exploring the 3Vs in a new product development process. *Journal of Product Innovation Management*, **34**(5), pp.640–658, doi.org/10.1111/jpim.12397

Jovanovi, U., Stimec A., and Vladusi, D. (2015). Big-data analytics: a critical review and some future directions. *International Journal of Business Intelligence Data Mining*, **10**(4): pp.337–355, doi.org/10.1504/IJBIDM.2015.072211

Kahn, W.A. (1990). Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal*, **33**(4), pp. 692-724, doi.org/10.5465/256287

Kar, A.K., and Dwivedi, Y.K. (2020). Theory building with big-data-drive research – Moving away from "what" towards the "why". *International Journal of Information Management*, 54, 102205, doi.org/10.1016/j.ijinfomgt.2020.102205

Laakso, M., and Kiviniemi, A. (2012). Document details – The IFC standard – A review of history, development, and standardisation. *Journal of Information Technology in Construction*, **17**, pp. 134-161,

Lagus, J., and Klami, A. (2021). Learning to lemmatise in the word representation space. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 31st May -2nd June, Reykjavik, Iceland, Linköping University Electronic Press, Sweden, Available at: https://aclanthology.org/2021.nodalida-main.25/, Accessed 22nd August 2021.

Laney, D. (2001). 3d data management: controlling data volume, velocity and variety META Group Research. Note, 6 p. 70

Le, S., Song, J., and Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computing Information Systems*, 51(1), pp.1-10**,** doi.org/10.1080/08874417.2010.11645444

Lepenioti, K., Bousdekis, A., Apostolou, D., and Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, **50**, pp.57-70, doi.org/10.1016/j.ijinfomgt.2019.04.003

Li, L., and Taylor, T.R.B. (2014). Modelling the impact of design rework on transport infrastructure construction performance. *ASCE Journal of Construction Engineering and Management*, **140**(9) doi.org/10.1061/(ASCE)CO.1943-7862.0000878

Lin, K.P., Shen, C.Y., Chang, T.L. and Chang, T.M. (2017). *A consumer review of driven recommender service for web e-commerce*. Proceedings of the 2017 IEEE Conference on Service-oriented Computing and Applications (SOCA), 25th-27th November, Kanazawa, Japan, pp.206-210, doi.org/ 10.1109/SOCA.2017.35

Ling, H., Ma, J., and Chen, C. (2017). Topic detection from microblogs using T-LDA and perplexity. *Asia-pacific Software Engineering Conference Workshops*, pp. 71-77, doi.org/10.1109/APSECW.2017.11

Love, P.E.D. (2002). Influence of project type and procurement method on rework costs in building construction projects**.** *ASCE Journal of Construction Engineering and Management*, **128**(1), pp.18-29, doi.org/10.1061/(ASCE)0733-9364(2002)128:1(18)

Love, P.E.D., and Irani, Z. (2003). Project management quality cost information system for the construction industry. *Information and Management* **40**, pp.649-661, doi.org/10.1016/S0378-7206(02)00094-0

Love, P.E.D. and Smith, J. (2019). Unpacking the ambiguity of rework in construction: Making sense of the literature. *Civil Engineering and Environmental Systems*, **34**, (1-4), pp. 180-203, doi.org/10.1080/10286608.2019.1577396

Love, P.E.D. and Matthews, J. (2020). Quality, requisite imagination and resilience: Managing risk and uncertainty in construction. *Reliability Engineering and System Safety*, **204**, 107172, doi.org/10.1016/j.ress.2020.107172

Love, P.E.D., Matthews, J., Zhou, J., Lavender, M., and Morse, T. (2018a). Managing rail infrastructure for a digital future. Future-proofing of asset information. *Transportation Research A: Policy and Practice*, **110**, pp.161-176, doi.org/10.1016/j.tra.2018.02.014

Love, P.E.D., Smith, J. Ackermann, F., and Irani, Z. (2018b). The praxis of stupidity: An explanation to understand the barriers to rework mitigation in construction. *Production Planning and Control* **29**(13), pp.1112-1125, doi.org/10.1080/09537287.2018.1518551

Love, P.E.D., Teo, P., and Morrison, J. (2018c). Unearthing the nature and interplay of quality and safety in construction projects: An empirical study. *Safety Science*, **103**, pp.270-279, doi.org/10.1016/j.ssci.2017.11.026

Love, P.E.D., Matthews, J., and Zhou, J. (2020). Is it too good to be to be true? Unearthing the benefits of disruptive technology. *International Journal of Information Management*, **52**, 102096, doi.org/10.1016/j.ijinfomgt.2020.102096

Love, P.E.D., Matthews, J., Ika, L., and Fang, W. (2021a). A rising tide lifts all boats, ignoring risks can sink them: The peril of rework in large-scale transport projects. *IEEE Engineering Management Review* **49**(2), pp.147-152, doi.org/ 10.1109/EMR.2021.3049158

Love, P.E.D., Ika, L. Matthews, J., and Fang, W. (2021b). Curbing poor-quality in large-scale transport infrastructure projects. *IEEE Transactions on Engineering Management*, doi.org/ 10.1109/TEM.2020.3031890

Low, S.P. and Yeo, H.K.C. (1998). A construction quality costs quantifying system for the building industry. *International Journal of Quality and Reliability Management* **15**(3), pp. 329-349, doi.org/10.1108/02656719810198926

Matthews, J. Love, P.E.D., Mewburn, J. and Stobaus, C. (2018). Building information modelling in 'practice': Views from a collaboration and change management perspective during construction. *Production Planning and Contol* **29**(3), pp.202-219, doi.org/10.1080/09537287.2017.1407005

McShea, S., Oakley, D., and Mazzei, C. (2016). The reason so many analytics efforts fall short. *Harvard Business Review*, 29th August, Available at: https://hbr.org/2016/08/the-reason-so-many-analytics-efforts-fall-short, Accessed 15th December 2020

Michaud, M, Forgues E-C, Carignan V, Forgues D, Ouellet-Plamondon C (2019). A lean approach to optimise BIM information flow using value stream mapping, *ITcon*, **24**, pp. 472-488, https://www.itcon.org/2019/25

Mehta, D. (2019). Highlight negative results to improve science. *Nature*, 4th October, doi.org/10.1038/d41586-019-02960-3

Namekawa, S., and Tezuka, T. (2021). Evolutionary neural architecture search by mutual *information analysis.* Proceedings of *2021 IEEE Congress on Evolutionary Computation (CEC)*, 2021, pp. 966-972, doi: 10.1109/CEC45853.2021.9504845

Newell, A., and Simon, H. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Ngo, J., Hwang, B-G., and Zhang, C. (2020). Factor-based big data and predictive analytics capability assessment tool for the construction industry. *Automation in Construction*, 110, 103042. doi.org/10.1016/j.autcon.2019.103042

Niu, J., and Issa, R.R.A. (2015). Developing taxonomy for the domain of ontology of construction contractual semantics: A case study of the AIA A201 document. Advanced Engineering Informatics, **29**(3), pp.472-482, doi.org/10.1016/j.aei.2015.03.009

Priven V., and Sacks, R. (2015). Effects of the last planner system on social networks among construction trade crews. *ASCE Journal of Construction Engineering and Management* doi.org/10.1061/(ASCE)CO.1943-7862.0000975

Rani, S., and Kumar, M. (2021). Topic modelling and its application in materials science and engineering. *Materials Today: Proceedings*, **45**, Part 6, pp.5591-5596, doi.org/10.1016/j.matpr.2021.02.313

Ranjan, J., and Foropon, C. (2021). Big data analytics in building the competitive intelligence of organisations. *International Journal of Information Management*, **56**, 102231, doi.org/10.1016/j.ijinfomgt.2020.102231

Rayner, S. (2012). Uncomfortable knowledge: the social construction of ignorance in science and environmental policy discourses. *Economy and Society*, **41**(1), pp.107-125, doi.org/ 10.1080/03085147.2011.637335

Robinson-Fayek, A., Dissanayake, M., and Campero, O. (2004). Developing a standard methodology for measuring and classifying construction field rework. *Canadian Journal of Civil Engineering*, **31**(6) pp. 1077-1089, doi.org/10.1139/l04-068

Rogge, D. F., C. Cogliser, H. Alaman, and S. McCormack. (2001). *An Investigation into Field Rework in Industrial Construction*. Construction Industry Institute, Report No. RR.153-11. Austin, TX.

Schütze, H., and Manning, C.D. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* **70**, pp.263–286, doi.org/10.1016/j.jbusres.2016.08.001

Strang, K.D., and Sun, Z. (2020). Hidden big data analytics issues in the healthcare industry. *Health Informatics Journal*, **26**(2), pp. 981-998, doi.org/10.1177/1460458219854603

Tian, D., Li, M., Shi, J., Shen, Y., and Han, S. (2021). On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach. *Advanced Engineering Informatics*, 49, 101355, doi.org/10.1016/j.aei.2021.101355

Tishby, N., Pereira, F.C., and Bialek, W. (1999). *The information bottleneck method*. Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, pp.368-377 Available at: https://www.cs.huji.ac.il/labs/learning/Papers/allerton.pdf, Accessed 21st August 2021

Tolman, F. (1999). Product modelling standards for the building and construction industry: Past, present and future. *Automation in Construction*, **8** (3), pp.227-235, doi.org/10.1016/S0926-5805(98)00073-9

Tsang, E.W.K. (2014). Case studies and generalisation in information systems research: A critical realist perspective. *Journal of Strategic Information Systems*, **23**(2), pp.174-186, doi.org/10.1016/j.jsis.2013.09.002

Turnbull, N. (2006). How should we theorise public policy? Problem solving and problematicity. *Policy and Society*, **25**(2), pp.3-22, doi.org/ 10.1016/S1449-4035(06)70072-8

Webb, C. (2017). Structural failure of precast-concrete span sets back Sydney metro job. Engineering News Record, 24th February, Available at: https://www.enr.com/articles/41504-structural-failure-of-precast-concrete-span-sets-back-sydney-metro-job, Accessed 26th February 2020.

Willis, T.H., and Willis, W.D. (1996) A quality performance management system for industrial construction engineering projects. *International Journal of Quality and Reliability Management* **13**(9), pp. 38-48, https://doi.org/10.1108/02656719610150605

Westrum, R. (2014). The study of information flow: a personal journey. *Safety Science*, **67**, pp. 58-63, 10.1016/j.ssci.2401.01.009

Van Steeg, G., and Galstyan, A. (2014). *Discovering structure in high-dimensional data through correlation explanation.* Proceedings of the 28th Annual Conference on Neural

Information Processing Systems,8-13th December, Montreal, Canada, Available at: https://arxiv.org/abs/1406.1222

Van Steeg, G., and Galstyan, A. (2017) *Maximally informative hierarchical representations of high-dimensional data.* Proceedings of the 18th International Conference on Artificial Intelligence and Statistics. 9th-12th May, San Diego, CA, PMLR **38**, pp.1004-1012, Available at: https://arxiv.org/abs/1410.7404

Yin, R.K. (2018). *Case Study Research and Applications*. 6th Edition Sage Publications, CA

Zhang, F., Fleyeh, H., Wang, X., and Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, **99**, pp. 238-248, 10.1016/j.autcon.2018.12.016

Zhong, B., Pan, X., Love, P.E.D., Sun, J., and Tao, C. (2020a). Hazard analysis: A deep learning and text mining framework for accident prevention. *Advanced Engineering Informatics*, **46**, 101152, doi.org/ doi.org/10.1016/j.aei.2020.101152

Zhong, B., He, W., Huang, Z., Love, P.E.D., Tang, J., and Luo, H. (2020b). A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, **46**, 101196, doi.org/10.1016/j.aei.2020.101195

Zhou, P., and EI-Gohary, N. (2016). Ontology-based multilabel text classification of construction regulatory documents. *ASCE Journal of Computing in Civil Engineering*, **30** (4) 04015058, doi.org/10.1061/(ASCE)CP.1943-5487.0000530

Zhou, P., and EI-Gohary, N. (2017). Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction*, **74**, pp.103-117, 10.1016/j.autcon.2016.09.004