CRANFIELD UNIVERSITY

Mohammed Hajer Alsubaey

A hybrid machine learning and text-mining approach for the automated generation of early warnings in construction project management

School of Aerospace, Transportation and Manufacturing

PhD
Academic Year: 2013 - 2017

Supervisor: Professor Charalampos (Harris) Makatsoris
May, 17

CRANFIELD UNIVERSITY

PhD

Academic Year 2013 - 2017

Mohammed Hajer Alsubaey

A hybrid machine learning and text-mining approach for the automated generation of early warnings in construction project management

Supervisor: Professor Charalampos (Harris) Makatsoris
May, 17

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD

© Cranfield University 2017. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

ABSTRACT

The thesis develops an early warning prediction methodology for project failure prediction by analysing unstructured project documentation. Project management documents contain certain subtle aspects that directly affect or contribute to various Key Performance Indicators (KPIs). Extracting actionable outcomes as early warnings (EWs) from management documents (e.g. minutes and project reports) to prevent or minimise discontinuities such as delays, shortages or amendments is a challenging process. These EWs, if modelled properly, may inform the project planners and managers in advance of any impending risks. At presents, there are no suitable machine learning techniques to benchmark the identification of such EWs in construction management documents.

Extraction of semantically crucial information is a challenging task which is reflected substantially as teams communicate via various project management documents. Realisation of various hidden signals from these documents in without a human interpreter is a challenging task due to the highly ambiguous nature of language used and can in turn be used to provide decision support to optimise a project's goals by pre-emptively warning teams. Following up on the research gap, this work develops a "weak signal" classification methodology from management documents via a two-tier machine learning model.

The first-tier model exploits the capability of a probabilistic Naïve Bayes classifier to extract early warnings from construction management text data. In the first step, a database corpus is prepared via a qualitative analysis of expertly-fed questionnaire responses that indicate relationships between various words and their mappings to EW classes. The second-tier model uses a Hybrid Naïve Bayes classifier which evaluates real-world construction management documents to identify the probabilistic relationship of various words used against certain EW classes and compare them with the KPIs. The work also reports on a supervised K-Nearest-Neighbour (KNN) TF-IDF methodology to cluster and model various "weak signals" based on their impact on the KPIs.

The Hybrid Naïve Bayes classifier was trained on a set of documents labelled based on expertly-guided and indicated keyword categories. The overall accuracy obtained via a 5-fold cross-validation test was 68.5% which improved to 71.5% for a class-reduced (6-class) KNN-analysis. The Weak Signal analysis of the same dataset generated an overall accuracy of 64%. The results were further analysed with Jack-Knife resembling and showed consistent accuracies of 65.15%, 71.42% and 64.1% respectively.

Keywords:

Risk management, Unstructured data, Construction project documents, Text mining, Early warning signal, Data mining, Artificial intelligent, Machine learning Naïve Bayes, K Nearest Neighbour, TF-IDF methodology, Key performance Indicators (KPI).

ACKNOWLEDGEMENTS

I am grateful to the Almighty God for the help and betterment throughout the academic years and during the project's progress. Moreover, my deepest gratitude to my project supervisor, Prof. Harris Makatsoris for his support and help in ensuring that I remained within the scope of this research, continuous support of my PhD study, as well as for his patience, motivation, and immense knowledge. His guidance always helped me during my work as well as during the write up of this thesis. I could not have anticipated having a better advisor and mentor for my PhD study. Also, I would like to express my sincere gratitude to my father Hajer Alsubaey for the moral and financial support during my studies. Moreover, I am also grateful to my brother Majid Hajer Alsubaie for supporting me spiritually throughout the write up phase of this thesis and my life in general.

Last but not the least, I would like to thank my friend Dr Mohammed Alqarni for the stimulating discussions, for the sleepless nights we worked together before any deadlines, and for all the good times we have had in the last five years. Also, special regards to Mousa Smaill for providing the necessary content in a timely manner as this ensured the completion of my project data without any setbacks. In the end, I am grateful to all those who were part of my work and assisted me and I wish them all the best in their lives.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AV	Actual Value
AOB	Any Other Business
CCPM	Critical Chain Project Management
CSV	Comma Separated Value
DICE	Duration, Integrity, Commitment, and Effort.
EWS	Early Warning Signal
EWI	Early Warning Indicators
EIA	Environment Impact Assessment
EVM	Earned Value Management
ER	Earned Rules
ERP	Enterprise Resource Planning
FSSMF	Future Signal Sense Making Framework
ICT	Information and Communication Technologies
KNN	K Nearest Neighbour
KDD	Knowledge Discovery In Database
KPI	Key Performance Indicators
MER	Minimiser of Reconstruction Error
MDP	Markov Decision Process
MOM	Minutes of Meeting
OCR	Optical Character Recognition
PAS	Project Planning Phases
PDF	Portable Document Format
PV	Planned Value
PCA	Principle Component Analysis
RFI	Request For Information
RMP	Risk Management Process
DSDM	Dynamic System Development Method
SDLC	System Development life cycle
SEWS	Strategic Early Warning System
SSIS	SQL server Integration Service
SME	Small and Medium Sized Enterprises
SVM	Support Vector Machine
WBS	Work Breakdown Structure

1 CHAPTER I: Introduction

1.1 Background

The complexity of construction projects makes them prone to a diverse range of failures leading to operational discontinuities such as delays, failed deliverables, staff shortcomings and pending milestones (Elmualim & Gilder, 2014). Project management teams are driven by individuals and organisations in various disciplines and departments ranging from raw material delivery firms to inventory control staff. Each of these organisations and individuals form critical operational units having their own role to play in various stages of a project's successful completion (Walker, 2015). Any discontinuity at any of these stages often have direct or indirect consequences on the sound completion of project's milestones. During the lifetime of a project, all such operational issues are discussed and documents in the form of memos, minutes of meetings, contracts, etc. However, critical information often recorded in these documents is often overlooked and missed by the staff responsible for its record keeping (Martínez-Rojas, et al., 2015). Established project management methodologies can only minimise the risk of failure but cannot guarantee successful completion. However, early prediction of future project trajectory can provide early warnings ahead of time and assist in the prediction of any significant deviations from the original project plans (Ulhaq, et al., 2017). For construction project management analysis of documents to extract meaningful indicators is currently not reported to have been addressed (Fan, et al., 2014).

The research seeks to develop a predictive early warning methodology for project failure prediction by analysing unstructured project documentation such as meeting Minutes. The research adheres to developing a project progress assessment methodology encompassing factors affecting project performance. In addition, it develops a method and algorithmic approach for the analysis of unstructured project documentation and extraction of key actionable information to allow inference of actual project progress rapidly. To validate the capability of this technique, the work also seeks to develop a prototype tool tuned against a series of case studies from the literature. Finally, it concludes by conducting an

empirical study demonstrating the approach. Such a capability requires rapid assessment of project documentation and reports. Moreover, such an assessment requires inferring potential future failures from unstructured information, and analysis of emotions and sentiments in the writing style of these reports.

To address this knowledge gap, this research proposes a novel, hybrid early-warning-sign classification methodology to predict project failure likelihoods. The aim of this research is to improve the extraction of these project early warnings via specifically defined semi-structured data-pairs and information islands. This goal is achieved via a Naïve Bayes classifier to identify critical data-pairs first predict various early warnings' classifications that are modelled over labelled data extracted from an expert-driven, questionnaire-based qualitative research baseline. The methodology also extends to a supervised KNN classifier to further improve the outcome via a TF-IDF classification technique of information extraction.

1.2 Challenges in construction project planning and control

Businesses are built around projects with people and processes as their core resources. A project is hence an activity with definitive start and end dates. To better understand any project's progress, the underlying planning and control, its key challenges must be properly understood before the actual plans are finalised. If not properly catered, these challenges may lead to delays or other substantial difficulties during the ongoing timescale of the project. According to Söderholm (2008), these challenges can broadly be identified into four distinct categories:

- **Lack of collaboration**

  In the current day and age, many projects have teams spanning globally. It is often the case that project board and management teams are not based in the immediate vicinity of the project. Regardless of the case, collaboration and communication is deemed paramount in a successfully running project.

Hence, and lack thereof, within a project plan or timeline must be identified ahead in time (Binder, 2016).

- **Lack of information granularity**

  To keep management phases as singular as possible, only relevant management information is to be shared with various teams (Pedrycz, 2014). Still, each group/team in project must be well-informed in all the details of its own boundaries of accountability.

- **Inadequate management engagement**

  One such, often miss-reported anomaly is that of ineffective financing mechanisms. For instance, a robust supply chain can only be maintained if there is a stable cost flow and payment system in place (Bertone, et al., 2006).

- **Poor planning for success**

  Lack of sustainability is often an overlooked aspect despite the project itself achieving success and growth in the earlier stages (Dagan & Isaac, 2015). Experienced managers tend to identify such shortcomings via their own experience when they go through the continuing project documents. However, pertaining to the massively high number of documents created in even small to medium-level construction projects makes it difficult for human experts to accurately identify hidden anomalies.

1.3 Characteristics of construction project

Figure 1: Characteristics of a construction project lifecycle

Construction projects involving multiple clients and stakeholders are characterised by a wide range of factors and responsibilities (Brady & Davies, 2014). Such projects tend to differ from standard projects due to their multifaceted nature as shown in Figure 1. These characteristics are substantially inter-related, for example, uniqueness of a project is strengthened by its diversity, dynamism, risk perception and intricacy. All these aspects of an international project differentiate it from standard projects and hence must be measured and modelled accurately to gain a better understanding of how that project works (Golini & Landoni, 2013). Yet, it is the project documentation that reports on these aspects during the running lifecycle of the project.

As project complexity plays a critical role in its proneness to develop completion-related issues, it may involve a variety of factors that are to be handled by the project manager (Caldas, et al. 2002). Hence, management skills and earlier experience of the manager as well as the team weighs more for a complex project (Köster, 2009). Many a times, discontinuities develop because of conflicts between various stakeholders involved in the project. As these conflicts often occur due to lack of communication there must be policies in place to provide a better communication and cooperation on geographically separate zones. Any deviations from planned correspondences must be identified on time and recorded (Floricel, et al., 2016). Any deviations from planned correspondences must be identified on time and recorded (Floricel, et al., 2016). In international projects, the impact of such communication, cost or time-related deviations is far worse since various socio-economic, political, and geographical aspects of one country are likely to have an impact on the stability of the project (Williams, 2016). New ventures tend to offer more challenging environments to operate in and are generally more likely to be subjected to new legislations particularly when a multinational project is concerned (Cicmil, et al., 2017). Moreover, in a construction environment, various standards might differ between different countries which may lead to delays in various project milestones as part of the project in a country may conflict with the laws such as data protection, user privacy and security of the other (Qazi, et al., 2016). A product of structural design which might be applicable and usable in one country might not be easily applied to another-one (McCombie & Jefferson, 2016; Wondimu, et al., 2016). Within an international construction project management scope, abrupt changes in a project phases may also have a variable level of impact depending upon the demographics of the project (Akil, et al., 2017).

Project lifecycles are nowadays documented via software packages to provide a better control and management. With the advent and improvement of computing platforms and the internet, it has further become easier to organise project plans, prepare and follow-up meeting agendas and track project continuity via software packages such as MS Project and Office. However, the information

entered in project meetings is generally unstructured and is merely aimed for human perception which is prone to forgetting and overlooking. Processing this information to pre-emptively extract future risks is an open area of research that present many challenges (Chilipirea, et al., 2017)

1.4 Extracting discontinuities from construction documents

Construction project management documents generally comprise of a combination of information aspects including textual and image-based content. A large portion of such text comprise of uneven entries which cannot directly be transformed into structured information due to its complexity (Braglia & Frosolini, 2014). Humans are well-accustomed to representations due to their well-developed cognitive ability to understand documentation cues such as table lines, bullet points and other text formatting signatures. Designing an information processing system to achieve this is a challenging task. This domain has extensively been investigated for the past two decades with the focus mainly on image processing and filtering mechanisms to improve document quality. However, cases where information cannot reliably be extracted from documents, direct extraction of information from image or PDF files is also commonly used (Zhang, 2014). However, due to the quality of such documents, an accurate categorisation based only on keywords is a very challenging information retrieval problem.

Based on the limitations discussed above, there appears to be a lack of effort in the domain of direct extraction of information from complex documents. An automatic document categorisation system for the European patent office has been proposed by Krier and Zacca (2002), which used fully labelled sample files to improve its identification accuracy. However, the technique cannot directly be used in cases where each document may belong to multiple classes. This is the case where each document may contain feature data originating, for instance, two separate labelled training documents with each labelled to a different class. Gomez (2014) presented a Minimiser of Reconstruction Error (mRE) technique used via the Principle Component Analysis (PCA) analyser which presented promising outcome. The domain of text-based document and media

categorisation has been extensively investigated in the application areas of health and wellbeing (de la Rosette, et al., 2012; Jindal, et al., 2015), video categorisation, social networking tweet clustering, web clustering, and analyses of unstructured data to identify hidden document clusters (Yan, et al., 2015; Cotelo, et al., 2016; Estruch, et al., 2006; Larsen, et al., 2008).

1.5 Research aim and underlying questions

The research aims to develop a predictive early warning methodology for the identification of project discontinuities by analysing unstructured project documentation such as project reports or Minutes. The project initially assesses the existing state of knowledge in this domain and then aims to propose a mechanism to address the knowledge gap in the domain.

In this research, a construction document analysis framework is developed according to the subsystems shown in Figure 2. This figure shows that the project performance and early warnings characteristics (structure, culture, and process) can be used to define the risk. A solid risk identification process could identify the impact of early warnings characteristics on project performance. This could be established by identifying the hidden knowledge about relationships between project failures and early warnings characteristics. For example, relationship between delivery process and project performance can reveal how the delivery process in a given project performance is impacting the required performance. With a proper implementation of EWS scoring process, the final early warning score model can be applied for project risk assessment.

Based on the research framework, the following research objectives and questions (research questions are written as RQ) shown in Figure 2 are explored for the research thesis.

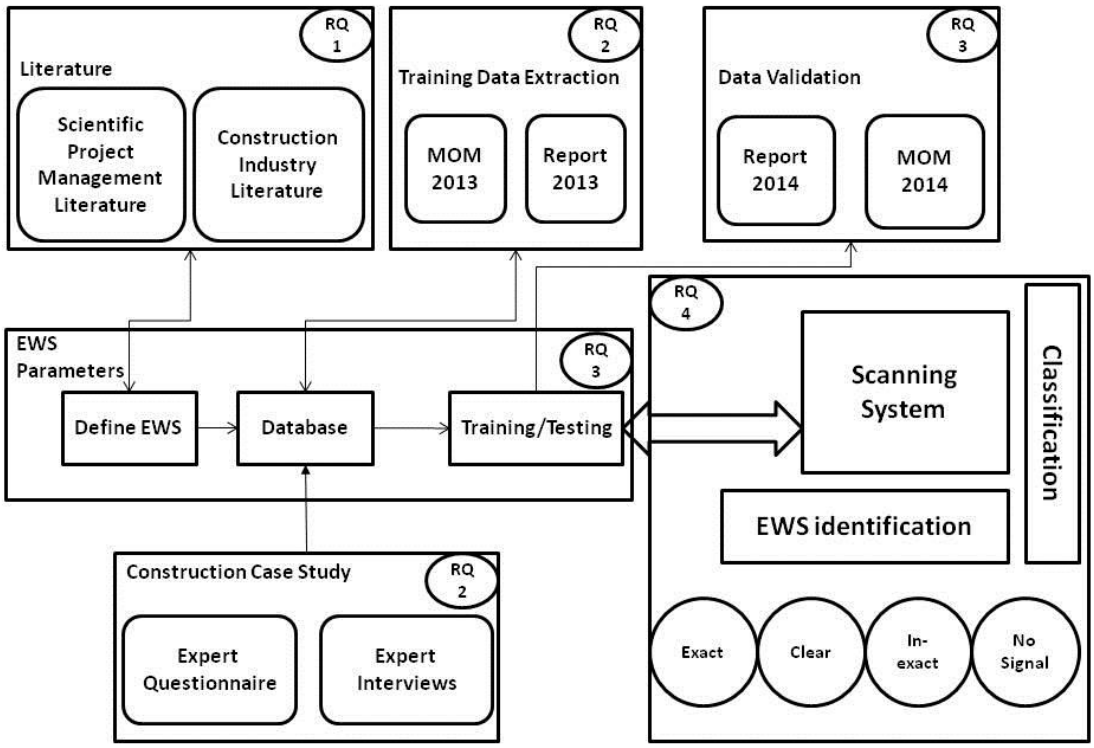


Figure 2: A research framework for early warnings assessment approach

Research Question 1: Could the information contained within management documents be used to identify early warnings from construction projects?

Research Question 2: How can the unstructured information present within construction project documents be extracted to train a document identification model?

Research Question 3: How can the information contained within the data be prepared and organised effectively to recognise various project discontinuity features?

Research Question 4: How effectively can the model be used to extract hidden signals and warnings from a wide range of construction management documents?

Based on the background discussed in this chapter and the research questions given above, the aim of this work is stated below:

"To develop an early warning prediction methodology for project failure identification by analysing unstructured project documentation"

Hence, the objectives of this work can be organised as follows:

Objective 1: To critically analyse the existing state-of-knowledge in extracting actionable information from construction project management documents

Objective 2: To develop a project progress assessment methodology encompassing factors affecting project performance

Objective 3: To develop a text mining methodology for the analysis of unstructured project documentation to extract key actionable information indicating potential most future failures

Objective 4: To develop an analysis tool for such risk identification and empirically evaluate its usefulness against various case studies

To address the abovementioned aims, objectives and research questions, the thesis is organised and documented as follows:

- Chapter 2 Literature Review: This chapter presents a critical literature overview of the existing state-of-the-knowledge in the domain of data/text mining, document and language processing, statistical and machine learning models. The chapter ultimately provides a baseline understanding of the existing knowledge gaps in the domain of construction management information mining and modelling.
- Chapter 3 Data and Text Mining Methodologies: This chapter presents and analysis the domain of data mining and the possibility of utilising the existing state of knowledge to improve the extraction accuracy of actionable information retrieved from management documents.
- Chapter 4 Development of a Naïve Bayes early warning identification classifier: This chapter proposes a novel probabilistic methodology trained

on a pre-labelled bag-of-words technique to facilitate early warnings of project failure.

- Chapter 5 Development of a KNN classifier for document identification: The technique proposed in this chapter is used for the extraction of “weak” project signals of failure from construction management projects
- Chapter 6 Results and Outcomes: The chapter presents an in-depth analysis of outcomes of Chapter 5 and Chapter 6.
- Chapter 7 Conclusion and Future Directions: Based on the findings from Chapter 6, this chapter presents a conclusive outcome of research. It also presents any future directions of this research.

1.6 Summary

This work primarily focuses on exploring the use of machine learning to analyse management files to predict anomalies in project timelines. Hence, the main aim of this work is to record and report on the overall progress of the project. As these documents are cornerstones of a project’s progress, the text contained within bear crucial information about the potential risks and possibilities of deviations from planned activities. A model trained with information extracted from similar case studies is expected to identify similar discontinuities from historical text data ahead of time thereby providing timely information to facilitate preventive and/or remedial actions to be taken.

2 CHAPTER II: A review of project management techniques and principles

In project management lifecycle, milestones and phases are generally marked by certain indicators that provide critical information about how the project will behave in the next few weeks or months. Such indicators are generally understood via performance metrics which are carefully designed to identify early warnings (Williams, et al., 2012). In project management, increasing the probability of achieving the underlying objectives is always considered a major issue. This involves aspects of time management, scope adherence, cost analysis, quality assessment, benefits realisation, team working and project organisation (Atkinson, 1999). To keep a check on project performance, regular feedback over the planning aspect is crucial to shuffle and rearrange resources to resolve the underlying problem and optimisation opportunities. Construction projects generally exhibit signs of early warnings during the entire lifecycle. These signs are evident in the form of various underlying issues that indicate developing problems within the projects. There are certain factors that characterise impact on future developments and can be used to predict any hidden indications of developing problems within the construction projects (Ansoff, 1975). Hence, Early Warning Indicators (EWI) can be employed as triggers that draw the attention of decision makers and stakeholders on developing risks before the warnings materialise into full-blown emergencies that may lead to project delays (Haji-Kazemi & Krane, 2013). Construction firms identify early warnings in their projects by various strategies among which one of the most notable one is the crisis management strategy by Igor Ansoff (Nikander, 2002). Another proposed option is following a mitigation approach which can determine and mitigate problems before any surprises are encountered (Soibelman, et al., 2002). In order to fully understand the scope of a project, there must be a good understanding of the business case and the realisation of the estimated delivery date. Generally, these aspects can only be measured once a detailed project plan along with the underlying work packages that are to be agreed on with the scheduled dates with the stakeholders. In everyday project management activities, meetings are an essential part of the

entire project organisation. Meetings are called to initiate, organise and re-focus project directions in order to achieve the ultimate goal in the most optimal way.

Project activities that may hold critical early warning information can generally be categorised as follows:

- Measure of feedback at reflective meetings (Senaratne & Ruwanpura, 2016)
- A number of work packages failing (Herbst, 2017)
- Approvals or disapprovals taking above average time (Alsubaey, et al., 2016)
- Supplier/production/contracting teams' performance (Kerzner, 2017)
- Comparison of rate of spend to original/agreed rate of expense (Asadi, et al., 2015)
- Staff turnover, skills and performance both at client and supplier sides (Dansoh, et al., 2017)
- Documented conflicts (Walker, 2015)
- Stoppages (Alsubaey, et al., 2015)
- Measure of claim submission (Herbst, 2017)

This chapter focuses on investigating the challenges in extracting actionable information from early warnings embedded in construction project management documents. At this stage, the chapter describes various knowledge gaps in the respective field and how the gaps that are directly covered by the research questions presented in the earlier chapter are to be addressed as a machine learning problem which is then discussed in Chapter 3. In doing so, this chapter discusses issues of managing project plans, and the underlying scopes and deliverables based on various early warnings. Hence, the chapter addresses on issues identification within the lifecycle of projects, challenges, review of best practices, and models investigated as part of the existing state of the art in project management.

2.1 Core reporting stages in project management

A project management lifecycle contains several steps that are capable of reporting embedded anomalies in a construction management lifecycle as shown in Figure 3 as adapted from (Lewis, 2006):

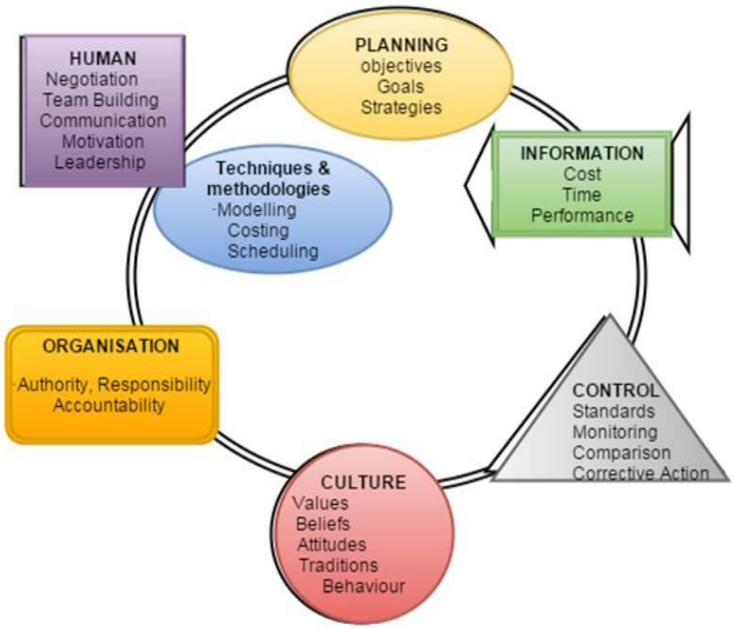


Figure 3: Core project management stages

According to Lewis (2006), project planning generally include how the objectives of the project are going to be achieved, how the path to any deliverables is to be set and the strategy to achieve both. However, as constructions projects are generally very complex and large due to the number of stakeholders and individuals involved. This leads to an organisational challenge for individuals managing overlapping activities and group responsibilities. Each project must be organised in a way that it grants authority and responsibility to relevant individuals which are then held responsible for their actions (Ahern, et al., 2014). However, not all individuals have the skills and expertise to document and follow-up on project updates. These updates may include critical aspects such as delay in inventory replenishment or staff support requests. The accountability of such responsibilities is monitored in the control phase which can broadly be categorised as the project management control aspect of the plan. Any such controls are covered via policies and

standards in-place. In the case of any unexpected anomalies, proper guidelines to take correction actions should be undertaken (Cooney, 2016). Nonetheless, despite the presence of these control guidelines, timely reporting by relevant individuals is often missed merely due to forgetfulness or inexperienced nature of the staff involved. There is always a particular “culture” of ethics and habits at work in every construction project. It plays a vital role in responsibility and accountability as individuals’ beliefs and values contribute significantly on her/her performance. How these individuals from a diverse set of backgrounds and cultures are integrated into a tightly knit team is what defined by Hovde (2014) to include the aspects of leadership, team building, negotiation, communication, and motivation. Hence, the human factors are possibly the most critical aspect as a project can only delivery if run and governed by individuals experienced in project deliveries. Therefore, every milestone and deliverable in a project is measured as function of cost, time integrated against performance which in turn are affected by weaknesses ingrained due to faults and deviations in the various human factors involved in the project. The information aspect entails and covers a detailed registration of finance and time-related aspects and their role in the continuation of the project. In addition, information also represents the details of non-material requirements such as specifications documents, shop drawings, and expediting schedules which are mainly used to pursue and push ongoing project disciplines (Cassidy, 2016). There are many technical frameworks in use that address various such aspects and challenges including modelling, costing and scheduling. Nonetheless, follow-up of any subtle and/or developing discontinuities in a project by human and technical factors governed by a brute force approach is a big challenge indeed. Such factors are often monitored and identified by means of triggers that operate on various performance measures of various project phases.

2.2 Designing and control in management projects

Design and control of projects primarily relies on the way different resources are planned and integrated with each other for example assigning human resources with specific skills to specific responsibilities while optimising their utilisation

during the project's running tenure (Matta, et al., 2014). However, not all the assignments can be allocated in parallel and there is always a level of interdependency between various tasks. It is the way various tasks and allocated and optimised that various uncertainties in projects are catered for. This phase forms a critical stage in a project's plan as unexpected issues do occur with little to no control over the time they happen. A few examples in this case are employee sickness, delivery delays, natural emergencies, inadequate timelines by contractors and economic factors. Various models of pre-emptive assessment and control or design issue prediction during a project's lifecycle have been investigated. Bai, et al. (2016) presented a queuing model which ensured maximum staff usage by contractors to avoid human resource wastage while still preventing skills/manpower shortage due to over-allocation. Similar work by (Bosch-Sijtsema & Henriksson, 2014) proposed a methodology to manage knowledge distributed in various aspects of a project (e.g. plans, design document, guidelines, letter-based communications and other electronic means) via various interactions such as meetings, emails and direct communications. However, tracking similar information in large and complex projects present another challenge of manageability of data. Organisation increasingly rely on digital and electronic means to handle high volumes of project management related data. Whyte et al (2016) presented a configuration management technique to handle vast amount of data generated and saved in large projects. The work investigated the limits of existing digital means. This aspect of increasing complexity also poses a problem for any system that aims to process a large number of relevant management documents in order to extract structured data. Similarly, focus has also been on maintaining a balance between risk and project control while identifying problems in the presence of uncertainty and equivocality (Sjödin, et al., 2016; Patanakul, 2014). A closer look into the focus of these paper indicate a major challenge in the design and implementation of a knowledge discovery/extraction system as the volume of unstructured data taken from larger projects' documents would be phenomenal. Hence, the design must incorporate a suitable technique to eliminate redundant and superfluous information.

2.3 Assessing project discontinuities via key performance indicators

Performance indicators are widely used to measure project lifecycle characteristics in any project (Mir & Pinnington, 2014). However, identifying the presence or lack of these indicators in a project is a cumbersome task. This process can be effectively addressed by analysing information and indicators hidden within project documents. Yet, understanding and identifying such indicators is an experience-driven task. Construction project performance in any of the management projects is measured under a range of criteria. However, according to Ika (2009), success in projects is still an abstract concept where various performance measures are reported to define how well a project has completed. The focus of these so-called key performance indicators (KPIs), in the construction industry has been to provide continuous improvement in the overall performance of the project. These indicators assist project managers to direct and control the way their project is being managed. Several metrics have been proposed in the literature in various areas of construction project management including cost, time, scope, quality and safety (Yun, et al., 2016; Love, et al., 2016; Demirkesen & Ozorhon, 2017). The indicators are broadly categorised into subjective and objective measures as reported by Cox et al (2003). The way these indicators affect a project performance varies greatly depending upon the nature of each project as well as the type of resources being allocated to a project.

2.3.1 Objective measures of assessing project performance

Objective measures can be identified directly and are often quantifiable. One example of an objective measure is that of construction time and speed. Timely completion of project deliverables are measurable via well-defined ontologies or evaluation methodologies (Diamantini, et al., 2016). According to Zavadskas et al (2014), variation in time and delays of planned activities or repeated extensions generally provide evidence of poor planning. Timely completion in project management is a key to successful project handover which can be measure by the frequency of time variation of various project activities.

Moreover, unit cost performance of any activity within a project is related to parameters such as energy efficiency, maintenance cycles, and renovation activities (Christen, et al., 2016). Also, occurrence of avoidable accidents provides a direct analysis of accident-proneness of a project's running environment (Moura, et al., 2016). As these aspects are quantifiable, learning from accidents, documenting and then minimising reoccurrence is the best way to avoid repeat incidents where a lower frequency can be used to report a better project performance in this domain. Similarly, the so-called Environment Impact Assessment (EIA) scores measure the extent to which a project's completion impacts on the surrounding and overall environment is measured. Despite the usefulness of these measures, each variable often depends upon many other variables which in the case of a medium-level project turns into a very complex metric. Hence, the scores are generally measured via statistical and AI techniques to advice on various aspects. Some well-known techniques in this domain include safety risk analysis via Bayesian classifiers (Zhang, et al., 2014), sustainable construction (Tabassi, et al., 2016) and environmental risk analysis (Olaru, et al., 2014).

2.3.2 Subjective measures to monitor a project lifecycle

In construction projects, KPIs are also termed as subjective measures that are used to monitor costs, track progress, measure client satisfaction, and understand project strengths and weaknesses. These measures originate from KPIs are non-functional aspects in a project lifecycle including quality, functionality, user, design and customer team satisfaction (Nicolaescu, et al., 2017). Unlike objective measures, it is more challenging to set criteria to measure subjective aspects of a project. The aspects are hence known as soft measures. However, these measures have been reported in the domain of construction management in areas such as queue performance assessment (Ehsanifar, et al., 2017), quality benchmarking measurement (Oppong, et al., 2017), construction waste optimisation (Wu, et al., 2017) and deliverable time optimisation (Zidane, et al., 2016). According to Oppong, et al. (2017), due to being subjective, factors such as stakeholder satisfaction, interaction of

objectives, and performance indicators are fuzzy in nature and hence, it is challenging to empirically test their performance. Similarly, the work detailed by Zidane, et al. (2016) and Wu, et al. (2016) focus on optimisation of certain measures. Despite work being done in the assessment of various project performance measures, the ability of soft-computing paradigm to identify and pre-emptively address any such risks have not so far been addressed.

These objective and subjective measures are also reported as measurement metrics to address aspects such as construction safety monitoring (Awolusi, et al., 2018), automation performance assessment (Kunz, 2015), sustainability and resilience in built environment (Marjaba & Chidiac, 2016), site layout planning optimisation (Xu & Li, 2012), staff performance analysis (Groen, et al., 2017), and overall community satisfaction assessment (Musa, et al., 2017). Despite the ability of these objective and subjective measurement metrics, they are more focussed on providing assessment and analysis metrics. For instance, a sudden staff shortage cannot generally be identified by an objective assessment of a resulting time-delay or a subjective assessment such as the “satisfaction of project clients” (Dao, et al., 2016). However, if these aspects are monitored overtime via recorded information such as real-time management documents, an on-time identification of anomalies would be predicted more efficiently as per the research objective of this thesis.

To further bolster this argument, Table 1 presents some information about various performance indicators taken from existing research. The table signifies the importance of planning, staff experience, communication, risk assessment/management and waste minimisation as some important performance factors in projects. However, measuring these aspects directly from project data is not a trivial task. For instance, “staff experience” can be measured via many variable factors and it is often not easy to understand what is effectively regarded as a skilled or keen staff member neither is it easy to provide an accurate assessment on what is deemed as staff shortage for a certain task. Alternatively, a model trained with expert feedback is envisaged to identify such risks from documents as and when they are stated by staff

members. Nonetheless, a better assessment of performance indicators makes it easier to tackle and assess various everyday challenges encountered in project planning and control (Brady & Davies, 2014).

Table 1 : Summary of various performance indicators reported in the literature on global level

Author/Year	**Country**	**Performance indicators**		
(Ali, et al., 2013; AlRababah, 2017)	Saudi Arabia	Planning duration	Staff experience	Safety
		Communication	Claims	
(Kumaraswamy, et al., 2017; Czarnigowska & Sobotka, 2013; Ajayi, et al., 2016; Olawale & Sun, 2015; Dagan & Isaac, 2015)	UK	Construction duration	Waste minimisation	Experience & knowledge transfer
		Project control	Relationship management	Safety
		Risk management	-	-
(Jiang & Wong, 2016; Lu, et al., 2015; Shaikh & Darade, 2017)	China	Corporate social responsibility	Lean construction	Time

Despite the domain of KPI-based assessment being extensively investigated, there is little evidence of using knowledge-base and predictive models to preemptively report any project anomalies. However, this knowledge transfer and its associated principles can be incorporated in an advanced model that is capable of then processing and analyzing new projects as a machine learning technique.

2.4 Causes of project failures and causation factors

A large number of reasons are cited to contribute to projects' failure (or success) (Pinto, 2014; Williams, et al., 2012; Asadi, et al., 2015). As planning is deemed central to any project, it is crucial to define what constitutes the success or failure of a project. Moreover, it is also important to breakdown larger tasks into sub-activities that are easier to handle/manage (Hirschman, 2014). Hence, if there is a lack of appropriate planning and/or inadequate processes in place, the project is likely to have a poor performance monitoring infrastructure (Adam, et al., 2017). For instance, if the deliverable 'X' has no concrete completion plan, there would not be an appropriate method in place to assess the delays present within that task.

Similar, tracking project progress also relies on appropriate documentation of the underlying processes. This is more of a process driven by the manager and it must include tracking of milestones to see if expected sub-completions are met properly (Alsubaey, et al., 2016). Documentation and tracking assists project managers into deciding at earlier stages on how to move/shuffle resources to address shortage of resources (e.g. skills, number of staff and materials) to address any delays from lack of these variables (Al Qady & Kandil, 2010).

Poor leadership also attributes significantly to projects' success or failures. If there is a lack of leadership on the board, executive or managerial level, it must be addressed as soon as possible (Alsubaey, et al., 2015). Moreover, other aspects have also been cited including staff training, cost estimation accuracy, lack of communication, and task prioritisation are reported to contribute substantially to a project outcome.

Most of issues related to the abovementioned aspects, are generally discussed in project meetings which are presumed to be held on time and under strict schedules (Schaufelberger & Holm, 2017). Hence, the paperwork prepared as part of these meetings tend to hold crucial information on the performance of the project (Caldas, et al. 2002). However, the way certain signals that indicate any deviations from planned project activities require succinct performance

assessment criteria. Several project failure factors have been analysed and reported in the literature.

2.4.1 Failure assessment criteria from project resources

Projects are generally known to fail due to problems in their planning, estimation or task implementation due to human factors. Per Cui and Loch (2014), three causes are known to contribute significantly to project failure:

2.4.1.1 Planning and estimation factors

Planning and estimation at the beginning of a project holds critical importance as any mistakes or overlooking at this stage is likely to create a ripple effect at the latter stages especially if iterative adjustments or revisions to these estimates are not made.

2.4.1.2 Implementation factors

If the required implementation methodology differs from what was more suitable for a project, such as lack of identification of project scope change, incorrect project methodology, and inability to incorporate major changes in project requirements, testing and/or inspections. An example of incorrect use of methodology would be that of using a waterfall model for clients who are expecting a tested product earlier in the lifecycle. It is a well-known fact that projects under the waterfall paradigm do not go through an iterative SDLC. For iterative builds, agile principles such as Scrum or more suitable (Cserhati & Szabo, 2014).

2.4.1.3 Human factors

Human factors may include in sufficient staff training, such as manager not appropriately trained to follow and organise a project. Moreover, poor communication is also attributed to human-originated project mishaps.

In construction project management, a delay is regarded as a condition where various stakeholders, particularly contractors, consultants and clients contribute to the delay in the project's set goals and objectives within the time and resource allocated on the original and agreed contract terms (Adam, et al.,

2017). Any such kind of delays generally lead to loss of productivity and results in work disruptions. In more serious cases, it often leads to parties blaming the responsibilities on each other which may lead to contract terminations or additional cost overheads to be met by those responsible. Nonetheless, in order to keep projects completing with minimum delays, it is important to understand the factors beforehand (Mubarak, 2015).

2.4.2 Factors linking project milestones and scopes

According to Kappelman et al (2006), early warnings within projects can originate from two separate aspects of people-based and process-related risks. The work documented a number of underlying factors that included weak project management, team commitment and resources. Most notably, the research indicated the role of lack of documented requirements in project failures indicating a major aspect of the proposed research of digging-out warning patterns from within project management documents. Weak documentation was also indicated to lead to lack of purchases and the subsequent lack of materials in the inventory. This issue generally resulted in delays.

2.4.3 Impact of staff training and skills

According to Williams et al (2012), staff knowledge, and training also had a direct impact on project delivery. The research focussed more on people's knowledge and communication skills and their role in project quality. Again, the communication part in turn adapted from various documents such as Minutes, project reports and memos. This indicates the importance of project progress information present within these documents.

2.4.4 Role of change requirements and lack of top management support during project kick-offs

According to Boehm (1991) in Kappelman et al (2006)'s work, frequent changes within the project during its various ongoing major phases have a crucial role in project delays and failures. He further pointed-out that lack of documentation to be the second-most prevalent Early Warning Sign with a mean score of 6.58 out

of 7 in terms of its role in project failures. In the same work, Schmidt et al (2001) reported lack of top management support to have a direct impact on project failures.

Despite the presence of these factors, any shortcoming can still be identified earlier in time if an appropriate identification model is in place. Yet, there exists a cause-and-effect relationship between procedures used and corresponding response from a project. Such relationships can potentially be identified by means of specific early warning signals originating at various stages of project lifecycles.

2.5 Extracting early warning signals from project management text

It is generally easier to identify clearly defined statements in project statements such as “project failure” or “ceased cash flow” compared to hidden/indirect signals such as missing milestone dates or staff absence. For instance, work done by Ansoff (1975) gives an example of the petroleum crisis of 1970s where the large oil firms were caught by surprise despite the advance forecast reports present on the managers’ desks that were largely overlooked. The work thus forms a baseline of data-mining research in project management where, timely analysis of early warnings contained within planning and design documents is critical contributed major research in the area of early warning where his work reported and analysed a wide range of problems within project management that could pre-emptively indicate unfolding issues (Nikander, 2002). The problems or issues investigated in his research included finance, project control, communication, performance, scheduling and delays. The work focussed on Igor Ansoff’s theory by analysing dependencies between early warnings, causes of problems, problems and responses. According to the research such information hidden in lines of contracts can be assigned a numerical weight based upon their frequency of occurrence (Ansoff, 1980). Similarly, Pang (2006) utilised SVM as a structural risk minimisation methodology and Rough Set theory to reduce noise to achieve 87.5% accuracy in the improvement of a binary early warning identification system. The basis formed by research

presented broadly identifies a number of research areas focussing on data mining techniques to gather information from project documents based on factors such as:

- Project performance in the presence of negligence.
- Delay causing factors.
- Training, staff skills and its impact on deliverables.
- Responsibility under set project plans.
- Staff resource backup in case of emergencies.
- Impact of supply chain disturbance.
- Inventory prediction and control.

However, in data mining, the majority of research either focuses on assessing the state of the project or advice on certain “adjustments” which can be made to minimise the identified early warnings (Alsubaey, et al., 2016). Per Kerzner (2013), there appears to be a directly proportional relationship between the probability of loss and the level of impact of project outcomes a higher combination of which generates a higher risk as shown in Figure 4.

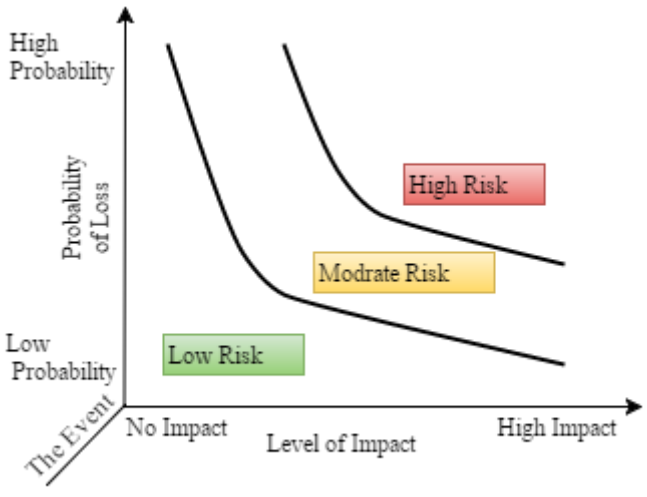


Figure 4: A proportionality diagram showing an increasing risk pattern against the probability of loss leading from project management failures

2.6 Measurement principles of project early warnings and indicator signals

The relationship between project delays and risks to various losses in project lifecycles does provide evidence in the form of various early warnings and indications as elaborated by Ansoff (1975) which was further extended by Klakegg (2012) for early warning in complex projects. Due to multifaceted nature of project in the current day and age, a standard project may demonstrate a wide range of early warnings depending upon the nature of the project including finance, staff, resource, material, and supply chain related issues. The challenge is to present a quantifiable measure capable of robustly integrating and modelling all these challenges in a bid to identify similar warnings from unseen documents.

2.6.1 Ansoff's theory of early warning systems in project management

The measure of risk in early warnings was originally extended by Ansoff where two unique approaches can be adopted to identify "weaknesses" in projects' lifecycles (Ansoff, 1975). These "weaknesses" are primarily based upon human logic based on experience. However, with the increasing complexity of the project and the set of rules government its quality, robustness and reliability, the identification becomes difficult. Statistically, it is possible to identify patterns or early warnings from complex datasets comprising of text corpuses such as Minutes of Meeting (MOM), presentations, emails, general correspondences, interviews and brainstorming. Based on the level of understanding and experience of engineers, managers and other project parties, it is statistically possible to train models that think and act like experienced human beings.

Text documents are the primary exchanges of information adopted in construction projects. The key elements for exchange include contracts, field reports and orders (Caldas, et al., 2005). Therefore, based on the structure of these documents, management becomes a challenge. Using a model based information extraction system, correlating the data models in the text document also becomes a daunting challenge. On the other hand, a manual approach of

establishing connections is impractical due to the high amount of documents stored. The existing systems do not provide room for the required integration. Therefore, an increased need for intelligent, search optimized approach emerges (Caldas, et al. 2002). Despite the availability of such systems, certain limitations are a hindrance to their performance. For example, in cases where words have multiple meanings, and where relevant documents do not contain the user-defined search terms, getting an exact match is a challenge (Al Qady & Kandil, 2010)(Al Qady& Kandil, 2010). Text mining becomes vital as it is used to denote all the tasks that involve analysis of large quantities of text documents and tries to extract possibly useful information (Caldas, et al. 2002). The results of the process denote a collection of documents stored in inter organizational systems. They can be used to improve information management and also to generate knowledge about the subjects contained in these documents (Soibelman, et al., 2002).

Text and data mining is a recent field of research which provides a soft-computing system with the capability to learn and identify patterns. This domain is extensively used in email spam detection, text-to-speech synthesis, etc (Chamatkar & Butey, 2015). On the basis of capabilities of these algorithms, this research integrated a statistical Naïve Bayes solution to mining early warnings from project management documents, particularly the Minutes of Meeting. The work will use construction projects' management as a case study to investigate the capability of Naïve Bayes to identify various early warning classes.

2.6.2 Analysing project lifecycles and symptoms for early warnings

Per Lientz (1995), project management issues and any symptoms of deviations from planned activities are embedded in the associated project documents. Ansoff's theory of weak signals state the very basis of these symptoms which can be identified within documents as and when reported by various members of the team. The awareness of these issues generally gets into the knowledge of various team resources which is then reported by means of management documents. For instance, delay of delivery of certain items are often lodged as

complaints or issues that are eventually discussed by the immediate line managers as they report on the routine meetings. These reports are recorded as special issues or under any other business (AOB) categories and hence indicate the actual harmful impact of any deviations way before they are in-effect detrimental to the overall progress of the project. Apart from the Minutes document, such indicators are not easy to track by average human beings due to their random and sporadically occurring nature. Per King (1987), various stages of such issues form part of conversation, contention and regulation. Hence, an organised track of all the project-related activities would increase the possibility of assessing problems ahead in time. Assessment various project stages to extract various roles responsible for failures and early warnings was investigated in detail by Williams et al (2012).

2.7 Extracting signals during project lifecycles

The so-called term “weak signal” was originally introduced by Ansoff (1975) which defined it as somewhat imprecise but early indications of future events of some importance. In project management, any such early indications are likely to provide valuable insights into any deviations from planned project goals (Lines, et al., 2015). The concept of weak signals is further explained by Mendonca et al (2012) to draw from ideas of logic and imagination, or, in the words of Smith et al (2011) as the changes that may not seem important at the existing stage but have a potential to turn into significant changes such as development anomalies, business threats, or technical deviations. These “weak signals” represent the initial signs of paradigm shifts, or future trends that a project may follow if not handled earlier-on.

Weak signals originated from the strategic planning and management context of businesses and companies with their utilisation remaining important in the strategic management of businesses for a long time. These signals are also often used to identify socio-political and societal changes (Holopainen & Toivonen, 2012). Hence, their association with linguistic and textual indicators can be safely assumed as most social, political and societal material originate from major information resources such as newspapers, and various other kinds

of electronic media. Further work in this domain by Kajava et al (2005) stated that due to the ability of these signals to define change, it can easily of used to shuffle the “definition of trends” with the “identification of threats”. The technical has also been used by Brynielsson et al (2013) to identify lone-wolf terrorists based on their activities. Any activities by such individuals are not concrete but they do leave subtle “traces” which are synonymous to “weak signals” that can be integrated and processed to predict future malicious acts.

2.7.1 Identification of weak signals from environmental factors

Forecasting systems are generally based on extrapolation of data based on models trained on past experiences (Ansoff, 1975). For instance, in any real-world phenomenon, a strategic discontinuity from a regular course of operation represents a threat or opportunity for the relevant organisation. However, understanding “weak signals” is a complex cognitive process that require implicit processing, scanning, interpreting and then establishing the overall situation. These sub-stages from Ansoff’s work were further extended by Ilmola & Kuusi, (2006) to address each as a unique mental model. The work was further extended by Kuosa (2010) who presented the Future Signal Sense Making Framework (FSSF). This framework presents the possibilities to extract and model critical project/system information to identify any hidden patterns and environmental behaviours.

Possibly the most commonly explored alternative to such framework has long been the utilisation of data processing and mining techniques that enable extraction, handling and processing of large amount of information from a wide range of mediums including the internet, scanned textual documents and conventional word-processed files (Decker, et al., 2005; Uskali, 2005; Tabatabaei, 2011). Extending the idea of internet-based environmental scanning for “weak signal” extraction, Thorleuchter et al (2014) developed a methodology capable of clustering signals at various points in time thereby enabling the identification of the most vulnerable stages of a project. The approach makes it possible to track and track such signals and provide better capability for strategic planning.

2.7.2 Existing models and procedures for weak signal identification

Humans tend to identify “weak signals” via their cognitive ability to process various real-world indicators, their nature and impact on future outcomes as well as based on their personal life experience. Weick & Sutcliffe (2001) state that organisations with high reliability predictive systems differentiate themselves from other based on their ability identify such weak signals and then take actions to pre-emptively mitigate any potential future effects. Such organisations are reported by Weick & Sutcliffe (2001) to iteratively update their assessment models to improve and optimise the identification of impending threats via these weak signals. Most of these organisations create simulations and training scenarios that help them better understand the situations that are custom-made to their own corporate situations. This knowledge is combined with best practice cultures, vigilance and responsiveness to any incidents.

The theory of “weak signals” also forms part of the Homeland Security Curriculum and is taught in several universities and colleges in the US (Bellavita & Gordon, 2006). The domain has also been taught in the long-range transport industry to truck drivers to be better aware of their immediate operating environments for dangerous events or conditions (Huang, et al., 2014). With regards to human training case of the Homeland Security experiments were not very positive predominantly because human beings tend to suffer from memory retention issues that inhibit their ability to remember specific and relevant details to model and identify threats. The theory has also been investigated extensively with several tools and methods proposed to understand, and identify their meaning (Gilad, 2003; Mendonça, et al., 2004; Mendonça, et al., 2009; Wiltshire, 2006). A well-known example of negative consequences of ignoring of weak signals has been reported in the case of Asbestos (Gee & Greenberg, 1896–2000). The most challenging aspect in the case of weak signal has been reported to be able to identify the correct scenario for which the “weak signal” act as a warning where different categorisations have been focussed on by (Mendonça, et al., 2012; Kaivo-oja, 2012; Schoemaker, et al., 2013). The work by Mendonça et al 2012 adapts a reverse approach where hidden and/or impending anomalies within project plans are identified because of the

presence of these so-called weak signals. This approach presents a potential and systematic way of sensing and recognising weak signals via organised search of project plan documents. Several communication models and theories have been presented in the literature that facilitate the understanding of these weak signals from complex project resources.

2.8 Early-warning concept in the construction project management context

Even though, in construction project managements, the majority of principles of generic project management still apply, the domain is substantially unique in its own sense. For instance, core project management aspects significantly depend upon the way human resources, materials and skills are intertwined with the core management principles discussed above including integration of project stages and assigned resources, scope management and its compliance with the project deliverables, human resource and its tasking to various project aspects, communication of tasks and responsibilities, the underlying risk of these aspect and ultimately the procurement of all the resources to deliver the project without delays. A substantial number of studies in the domain of identification of aspects directly relevant to the failure of constructions projects have been reported in the literature.

Extending Ansoff's work resulted in the setting up of individually assigned early warning and risk indication from project managers (Ansoff, 1975). That is, based upon management-driven feedback, systems can be tuned to detect developing errors in a project. According to the original work on the so-called early warnings, the theory of weak signals has frequently been used in communications, military science, financial forecasting, and lifecycle management (Williams, et al., 2012). These weak signals can be used to predict early symptoms thereby saving overheads that normally occur when problems happen in project management lifecycle.

2.8.1 Categories of early warnings in the construction domain

Construction projects are generally reported to have three categories of crisis with their project management aspect. The majority of these crises are not addressed until they are encountered (Williams, et al., 2012). Sudden crisis appear without any warnings. An example of this could be staff emergency sick leaves which may have a profound impact on the delivery of the project if a proper replacement is not available. Periodic crises occur in cycles where an example could be the repetitive delay of materials to the construction site by a particular contractor. This repetitive behaviour may have a snowball effect on the performance of other departments how use the material to complete their tasks (Al Qady & Kandil, 2014). The third and less likely type of crises are those that, despite having a low probability of occurrence, if occur, may result in a serious impact on the projects delivery date. For instance, an uncontrolled climate such as high winds which may have a direct impact on the cranes' stability as they operate on the construction site. Incidents have been reported where, due to unexpected weather, even loss of life has been reported in addition to project delays as a result of construction equipment failure and hence loss of planned task mobility (Lee, et al., 2016). Such incidents require highly skilled project management teams to recognise the EWS and respond accordingly (Mintzberg, 1998; Mintzberg, 1987). However, these studies mainly focus on governance frameworks instead of developing a methodology to identify the early warnings. The focus of these studies has primarily been on assessing projects and hence leading to identifying the EWS. Three studies within the UK discovered the relationship indicating that project assessments embedded within the governance structure to effectively identify EWS.

2.8.2 Relationship of early warning types in various projects

Generally, project performance is used as the main indicator of identifying any early warnings of a projects impending failure. However, performance lapse generally occur quite late in a project's lifecycle and it is generally too late to act upon any "early warnings" at these stages as it has already been too late to repair any damage. For instance, performance lags in Gantt charts or project

plans are identified by missing deliverable deadlines. Such slippages occur due to factors that have already matured by now. Hence, the core rationale of EWS is to be able to identify them ahead in time before any failures actually occur. Hence, according to Kappelman et al (2006), the idea of EWS is to identify the critical areas and aspects that lead to these early warnings and consequently failures in the long run. Nonetheless, it is the people and process-related aspects that have the worst impact on project failures. For instance, behaviour-related performance measures used to evaluate a project's performance are the deciding variables into a project's success or failure (Syamil, et al., 2004). For instance, employee's average reporting time every day is one such behaviour that indicates the overall running performance of the project. Moreover, collaborative processes such as routine project meetings and discussions are also reported to have crucial early warning embedded within (Hoegl, et al., 2004). Also, human-related early warnings should also not be overlooked. For instance, staff annoyance or unease on the migration of the company inventory control system to the cloud may be an early warning that the project may face a negative impact from individuals not willing to adapt (Dulewicz, 2000). In addition to such fears, the so-called "gut-feelings" shown by the staff and raised in project meetings may represent any developing threats to a project's failure as reported by (Nikander, 2001; Nikander, 2002).

Despite the presence of these early warning signs, the identification of such during a running project still remains a challenge. A possibility of learning from previous projects has been reported by Williams et al (2012) but has been reported to be rarely effective. The reason behind this could be due to the diverse range of keywords and other indicators that lead to such early warnings. However, it is advised that learning from best practices that are common to the majority of projects is more likely to result in better indicators. Based on these facts, it can be assumed that project artefacts holding crucial management and planning information can be used to identify early warnings.

2.8.3 Existing case studies in early warning extraction systems

Williams et al (2012) have reported 8 cases in a diverse range of areas where early warning signs were successfully identified. Out of these cases, half were in construction, civil/development and industrial development areas with complexity ranging from medium to high. Based on these case studies, the early warnings can be found at three stages of a project at early setup, early stages or during execution as shown in Table 2.

Table 2: Distribution of EWS at various stages of a project (Williams, et al., 2012)

	EWS report at project setup	**EWS during early stages**	**EWS during project execution**
Assessment	Unclear justification of undertaking the project	Absence of a good business case	Lack of documentation
	Poor project description and specification based on Program of Work	Lack of harmony and collaboration between various stakeholders	Lack of expedited tracking and scheduling
	An unsound business plan	Lack of project responsibility	Lack of project responsibility assigned to appropriate individuals
	Poor requirement specification	Lack of communication – letters and memos indicating actionable issues	Lack of commitments to making decisions resulting in frozen actions and unmet deadlines

		Over-reliance on contractors leading to delivery and resource management problems	Inexperience workforce and/or contractors

The signs stated in Table 2 can then further be categorised via relevant keywords that can further be sub-divided into two groups of “thorough assessment” and “based on gut-feeling” which is further elaborated in Table 3.

Table 3: Important signs taken from case studies as presented by Williams et al (2012).

Keywords representing direct EWS indications	**Keywords representing indirect human behaviour**
Information/data missing in numbers and/or quantities	Hesitation between stakeholders and actors
Lack of assessment and/or appropriate documentation	Unfriendly working environment
Delays in plans and report completions or non-clarity of documents	Poor needs assessment
Slipping milestones and/or missing activity definitions	Inconsistent agenda documents
Absence of a pragmatic governance framework	Lack of trust, questions and inconclusive and/or uncertain decisions

2.8.4 EW system exploitation as a project risk identification system

The thesis of the study is that we often fail to have pre-warning of project failure clearly the method of project assessment are often not successful in picking up early warning signals. Several such EW factors are reported in the literature Table 4. According to Williams et al (2012) three areas are studied.

Table 4: Selection of various classes based on secondary research

Early warnings	References
Lack of documented requirements	(Kappelman, et al., 2006)
Lack of changing control process being implemented	(Schmidt, 2001)
Lack of effective schedule planning and/or management.	(Kappelman, et al., 2006)
Lack of key stakeholder support	
Lack of understanding of new project	(Nikander, 2001)
Lack of trust	
Lack of project team commitment	
Lack of contact with client	
Lack of resources	
Lack of the efficiency of work initiation	
Lack of the initial information	
Lack of processing decision-making	
Lack of processing additional information	
Lack of making purchases	
Lack of valid and correct revisions	
Lack of top management support or	(Schmidt, 2001)

commitment to the project	
Lack of stable milestone deliverables and due dates	
Lack of keen commitment to the project scope and schedule	
Lack of project team required knowledge/skills	(Barki, et al., 2011)
Lack of planning and estimation documentation	(Jones, 2004)
Lack of due diligence on vendor(s) and team members	(McKeeman, 2001)
Lack of stable project requirements	(Boehm, 2009)
Lack of good project communication	(Kappelman, et al., 2006)
Lack of project charter document at early stage of project	
Lack of gathering requirements via joint project design	
Lack of studying major requirements before the project kick-off	
Lack of metrics tracking process	(Jones, 2004)
Lack of budget estimated by the project team	(Kappelman, et al., 2006)
Lack of contingency budget	
Lack of stable organizational environment	(Schmidt, 2001)
Lack of interface management process to the project requirements	(Alsubaey, et al., 2016)
Lack of understanding of new data base	(McFarlan, 1982)

system capabilities	
Lack of selected project suppliers' knowledge/skills	(Alsubaey, et al., 2016)
Lack of congruence among projects requirements	

2.9 Extraction of indicator factors from to project documents

It must be clearly understood that much of engineering and construction projects are planned on fast-tracked design and management schedule. Due to poor project control and process management a large number of such projects tend to fail as they do not effectively integrate a sound control process. The phenomenon is reported in several studies with a focus on issues of scope definition, integrated management and services and a sound quality control process.

Despite balancing various tasks and their cost-benefit outcome, there is always a balancing line that must be drawn to optimise the outcome of management plans on project timelines (Kennedy, 2013). Time and cost is generally the trade-off that is sought to be optimised in project management schedules (Suliman, et al., 2011). Other cases also report another conjunction with time and cost where quality must be treated as a factor as well (Liberatore & Pollack-Johnson, 2009). The case of pre-emptively identifying time or cost loss during the running of a project is likely to improve the overall adjustments in a project and avoid missed planned schedules and or management tasks.

Delays are not only induced by working management teams in a project. Often, it is the continued changes or feedback delays that results in overall project delays. A major challenge in this aspect is where different stakeholders have conflicting requirements. For instance, one stakeholder might prefer buying all the resources from one supplier whereas the other is not willing to store

resources in storage for long. Such conflicts might result in indirectly induced delays (Ramanathan & Narayanan, 2014).

As discussed above, continued change requests also result in substantial project delays occasionally resulting in the change of the entire scope of the project. Such indications are generally reported in regular communication and are also discussed in project meeting minutes. Scope management and change control process documents are often used in project management to tackle client-change of requirement issues (Alp & Stack, 2012). Moreover, behavioural aspects measured via probabilistic analysis are also reported in the literature to be used to identify potential stakeholder-originating change-initiated delays based on PERT/CPM methods (Min & Shou-rong, 2013).

Team-level commitment to responsibilities and tasks is generally dependent upon a number of factors including training level, education, and overall experience of the staff (Mench, 2002; Andrey, 2015; Lei, et al., 2007). These aspects can generally be identified in documents including HR press releases, training requirement documents and advertisements (Zaghbib, et al., 2012).

Change management is one factor in construction project lifecycle often resisted by the staff. Any new process integrated within the running time of a project is likely to be countered by the staff employed which in-turn impacts on the regular deliverable and milestone deadlines. This behaviour is often witnessed in group-generated management documents such as meeting minutes and email-based communication. However, another alternative to minimise resistance to change can be to integrate information on expected changes within the project documentation (Viljamaa & Peltomaa, 2014).

Resources bear a large and diverse context in construction project management ranging from materials to humans. Sudden shortages in inventory and/or staff absence-related skill shortage is likely to reflect in memos sent by managers and HR departments. The aspect of improving on human resource management effectiveness via sub-contracting has been reported by (Othman, et al., 2011). On the material resource side, resource-constrained scheduling study has been reported by (Trautmann & Baumann, 2009). Further building on

the idea of constrained scheduling, a novel approach to establish knowledge acquisition and sharing between contractors have also been proposed by (Hu, 2008).

A summary of several issues leading to project failures and the types of documents involved are further elaborated in Table 5. The majority of issues are reported to originate from aspects including ambiguous goals, unclear deadlines, weaker control processes as well as ineffective communication and stakeholder management.

Table 5: Relationship between various project management problems with keywords found in management documents

Issues leading to project failures	**Factors potentially indicating signals that may result in discontinuities**	**Document source indicating hidden signals**	**Sources**
• Level of understanding of project requirements • Uncertain initial requirements • Standard of requirement specification	Unclear deadlines, clarity of requirements, confusing context, lack of agreement, cost estimation problems, contractor integration	Project requirement specification, client requirement collection documents, non-material requirement	(Cao, et al., 2002; Yusof, et al., 2016; Baghdadi & Kishk, 2015; Sedita & Apa, 2015)
• Changes in control process implementation • Change management process integration • Deviation from project scope & schedule • Abrupt changes in project schedule • Minimising changes via information integration	Scope definition, integrated process and quality control measures and expediting management tracking activities, change requests, scope changes, resistance to changes	Process control documents and quality guidelines, project plans and contract documents	(Alp & Stack, 2012 ; Weihong & Mingming, 2009; Min & Shou-rong, 2013; Zhao & Xue, 2010; Lines, et al.,

			2015; Viljamaa & Peltomaa, 2014)
Missed planned schedules and/or management tasks	Time, cost and quality	Project plans	(Kennedy, 2013)(Liberatore & Pollack-Johnson, 2009) (Suliman, et al., 2011)
Delays on stakeholder side	Conflicting requirements and objectives	Project requirement specification	(Ramanathan & Narayanan, 2014)
• Trust between parties • Communication with clients		General communication and letters	
• Weaker team-level commitments and diligence to tasks • Extent of team experience and skill level • Measure of new system's understanding of the staff	Skills, experience, training and education of staff members Impact of individual contractor specialisations	Training brochures, HR staff press releases, generated training completion certificates, manpower schedules	(Mench, 2002; Zaghbib, et al., 2012; Lei, et al., 2007; Andrey, 2015; Arashpour, et al., 2015)
• Poor resource management • Contingency budget estimation • Purchase and material procurement processes	Shortage of staff and required skills Lack of appropriate training Resource-constrained scheduling	Communication memos and emails	(Othman, et al., 2011; Trautmann & Baumann, 2009; Hu, 2008)

	Backup plans on emergent staff shortages		
• Poor management hierarchy • Decision making responsibilities allocation	Incorrectly assigned management responsibilities, lack of experience of staff having managerial duties	Position charge document	
• Project plan documentation • Planning and estimation documentation • Charter documentation at project start-up	Frequency of updates, number of changes, repeated updates of project activities within documents	Contracts POs Bid documents Schedule data Project diaries Change orders Correspondence Cost reports and estimates	
• Number of conflicts occurring within the project • Extent of discrepancies within projects	Poorly met resource requirements, inappropriately fulfilled material needs, insufficient resources to achieve planned project goals, initial cost estimation	Contract documents, project requirements, construction bidding document	(Arashpour, et al., 2015)

2.10 Summary

Early warnings during a project's lifecycle can be identified based on certain documented indicators that are directly associated with KPIs in the construction project management domain. On the leadership level, these indicators can be derived from documents detailing organisational control such as company hierarchy, top management support and governance structure. Moreover, timely actions and decisions during a project's running can also be identified from certain keywords and/or terms used in the project. On the human resources side, such early warning information can also be extracted from the overall team culture and commitment to project scopes.

There are a diverse range of documents generated during a project's running lifecycle that provide critical information such as minutes of meeting, project report, resource configuration management, inventory specification and documentation, procurement execution & control, project strategy and plan and effective supply chain management. Moving to extracting team performance, relevant early warning identification can be done via project scope, work-breakdown structure, scope and change management, scheduling and time management, and staff skills assessment and training.

The problem highlighted in this chapter highlights the importance of information extraction from management documents as and when they are produced at various stages of a project. Table 5 indicates the importance of these documents to train an intelligent system to then identify similar problems in ongoing projects. The domain of mining actionable information from documents and other media sources has been widely investigated. Therefore, the next chapter focuses on explaining how project discontinuities can be mined from management documents. In doing so, the chapter presents various data formats encountered in management documents along with the associated challenges in discovering knowledge from them. The chapter also presents different machine learning techniques employed to facilitate and transform this knowledge discovery into actionable information.

3 CHAPTER III: Mining actionable information and Machine learning

Exploratory text mining in unstructured project management documents is a process used to extract novel knowledge (Otsuka & Matsushita, 2014). In the project management domain, data and text mining techniques have long been used to extract useful such novel information in order to advance the state-of-knowledge of the underlying project's management lifecycle. Information in critical project management documents are frequently been analysed and modelled to extract early warnings in the domains of fraud prevention (Yongpeng, et al., 2014), email/web analytics (Dey, et al., 2013), financial analysis (Keqin et al., 2015), and socio-political impact analyses (Wex, et al., 2013).

Extending unstructured data into useful information has often been modelled via various statistical machine-learning techniques. Techniques such as Support Vector Machines, Artificial Neural Networks, Fuzzy Inference and Markov Chains are used to train text-mining models via labelled datasets (Wang & Chu, 2010; Larouk & Batache, 1995; Indhuja & Reghu, 2014; Bavan, 2009; Merkl & Schweighofer, 1997; Karaali, et al., 1998; Cerulo, et al., 2013). These techniques are used to predict and classify embedded information about the document or the relevant project based upon previously trained and labelled datasets. Yet, statistical machine learning techniques often fail to model and extract information from morphological structure of the text. For instance, it is not always possible to gain a word-level understanding of unstructured data as sentence-level structuring provides a deeper understanding of the message encoded in the document. However, at the same time, discrete word-pairs do contain crucial information particularly when semi-structured documents such as memos and meeting Minutes are concerned. For instance, a past-date of three weeks ago specified as a word-column pair "Delivery – 23/10/2015" indicates a delay of three weeks if the current date is 15/11/2015. However, memos can still have statements such as "the procurement is still not done due

to the required parts not being delivered" which cannot only be understood on word-level and a temporal modelling must be considered.

Based on the abovementioned limitation, this chapter presents a hybrid text classification methodology with a focus on text mining from construction project management documents. In corporate sector organisations, text mining is generally used to improve decision support on organisation activities and improved understanding of processes. The phenomenon is used to improve the understanding of Knowledge Management (KM) systems in order to help professional improve their skills and competencies (Verma, et al., 2015). Particularly, any project management documents like Minutes tend to contain information that possess critical information about how well the project is going-on. However, the documents tend to contain associated data-pairs and text-islands holding information that may be exploited to identify certain early warnings with a degree of confidence.

This chapter also covers the knowledge extraction and information gathering aspect of this research which reviews the existing state of knowledge of how unstructured information read or scanned from management documents can be transformed into actionable project data such as pre-emptive warnings. In order to address this problem, a detailed review of existing data and text mining systems will be presented. This will include information and knowledge extraction systems also known as Knowledge Discover (KD) systems. Before further getting into this problem, the chapter will explain the generic document structure of project documents in the construction industry and the measure of their complexity in extracting useful information. Finally, this analysis will lead to an in-depth analysis of machine learning techniques used for data extraction, KD and quantification paradigms that potentially lead to the proposed early warning prediction methodology.

3.1 Information systems and knowledge management

Project management documents tend to contain both structured and unstructured data, which are tightly correlated and embedded in project management documents. As the computer-based knowledge management is

becoming a norm, proper organisation and control of such documents is becoming a challenge. Moreover, in the present-day age of information systems modelling, extracting useful information from such documents is an active area of research (Mao, et al., 2007).

In the context of construction project management, data and knowledge generally derive from a number of interrelated elements that include human resources, cost, materials, planning and design. A combination of these factors makes it difficult to assess and analyse any risk factors within project plans. Kim (2008) proposed a technique commonly regarded as the Knowledge Discovery in Databases (KDD) to select a set of these processes that have the most impact on the performance of the on-going process. The methodology used a Back Propagation Neural Network algorithm to train against a set of databases and relevant classification labels to train a model to identify the cause of project delays. The main technique was compared with different machine learning algorithms including Bayesian network, Correlation matrix and Factor analysis and reported the Back-Propagation training approach to have the best identification accuracy.

An in-depth analysis of research done in the domain of project and human factors management was done by Nagai et al (2009). The work presented an analysis of various machine-learning approaches that could be used to model and identify project management activities as summarised in Table 6.

Table 6: A summary of various AI approaches used in customer-oriented project management activities

Project managements aspect mapping	Sub-areas focussed	Algorithms and/or models employed
Early warning identification	Response against early warnings	Ansoff's management model (Haji-Kazemi, et al., 2015)
	Accident prevention and safety management	Fault tree analysis (Dokas, et al., 2009), Nearest neighbour (Tseng, et al., 2007), fiber Bragg grating system (Ding, et al., 2013), Pareto analysis (Lee, et al., 2015)
	Information collection, sharing and	Hybrid data fusion model, ((Ding, et al., 2013), Decision tree

	communication	analysis (Chi, et al., 2012)
	Decision support	(Haji-Kazemi & Krane, 2013), Rough set and SVM (Pang, et al., 2006)
Resource utilisation	Multi-skilled staff utilisation	Process integration (Arashpour, et al., 2015), knowledge discovery processes (Erohin, et al., 2012), 4D modelling (Wang, et al., 2004)
Supply chain control	Multi-project management	Automated resource allocation (Ponsteen & Kusters, 2015)
Best practice modelling	IT integration, contracting, quality management, problem solving and cost reduction.	Project cost growth via 3DCAD(Chi, et al., 2012), quality management system modelling (Ingason, et al., 2015), PPP Best Practice (Gordon, et al., 2013), contractor to stake-holder problem solving (Handford & Matous, 2015), contractor performance analyser (Proverbs & Holt, 2000)

3.2 Structure of various construction project documents

It is generally importantly in a medium to large construction project management initiative to automate as much of its management infrastructure as possible. Large construction enterprises global tend to have specialised management support systems such as ERP II class (Hoła & Sawicki, 2014). However small and medium sized enterprises (SMEs) generally do not use such systems due to related high costs. A knowledge map was generated in the work of Hola et al (2014) which reported on the crucial areas that encompassed various types of information contained within specific documents. The work indicated a number of critical areas generating various document types as follows:

- System and environment: Any activity related to the system and environment generated documents related to scope and aim of activity and the organisation structure. These aspects give information about important factors such as company chain of command, accountability and HR-related issues (Kummamuru, 2014).
- Assets and resources: This category generally contain documented information about human resources such as total staff employed and requirement assessment of resources. Building and installation design plans

and technical reports on production details are also documented under this category (Azhar, et al., 2015).

- Process documents: The processes aspect cover paperwork generated as part of aspects of main, supporting and management processes. Moreover, this phase also generates a large number of other documents during the project's lifecycle including project preparation, design and execution (Porwal & Hewage, 2013). This may also contain tenders raised, contracts made and issued and other negotiation documents.
- Analysis and correction resources: During the lifecycle of any project, analysis is always a critical path. Moreover, due to unexpected issues and problems, a number of corrective documents are generated.
- Lessons learnt: As a result of certain actions taken to mitigate or resolve issues, document covering lessons learnt are generated. These documents contain potentially indicative information of certain delays in a project for e.g. delay in a deliverable due to a delay in the shipment of a part as a result of certain export or import restrictions (Herbst, 2017).

Based on various types of information contained within, the documents can be clustered into categories. Al Qady & Kandil (2014) reported on the development of a text classifier which clustered topic-specific documents in groups organised via certain keywords such as relocation, site and approval. On one hand, though this technique was capable of isolating documents based on categories, it did not address the issue of extracting organised/structured information from the text contained within these documents.

3.3 Structured and unstructured text

Construction industry employs both structured and unstructured means of communication to forward various actions and decisions materialising during meetings and work-related correspondences. The advent of advanced computing and information retrieval technologies and big data handling techniques have indeed revolutionised the way in which information is handled and processed (Berger & Doban, 2014). The same applies to the construction industry as well, which generates a large amount of project-related data which

must be organised to facilitate project control and speed-up access to required information (Chen, et al., 2016). A technique to extract information via document-integrated meta-data by applying a Request for Information (RFI) has been reported (Mao, et al., 2007). This enabled a better organisation and understanding of different construction document types and also facilitate the categorisation of files via certain indicators present within those files. Similar work presented a text information integration methodology which developed a 5-step model as follows (Soibelman, et al., 2008):

- Document preparation – This stage read entire project documents to prepare a Vector Space Model of the document. At its first stage, this model enabled document indexing which eliminated irrelevant words such as “is”, “the”, “which” or “for” while keeping frequency information of other words. In the second step, each word is assigned a “term weighting” based on various factors such as repetitions and frequent use (Bassett & Kraft, 2013). Finally, a “similarity coefficient” is used to measure the so-called similarity between various terms. A few well-known coefficients in the literature are cosine Jaccard and Dice (Soibelman & Kim, 2002).
- Model preparation - This stage creates a model specific to the project at hand which will potentially lead to the classification stage covering the decision of various types of classes in which early warning can be categorised (Kabakchieva, 2013).
- Classification model - This stage provides information about various types of classifications into which project documents will be categorised. Again, this will depend upon the type of project at hand (Bijalwan, et al., 2014). For instance, in the case of an early warning identification system, a classification model is likely to generate a system to predict the class of early warning embedded in a specific construction document.
- Retrieval and ranking - This stage in general is used to predict its extent of similarity to a certain class (Jiang, et al., 2013). For instance, if a document is a meeting minute where the meeting was held to discuss lean development in order to minimise wastage. A retrieval and ranking system for such a document will provide a similarity measure of this document to

various early warnings such as the measure of task completion (Al Qady & Kandil, 2013).

- **Association** – This stage is the final step in assigning a document to a certain category or class. At this stage, a project document shows its association to one (or more) of the factors indicating problems with the project (Sheffield & Lemétayer, 2013). For instance, a minute document may be associated to project delays to the fact that it is related to a meeting held to minimise project delays due to lack of skilled staff or employee absence resulting in shortage of manpower.

The abovementioned traits can further be elaborated in relation to the investigation performed in this work which is that of the measure of early warnings in project management via intelligent document text mining. The overall process is effectively a 5-step system starting from document pre-processing, project-specific model preparation and classification to retrieve and ranking and finally association of each document in various warning categories as shown in Figure 5. An example of the process of extracting unstructured content from construction documents is further shown in Figure 6 where significant words are clustered into certain feature groups that are then used to model and identify other documents into various identification classes.

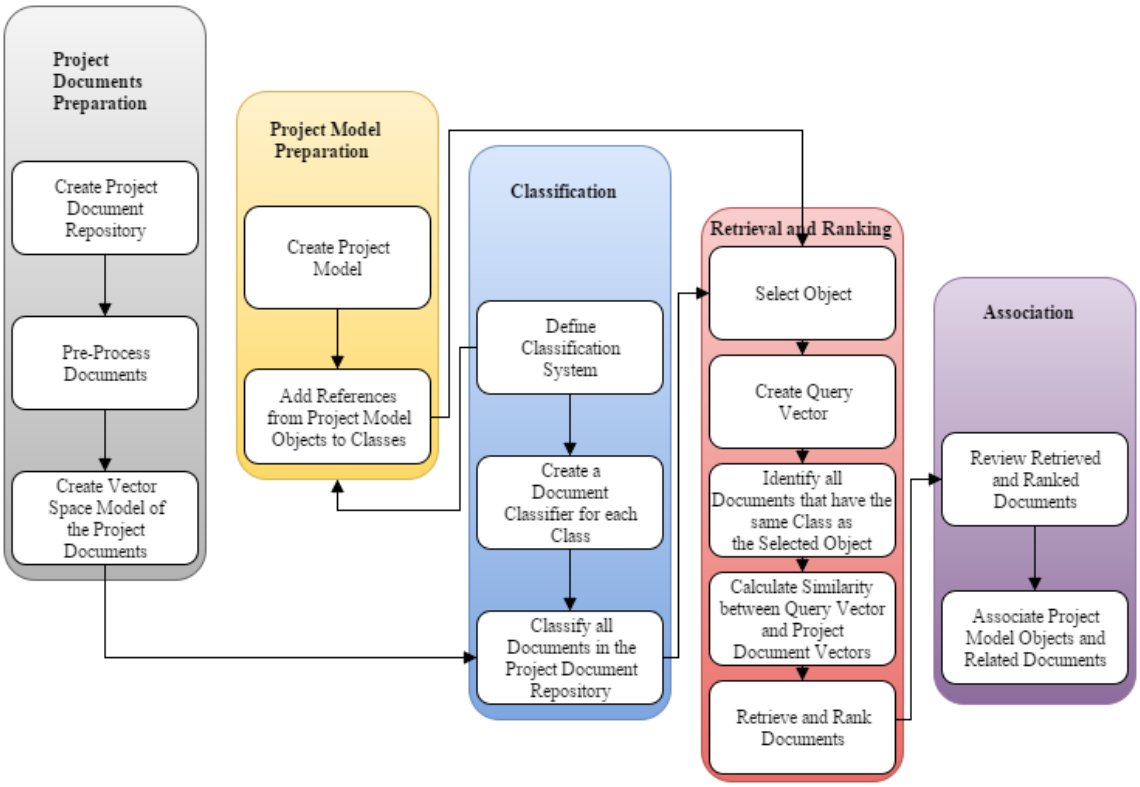


Figure 5: A proposed 5-step early warning document analysis architecture

ITEM	SUBJECT	STATUS	DUE DATE	ACTION BY
1	SAFETY			
1.1	Facilites: S.O said that a new logistic manager is appointed and is closely working with site to complete all the outstanding isuess. Mr. will be the point of contact from S.O to resolve the pending issues. S.O is aggressively persuing to complete the punch list items.	Open	12-Aug-15	
1.2	Road: G.A.PMT request S.O to provided an action plan and report on any noted improvements as a result of the plan to correct concern of drivers leaving the site and not complying with the Regulatory Signs. S.O said that they have increased the number of safety to monitor and advised compnay workers .	Open	7-Jan-15	
1.3	Scaffolds: G.A expressed its concern with scaffolds being built on site not in compliance with the G.A.I. and scaffold handbook requirements. S.O stated that it would mobilize RGS to construct scaffolds on site and would submit a light weight design for rolling scaffold as an alternative product for review. S.O stated that RGS will be mobilized on site	Open	13-Jan-15	

Figure 6: Early warning retrieval from construction documents based on unstructured content attributes (Ithra, 2009)

Existing computing techniques have presented ways of transforming raw textual documents into organised and highly structured data formats such as comma-separated-values, XML files, database tables and NoSQL systems (Dong, 2016). These paradigms offer highly integrated data formats which can be queried to retrieve information about anything happening in lifecycle of a project including documentation, resource allocation, skills and training and finance.

A standard document contains unstructured data which is processed via several steps to make it suitable to be used train a text mining classifier (Witten, et al., 2016):

- Extracted raw information - This information generally contains sections of text within documents which are extracted from various document sections. Once converted, the data extracted in this way is presented in raw and unorganised form and hence it is difficult to gain much understanding from it (Zhai & Massung, 2016).
- Feature extraction - Once the raw data has been read, it is then grouped based on a rule-base or existing, labelled training data which may contained cluster of words mapped into certain textual classes (Ganz, et al., 2016)
- Fact organisation - Based on feature map, a family of facts are then combined into distinct classes that represent various categories (Maynard, et al., 2012). Care must be taken at this step to avoid similar classes which may lead to inter-class bias due to similarity of input text.
- Text mining - Finally, the trained classifier obtained in the previous stage is used to identify/tag unseen data into useful categories (Williams & Gong, 2014).

The entire process mentioned above is a type of knowledge discovery system that can acquire, train, model and then identify unseen text data into information that can potentially be used to predict anomalies in project plans in time. The process is further elaborated in Figure 7 that show various data sources, methods and machine learning techniques used to design an intelligent text mining system.

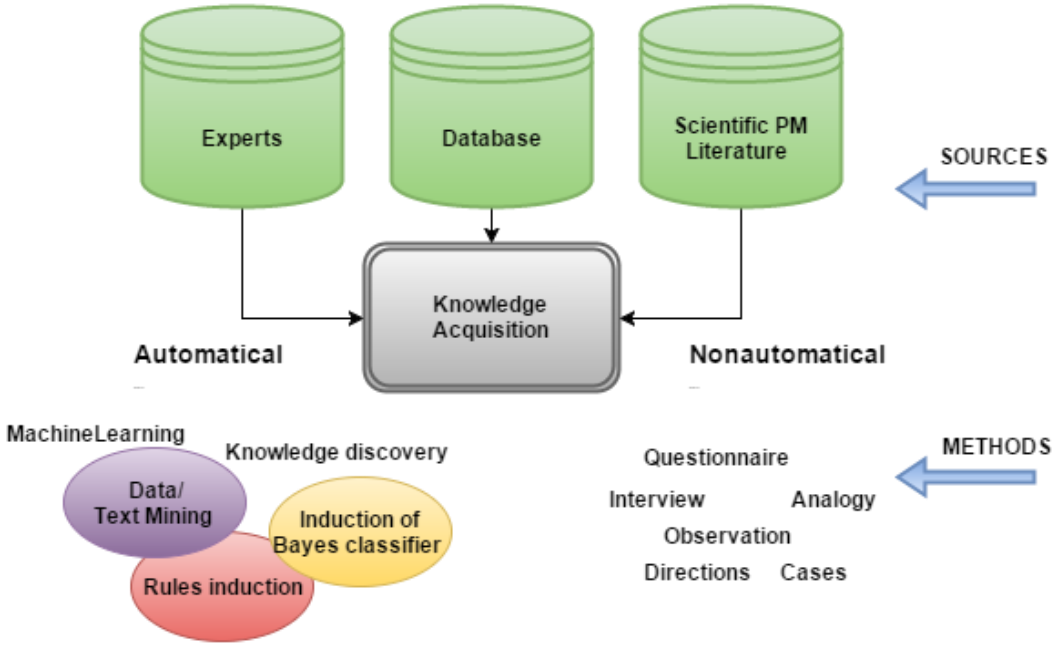


Figure 7: Various techniques of data mining and knowledge acquisition (Gajzler, 2010)

The illustration shows three main sources of information to train such a data training model including qualitative information from experts, databases containing ground-truth information and scientific literature to map text to certain management classes. Represent to distinct approaches to document classification. The simplest approach is a non-automated method which may take assistance from qualitative methods such as questionnaires, interviews and case study analyses to classify the wellbeing of a management project. The automated technique covers artificially intelligent mechanisms that use probabilistic and statistical relationships between various labelled text datasets to train a model that is capable of classifying unknown/unlabelled text sources.

3.4 Data mining techniques in project management

Data mining offers promising ways to expose hidden patterns within a large amount of data, which can be used to predict future behaviours. It is a well-known area of research aimed at processing unstructured information from documents, web-pages and other resources in a bid to gain a better understanding of the system being managed around that document (Ur-Rahman & Harding, 2012). For instance, a general human being can use his/her power of perception to read information placed in a meeting Minutes file

and understand any hidden indicators of delay (Ding, 2016). Any results achieved for project delay predictions via such data mining techniques would help managers categorize risks associated with execution stages allowing them to identify delays before they happen and their main causes rather than minimize the consequences as well as measure the progress to success. Establishing predictions and thus assumptions engage managers to chase specific information and more importantly define their value, which in turn help to determine the areas of concerns in which the managers should be willing to improve. Delay has significant effect on completion cost and time of construction projects. Delays can be minimized when their causes are identified. Knowing the cause of any particular delay in a construction project would help avoiding the same (González, et al., 2013). Although various studies have been undertaken to identify the factors affecting the causes of delays, since the problems are rather contextual, the studies need to focus on specific geographical area, country, or region. For instance, a major criticism of the Middle East construction industry is due to the growing rate of delays in projects delivery. A study in London presented new project delay prediction model with the use of data mining. The researchers identified the causes of construction delays as seen by clients, consultants, and contractors, and then combined the top delay factors with secondary data collected from an on-going project and built a model using Weka software to predict delays for similar projects (Asadi, et al., 2015). Their research explored different models, various classification methods and attained interesting results on predictability of project delays by measuring the performance of the models using real data. The results for two classification algorithms showed an average test accuracy of 76% when classifying delays from a combination of real-project monitored data and one of the top delay factors, for 34 months' duration. The best algorithm for the data used from construction project is Decision Tree (J48) classifier with accuracy of 79.41%. Naïve Bayes classifier has lowest accuracy and higher error rate compared to J48. Despite the small differences in readings obtained for the two classifiers, the results suggested that among the machine-learning algorithm tested, J48 classifier has the potential to significantly improve the conventional

classification methods for use in construction field. One major aspect in project lifecycles that can be extracted from the underlying documents is the level of uncertainty about certain things (Harrison & Lock, 2017).

3.4.1 Extracting uncertainties from project documents via ques and indicators

Project managers tend to review project documents in a bid to understand and identify uncertainties that may lead to future shortcomings in the project. Such an earlier gain improved control and neutralised project threats before they materialise into more problematic issues (Amarasiri, et al., 2013). Moreover, based on their experiences, managers use ques and indicators via a risk management process (RMP) in order to document threats, quantify uncertainty and improve the overall risk efficiency of the project (Bromiley, et al., 2015).

In everyday project management activities, documentation holds a critical aspect. Every phase and responsibility within a project is documented to enable standardisation, audit and plan. For instance, a project plan is the first thing that is prepared along with resource allocation, milestones and deliverable setup. These phases are important for the management team to allow timely decisions and adjustments on various phases of an ongoing project (Coombs, 2014). Likewise, contracts, purchase orders, Minutes, project reports and other similar documents form part of a project to assist in providing a strong and management baseline for project managers for consultation. Such documents are crucial in establishing the measure of risk to which various phases of a project are subject to. Documentation is particularly important as it deals with and measure uncertainty in project design, activities, responses, and decision support (Wise, et al., 2014). Such documents provide support in a number of following aspects:

- Clear thinking: Any documented facts allow managers to recall lessons learnt and previously adopted best practices.
- Succinct communication: Any written aspects of a project are a lot easier to communication and hence minimise the level of misunderstanding between stakeholders and the management team (Halter, 2015). Moreover, it also

provides a measure of accountability for various personnel within a team by assigning duties and responsibilities. For instance, in a project plan, a particular deliverable may form 30% responsibility for a particular team member and hence its allocation and completion would generally be bound to that specific individual and his/her time allocation for the project. Similar, other aspects may include financial consequence accountability, allocation of shared resources, documentation of threats and opportunities and any other assumptions (Davison, et al., 2014).

- Familiarisation: Any documented also includes reference details for new member and also serve as a training tool for the staff.
- Record of decisions: A number of such documents are used to record decisions within projects as well. Hence, any historic malpractices or decision leading to project losses are also recorded in such documents (Haidar, 2015).
- Knowledge base: Moreover, corporate knowledge is also often recorded in various documents to assist other teams with the lessons and practices recorded in such documents. Hence, any deviations from set milestones are recorded in such documents as a factor which can be referenced by future project teams (Duffield & Whitty, 2015).

A framework for data acquisition and parsing: Data acquisition in document process and management systems originate from a well-known field known as “data mining”. The concept is used very often in a domain where data is extracted from one of more database to generate more information (Witten, et al. 2016). One example commonly used in this domain is that of text mining from license plate recognition cameras for vehicle identification, TV highlights reading, optical character recognition (OCR), old document archiving, etc. The majority of these documents comprise of unstructured data which is read and understood by human beings due to their outstanding ability to differential various text sections from image and other data via visual ques and other area of interest segmentation capabilities (Ramisch, 2014).

Due to the importance of extraction of unstructured data and its transformation into structured formation, text mining for management documents present a significant area of research due to the complexity of management documents (Jin, et al., 2015). Hence, this research primarily focuses on the exploration, design and development of techniques to improve management document categorisation via text mining.

These phases are important for the management team to allow timely decisions and adjustments on various phases of an ongoing project. Likewise, contracts, purchase orders, Minutes, project reports and other similar documents form part of a project to assist in providing a strong and management baseline for project managers for consultation. Such documents are crucial in establishing the measure of risk to which various phases of a project are subject to (Alhawari, et al., 2012).

A large fraction of companies tends to miss their planned goals due to issues arising from budget and schedule overruns (Tsai, et al., 2009; Legodi, 2010). Many projects fail completely due to the inability of their management to realise hidden problems and risks within the projects' trajectory (Baghdadi & Kishk, 2015). The problems mainly identified in such failures primarily include poor project scopes, abrupt scope changes, poor budgeting, inefficient planning and/or scheduling and poor selection of management teams (Ratsiepe & Yazdanifard, 2011). According to Harding (2012), inadequate project support, improper risk management, poor stakeholder consideration and inappropriately selected contractors and engineering firms also play a major role in project failures (Harding, 2012). With the automation and electronic-control of project management with the advent of information and communication technologies (ICT), knowledge extraction for risks and impending project failures have taken a completely different aspect commonly regarding as information retrieval, text and data mining. The domain uses advanced text and information processing algorithms to gain novel information about the project and is commonly known as data and/or text mining (Feilmayr, 2011).

3.5 Application of machine learning in data mining

Earlier information and data modelling techniques in project management context relied predominantly on human experts who performed manual classification of documents. However, with the massive increase of management-related documents, this has become in infeasible and prone to mistakes. Caldas et al (2002) integrated unstructured information into a machine-learning approach to develop a document classification system. The work utilised a construction project database as a case study and integrated topics extracted from specifications, meeting minutes, information requests, change orders, contracts and field reports to implement the underlying classification system. Extending on this work, extensive research has been done focussing on various sub-areas of automated information extraction in the project management industry. Since this work, machine-learning algorithms have increasingly been investigated and employed in information retrieval and data mining. Work in project management data mining has focussed on the automated classification of various linguistic documents where work done by Caldas et al (2005) has presented and compared SVM, ROCCHIO, IBM, Naïve Bayes and KNN algorithms to classify documents into various categories such as schedule, demolition, and conveyance categories as shown in Table 7.

Table 7: Comparative analysis of existing text mining algorithms

ML	SVM	ROCCHIO	IBM Miner	Naïve Bayes	K-Nearest Neighbours
Score	91.12%	64.71%	60.95%	58.82%	49.11%

Artificial neural networks (ANN), support vector machines (SVM), Naïve Bayes and k-nearest neighbour classification approaches have been applied to address several data mining problems. Clustering approaches have been used in clustering documents in a supervised manner especially when the possibility of labelling is not feasible. Segmentation of high dimensional data based unsupervised feature selection was undertaken by Li et al (2014). Though the objective of this work can be more applicable on segmenting and behavioural

modelling, it draws parallel to identifying various working practices in the management industry. Most importantly, the research directly relates to future extensions into Naïve Bayes Trees and supervised clustering techniques to generalise and automatically segment behaviours in a large and diverse set of practises undertaken in construction management. One such work has been reported in Almazán et al (2014) that utilises hand-written word spotting in a common vectorial subspace as a nearest neighbour problem. Though the work focussed in creating associations between writing styles and document images, a similar context can still exist if attempts are made to create associations between textual figures used in documents and the document type. Similar attempts to identify sentiments, belief propagation and even extracting non-functional information (e.g. qualities and performance) have reported in the literature by Schouten & Frasincar (2016), Zeng et al (2013) and Slankas & Williams (2013). What is common in these three researches is the fact that they have utilised domain-specific knowledge of various datasets using semi-supervised techniques such as Linear Discriminant Analysis (LDA) and Naïve Bayes multinomial clustering mechanisms. The domain is further extended on temporal information where sequential data can be used to improve the contextual understanding of a document. Work by Emonent et al (2014) reported using nonparametric Naïve Bayes classification to describe motifs and their occurrences in documents. Utilisation of context independence of input variables used in Bayes classification makes them an ideal candidate to be used in situations where there are multiple, mutually irrelevant factors such as level of inventory, staff availability and/or number of rejected materials. Yet, when combined, these factors affect the overall likelihood of failure of a project in a probabilistic manner. KNN algorithms have extensively been investigated in document and data classification based on clustered feature sets. Work has focussed on text categorisation (Soucy & Mineau, 2001; Rathore, et al., 2013), integration of supervised feature selection (Basu & Murthy, 2012) , integration of KNNs with ANNs to leverage supervised dataset training (Adeniyi, et al., 2016) and multi-class document categorisation (Vijayan, et al., 2017).

Based on the existing research state discussed above and a review work by Ko and Cheng (2007), the overall text-mining research at present focuses on the following areas:

- Improvement and exploration of feature selection methods.
- Reduction of training and testing time of classifiers.
- Filtering specific keywords to detect early warnings.
- Use of semantics and ontologies for the classification of documents.
- Extracting meaningful information such as factors and trends in business, marketing and financial trends in project management documents.

Despite this work, extracting terms, textual landmarks and hidden warnings from text is still an active area of research facing numerous challenges. A soft-computing approach to parsing and processing such documents for hidden indicators is itself a challenging task primarily since a large number of these documents contain a mix of raw, tabular and figure-based data which cannot accurately be interpreted by straightforward text processing and mining algorithms (Bilal, et al., 2016).

3.6 Core components in NLP modelling

Natural language processing is utilised to contextually understand the nature and content of documents. This is generally addressed in two capacities.

3.6.1 Natural Language Understanding (NLU):

This step involves the mapping of given input in natural language to useful depiction. The stage also includes the analysis of different aspects of the language (Witten, et al., 2011).

3.6.2 Natural Language Generation (NLG):

The generation stage involves the production of meaningful phrases, and sentences in the form of natural language via internal representation. This process involves text planning, sentence planning, and text realisation (Witten, et al., 2011).

3.7 Problems arising from automated knowledge extraction in the construction domain

Continuing with the challenge of handling large dataset extracted from a high volume of documents in order to identify early warnings present a problem with the volume of the data. Focussing on the construction industry, the majority of such documents are text-based. Existing systems for the classification of such documents in the construction industry are mainly manual (Tixier, et al., 2016). A prototype for document identification system with a focus on using supervisory classifiers was presented by Caldas et al (2002). The work however focussed mainly on document classification based on the text present within. In the existing scenario, learning and categorising from massive datasets has increasingly becoming a computational challenge despite all the advanced in the computing infrastructure. Hence, the focus currently is on computationally efficient extraction of information from large datasets. Recent work in this domain has focussed on aspects of attribute-based keyword selection for data classification (Silva, et al., 2015), parallel ontologies (Li & Sima, 2015), high-dimensional data analysis (Sohrabi & Barforoush, 2012) and utilisation of machine learning algorithms to handle imbalanced document data (Pérez-Godoy, et al., 2014; Bao, et al., 2016).

3.8 Challenges in analysing unstructured data

In addition to the data-related challenges, aspects of process-related issues including how to capture data, integration, transformation and finally the selection of the most suitable models are many challenges as quoted by (Sivarajah, et al., 2017). Moreover, with the increasing richness of printed documents with the improvement in printing techniques, the majority of such documents contain a combination of multimedia and image-based content which cannot directly be processed without using a reliable optical character recognition (OCR) software. Tabular data further increases the complexity of such documents as it becomes immensely challenging to relate rows and columns of documents into meaningful relationships without human intervention. Text and data mining algorithms facilitate an intelligent way of

extracting data relationships while reducing the margin of error in textual interpretation (Al-Shameri, 2012).

3.9 Data mining classification algorithms in the project management domain

Use of AI and machine learning algorithms to process and classify documents for meaningful indicators is a well-explored area of research in text and data mining. Two main approaches are used to extract information from documents for classification purposes namely: supervised and unsupervised (MartíN-Valdivia, et al., 2013). In the former (supervised) methodology, existing data input output pairs are needed to label the classes in order to use the underlying model to identify unseen data into those classes. The labelled data in such cases is also known as the “ground truth”. In the unsupervised case, the training data is not labelled via human intervention at all (Dau, et al., 2016). So, it is on the algorithm to cluster the information into various groups for future clustering of test data into one of these clusters. Data or text mining is one of such domains which contain massive information contained within various corpuses (such as minutes of meetings, contracts and CAD drawings) (Hsu, 2013). However, this information is largely unlabelled and hence cannot directly be used to process and understand unseen documents for the identification of various early-warning indicators.

3.9.1 A text mining problems generally involves the following four stages

- Classification - The classification problem aims at identifying various groups within a dataset which are then used to classify new data. Email spam classification is one such example where a number of techniques are commonly used including Naïve Bayes, Nearest Neighbour and Artificial Neural Networks (ANNs). This type is primarily used in supervised learning techniques (Chamatkar & Butey, 2015; Zurada & Fife, 2006).
- Clustering - Clustering on the other hand, performs unsupervised categorisation of data based on unique centroids and the clusters of those

centroids. Hence, unseen data is classified into a class whose centroid is closest to the test data value.

- Regression - Similar to classification, regressions is generally supervised and aims to fit a function modelling the data with least error. This type is most commonly reported in ANNs. This aspect is commonly used in defect report mining in software engineering (Jindal, et al., 2015), academic writing grading (Lam, et al., 2010), and email processing (Liu & Lee, 2015).
- Associative learning/rule-based learning - The category looks for typical learning rules which may be tuned and improved via other machine learning algorithms to develop a rule-based classification system. Fuzzy inference systems (FIS) and hybrid artificial network-based ANNs are two commonly reported types (Kadir, et al., 2011; Wen-Tsao, 2008). FIS/ANFIS systems are commonly reported in business and environmental prediction systems for early warnings.

3.10 Probabilistic models

In the text mining case, a model is regarded as an observed data corpus on which the mathematics of probability theory is applied to express all the uncertainty and noise. Hence, based on inverse probability, unknown/unseen datasets can be predicted to adapt the underlying model and hence make predictions. Naïve Bayes has been reported in various applications of data classification including spam filtering, search predictions, text classification, and hybrid recommender systems such as those used in video suggesting websites (Isa, et al., 2008; Almeida, et al., 2009; Puntheeranurak & Sanprasert, 2011).

3.10.1 The Bayesian approach for identification can be explained in general as follows

Training samples are observed incrementally and hence, each sample increases or decreases the probability of hypothesis instead of eliminating it. The hypothesis can be determined through a combination of prior knowledge combined with observed (training) data (Choi, et al., 2013). The method can therefore be used to probabilistically predict the hypotheses. Hence, new data

samples can be classified by integrating multiple hypotheses based on their probabilities (Kruschke, et al., 2012).

In other probabilistic approaches, Helmholtz Principle from Gestalt theory is used to extract and predict patterns from small document, text summarisation, text change detection and document segmentation (Dadachev, et al., 2012; Ganiz, et al., 2015). Similarly, Latent Variable Models (LVMs) is used particularly time-series-based extraction, topic modelling, and web-page analysis. (Biro, et al., 2008; Anupriya & Karpagavalli, 2015).

It is well understood that text mining and classification includes labelling pre-determined categories to textual resources. However, training probabilistic models with large amount of textual data is a computationally infeasible process. An approach to handle this is to use a dimensionality reduction methodology such as principle component analysis, the well-known J48 decision tree or decision tables (Jaffali & Jamoussi, 2012; Kondor, et al., 2013; Weiss, et al., 1999). Decision tables are known to improve classification accuracies a lot better than Naïve Bayes and J48 classifiers (Gaurav, et al., 2004).

3.10.2 Hybridisation of semi-supervised/unsupervised text mining techniques

There have been other less frequently used techniques for information mining in general. Reinforcement learning is an extended type of semi-supervised learning methodology which forms its basis on maximising cumulative award of certain decisions (Kacprzyk & Pedrycz, 2015). Similar to standard Markov chains/Markovian theory, this technique utilises dynamic programming principles of the Markov decision process (MDP) of maintaining a balance between search of the best solution and the existing knowledge of that solution. The technique to date has not directly been reported in text mining however the focus has been on sub-areas of data mining such as pattern modelling, self-reinforcement to improve data identification, and online analytical processing (OLAP)(Vieira, et al., 2010; Pi & Zhang, 2014; Kaya & Alhajj, 2005). Additionally, Boosting algorithms'-based classification techniques have

particularly been focussed on big-data analytics and query-based text retrieval system (Polig, et al., 2014). On the evolutionary heuristics side, information extraction techniques have been reported with genetic algorithms (GA) for high level knowledge discovery, optimal word selection for text pattern searching, large scale/big-data/information clustering via optimised classification functions (Atkinson-Abutridy, et al., 2003; Atkinson-Abutridy, et al., 2004; Amarasinghe, et al., 2015; Wang, et al., 2012).

Extending on the clustering techniques discussed above, K-nearest neighbour (KNN) has extensively been discussed as one of the prime, unsupervised searching mechanisms for extracting unusual signatures from textual documents. Clustering for certain patterns or indicators based on Mahalanobis distance have been reported (Suli & Xin, 2011). Similarity search mechanisms based on KNN for dimensionality reduction in text classification have been reported to eliminate irrelevant textual documents and improve the overall complexity (Kim, et al., 2012). Other extended techniques have been used as weighted KNN (Fang & Qingyuan, 2010), Class Core Extraction (CCE) (Yu & Zhang, 2009), academic document filtering (H, et al., 2012) and bug-tracking in software program documents (Chaturvedi & Singh, 2012).

However, the majority of this work has focussed on clustering topics in comparison to the objective of the proposed approach which focussed primarily on extracting hidden indicators within text documents. Moreover, the technique proposed is semi-supervised in a sense that it already contained critical information about the document itself as it is already aimed to be labelled via expertly driven questionnaires from field experts.

3.11 Algorithms addressing success and failure factors

Based on the literature review of various AI techniques used in text and data mining industry, it is well-understood that most of techniques focus on labelling and extraction of generic project management documents. In construction project management domain, this approach cannot directly be assumed as different projects tend to have a diverse set of performance analysis criteria. Particularly for the Middle Easter construction management paradigm, the

baseline must form on the feedback provided by the construction industry engineers and experienced managers (Asadi, et al., 2015). Hence, this approach forms its training corpus from structured/semi-structured interviews held with staff having years of experience in the construction industry domain.

Different AI algorithms present a varied measure of handling of the text mining problem at hand in this research. ANNs tend to address noisy data quite well. Hence, for documents with missing information, ANNs offer a better solution (Huang, et al., 2013). Yet ANNs suffer from issues of over-fitting and unequal class bias. SVM are well-known to address dual classification problems and are capable of better classification even with smaller datasets (Cheng & Hoang, 2014). However, the main shortcoming of conventional SVM is in its inability to handle/classify more than two classes. Missing data also tend to present a substantial challenge in accurate modelling of text classification problems (Kersting & Železný, 2013). Several techniques to cater this problem is presented in the literature. Kalman filtering is a well-known technique that tends to address missing data issues particularly in temporally spread data problems. Other potentially useful solutions include conditional random fields, hidden Markov models and time-series ANNs (Candy, 2016; Tascikaraoglu & Uzunoglu, 2014). Most recently, ANNs have focussed on text mining problems of foreign exchange & market prediction, self-organising maps for text hierarchy generation, online malware detection and processing of multi-lingual online web documents (Kumar & Ravi, 2016). Due to the inherently time-series nature of HMMs, this genre of algorithms has focussed on problems including text sequence recognition, text mining from biomedical sequences, natural language processing, and hand-written document analysis (Zubrinic, et al., 2012). However, text extracted from documents is generally ridden with noisy and uncertain data features. Hence, this aspect further leads to regressive and extrapolating techniques such as Kalman filters (Speicher, et al., 2013). The overall picture of existing research shows a substantial level of work being done in prediction and identification of certain features within documents.

3.12 KD application to the risk identification process

As discussed above, KD systems for risk identification in the construction project management document is aimed substantially on the discovery of hidden risks in certain management aspects as discussed in Table 2. The underlying principle is effectively to transform unstructured data into a semantic knowledge base which can be organised and queried at latter stages of the project management process. The risk entity in such systems can be modelled in a variety of ways as reported in the literature. A lot of reported work focuses on manually labelling data to tune and optimise machine learning models thereby enabling them to predict hidden anomalies from documents based upon these models.

3.13 Naïve Bayes classification in machine learning and data mining

It is generally the responsibility of project managers to assess project plans for potential deviations such as delays in milestone achievements, impacts of deliveries of project deliverables and thus meeting deadlines. A lot of time is thus spent searching for valuable information from archived hard copies such as meeting Minutes, memos, and other documents containing unstructured data. According to (Mao, et al., 2007), decisions made on information extracted in this way is prone to human error and may lead to judgmental errors due to the inherent inability of humans to analyse swathes of management document data. Bayesian analysis has routinely been used to extract information and factors from documents based upon trained models. However, the majority of work focuses on human factors or expert knowledge integration to enable classification. (Raghuram, et al., 2009)'s work focussed on retraining the network based on new information where the initial model is constructed via expert knowledge. However, in classification based on retraining, the issue of incomplete data has always been a challenge. Gunasinghe & Alahakoon (2010) presented a fuzzy ARTMAP and back-propagation-based approach to fill lack of input information. However, Bayesian identification generally identifies discrete keywords based on their probabilistic likelihood to belong to certain classes.

This poses a challenging in cases where information is not only hidden in keywords but in complete sentences. For instance, for a sentence found in a meeting Minutes document given below:

***Sentence**: There is a good level of delay in delivery of material*

If the sentence given above is ranked based on positive, negative and neutral class associations, then words such as good or delivery can be taken as positive. However, if the sentence is taken as a whole, it indicates an early warning into possible delays in the supply chain. This characteristic of text documents presents a time based nature of words. For instance, sentences with most words being common can still belong to different categories as follows in

Table 8: Examples of Early warning scoring and classification system for identification Minutes items

Items from Minutes of meeting	Classification	EWSs score
There is a good level of delay in delivery of material	Negative	Low “materials” score
There is a good level of coordination in the delivery of material	Positive	High “scope of material delivery” score

Table 8 presents a well-known data/text-mining problem where only words cannot be used to completely understand the hidden information within a document. Time-based analysis has commonly been used in this domain where machine-learning classifiers are used to learn from complete statements in documents, which are labelled by relevant field experts. For instance, Markov chains are commonly used to identify “islands of information” in free text (Cerulo, et al., 2013).

3.14 Data mining techniques for risk scoring model setup

Risk scoring in construction project management is a well-explored area. For instance, risk scoring for investment decision making based on modified risk adjusted discount rate for different construction periods has been reported by (Mao, et al., 2009). Fuzzy logic and Failure Mode Effect Analysis (FMEA) has

also been used to assess risk based on impacts of time, cost, quality and safety (Mohammadi & Tavakolan, 2013). Similar scoring mechanisms have been reported in credit behaviour scoring systems, indexing of multimedia documents, ontology definition in construction risk knowledge management, scoring parameters in capped tendering, and assessment of environmental impacts in construction projects (Neto, et al., 2016). In the proposed research, this concept is extended to assign scores to various word/sentence/segment pairs via the well-known “bag of word” and TF-IDF systems which are discussed in the detail in the later chapters.

3.15 Knowledge gap in the existing research

Much of research in the domain of document analysis has been limited to document or scene categorisation which means that extracting multiple cues from various sub areas of documents has not been investigated to greater depths (Kaur & Singh, 2016). For instance, social and other web-originated media has been generating an enormous amount of unstructured information which has not yet been annotated to various categories (Njadat, et al., 2016; Kumar & Ravi, 2016; Tsirakis, et al., 2016). Similarly, unstructured information mining in medical documents via various document clustering methods have been used to identify a limited number of emotion categories such as positive, neutral and negation behaviours (Zeng, et al., 2016). There has been some work reporting multiple categorisations of text within document pages (Carlos Gomez), handling and classifying multi-lingual documents (Chih-Tien, et al., 2011; Zhang, et al., 2011), opinion mining via NLP algorithms (Sun, et al., 2017) financial data mining (Kumar & Ravi, 2016), text-mining for natural disaster prediction (Goswami, et al., 2016), and in criminal identification via identity-theft documents (Nokhbeh Zaeem, et al., 2017). Despite extensive work done in various document sub-indicator extraction, the field of construction project management has largely been left unexplored. There has been attempt at document categorisation like discussed above, however, a comprehensive analysis of construction-related documents and their role in modelling various

project lifecycle aspects such as inventory control, potential delays, and lack of resources has not been extensively and holistically explored.

3.16 Proposed training system for EW signal extraction

The initial scanning of management documents will generate a large amount of text. This text must be parsed into a suitable format that could be analysed and processed by an AI machine. The initial aim of this project, hence, is to be able to classify “weak signals” from management documents.

- **Phase 1:** Information extraction via qualitative assisted ground-truth preparation

This stage handles data that is to be associated with “weak signals” based on the nature of information hidden. The information hidden can ideally be understood either via an existing AI model or via expert input. Based on the literature review, there is currently no suitable AI model available that could be used as a benchmark to identify early warnings in construction management documents. Therefore, in order to train a machine learning model, it was necessary to generate a ground truth for the initial data via a qualitative assessment of project management documents. The aim of this qualitative approach is to assist in gathering and labelling text regions within management documents along with their association to certain signals. This stage was hence aimed at pinpointing certain signal segments within documents to have a potential to contain early warnings. At the end of this phase, the following information was made available:

- Result of the outcome of a semi-structured questionnaire to gain “ground-truth” on signals identification certain EWs by generating the following:
- A corpus of sentence input/out pairs from Minutes of Meeting documents with their association against the extent to which a signal was present within
- Based on the abovementioned corpus, an AI classifier capable of identifying these warnings via an unsupervised methodology is then proposed

- **Phase 2:**

The second phase takes feedback from the same questionnaire where the experts further identify/group certain keywords to identify their association with

concrete early warnings. These early warnings have already been agreed in this chapter via secondary research

- Based on the weak-signals, this stage further extends to identify clear warnings on the type of warning the project is most likely to face
- Once, the secondary corpus containing a concrete EW ground-truth was available, it was used to train a supervised AI algorithm in order to predict direct early warnings construction project documents

- **Phase 3:** This phase ultimately presents a summary of presence (or absence) of signals as well as subsequent EW types in order to assist the designers, planners and stakeholders to take corrective actions.

Chapter 2 primarily discusses the project management aspect of this research whereas this chapter addresses the application of AI to the document processing problem introduced in Chapter 1.

3.17 Summary of research gap

During any project's management, many diverse documents are generated which can be processed via an AI parser algorithm. These range from general-purpose letters to key meeting-related documents containing crucial decisions. Most of information in these documents is highly unstructured where 70% - 80% of information is classed as unstructured textual information which is difficult for an AI parsing algorithm to understand (Gajzler, 2010; Shariff, 2011). Hence it is difficult to understand the scope of it without generating its AI model representing the underlying text mining model. However, the information contained within these documents may bear format that could directly be read by a text-reading algorithm such as images and tabular data. Therefore, this information cannot directly be transformed into structured data pairs or tuples. A number of data and text mining techniques have been reported to handle image-based documents. However, current research does not focus on this aspect of the research. Hence, the large amount of information present within such documents created as part of meetings, brainstorming sessions and generic discussions in the form of memos, minutes or reports.

The overwhelming availability of unstructured information in such management documents presents a huge challenge for text mining industry as most of research has focussed on extraction of shorter highlights such as TV news lines, voice-to-text systems, document identification, etc. However, changing this textual information to an organised set of indicators is an open and widely explored problem. Moreover, in general documents, text is arbitrarily organised and hence it is very difficult to identify certain relationships such as association of values to variables in tables. This shortcoming presents a potential knowledge gap where structured/useful information to understand hidden early warnings is derived from construction management documents. The most recent and relevant work in the domain of extracting early warnings has been reported by (Nikander, 2002) which extends on the application of Ansoff's theory to identify early warnings from text data.

Further extending on the original idea of Ansoff's theory of weak signals, it is possible to detect discontinuities or strategic surprises in organisational management (Ansoff, 1975). The theory is reported as the Strategic Early Warning System (SEWS) theory which is aimed at identifying operational and strategic anomalies within enterprises. The proposed work utilises this core principle that discontinuities do not occur suddenly, and it is possible to detect hidden issues by scanning the overall environment of a company or that of a project's management.

The core principle followed is to first focus on a Boolean identification of "risk signals" indicating the presence or absence of any developing discontinuities or risks within the project. The phase was assisted by expert respondents to the survey who pinpointed various keywords and their association. The normal-class association was prepared by using documents that were tagged as normal by the respondents. The second stage was based on the principle of identification by training where a model, again based on questionnaire input, was used to train a supervised and probabilistic AI Naïve Bayes model. The in-depth design and implementation of both of these models is presented in the next chapter.

4 CHAPTER IV: A hybrid Bayesian classifier to improve early warning identification

Text and information mining systems have extensively been explored in the statistical and probabilistic machine learning domain. Bayes Classification is a probabilistic belief system based methodology that draws its name from its ability to predict likelihood of classes based upon the previous presence in a test set. This means that the more sample frequency a certain class has the more probability a sample closer to it will have to belong to that class. So, in Bayes interpretation, the identification of a certain data value is based upon the probability measuring the degree of belief.

In construction project planning, the degree of belief may come from a wide and diverse range of factors that may include better communication, shortage of materials and lack of financial resources, poor supervision, contract variations and improper design (C.Tung-Tsan, 2010; S. R. M. Nasir, 2012). However, this research work primarily focuses on meeting minute's files to predict anomalies in project timelines. The main aim of meeting minute's files is to record and report on the overall progress of the project. As these documents are cornerstones of a project's progress, the text contained within bear crucial information about the potential risks and possibilities of deviations from planned activities. For instances, higher frequency of certain words within such files may give an indication of an unexpected and developing risk such as "lack of financial resources". Based on the inherent ability of Naïve Bayes classifier to learn from prior (previous) probabilities of data, the algorithm is a likely candidate that can be used to evaluate hidden early warning within such files. For example, based on our initial assumption, a certain project may twice as likely of delay due to the lack of resources then due to lack of staff skills or training. So the degree of belief at the initial part would be 50%. However, we may get additional project files containing more evidence of why other similar projects failed and this may lead us to have an increased belief of 70% that the current project failed due to lack of resources (or any other class).

Nikander & Eloranta (2001) indicated a number of early warnings that were adapted in this research as a set of classes to be identified (Nikander, 2001). These classes are shown in Figure 9 where based on a structured data system (CSV), a machine learning classifier is used to train a Naïve Bayes model as shown in Figure 8 . This trained model is then used via a PDF parsing algorithm to identify text patterns from MOM files into one of these four categories. The proposed methodology presents a dual Naïve Bayes classifier to report two unique early warning categories present in construction project plans and the other relevant documents such as the MOM documents. The technique is unique in a sense that the Bayes classifier only considers the textual context and ignores the numerical data present within these PDF files. As it is human nature to judge patterns via numbers and not text frequencies whereas the numerical data can be studied by average human beings. On the other hand, it is substantially difficult for engineers to identify and isolate early warning patterns from textual information. As the variables used to predict such patterns are used independently of each other, combining qualitative and quantitative information would not have any inter-variable relationship due to the “naïve” nature of the methodology. Hence, this concept forms the underlying core of this research.

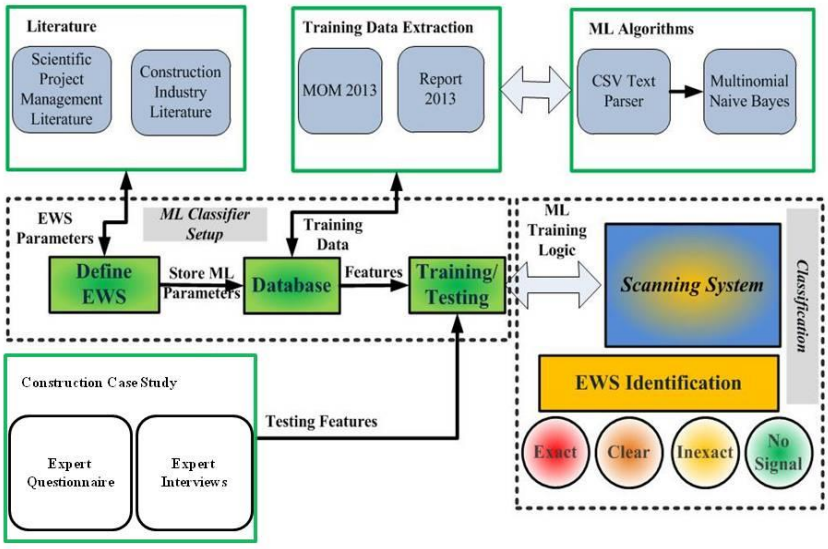


Figure 8: A design science approach based research framework for early warnings assessment approach

The first category computes the likelihood of each sample project file to belong to the following five classes. The factors or classes used to evaluate this model are as follows:

- $R_{ks(k)}$ – Lack of team requires knowledge and skills
- $k_{\boldsymbol{ms(k)}}$ - Lack of keen commitment to project milestones and scopes
- $M_{\boldsymbol{p(k)}}$- Lack of making purchases
- $M_{o(k)}$ - Lack of onsite materials
- $M_{r(k)}$ – Lack of resources

Where ($\boldsymbol{k}$) represents the group of data to which each class belongs

The posterior probabilities of these classes, given a data sample can be formulated as follows:

Posterior probabilities: **(1)**

$$p(R_{ks(k)}) = \frac{P(R_{ks(k)})p(K_{ms(k)}/R_{ks(k)})p(M_{o(k)}/R_{ks(k)})p(M_{o(k)}/R_{km(k)})p(M_{r(k)}/R_{ks(k)})}{evidence}$$

Where the evidence for equation (1) is the summation of all the prior probabilities for each class. Moreover, the evidence in this case can be ignored due to being a normal distribution constant.

Therefore, the Naïve Bayes probability model can be drawn as shown in (2) and stated in (Schmidt, 2001):

Naïve Bayes probability: **(2)**

$$P(R_{kS(k)}) = arg \max_{k\in\{1,\ldots,K\}} p(K_{ms(k)}) \prod_{i=1}^{N} p(x_i/k_{ms(k)})$$

As the proposed research is based on a multi-frequency count, it uses a multinomial Naïve Bayes algorithm where the feature vectors are represented

as word frequencies from a textual dataset with which specific events are generated as multinomial probabilistic models (Getoor, 2007; Kim & Chang, 2007):

Multinomial probabilistic models: **(3)**

$$P_1, P_2, \dots, P_n$$

In (3), p_i is the probability that k multinomial (due to a multi-class case) occur. Therefore, for each project plan text file in the proposed case, $\mathrm{T} = \tau_1, \tau_2, \dots, \tau_n$ be a text frequency histogram with $\boldsymbol{\tau_i}$ counting the number of times word (event) i was observed in a sample/training text file. This model is generally used in document classification and can therefore be used in text mining as well where events represent the frequency of occurrence of words in a single text file. Therefore, the likelihood of observing a word histogram $\boldsymbol{\tau}$ can be given as shown in (4):

Likelihood of observing a word histogram: **(4)**

$$p\left(\tau / K_{ms(k)}\right) = \frac{\sum_i \tau_i!}{\prod_i \tau_i!} \prod_i p_{ki}^{xi}$$

Based on (4), in log space, the multinomial Naïve Bayes classifier can be expressed as follows shown in (5) and (6):

Multinomial Naïve Bayes classifier I: **(5)**

$$\boldsymbol{log_p\left(K_{ms(k)}/\tau\right) = log_p\, C_{k(k)} + \sum_{i=1}^{n} \tau_i\, log_p\, ki}$$

Multinomial Naïve Bayes classifier II: **(6)**

$$log_p\left(K_{ms(k)}/\tau\right) = b + w_\tau^T x$$

4.1 Feature data formatting and labelling

In order to process unstructured data into a proper text classification system this methodology presents a multi-level approach where the initial raw data is taken

from project management files which are the minutes of meetings taking place during various stages of the project. The underlying rationale is to extract structured information which can then be used to train a machine learning model. The details of this methodology are further explained in Figure 9 which are given listed as follows as well:

- Unstructured data (PDF minutes of meetings)
- Data extraction
- Syntactic and semantic analysis
- Data labelling and training
- System modelling
- Unseen document classification

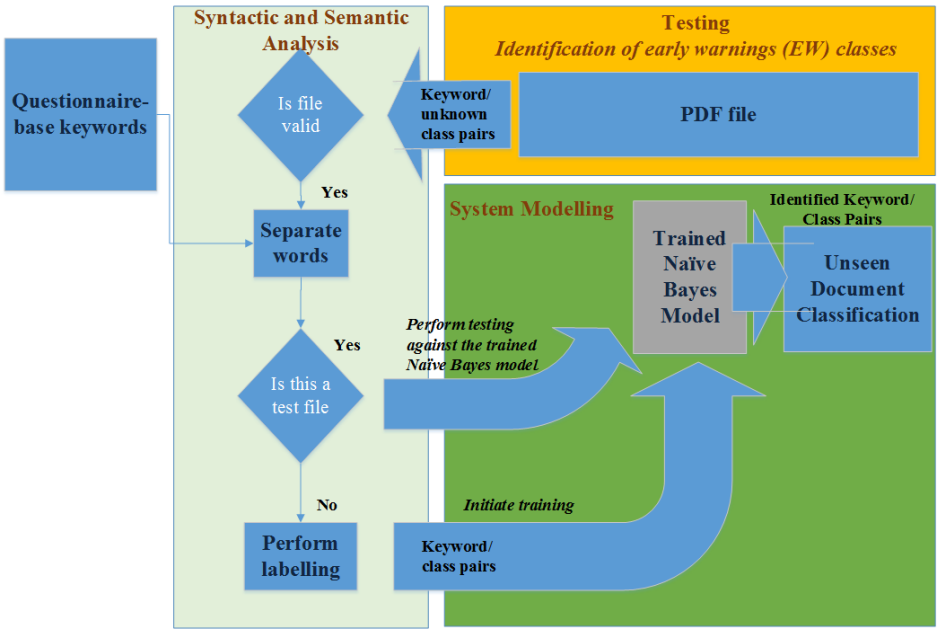


Figure 9: Core text mining stages in the proposed EW identification system

4.1.1 Unstructured data processing

At this stage, a large amount of raw data is obtained via a PDF parser programme. The program scan through all the PDF files which are text and image-based scans of minutes of meetings. It then discards the image-based documents as implementation of image-based parsing also known as optical character recognition (OCR) is not part of this research.

The test data is based upon 166 PDF project (MOM) files (1.7 GB), through which a total of 7470 item of text were extracted. The lines were then evaluated against the trained Naïve Bayes classifier for two separate classification categories.

The training data was extracted from comma-separated-value (CSV) files that contained, along with other indicators, the MOM text, project report and the early warning categories for each of these MOM text samples. The category trained from a set of universal “lacking” or “lack-of” indicators extracted from (training) data extracted from MOM files from the year 2013 taken from construction-related documents. This classification category focuses more on the textual context of the entire files. That is, each file represented a single, most influential impact on a project’s plan. Therefore, the main objective of this Naïve Bayes classifier was to read the data from each file and learn the hidden classification from a set of factors given below:

4.1.2 Data extraction: Information parsing

The text-based documents however are used to save sentence-level information. This information however, still contains a large amount of irrelevant information such as basic words (the, is, then, etc), page/section numbers and captions which must be removed to reduce computational overhead of the parsing algorithm. A straightforward approach would have been to use a distance matrix where simple Euclidean distances between various class clusters would be enough to identify the association of a test text string into one of the classes. However, it is unlikely to be an effective model due to cross-category occurring of the so-called common words discussed above such as “the, a, an”. For a discrete word-level classification technique as discussed in this chapter later-on, a simple bag-of-word approach will work which can completely discard such common words based on the training database used. The approach, via Bayesian classification is discussed in this chapter along with any shortcomings that appeared due to the inherent Bayesian weaknesses as a result of over-reliance on the training data only.

4.1.3 Data preparation: Syntactic and semantic analysis

The data this obtained is then weighted via a frequency table based on each word's occurrence. This data is still unmapped as it is not associated to any of the early warning classes discussed earlier-on in this work.

4.1.4 Data labelling: Preparation of ground-truth for model training

Based on the quantitative responses of the questionnaire as shown in Table 39, at this stage, each word is associated to a certain class as extracted from the so-called bag-of-words database. This stage produces a large database of word-class pairs which are to be used for the training and testing of the underlying machine learning models.

4.1.5 System modelling: Training a model via a machine learning classifier

The training dataset for this research was labelled manually into a ground-truth database which was labelled into four of the stated classes i.e. no-signal, in-exact, clear and exact. This labelled data was used to train the Naïve Bayes classifier. The labelling was done by the author on the basis of qualitative research with field experts. A framework for the proposed system model is shown in

Figure **10**.

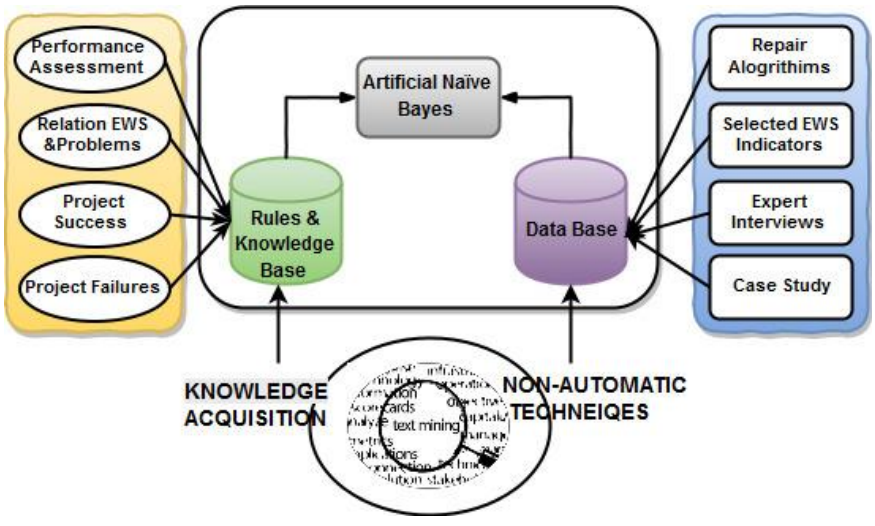


Figure 10: A hybrid Naïve Bayes classifier for early warning training and identification construction projects (Gajzler, 2010).

- **Prediction of textual classes**

The prediction of textual classes was based on a trained Naïve Bayes classifier which mapped the maximum likelihood data to each of the early warning classes.

- **Implementing a classification system**

The classification system was developed in Encog C# API which provides a number of machine learning algorithms to assist and extent with data and text mining techniques.

- **Feature extraction from construction management documents**

A PDF-parser engine was developed to scan documents from top to bottom with core keywords extracted. A possibility at this stage was to use an unsupervised (automatic) clustering mechanism to extract classes. However, this approach would still have relied on a manual labelling of each cluster. Hence, manual selection and labelling of EW features was used to train the model.

- **Improvement techniques via ensemble classifiers**

The possibility of integrating and extending this classifier can be explored. For instance, being discrete in nature, multiple Bayes models can be temporally connected and used to train an ANN classifier.

4.2 Naïve Bayes classifier for EW classification

Naïve Bayes classification is extensively used in text mining and Big Data analysis. The algorithm is well-known for its capability to identify the likelihood of patterns based upon the historic occurrence of previous data samples. Extensive work has been done in the test-based classification of images, document text normalisation, bug-report prediction and dictionary translation.

The majority of research at present focuses on learning from word frequencies or the so-called bag-of-word approach (L. Yuan, 2010 ; Lv & Liu, 2005).

The proposed methodology in this Chapter thus utilises the capability of Naïve Bayes to learn from word frequencies in specific documents to belong to certain early warning categories. The framework is further elaborated in **Error! R eference source not found.**. For instance, documents indicating the lack of materials are likely to have words indicating inventory anomalies such as “delayed purchases” or “pending financial approvals”, etc. The technique is further explained in the next section.

- **Step 1:** Conduct questionnaire-driven qualitative analysis to extract terms/keywords indicating early warnings and their association with certain early warning classes
- **Step 2:** Use the corpus obtained in Step 1 to label meeting Minutes documents
- **Step 3**: Based on the ground-truth documents obtained in Step 2, train a Naïve Bayes classifier
- **Step 4:** Divide the labelled data into training and testing sections and evaluate the classifier for various “early warning” classes obtained in Step 1
- **Step 5:** Assess the algorithm outcome based on True Positives (TP) and Negatives TN

4.2.1 Naïve Bayes model for extraction of early-warnings

In construction project management, a large number of documents are prepared during projects’ lifecycle to track, plan and control the progress of ongoing work. The text data presented in these document contain a cause-and-effect relationship between certain keyword combinations that can be learned and then used to train a model. The model can then mine similar information from unclassified/unlabelled construction documents and identify early warnings hidden within these documents. For instance,

1. The text “unavailability of the required human resources” can indirectly predict the “lack of required human skills” early warning for a project.

2. This early warning can be used to pre-emptively provide additional staff at the early stages of a project to avoid a project delay

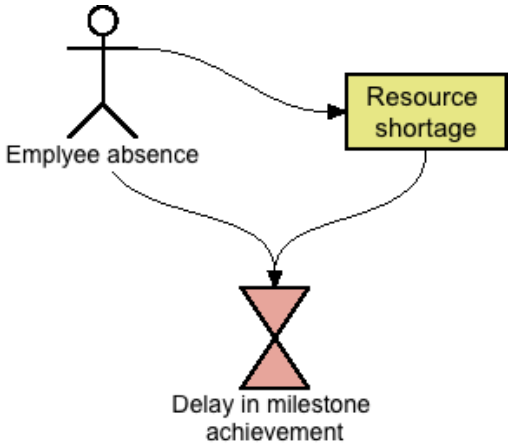


Figure 11: Bayesian belief network for a simple early warning identification

Table 9 shows an example probabilistic truth table where the possibility of two case combinations are shown along with their likelihoods. For instance, the likelihood of employee absence and resources occurring at the same time is generally a very unlikely case. However, such associative probabilities can be adjusted based on model training data containing employee absence and resource shortage statistics.

Table 9 : A probablistic truth table showing initial probabilities of a model to have chances of two aspects (employee absense and resource shortage) occuring in various combinations

Employee absence	Resource shortage	True	False
F	F	0.02	0.98
T	F	0.56	0.44
F	T	0.39	0.61
T	T	0.99	0.01

The technique draws its concept from the baseline Bayesian belief network that identifies various word-to-word relationships. As shown in Figure 11, for a simple two-event case, the outcome class “Delay in milestone achievement” in a project may be due to two reasons being either the employee being absent or resource being short. Each state will have a different probability to have resulted in the outcome class. The situation can be modelled as a Bayesian network with two probable values of truth attributing to the delay happening and false being not happening. This joint probability function for this case can therefore be defined as shown in equation (7):

Bayesian network: **(7)**

$$P(\gamma, \delta, \kappa) = P(\gamma|\delta, \kappa) \times P(\delta|\kappa) \times P(\kappa)$$

Where the **γ, δ** and **κ** stand for delay, employee absence and resource shortage. This model is capable of answering a question:

“What is the probability of delaying a project when as a result of resource shortage?”

Based on the conditional probability formulation, the anatomy of testing a simple statement from a project document for impending project failure can be done as given in Equation (8):

Table 10: Measures of various (supposed) statistical values depicting defining occurrences of project delays due to various project lifecycle issues.

	Project Delay of 1+ month (A) (35%)	No Project Delay (Not A) (65%)
Employee absence $\boldsymbol{X}$: Manpower below the stated requirement	66%	7.3%
Resource shortage $\boldsymbol{\kappa}$: Urgent materials not in inventory	10%	92.6%

Probability function: **(8)**

$$P(A|X) = \frac{P(X|A)P(A)}{P(X|A)P(A) + P(X|not\ A)P(not\ A)}$$

Equation (8) can further be defined as:

$P(X|A)$: Chance of employee absence X in case of a project delayed by 1+ month A (In this case 66%)

$P(A)$: Chance of employee absence A (In this case 35%)

$P(not\ A)$: Chance of not having even one employee absent $A\sim$ (In this case 65%)

$P(X|not\ A)$: Chance of project delay given that you did not have any employee absent (7.3% in this case)

Naïve Bayes rules: **(9)**

$$P(A|X) = \frac{0.66\ \times 0.35}{(0.66\ \times 0.35) + (0.073\ \times 0.65)} = \frac{0.231}{0.7055} = 82.95\%$$

Therefore, based on the delays given in row 2 Table 10, the chance of delaying project by 1+ month $P(A|X)$ given employee absence X to be 66% calculates out to be 82.95% per the Naïve Bayes rules as calculated in (9). Using this concept, the proposed methodology trains a Naïve Bayes model on the basis of an expertly-fed word corpus against a set of early warning classifications as described in the latter sections.

4.2.2 Data corpus preparation and modelling

The model training part of this work forms its basis from information gathered via expert input. An initial text extraction mechanism based on the following three mechanisms:

- A generic Naïve Bayes classifier based on document parsing from management documents (Alsubaey, et al., 2015).
- A generic Naïve Bayes classifier based on expert-defined ground-truth keywords/class combination.

- An extended Naïve Bayes classifier based on labelled Minutes document database.

Table 11: Elaboration of the sample size used for ground-truth training data preparation from user questionnaires

Sample item/answer	**Reductions in the quality of the constructed facility**
Professional association	Project manager
Experience in field	Up to 20 years

A detailed data parsing procedure is given below:

Step1: Data cleansing for word association extraction

In standard word documents, each word can be associated to a range of meanings depending upon its context as shown in Table 11. For instance, particular words such as “low”, “shortage”, “less” can generally be attributed to a situation indicating a negative meaning from a sentence. However, the documents often contain word associations or tuples, which provide better context when used in pairs or rows. Once such example of an item with a description and expected due data is shown in Figure 12 where a due-date earlier than the current date for a subject stating word pairs such as “outstanding issues” indicate an impending delay in project goal achievement.

Item	Textual Information ⇩ Subject	First Raised in Meeting N°	Due Date	Action By
4.2	**Workshop for Schedule issues :** - It is decided to conduct a workshop for sorting out the outstanding issues on the schedule. proposed that these workshops will start from next week onwards.	186	10-Mar-14 ⇧	
4.3	**Progress:** - Progress achieved in various areas are discussed and PMT expressed their concerns on the slippage from 7.2 targets. a) highlighted the gap of 23,000 m2 between the duct installed and duct insulated. b). Gap of 800 lm between the CHW pipe installed and insulated. c). Slow progress of fire fighting pipe installation.	187	Numerical Information	

Figure 12: Unstructured data - Difference of textual and numerical information and their complex association in project management documents (Ithra, 2009)

Construction project management documents are unique in a sense that the language used is generally different from standard management documents. For instance, the documents use technical terms such as “letter of credit”, “bidding contract”, “list of requirements” and/or “change orders” which are frequently accompanied with meaningful numerical values such as dates, and quantities which provide additional information on delivery schedules, inventory statistics or other measurement parameters. All these values can be used in both negative and positive capacities and these parameters play a crucial role in such an assessment. For instance, a high number in a “change order” may mean an indication of lack of stable project requirements. This effectively changes the scope to text and numerical pairs to be modelled together as shown in Figure 12 where unstructured data from column two must be used in conjunction with numerical information from the Due Date column to provide a better assessment capability.

Step 2: Term/word extraction from management documents

The most important step in the model training is to convert unstructured data to structured information. In Figure **12**: Unstructured data - Difference of textual and numerical information and their complex association in project management documents Figure 12, row 2, column 3 contains free-form text under the “Subject” category that must be converted into meaningful keywords. In Natural Language Processing (NLP), term extraction tools such as ‘StanfordNLP’, MALLET, ‘OpenNLP’ or “Microsoft Sql Server Integration Services (SSIS)” are used for term lookup. However, the majority of these tools are limited to single words and do not extract wider, sentence-level contexts. In order to train a viable model to identify early warnings from similar management texts, an expert-driven word corpus was created to train the Naïve Bayes model for early warning identification. A hierarchical PDF-parser was developed to extract nouns, verbs and adjectives from sentences extracted from Minutes meeting documents while eliminating unnecessary articles such as “the”, “a” and “then”. The terms thus extracted from Figure 12 are shown in Table 11 Sample extraction of early warning terms from a Minutes document shown Figure 12

with the occurrence of each depicted as a Score (Column 4). The Item “Id” is the serial-numbered items discussed in Meeting Minutes. Terms are extracted as nouns, adjectives and verbs from parsed sentences. For instance, for the sentence “slow progress of firefighting pipe installation” will generate three verbs uniquely expressing actions or occurrences that are “progress”, “firefighting” and “installation”, with adjective “slow” used to further indicate the negative nature of the case made for the associated noun “pipe”. Hence, in Figure 12, is transformed into structured data as shown in Table 12:

Table 12: Sample extraction of early warning terms from a Minutes document shown in Figure 12

Item Nu#	Terms	Type	Score	Date
4.2	Outstanding	Adjective	1	10-Mar-14
4.2	Issues	Noun	2	10-Mar-14
4.2	Schedule	Noun	2	10-Mar-14
4.3	Progress	Verb	3	10-Mar-14
4.3	Expressed	Verb	1	-
4.3	Concerns	Noun	1	-
4.3	Gap	Verb	1	-
4.3	Fire fighting	Verb	1	-
4.3	Pipe	Noun	1	-
4.3	Installation/Installed	Verb	2	-
4.3	Slow	Adjective	1	-
4.3	Insulated	Adjective	1	-

Step 3: Analyse word association in management documents

The overall meaning of a sentence structure will broadly depend upon the entire neighbourhood of the sentence. For instance, “required skills” will indicate a shortage of an aspect necessary for a project. On the other hand, “there is no shortage” actually demonstrates the availability rather than unavailability. Therefore, the proposed technique draws on a word-pair model where it first looks into a basic word classification and then creates a sentence-level word map to improve the overall context.

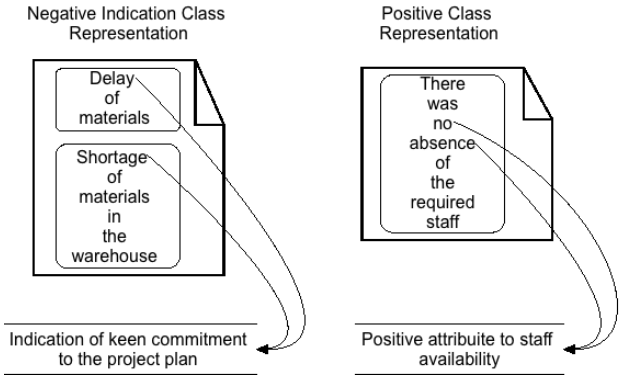


Figure 13 : A sentence-level word map representing two different classes

As shown in Figure 13, a frequency model can be derived to demonstrate sentence-level word-pair class association. The association can be taken as a bag-of-words type of model where each bag would representation a certain class as shown in Table 13 and the Appendices Table 39.

Step 4: Modelling data preparation for Naïve Bayes training

In order to prepare a Naïve Bayes early warning identifier, a text tagging system for PDF documents was necessary to establish usage relationships between certain words and terminologies used in meeting Minutes, drafts, memos and other documents. The initial term-to-early-warning relationship was developed based on a set of 20 questions. This survey was posted to individuals belonging to individuals with experience ranging from 5 to 20 years in construction project management domain. A total of 13 individuals responded to these questions

with 15 questions shown in Table 39 with the resultant potential impacts cited in Table 14. The sample responses from random respondents, their extracted terms for model training, scores and the relevant class (risk/warning) are shown in the last column. Respondents were allowed to state any class to the questions which were then manually filtered for minor deviations in the meaning and generated a total of 12 unique early warnings as follows in Table 13:

- Keen commitment to project scope **(KS)**
- Key management support **(MS)**
- Manpower resource **(MR)**
- Team required skills **(RS)**
- Stable project scope **(SS)**
- Team's required knowledge **(RK)**
- Keen commitment to project milestone **(KM)**
- Stable project responsibility **(SP)**
- Making purchases **(MP)**
- Stable project milestone **(SM)**
- Materials on site **(MO)**
- Stable project requirements **(SR)**

Table 13: A selected list of questions used in the interview to identify text trends indicating early warnings in project management documents

Selected Questions	**Sample answer from one of the respondents**	**Extracted terms for Naïve Bayes training** ***Term (Score)***	**Class identified by the expert** **(Class code)**
How do you think specific staff delay affects different aspects of a project? Please name various	delay in finishing the tasks assigned, delay in response and	VB: Delay (3), VB: Finishing (1), VB: Assigned (1), VB: Response (1),	Keen commitment to project scope (KS)

types of delays?	not attending meetings	AD: Not Attending (1) CN: Tasks (1), CN: Meetings (1)	
How supply chain problems can be identified in management documents?	untracked long lead item	VB: Untracked (1), AD: Long (1) CN: Lead (1), CN: Item (1)	Key management support (MS)
Explain the impact of lack of scope of personnel representing the managements to have to the overall efficiency of the project?	idle workers whereas the shortage of useful staff generally lead to tasks not completed on time	VB: Idle (1) CN: Workers (1), VB: Shortage (1), AD: Useful (1) CN: Staff (1), CN: Tasks (1), VB: Not completed on time (1)	Manpower resource (MR)

Table 14: Indicating associations between certain questions

Questions	Keywords	Impact
Key management support	Frozen action	Delay management and engineering decision have a more serious effect on the planning and design aspect of the project
Keen commitment to	Inconsistent supply chain hold-ups	There may be hold-ups causing surpluses or shortages of

project milestone		materials due to unmanaged supply chain issues
Keen commitment to project scopes	Inflexibility in how to control their projects	Sponsor with unclear expectations and schedules
Stable project requirements	Contractor asks about things that are explained in the bid documents	Discrepancy in the project requirements
Team required knowledge/ skills	Repetitive disapproval or rejected	delay in document issuance and lack of properly documented guidelines
Materials on site	Poor inventory control	L/C and down payment excessive overtime
Manpower resource	"Mr X has taken leave of absence due to no reason	Negative impact on the delivery of the project so a replacement resource is required"

4.2.3 Controlling the independence assumption

The independence assumption here forms the basis of how much the underlying Bayes model depends upon other variables. In the current case, each classification is only dependent upon the occurrence of identify keywords and their frequencies. For instance, the occurrence of keyword combinations such as “inventory delay”, “skill lack” or “inability to reach milestones” relate to how the document is classified within certain categories. This characteristic however indicates a weakness in a discrete word-based Bayes classifier which is the occurrence of common words such as “the”, “they”, “a” and “an”. In the currently presented case, these are eliminated via a data cleansing approach. An extended approach to reduce the weights of such words is proposed in the next chapter.

4.3 Core Naïve Bayes algorithm

Table 15: Example 'n' keyword used for EW modelling for a reduced 2-class case

Keyword Group '1'	Keyword Group n	Classes
Contract	Pending	MS
Untracked	Lead	MS
Workers	Shortage	MS
Drawing	Meeting	MS
Unfinished	Pending	MS
Increased	Requirement	RS
Issuance	Pending	RS
Lack	Absence	MS
Failure	Restrictions	MS
Overdraft	Finance	MS
Serious	Change	MS
Unwarranted	Absence	RS
Continuous	Pending	RS
Inventory	Shortage	RS

Table 15 shows a 2-class EW identification case for the sake of simplicity as follows:

$$P(MS) = \frac{9}{14} = 0.6428$$

$$P(RS) = \frac{5}{14} = 0.3571$$

Based on Table 16, a frequency table is then developed the details of which are given below:

Table 16: Example frequency group calculation for Keyword Group 'n'

		EW	
		MS (9)	RS (5)
Keyword Group n	Pending	2/9	2/5
	Shortage	1/9	1/5
	Absence	1/9	1/5

Table 17: Naïve Bayes Likelihood table for Keyword Group 'n'

		EW		
		MS (9)	RS (5)	
Keyword Group n	Pending	$P(x/c) = P\left(\frac{Pending}{MS}\right) = 2/9$	$P(x/c) = P\left(\frac{Pending}{RS}\right) = 2/5$	$P(x) = P(Pending)$ 4/14
	Shortage	$(x/c) = P\left(\frac{Shortage}{MS}\right) = 1/9$	$P(x/c) = P\left(\frac{Shortage}{RS}\right) = 1/5$	$P(x) = P(Shortage)$ 2/14

	Absence	$(x/c) = P\left(\frac{Absence}{MS}\right) = 1/9$	$P(x/c) = P\left(\frac{Absence}{RS}\right) = 1/5$	$P(x) = P(Absence)$2/14
		$P(MS) = 9/14$	$P(RS) = 4/14$	

Table 17 shows the posterior probability $P(c/x)$ calculation which, for the class RS is calculated as $P(RS/Pending)$ which is defined as the probability of class RS in the presence of the Delay keyword as follows:

$$P\left(\frac{MS}{Pending}\right) = \frac{P\left(\frac{Pending}{MS}\right) * P(MS)}{P(Pending)} = \frac{0.222 * 0.6428}{0.2857} = 0.499$$

Hence $\mathrm{P}\left(\frac{\mathrm{MS}}{\mathrm{Pending}}\right)$ presents the Naïve Bayes probability calculation which is also describe as the probability of a certain class (RS in this case) if the keyword Delay is found in the document. Hence, based on the probabilities of all the keywords, the likelihood of any class can be calculated.

4.3.1 Evaluating an optimal Bayes classifier

In ideal circumstances, Bayes classifier relies on the occurrence of specific likelihood signatures trained by the occurrence (frequency) of various keywords. An optimal Bayes classifier should therefore generate minimal false positives.

4.3.2 Extending the Naïve Bayes classifier

A detailed set of results are presented in the Results and Findings chapter. The technique predominantly utilises discrete words and their association against a manually developed dataset labelled from a questionnaire response from experts. The technique however is dependent upon common word (data cleansing) discussed in Section 4.7.4. Hence, the classification relies on separate words where different the words contained context cannot be made based-sentence. For example, the statement “no delay in delivery of materials”

and the delay is not with the delivery if materials but there are no materials on site" may be classified into a signal classification. The first type indicates a classification into "no delay" and "no inventory" whereas a classifier trained on individual words may not give due importance to the place where the word "no" exists. Hence, the Bayes classifier must be extended to a non-discrete and continuous (time-series) fashion capable of extracting sentence-specific information.

4.3.3 The model training framework

The overall Naïve Bayes model presents an early-warning analysis framework. This framework using a training corpus to associate probabilistic mapping to each of the word appearing in a text document. Based on the cumulative likelihood of these words, the association of each file is generated as shown in Figure 14.

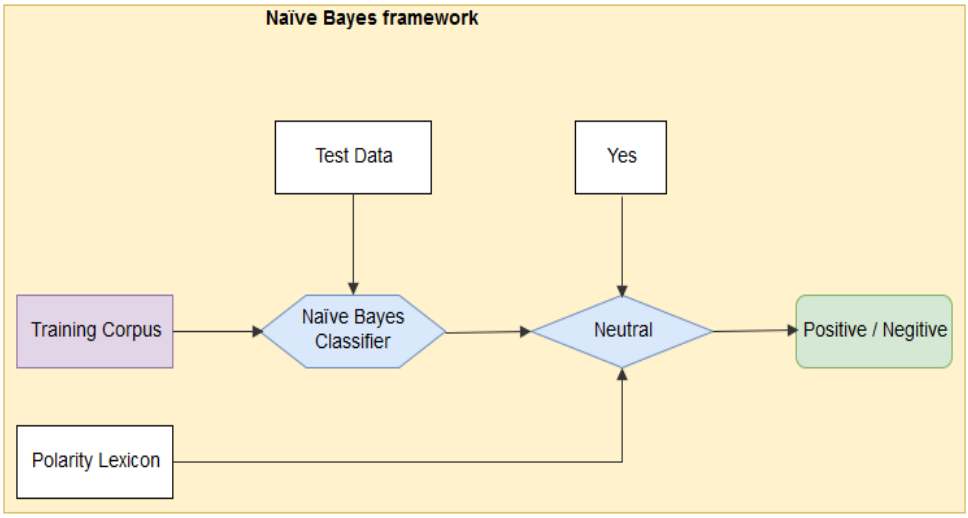


Figure 14: Framework of Naïve Bayes model for early warning modeling

4.3.4 Algorithm complexity

The baseline algorithm proposed in this research conforms to a linear computation time as the core of it performs a linear search for the keyword parsing section which has a O(1) complexity. This is since the overall data structure is organised as a C# Dictionary which is essentially a doubly linked list implementation. The cross-keyword Naive Bayes classification algorithm has a

complexity of O(m) type where ‘m’ is the number of keywords being cross-matched. It basically builds a map of counts represented by ‘m’ where:

Algorithm complexity: **(10)**

$$m = \sum_i nonzero(x_i)$$

In the above equation, $nonzero(x_i)$ is the number of nonzero features in the element i.

4.4 Summary

The technique addresses an information mining approach from project management documents to extract early warnings via a supervised Naïve Bayes machine learning algorithm working on the principles of probabilistic mapping of class feature vectors from outputs. The underlying approach targeted two core problems in construction data mining:

1. Identifies early warning on individual word usage and then comes up with a cumulative probabilistic distribution for each class. The class with highest probability is then selected as the identified category.
2. Identifying hidden word-pair relationships to improve sentence-level risks. This is where two words are paired due to their intrinsically similar association. An example is the word “does not” where each of these two may be eliminated during the data cleansing process however when used together, they present an important terminology.

The approach is novel in a sense that it improves on discrete word-level early warning extraction based on a supervised, expert-labelled project management document database. The approach was evaluated against 12 unique aspects that indicate early warnings including lack of management support, requirement delays, skills & material shortage and planning weaknesses. The underlying principles of extracting associated data-pairs was modelled via keywords extracted through qualitative information from an expertly guided questionnaire.

An extension to this work is presented in the next chapter where sentence-level warnings on two separate categories of document classification is presented and analysed. A more direct approach to evaluate common words and relevant/uncommon words under various weights to minimise the bias they induce in raw document classification systems is proposed in the next chapter.

5 CHAPTER V: A “weak signal” identification system

This chapter forms a basis to present the later part of the design of an intelligent signal identification system to warn of any discontinuities or anomalies indicating potential future problems within a construction project. The methodology forms its basis on little or no information on the project itself. Hence, it is necessary to gather as much information as possible on the normal state of the project. Nonetheless, in order to provide a baseline to train an AI algorithm for such a problem, a qualitative assessment was deemed necessary. Qualitative analyses are generally performed to extract information about any field of research. In the current, this process is used to extract knowledge on how certain texts or keywords within construction documents indicate a hidden problem. In order to implement this approach to predict with high reliability and minimal false positives, a K-nearest neighbour unsupervised clustering technique was proposed and evaluated. This chapter presents a detailed and critical analysis of this technique.

5.1 Qualitative analysis of construction early warnings

Based upon the study presented, the core challenge had been to understand and ascertain the ground-truth of what certain keywords and sentences within a document lead to various signals types potentially indicating any risk. The risk in this context is termed as a condition present within the project, as evident from its management documents that may lead to a discontinuity, failure or just delays.

The best reference to be used to train an AI model so that it identifies such “weak signals” is to learn from previous documents in similar contexts. However, understanding the association of a document to a certain “weak signal” cannot be judged by an average person. Hence, the proposed work uses semi-structured interviews as a mean to label certain areas (e.g. keywords and sentences) in management documents. Based on the literature survey in **Table 18** presents the set of critical factors that contribute to various project

failures. There is a wide range of clustering techniques used in the literature depending upon the type of clustering used.

Table 18: Keyword-to-signal association drawn via a set of questionnaires

Variance	Questions	Keywords	Relevant "EWS"	Relevant "weak signals"
-2.30	What effect does activities of PO placement issue under signature or processing not being met on time has on the continuation of a project?	Purchase order unsigned yet due to quotation review	Lack of making Purchases	clear signal
- 7.91	Please state the reasons of quality testing failures and the factors that play any significant role in such failures e.g. employee skills, experience or knowledge	QA / quality assurance disapproved	Team Required skills	exact signal
-13.77	Please state which activities and how their repetition affects a project's successful completion	Compliance not handed over yet	Commitment to project milestone	exact signal

-4.24	Please state how requests such as that of trainers invited to assist workers-on-site have any positive or negative impact on the progress of a project	Need special experts	Lack of Manpower	clear signal
- 13.23	Do you think a change in orders has an impact on the underlying contract conditions	Large change order can disturb the project progress and can be indicated via the following keywords: Change, altering, update, extend and/or delivery schedules,	Stable project requirements	exact signal
-9.38	What impact do you think experienced to top staff absence (e.g. on leave without hand over) has on the project's progress?	It might take a long time for other staff to take over but there is also a big role for the management to control/support	Key management support	Clear signal
2.2	state any	Must be	Lack of stable	No signal

	mechanisms that may support construction activities such as material and equipment e.g. any standards that may be required to create sound architecture specifications	standards to ensure lessons learned are recorded and briefed in future to prevent any mistakes from being repeated	project responsibility	

5.2 Handling of cross-class common words

As discussed in the earlier chapter, the most common problem encountered while training text-classification system is the fact that training data for each class contains an unlimited number of so-called common words such as “the, their, an, in, for, from, a, to, is”. The fact that these words can appear without any restrictions in any class’s training data, severely undermines the ability of a classifier to differentiate to completely separate text-groups to two distinct classes. For KNN, which is the predominant technique used in this chapter, the centroid of each class depends upon the weight of each word present in a cluster. Hence, an unhandled occurrence of similar common words in each cluster would significantly create a cross-class bias (Bijalwan, et al., 2014). Simple Euclidean distances are hence unlikely to be effective to model class boundaries and a much higher representation such as a Vector Space Model (VSM) should be used to handle the problem of common words. In VSM, each document is represented as a string of words as given in (10):

Cross-class common words: **(11)**

$$Count\ (W_1), Count\ (W_2), \dots, Count\ (W_N)$$

In (11), ‘N’ indicates the N^thword in a sequence of words present in a single group. For a number of groups like these, each word is given a certain weight depending upon its importance in that group. For instance, the work “skills” is given more weight compared to any of the common words discussed above. The technique is already reported in the work by Gowtham et al (2014).

5.3 Implementation of the TF-IDF model for word numerical representation

The term TF-IDF stands for term frequency-inverse document frequency model which basically provides a degree of importance to each word. The proposed approach forms a term-document matrix where the number of documents are equal to the number of classes representing various early warnings with each document containing N terms. The outcomes are plotted as a matrix where each word in the matrix corresponds to the importance (or weight) of that word in the respective document. These weights are represented as numerical values where for a document δ and a term τ the underlying TF-IDF model can be elaborated as formulation (12) and stated by Gowtham et al (2014):

TF-IDF model **(12)**

$$M_{TF-IDF}(\tau, \delta) = tf(\tau, \delta) * idf(\tau)$$

In (12), $tf(\tau, \delta)$ represents the frequency of term τ and $idf(\tau)$ represents the inverse document frequency. Hence, the term frequency can be defined as shown in (13):

Term frequency: **(13)**

$$TF(\tau) = {frequency(\tau, \delta)}\Big/{\sum \tau}$$

In (13), the term $frequency(\tau, \delta)$ represents the frequency of the term τ in the document δ and $\sum \tau$ represents the total number of that term in the same document.

For example, for the following sentence:

"*This is a difficult theme to process in a long text present in **the** engineering document whereas **the** other sentences are a lot simpler. **The** document is not simple enough*"

For the above example, the TF of term "the" can be calculated as shown in (14) and (15) for two different words and their resultant, calculated value:

TF of term "The": **(14)**

$$\boldsymbol{TF(\tau = \text{"}the\text{"}) = frequency(\tau = \text{"}the\text{"}, \delta) \Big/ \sum \tau = 3 \Big/ 30 = 0.1}$$

TF of term (different words) **(15)**

$$TF(\tau = "engineering") = frequency(\tau = "engineering", \delta) \Big/ \sum \tau = 1 \Big/ 30 = 0.033$$

Table 19: Representation of Augmented Frequency value for each word in the Minutes of meeting

	1	2	3	4	5	6
The	*					
A	*					
An	*					
Contract			*			
Skills						*
Management					*	

5.4 KNN signal identification technique

As this research focusses on word combinations that form vectors, KNNs are well-known at differentiating between such vectors or clusters via a norm function to calculate distance. Hence, the proposed technique uses KNN as the grouping methodology to isolate various sections of a document into certain classes with each having a unique signal indicator value.

The objective of the clustering algorithm is to partition n-objects into k-clusters. For the case of weak signals, each document is divided into ‘k’ clusters of distinction based on the frequency of warning words present within the document. As the frequency is not known in advance, the best number of clusters is also not known a-priori. The objective ultimately is to minimise the intra-cluster variance.

KNN steps for document signal identification as follows and shown in Figure 15:

- Process Minutes
  - Process words into frequency weighted (TF-IDF) bags
  - Categorise weak signal type
- Assign to experts for labelling
- Perform clustering based on KNN
- Take unseen Minutes documents
- Search for keywords saved in the TF-IDF bag-of-words corpus
- Categorise the document’s signal type based on its nearest classification cluster

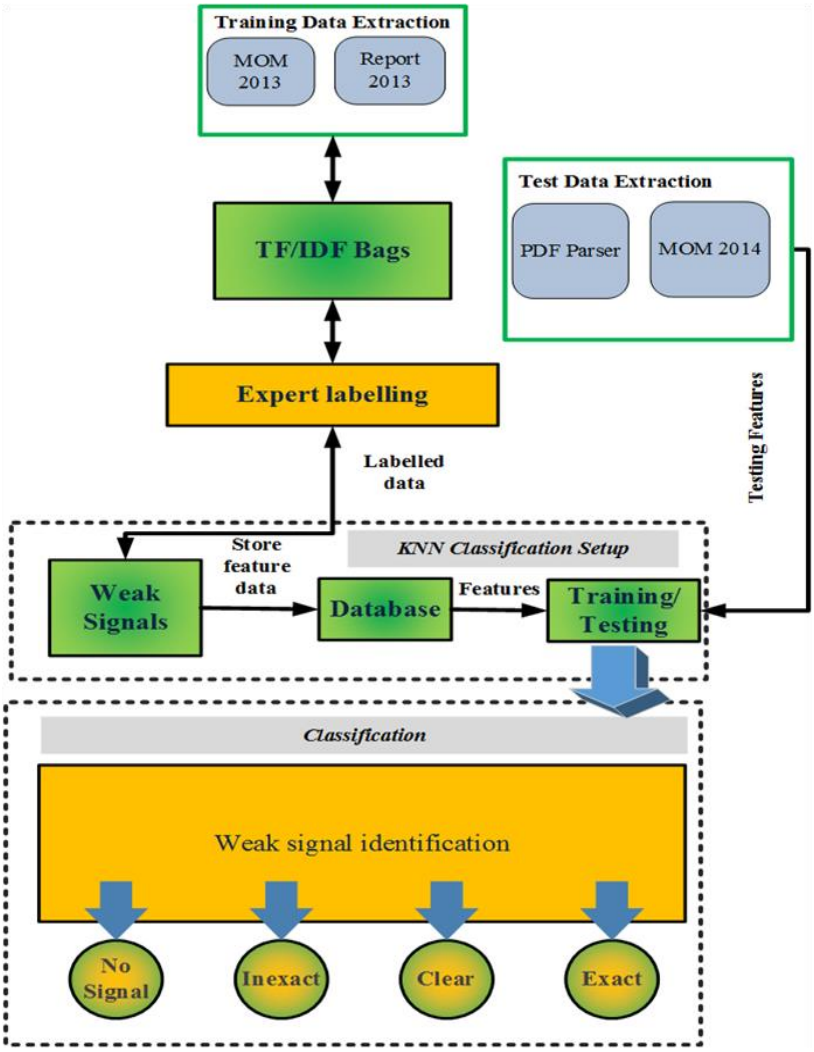


Figure 15: KNN framework for document signal identification

5.4.1 Feature selection with KNN

The proposed approach in the presented section aims to find unique clusters of keywords representing various levels of signals as listed and classified below based on Nikander's work (2000):

- **Exact warning:** This category will precisely indicate each warning with clear numbers.
- **Clear warning:** This category will identify warnings via textual indicators but it would not be possible to give a number to those warnings
- **Inexact warning:** This is an uncertain warning level which will draw from keywords subtly indicating problems where the source is detectable but the information is very inexact

- **No warning:** In the case of a normal project running condition, ideally, there should not be a warning at all as represented by this class

In the case of management documents particularly reports and Minutes, any updates are generally accompanied either by values or simple statements. Each statement made tend to represent its source which is a keyword represent an important issue, resource or event. For instance, a keyword “staff” is an important resource but in itself, it cannot be understood to carry any hidden signals. However, when combined with a verb which is termed as an action, state or occurrence, it gives a completely different meaning to the entire situation as shown in Table 20.

Table 20: Table representing various levels of “weak signals” in documents described based on availability of data

Example 1					
Planned delivery	Actual Delivery	Expected Delivery Date	Actual Delivery Date	Signal	Class
42	38	5	6	Exact Warning	Lack of skills
Example 2					
Total staff required	Actual Staff available			Signal	Class
13	11			Exact warning	Lack of staff
Example 3					
Total staff required	Actual Staff available	MOM Text		Signal	Class
-	-	Staff shortage, lack of availability		Clear warning	Lack of Manpower
Example 4					
Total staff required	Actual Staff available	MOM Text		Signal	Class
-	-	Staff is looking for a holiday		Inexact warning	Lack of keen commitment

Table 21: Task delay representation used to calculate the WS-Index (16) as shown

Delay in delivery time – Per day						
	0 task delay	1 task delay	2 task delay	3 task delay	4 task delay	5 task delay
Task 1: PO Placed	1	1	1	1	1	1
Task 2: Mobilization	1	13	13	13	13	13
Task 3: Earthworks	-4	-4	11	11	11	11
Task 4: Concrete	-3	-3	-3	-3	-3	-3
Task 5: Waterproofing	-1	-1	-1	-1	-1	-1
Task 6: Masonry	-1	-1	-1	-1	-1	7
Task 7: Mechanical	-2	-2	-2	4	4	4
Task 8: Electrical	-3	-3	-3	-3	-3	-3
Task 9: Finishes	1	1	1	1	5	5
Task 10: Structural steel	1	1	1	1	1	1
WS-Index	-4.77	4.90	56.11	72.85	86.73	111.80

For instance, a “staff” noun used with a verb “lack” could indicate a clear warning. However, based on Nikander’s representation, the verb lack cannot still be given a number and hence it is not an “exact warning”. The different situations representing these warnings are further elaborated in Table 20. The values at the bottom of each column represent the product of variance and

average for each column. This index is derived to show the tasks with highest delays to have highest index values due to their large variance.

Extending on Nikander's technique, an exact warning in the current work indicates planned and actual deviations in certain goals. For instance, as shown in Table 20, Example 1, a table in textual documents may indicate the number of days required for planned delivery and the actual delivery days taken by a supplier. Similarly, if the figures shown are not in number of days but in terms of variance, then a substantially high difference in a variable such as required resources, average number of days or patterns in late deliveries can be used. Most MoM documents in construction management use variance to represent anomalous behaviour as it shows inconsistencies in tabular data.

The proposed Variance index calculation shown in this table is meant to be used to measure the level of inconsistency in addition to any higher values being penalised (Chen & Luo, 2014). For instance, if there is a higher number of items continuously reported in the inventory, the index must show a higher penalty based on the following formulation in (17):

Variance **(16)**

$$S_{WS} = (\sigma^2 + 1) * \frac{\sum_{i=1}^{10} D_i}{N}$$

In (17), σ^2 represent the variance over a 10-day value period, $\mathrm{D_i}$ represent days and N is the total number of values. The outcome of the $\mathrm{S_{EW}}$ index is shown in Table 21 where it can be observed that an unused inventory generates a very high value representing a representative value for the "Exact signal". Similar case with "No-Signal" can be seen for "Daily manpower shortage" where the number of employees absent has never been more than 1 or 2 leading to a very small value. Similarly, for the time-dependent delay shown in the first column for "delay in delivery" there is only one class reported to have a large delay of 13 days whereas the remaining task were in-fact completed on time. The proposed index handled this very well as the overall value of 4.9 is substantially lower than any other categories. The case of "delay in delivery time" based on the

proposed WS-index was further investigated by inducing task delays as shown in Table 21 and Figure 16. Analysing the table, the first column only reports one major delay of 13 days, this increased to two separate delays of 13 and 11 days for Tasks 2 and 3 which increased the WS-Index from 4.9 to 56.11 which is a 10+ times increase as more than one task delay cannot be ignored as an outlier. This situation is further expanded for 3, 4 and 5 task delays with the index robustly indicating an increase to 111.8.

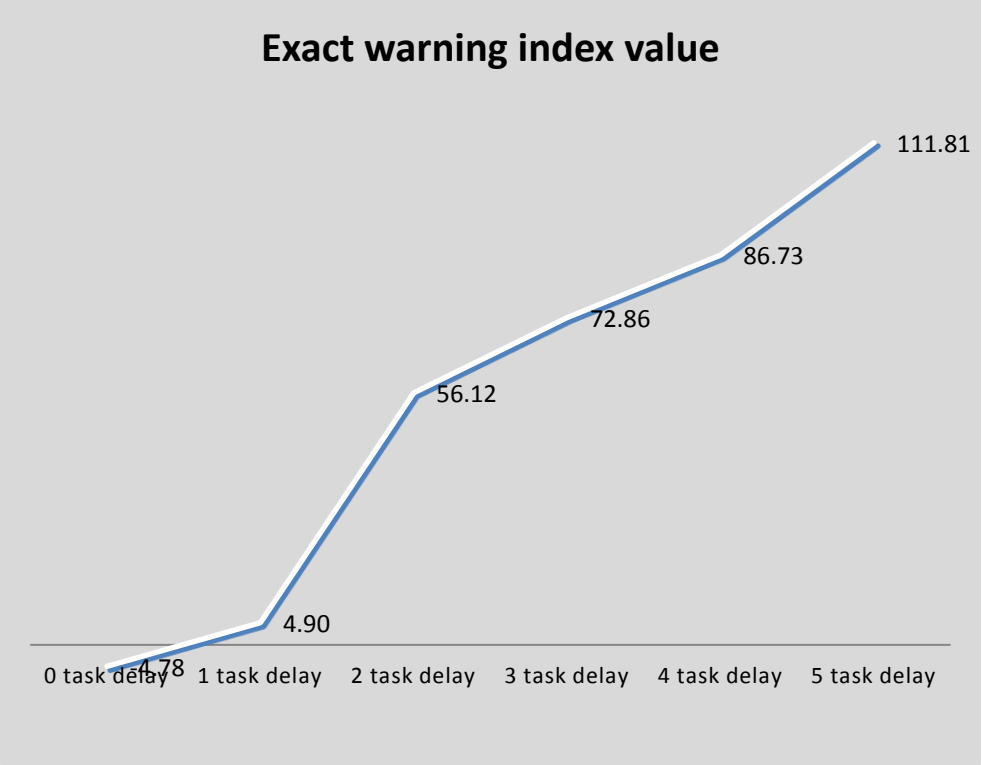


Figure 16: Exact warning index value increment as more Tasks in a project timeline report delays

5.4.2 Formulating Exact warnings from MoM data:

A “No-signal” case can also be seen in Table 21(0 task delay column) in the case of “delay in delivery time” class where either all the tasks were completed on time or before the stated deadline date represented as negative values. The resultant SW-Index hence shows a ‘-ve’ value.

5.4.3 Case of Inexact and Clear Warnings:

Clear warnings are identified merely by the presence of keywords that are semantically important; appear in the specific document and their TF-IDF frequency map. The closest keyword to the centroid of the relevant class cluster is identified as its class. However, since the word is not “paired” with a value. For instance, a delay keyword without a value, it is categorised within the Inexact category. On the hand, if the word/sentence belongs to a precise group such as “extreme shortage of manpower”, it is categorised in the Inexact classification. The overall outcome of TF-IDF-based frequency analysis is analysed and explained in the next “Results and Findings” chapter.

5.5 Core K-nearest-neighbour algorithm

The algorithm uses a distance function between two (or more) classes using a distance function $d(x, y)$ where $x\ and\ y$ are cases comprising of N features. Hence the distance features are calculated as follows:

KNN algorithm classifier: **(17)**

$$d_E(x, y) = \sum_{i=1}^{N} \left| x_i^2 - y_i^2 \right|$$

Based on the distance equation, the Weak Signal identification algorithm can be elaborated as follows:

- Store M Weak Signal classes in vector: $v_i = \{v_1, \ldots, v_M\}$
    - Go to case C^i in the word dataset
    - If C is not set or $c < d\left(v, C^i\right)$: $c \leftarrow d\left(v, C^i\right), t \leftarrow o^i$
    - Iterate until the entire dataset is scanned
    - Store v in vector c and t in vector r
- Calculate the arithmetic mean across r as follows:

$$r = 1/M(\sum_{i=1}^{M} r_i)$$

- Return r as the closest matching class in Weak Signal identification

5.5.1 KNN formulation example

In the case shown in Table 22, the KNN-based clustering can be used as follows as stated by (Harisinghaney, et al., 2014):

$$q_1 = \{KeywordGroup1 = Contract, KeywordGroup2 = Pending\}$$

The above equation is then used to investigate the nearest neighbours as follows:

Mapping the output values MS and RS to 0 and 1 would generate the output vector as follows:

$$r_{q1} = \frac{1}{M(\sum_{i=1}^{M}\{0,1,0\})} = 0.3333$$

Hence the result can be mapped to the real class situation as follows:

$$r_q = \{if\ r_{q1} \geq \frac{1}{2} then\ Yes\ else\ No$$

Table 22: Example 'n' keyword used for EW modelling for a reduced 2-class case

Keyword Group '1'	Keyword Group n	Classes
Contract	Pending	MS
Untracked	Lead	MS
Workers	Shortage	MS
Drawing	Outstanding	MS
Unfinished	Pending	MS
Increased	Requirement	RS
Issuance	Pending	RS
Lack	Absence	MS
Failure	Restrictions	MS

Overdraft	Finance	MS
Serious	Change	MS
Unwarranted	Absence	RS
Continuous	Pending	RS
Inventory	Shortage	RS

5.6 Summary

This chapter presents a detailed documentation of the TF-IDF KNN methodology for two different class-group-based text/data mining cases. The initial results showed a high cross-class confusion which lead to a high false positive rate. This was found to be due to ground-truth feature similarities as a result of common keywords between various classes. These groups were eventually combined via pairwise matching which resulted in a reduced number of classes. The outcome was evaluated against the four categories of weak signals. As further training and tests were tested with a reduced-class-model, a substantial increase in classification accuracy was reported which has been discussed in the next chapter.

6 Chapter VI: Critical Analysis of Results and Findings

The work approaches the problem of text mining in construction management in two distinct domains covered in Chapter 4 and Chapter 5. The chapters uniquely address issues of extracting various classes of “early warnings” (Chapter 4) and “weak signals” (Chapter 5) to forewarn project managers of impending project issues. This chapter provides a detailed and critical analysis of the result obtained while addressing the research methodologies presented in both the chapters.

6.1 Case study data employed to analyse results

To analyse the practicality and functionality of the proposed approach and design system, the approach is applied to one of the mega construction and landmark projects situated in Saudi Arabia sized (10000 m2). The *project* is complex, architecturally and *construction*-wise due to being subject to changes, performance outcomes, interdependencies, over disciplines and unstructured data sources. The project contains diverse cultural facilities, including an auditorium, cinema, library, exhibition hall, museum and archive. The auditorium seats 930 visitors and provides for a wide range of events ranging from opera, symphony concerts, musicals and lectures etc. The library is expected to become a centre of learning while containing some 200,000 books on open access and catering for all ages and categories of users. The project is addressed as the “construction case study” in this thesis.

6.1.1 Survey data preparation against EW model class categories

To train the model against the usage of specific terminologies, a set of 20 questions were prepared to help with mapping keywords against class models. A survey questionnaire was prepared containing 15 questions and their associated impacts on certain EW classes (Appendix A.2). Respondents could map any classes to the questions which were then manually filtered against minor deviations in the meaning. This step generated a total of 12 unique early warnings.

6.1.2 Outcome analysis of “weak signals” in MoM data

This section addresses the outcome of TF-IDF and KNN approach to extract and identify “weak signals” from the construction management MoM documents. To evaluate the effectiveness of the proposed methodology, a real-world case of “weak signal” identification was evaluated against the algorithm. In order to evaluate the performance, the trained algorithm was tested against selected MoM documents that actually had reports of eventual problems at the later stages of the project lifecycle. Hence, the case tested in MoM construction document had information representing various ongoing problems such as staff shortages, delivery delays or lack of resources reported. Hence, any such abrupt changes such as manpower shortages, delays represented in number of days, inventory surplus were penalised for higher values further shown in Table 20. The signals to be sought in these documents, as discussed in Chapter 5, were categorised into 4 classes with the case of the most relevant signal, the “Exact signal”, exemplified in

Table **23**. The table utilises a sum of variance and average as an exact calculation formulation which increases the penalty in direct proportionality as the all the factors report and increasing delay. The penalty is hence zero if there is not a single delay found in the entire document in any of the classes/categories. However, before performing an index calculation, as per the definition for exact signals, the information must be extracted from the underlying MoM files in a manner that the data is robustly paired with its associated value.

Table 23: Proposed “Exact signal” index calculation

	Delay in delivery time		Daily manpower shortage		Inventory surplus (number of extra items)
Task 1	1	Day 1	1	Day 1	1
Task 2	13	Day 2	0	Day 2	1

Task 3	-4	Day 3	0	Day 3	22
Task 4	-3	Day 4	0	Day 4	19
Task 5	-1	Day 5	0	Day 5	12
Task 6	-1	Day 6	0	Day 6	10
Task 7	-2	Day 7	1	Day 7	2
Task 8	-3	Day 8	2	Day 8	2
Task 9	1	Day 9	1	Day 9	2
Task 10	1	Day 10	0	Day 10	2
	4.90		0.75		472.95

6.1.3 Extracting and pre-processing datasets from management documents

The data from management document is extracted along with any numerical values of each MOM document. The numerical values from the MoM files are manually tagged as the columns and rows are presumed not to change during the entire project management session. This step is performed to make data extraction an easier task from the complex project management documents.

The approach does not use the data cleansing approach proposed in the previous chapter. Alternatively, as it employs a so-called TF-IDF approach to set importance and frequency-based weights, it manages to only extract relevant words that have a high weightage.

6.1.4 The proposed KNN algorithm for signal identification

The underlying principle of this extension is to use the unsupervised capability of this work to extend and identify hidden sentence-level patterns. The representation of each of these words in any document can therefore be done either as existing or non-existing Boolean identifier.

- Boolean representation of documents

Each document in a training or testing corpus can be identified based on a binary vector of words where the presence of a particular word is represented as “true” or 1 whereas the absence of that word is represented as “false” or 0. For instance, if it is presumed that the corpus contains the following 5 keywords as an example Table 24, then if a document contains only the highlighted words, the Boolean vector would be represented as the one showed in Table 24:

Table 24: Sample keywords assumed to be present in a document

Resources	Staff	Delay	Lack	Shortage

Table 25: Binary representation of the keywords shown in Table 24

Resources	Staff	Delay	Lack	Shortage
1	1	0	1	0
2^4	2^3	2^2	2^1	2^0
16	8	0	2	0

The binary representation thus shown in

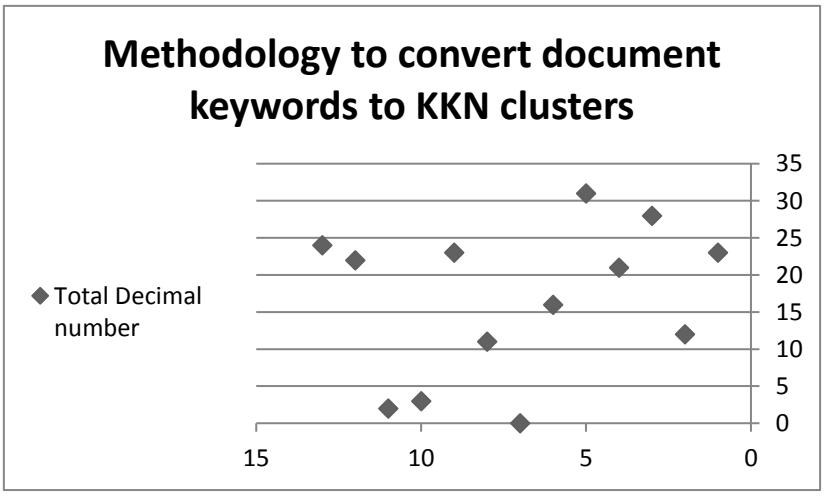
Table 25 will generate a decimal number 26. This number will represent a single sample representation in two axes plot as shown in Figure 17. Now, for a case of multiple documents, to calculate the cluster representation of each, can be seen in Table 26:

Table 26: A binary-to-keyword clustering technique extended to KNN classification

	Document number	Resources	Staff	Delay	Lack	Shortage	Total Decimal #
Training	1	1	0	1	1	1	23
	2	0	1	1	0	0	12

	3	1	1	1	0	0	28
	4	1	0	1	0	1	21
	5	1	1	1	1	1	31
	6	1	0	0	0	0	16
	7	0	0	0	0	0	0
	8	0	1	0	1	1	11
	9	1	0	1	1	1	23
	10	0	0	0	1	1	3
Test data	11	0	0	0	1	0	2
	12	1	0	1	1	0	22
	13	1	1	0	0	0	24

Figure 17: A technique to map document keyword frequency to various class-clusters to be manually labelled based on expert data.



6.1.5 Classification of Minutes files based on TF-IDF-KNN modelling

It must be noted that in the earlier Bayesian/Early Warning chapter (Chapter 4), the work predominantly focussed on common word elimination by selecting keywords to each class and eliminating remaining common-use words which, if exist, are likely to increase cross-class bias.

6.1.6 Assessment of initial 12 classes from the Bayesian model:

The initial tests for the KNN classifiers utilised a similar number of classes as used in Naïve Bayesian classification as shown in Table 27. The outcome used ground-truth based on keywords allocated to each class by experts via their questionnaire responses. A very large number of false or misclassified association were encountered during this run. The highest percentage of false positives originated from the "Keen commitment to project scope" class. The highest proportion of misclassification was attributed to 20% (5) classes incorrectly identified in the "Stable project responsibility" class and 16% (4) classes incorrectly identified towards the "Stable project requirements" class. When the latter two classified are analysed as shown in

Table 25, for the "Stable project responsibility" class, 15.78% (3) of 19 "MoM" files were miss classified as "Keen commitment to project milestones" and 31.57% (6) of 19 "MoM" files were incorrectly classified as the "Stable project requirements" class. Moving further ahead to the "Stable project requirements" class itself, an overwhelming 22.72% of all its files were miss-classified as "Keen commitment to project scope" class and 18.18% were miss-classified as the "Stable project requirements" class.

Similarly, the analyses for other classes also formed the so-called miss-classification clusters where a set number of classes were miss-identified between each other. When this issue was closely evaluated, it was understood that these classes contained many cross-existing texts. For instance, it was observed that expert inputs frequently used common terminology to represent similar classes. For instance, the classes "Stable project responsibility" and "Stable project requirements" both contained words such as assignments, allocations, milestones, deliverables and timelines. Unlike the very common words that were discarded via the "bag of words" approach, these words cannot be eliminated from any of these classes as that would create a substantial bias towards one class. Hence, it was decided to combine such classes where the keywords formed a union of at least 25% keywords i.e. 1/4th of all the words used between any two or more classes were found to be common. This did not include standard words which were already discarded such as 'a, an, the, their,

that, etc.’ .The closeness of files was measured by initially selected pairs and then adding all their representative keywords in a word-pool. Afterwards, if the total number of words from one class formed at least 25% of the entire word pool, the two classes were selected for merging.

Table 27: A confusion matrix representation cross-class false positives and the overall accuracy based on the 12-class model used in extended Naïve Bayes classifier.

EWS	(SS)	(MS)	(MR)	(RS)	(KS)	(SM)	(RK)	(KM)	(SP)	(MP)	(MO)	(SR)	Total files	Percent Accuracy
Stable project scope **(SS)**	11	0	0	1	2	0	0	0	5	2	0	4	25	44.00%
Key management support **(MS)**	**1**	7		0	0	0	0	1	0	0	0	**1**	10	70.00%
Manpower resource **(MR)**	0	0	19	4	0	0	0	0	0	0	0	0	23	82.61%
Team's required skills **(RS)**	0	0	6	16	0	0	13	0	0	0	0	0	35	45.71%
Stable project milestone **(SM)**	0	0	0	0	0	14	0	7	0	0	0	1	21	66.67%
Keen commitment to project scope **(KS)**	2	0	0	0	27	0	0	19	2	0	0	0	50	54.00%
Team's required knowledge **(RK)**	0	0	5	11	0	0	23	0	0	0	0	0	39	58.97%

Keen commitment to project milestone **(KM)**	0	0	0	3	7	10	0	19	0	2	0	0	31	61.29%
Stable project responsibility **(SP)**	3	0	0	0	0	0	0	3	7	0	0	6	19	15.79%
Making purchases **(MP)**	2	0	0	0	0	0	0	0	0	13	6	0	19	68.42%
Materials on site **(MO)**	1	0	0	0	0	0	0	0	0	7	11	0	19	57.89%
Stable project requirements **(SR)**	5	0	0	0	0	0	0	2	4	0	0	11	11	45.45%
													281	54.92%

The filtered numbers of classes were finally shortened to a total of 6 categories as shown in Table 27. The table also shows a marked increase in accuracy when these classes were evaluated against unseen model data which is further presented in

Table 28. Moreover, both the tables represent outcomes based on the TF-IDF approach that also assigns reoccurrence-based weight to each class. The overall improvement can be seen to only show a 36% inaccuracy in the “Commitment to project milestones” class. Since, the two classes “Stable project responsibility” and “Stable project requirements” were now merged and formed part of the class, the highest number of false positive originated towards the “Lack of making purchases” class which is only 18.18% or two text files. Similarly, the second highest false positive rate originated from the “Lack of making purchases” category where there was a 31% miss-classification towards the “Materials on site” class.

An overall assessment of these accuracies provides 54.92% accuracy for the 12-class case which improved to 72.78%. The highest classification accuracy was that in the “Lack of manpower” category where the overall accuracy improved from 82.61% to 83.33%.

The main challenge at this stage had been due to the extensive similarity of classes as can be seen in

Table 28. The overall accuracy shown in this table was noticed to be extremely low. However, the false positives had a clustering behaviour where many miss-classifications were actually concentrated into similar categories. For instance, in Table 27, for classe “Keen commitment to project milestone” towards class “Stable project milestone” and “Keen commitment to project scopes” both had a cross-class miss-classification of 32% and 22.58%. Similar patterns were also found in other classes.

Table 28: Accuracy of TF-IDF KNN approach against various classes

	Identification accuracy	Test files representing each class (as labelled by experts)
Keen Commitment to project milestones	64%	11
Key management support	70%	10
Lack of making purchases	69%	13
Lack of manpower	83%	6
Stable project scopes and requirements	73%	11
Team's required knowledge and skills	78%	9
	Total	60

6.2 Critical analysis of false positives and the confusion matrix

To understand the underlying reason, the representative keywords for all the classes as taken from the expert questionnaire responses were analysed. It was found that a large fraction of these classes had cross-usage of specific keywords. It was observed that several these classes had shared keywords which created a cross-class bias leading to an increase in cross-class false

positives and the subsequent inaccuracies. To improve the overall accuracy, it was hence crucial to combine the classes which the contextual similarities.

To reduce the overall cross-class similarity, the classes were cross-compared in a pair-wise fashion. Common word percentage for each class to a common word pool was the calculated as shown in next Table 29. In the table, row 1 and row 2 represent classes with common words. To compute the similarly of these classes, class matches were summed and then divided from the total number of words representing the actual pool. The percent similar can hence be seen in the last column where the total matching keywords i.e. 4 against the overall size of the pool i.e. 16 gives a similarly score of 25%. Similarly, scores for other cross-class similarly comparisons are calculated as calculated to be 43.75% and 12.5%. If the similarity is above 10%, the classes are combined in a single class representation further shown and discussed in Table 30. The ultimate outcome after this class merger is shown in Table 30.

Table 29: A short example of cross-class common word percentage calculation

Stable project responsibility	pending	responsibility	failure	lack	shortage	expedite	cancel	milestones	Total words in both class	16
Stable project requirements	pending	requirement	milestones	deliverables	resources	time	delay	success	Words common in the first class from the pool	25.00%
Stable project requirements	pending	requirement	milestones	deliverables	resources	time	missing	success	Total words in both class	16
Stable project scope	deadline	Pending	Progress	time	missing	delivery	task	missing	Words common in the first class from the pool	43.75%
Stable project responsibility	pending	responsibility	failure	lack	shortage	expedite	cancel	milestones	Total words in both class	16
Stable project scope	deadline	pending	Progress	time	missing	delivery	task	missing	Words common in the first class from the pool	12.50%

Table 30: A confusion matrix representation cross-class false positives and true-negatives via the proposed TF-IDF KNN approach

	Commitment to project milestones	Key management support	Lack of making purchases	Lack of manpower	Stable project scopes and requirements	Team's required knowledge and skills	Total	Percentage	False positives
Commitment to project milestones	7	0	2	0	1	1	11	64%	36%
Key management support	1	7	0	0	2	0	10	70%	30%
Lack of making purchases	0	0	9	0	4	0	13	69%	31%
Lack of manpower	0	0	0	5	0	1	6	83%	17%
Stable project	3	0	0	0	8	0	11	73%	27%

scopes and requirements									
Team's required knowledge and skills	2	0	0		0	7	9	78%	22%
								73%	27%

The final number of classes were reduced from 12 to 6 as shown in Table 30 based upon the class-level ambiguity created by a redundant, similar classes as elaborated in Table 27 and Table 28. Based on this, the KNN algorithm was once again trained and then evaluated against a similar set of test MoM documents. A substantial improvement in the outcomes are shown in Table 30. The worst performing combination with the new training model and labelled dataset showed a 36% false positive for the “Commitment to project milestones” class. However, only 2 files in this category, making 18.18% were miss-classified in the “Lack of making purchases” categories. The remaining two false positives originated/identified “Stable project scopes and requirements” and “Team's required knowledge and skills” classes with a false positive rate of 9.09% (1 MoM file). The overall accuracy improved from 52.86% to 71.66%. It must be noted that the total number of files used in latter case were reduced from 262 to 60 on an arbitrary selection case. Out of 60 files selected, only 17 were miss-identified to other classes. A further comparison of the two varied class cases can be seen in the surface maps shown in Figure 18 and **Error! Reference source not found.** which show multiple peaks for each of the classes indicating high cross-class confusion. The case is improved substantially in Table 30 due to a better class distinction created because of cross-class shortening.

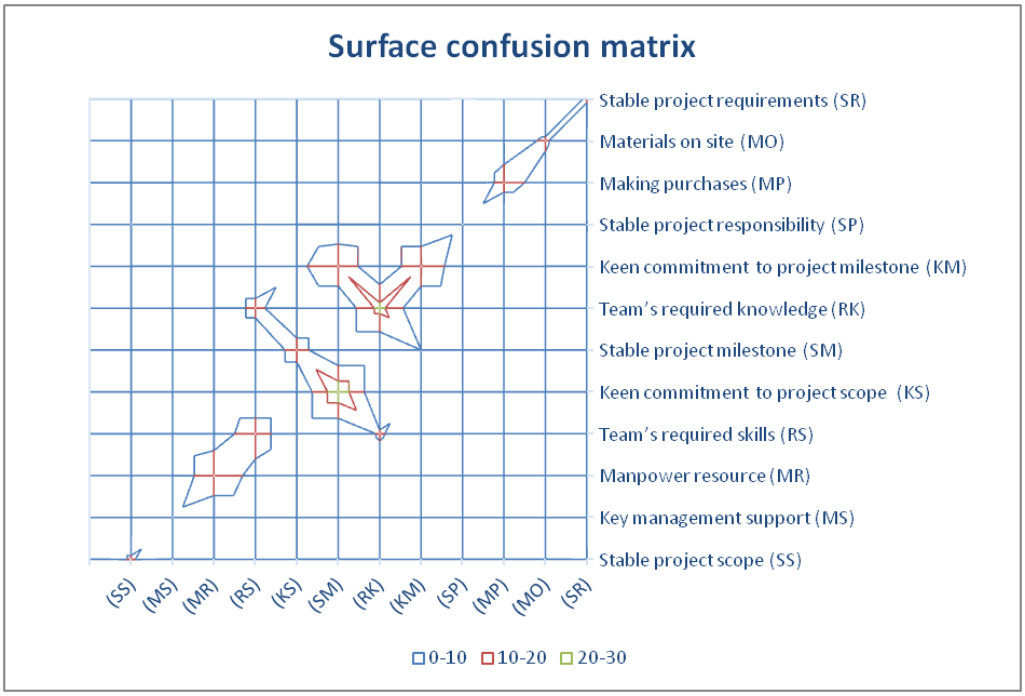


Figure 18: A surface confusion matrix mapping where more-than one peak shows a high cross-class similarity

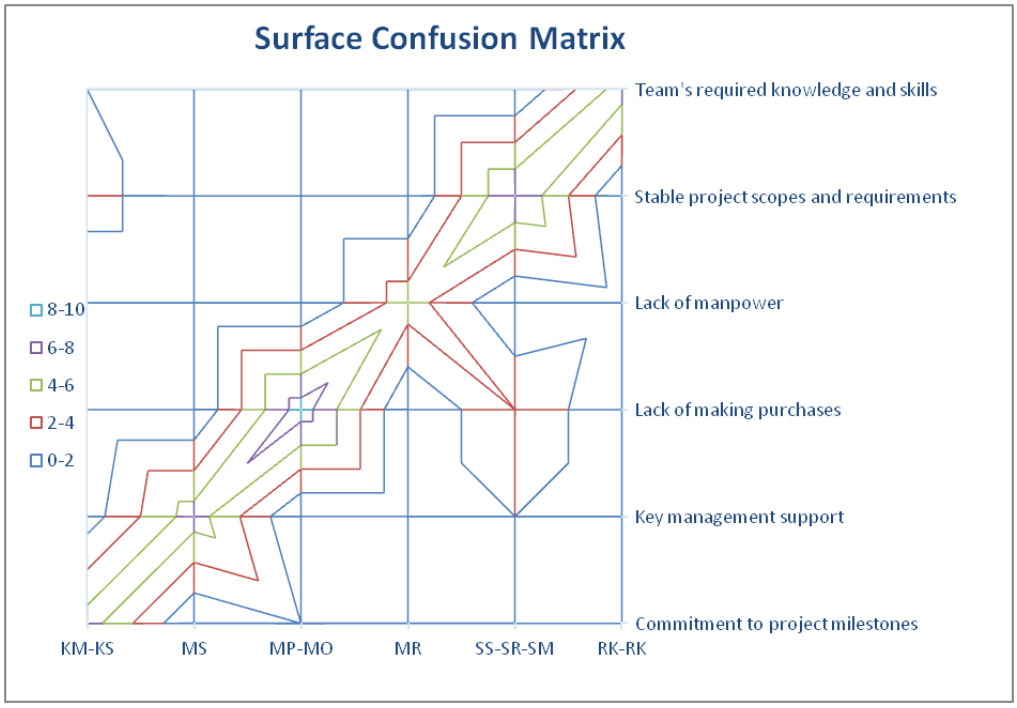


Figure 19: A surface confusion matrix mapping showing a diagonal consistency (single peaks only) as an evidence of superior cross-class accuracy compared to Figure 18

6.3 Early warning and signal identification via the Naïve Bayes classifier

Two separate Naïve Bayes classifiers were evaluated. The confusion matrices of the outcomes against the 46 test files are presented in Table 31 and

Table 32. Table 31 presents an inner-level representation of each test file. The objective behind this classifier was to dig deep into the text present in each file and analyse each sentence present within to extract its association to one of the four EWS categories.

The classification outcome currently presented a no-signal classification to the majority (86.95%) of test files. This outcome can be attributed to the fact that Naïve Bayes is primarily a distinct variable classifier that operates on independent variables bearing no association with each other. This makes it harder for this algorithm to detect meanings hidden deep inside complete sentences instead of words.

Table 31: Confusing matrix presenting false positives against for various "weak signal" categories

	$No - signal$	Inexact	$Clear$	$Exact$
$No - Signal$	40	0	0	0
$Inexact$	0	0	0	0
$Clear$	0	0	100% (6/6)	0
$Exact$	0	0	0	0

Table 32: Confusing matrix for "early warning" cross-class false positives

	S_k	C_k	P_k	M_k	R_k
S_k	3	0	0	0	0
C_k	0	6	0	0	0
P_k	0	0	0	0	0
M_k	0	0	0	37	0
R_k	0	0	0	0	0

Based on the outcome shown in

Table 32, most classification out of the 46 training files was classified and attributed to "lack of materials". In total 80.43% of data files indicated a lack of materials which demonstrates the impact of poor inventory control and management on the overall completion of construction projects. Moreover, the second highest contributing class was "Lack of keen commitment to project milestones and scopes" which also demonstrated the importance of adherence to project plans and proper organisation. Lack of team requires knowledge and skills was the third classified and contributed to just 6.5% of project test files. This still indicated the importance of staff training and development in the improvement of a construction project's development lifecycle. Furthermore, the file-level categorisation of test files is further shown in Table 33.

Table 33 File-level categorisation of "weak signal" and EWS classification

Filename	EWS	True Classification	Observed Classification
218-2014	M_k	No-signal	Clear
219-2014	C_k	No-signal	No-Signal
217-2014	S_k	No-signal	No-Signal
215-2014	R_k	No-Signal	No-Signal

6.3.1 System usefulness and user satisfaction validation

To assess the system for its measure of achievements in terms of the stated research questions, aims and objectives, several models have been presented in the literature. Most these models focus on real-world industrial cases like products and systems. While analysing these models, it must be understood that the core expectation from the tool is to measures its usefulness as well as the satisfaction (of its performance) to its users.

6.4 Outcome of the hybrid Naïve Bayes classifier to improve early warning prediction

The section reports on the outcomes obtained from the discretely trained methodology via the Naïve Bayes classifier based on word-level frequencies. To train the supervised Naïve Bayes training model, the data was initially taken directly from questionnaire responses, which was then used to label Minutes' text. The questions asked experts and engineers from the field of construction project management to indicate specific words used in management documents and the level of risk and warning they attributed towards a situation. For instance, delay in clearance of funds for any purchases may directly attribute to any impending project failures due to lack of project responsibility. Similarly, delays in the completion of project deliverables may be indicated by keywords such as "staff unavailability", "key skills shortage" or "lack of personnel availability". The association so such groups can potentially provide better training of a model compared to a manually labelled project meeting minutes.

70% of this textual information was then used to train the model whereas the remaining 30% was used to assess the outcome of the classifier.

6.4.1 Data consolidation and analysis

As before, a total of 46 PDF documents were labelled based on word-labels from questionnaire responses to extract indicators based on resulting delay types. For example, a Minutes document containing discussions on staff shortage leading to delays due to lack of manpower was used to extract critical sentences containing relevant information. A total of 23 sentences stating keyword pairs such as “employee unavailability”, “manpower shortage”, “frozen action” and “unexpected absence” were fed to the classifier and the outcome were reported as shown in Table 34.

On application of the classification model, the sentences that were correctly known to identify the early warnings were regarded as accurate hits whereas any other matches were termed inaccurate. The best classification accuracy was found to be for the early warning of MP (Making Purchases) class with an accurate identification of 79.46% words. The category only reported 18.75% text terms to be miss-classified as Materials on site (MO) and 1.78% incorrectly identified in the PR (Project Responsibility) category. The lowest accuracy match was that of Teams’ Required Knowledge class where 24.73% of the terms were misclassified as (lack of) Manpower Resource (MR). The overall accuracy in prediction came out to be 68.06%.

Table 34: A confusion matrix elaborating on the TN and TN accuracies of the 12 of early warning classes

Class indices	KS	MS	MR	RS	SM	RK	KM	SP	MS	MP	MO	SR	Total	Accuracy
KS	27	5	0	0	6	0	0	3	0	0	0	0	41	65.85%
MS	11	46	1	0	3	0	0	4	9	0	0	0	74	62.16%
MR	0	0	75	22	0	12	0	0	0	0	0	6	115	65.22%
RS	0	0	4	91	0	5	0	0	0	0	0	22	122	74.59%
SM	14	2	0	0	79	0	7	0	13	0	0	0	115	68.70%
RK	0	0	23	5	0	56	0	1	0	4	1	3	93	60.22%
KM	7	1	7	0	2	1	45	0	6	0	0	0	69	65.22%
SP	17	8	3	0	4	0	0	77	14	0	0	0	123	62.60%
MS	11	7	1	0	0	0	0	23	117	0	1	0	160	73.13%
MP	0	0	0	0	0	0	0	2	0	89	21	0	112	79.46%
MO	0	0	0	0	0	0	0	0	0	28	73	0	101	72.28%
SR	0	0	0	17	0	0	0	0	0	0	0	35	52	67.31%

6.5 Evaluation against the model feature space via K-fold cross-validation

To evaluate the effectiveness and appropriateness of the training data against a diverse range of data types, a 5-fold cross validation technique was used. In this case, the original 45 of 46 training files were used for training were divided into 5-groups (folds) with 9 files in each group. The training process was then performed on each single group of these 5 sets whereas the tests were then performed against the remaining 4-groups (40 files). This was done to ensure the ability of the proposed techniques to handle generic datasets. The results are shown in Table 35 and Table 36. It must be observed that Table 35 only presents Early Warning prediction results as the underlying model was only trained for this category.

Table 35: Model sensitivity analysis for early warning prediction via the Hybrid Naïve Bayes classifier

Hybrid Bayesian Analysis (EW Prediction)						**Total testing files**
Cross-validation Run 1	Group 1 (Training)	Group 2	Group 3	Group 4	Group 5	**40**
Files Correctly Classified	9 training files	6	7	6	7	65
Cross-validation Run 2	Group 2 (Training)	Group 1	Group 3	Group 4	Group 5	
Files Correctly Classified	9 training files	6	8	6	6	65
Cross-validation Run 3	Group 3 (Training)	Group 1	Group 2	Group 4	Group 5	
Files Correctly Classified	9 training files	7	7	6	8	70
Cross-validation Run 4	Group 4 (Training)	Group 1	Group 2	Group 3	Group 5	
Files Correctly Classified	9 training files	8	8	5	8	72.5
Cross-validation Run 4	Group 5 (Training)	Group 1	Group 2	Group 3	Group 4	
Files Correctly Classified	9 training files	7	6	8	7	70
				Average accuracy (%)		68.5

Table 36: Model sensitivity analysis for early warning and weak signal prediction via the Hybrid KNN classifier

KNN Analysis (EW Prediction) - 6-class-case						**Total testing files**
Cross-validation Run 1	Group 1 (Training)	Group 2	Group 3	Group 4	Group 5	**40**
Files Correctly Classified	9 training files	6	8	8	7	72.5
Cross-validation Run 2	Group 2 (Training)	Group 1	Group 3	Group 4	Group 5	
Files Correctly Classified	9 training files	9	7	7	6	72.5
Cross-validation Run 3	Group 3 (Training)	Group 1	Group 2	Group 4	Group 5	
Files Correctly Classified	9 training files	9	4	6	9	70
Cross-validation Run 4	Group 4 (Training)	Group 1	Group 2	Group 3	Group 5	
Files Correctly Classified	9 training files	5	8	7	8	70
Cross-validation Run 4	Group 5 (Training)	Group 1	Group 2	Group 3	Group 4	
Files Correctly Classified	9 training files	7	6	8	8	72.5
				Average accuracy (%)		71.5
KNN Analysis (Weak Signal Prediction)						**Total testing files**
Cross-validation Run 1	Group 1 (Training)	Group 2	Group 3	Group 4	Group 5	**40**
Files Correctly Classified	9 training files	6	6	7	7	65
Cross-validation Run 2	Group 2 (Training)	Group 1	Group 3	Group 4	Group 5	
Files Correctly Classified	9 training files	6	6	7	6	62.5
Cross-validation Run 3	Group 3 (Training)	Group 1	Group 2	Group 4	Group 5	
Files Correctly Classified	9 training files	6	3	6	9	60
Cross-validation Run 4	Group 4 (Training)	Group 1	Group 2	Group 3	Group 5	
Files Correctly Classified	9 training files	4	8	6	8	65

Cross-validation Run 4	Group 5 (Training)	Group 1	Group 2	Group 3	Group 4	
Files Correctly Classified	9 training files	6	5	8	8	67.5
				Average accuracy (%)		64

On the other hand, Table 36 presents both TF-IDF-based Early Warning prediction outcome for a classifier labelled and trained with 6-class classification and the prediction of Weak Signals as well. In the earlier case, the cross-validation technique shows a significant performance similarity for all 5 groups with only 9% variance for the percentage accuracies shown in Table 35, 1.5 for the (EW Prediction) - 6-class-case in Table 36 and 6.5 for the Weak Signal prediction case shown in Table 35. The low variance values present a highly coherent dataset with no outlying behaviour between various k-fold group combinations.

6.6 Evaluation against the model feature space via Jack-knife resembling

Having evaluated the dataset against a 5-fold cross-validation mechanism, the data was further subjected to a Jack-knife testing technique to further ascertain the robustness of the training and testing datasets. To make the process completely random, the datasets were tested via a random cut-off point generated to divide a dataset into two partitions for training and testing purposes. As the total number of files were 46 (not 45 which was number used due to being a multiple of 5), the random number of generated between 0 and 1 and then the value was multiplied to the total number of files and then rounded to the nearest whole number. For instance, if the random number was 0.37, after multiplying it from 46, the result was 17.02 which was then rounded to 17. Having done this, another random number was generated as the starting index to start the training partition in the dataset. So, if the next number generated was 0.76, after multiplying it by 46, the result would be 34.96 rounding-off to 35. Hence, the starting index for the training data partition came out to be from the 35^{th} file to 46^{th} file (equalling to 12 out of 17 files) and then 1^{st} file to 5^{th} file (equalling the remaining 5 files). This process ensured a more diverse selection of dataset instead of always selecting the first (or last) partitions of data from the dataset. The dataset was run once against each of the three prediction techniques i.e. Hybrid Bayesian, KNN analysis of EW prediction and KNN analysis of Weak Signal prediction. The overall performance of all three dataset partitions generated accuracies quite like the ones observed in the previous K-fold validation techniques. The overall accuracy for KNN-based EW prediction was still better compared to the Hybrid Bayesian analysis. The Weak Signal classification showed a stark similarity against the K-fold classification outcome indicating reliability of the underlying algorithm to have a prediction accuracy of around 64% (via K-fold) to 64.1% (via Jack-Knife) in Table 37.

Table 37: Jack-knife resampling testing based on dual-dataset partition driven by randomised slicing

Hybrid Bayesian Analysis (EW Prediction)				
Jack-knife run 1	Training group	Testing group	Training file start position	Training file end position
	18	28	41st file - 46th file	1st - 11th file
		Total files	6.00	11.00
		Correctly identified	4.00	7.00
		Percent accuracy	6.66	63.63
		Overall accuracy	65.15	
KNN Analysis (EW Prediction) - 6-class-case				
Jack-knife run 2	Training group	Testing group	Training file start position	-
	32	14	21st file - 34th file	-
		Total files	14.00	-
		Correctly identified	10.00	-
		Percent accuracy	71.42	-
		Overall accuracy	71.42	-
KNN Analysis (Weak Signal Prediction)				
Jack-knife run 3	Training group	Testing group	Training file start position	Training file end position
	24	22	38th file - 46th file	1st - 13th file
		Total files	9	13.00
		Correctly identified	6	8.00
		Percent accuracy	66.67	61.53
		Overall accuracy	64.10	

6.7 Transferability of results to different document types

The outcome analysis presented above trains and tests the three models on Minutes (MoM) data. To assess the viability of these algorithms against a number of other documents, the model originally trained on the 46 MoM documents was then used to evaluate its ability to identify warnings from other construction management document types. These files were sourced from live construction projects where the project managers were asked to label each document against the 6-class presented above and the Weak-Signal case as well. The model used was still based on the MoM data where the test files were then evaluated against the ground-truth labels as advised by the expert users (project managers). The outcomes of these algorithms are shown in Table 38. The results were a mix where the memos had below average prediction accuracy for Naïve Bayes and KNN weak signal prediction case but performed well for the KNN Early Warning case with an overall classification accuracy of 71.42%. The KNN-EWW case also performed extremely well for Bid Contracts and Expediting Reports with accuracies of 80 and 81.81%. The overall 73.31% accuracy shows the suitability of this genre of algorithms compared to the Hybrid Naïve Bayes classifier which had an overall accuracy of only 60.19% with Bid Contracts and Engineering Sheets among the worst performers. The KNN analysis of Weak Signals performed particularly poorly for Memos with an accuracy of 42.85% as shown in Table 38.

Table 38: Impact of change of document on the ability of a model trained on MoM data

		Hybrid Bayesian Analysis (EW Prediction)		KNN Analysis (EW Prediction) - 6-class-case		KNN Analysis (Weak Signal Prediction)	
	Total files	Correctly identified	%	Correctly identified	%	Correctly identified	%
Memos	7	4	57.14	5	71.42	3	42.85
Bid contracts	5	3	60.00	4	80.00	3	60.00
Engineering sheets	5	3	60.00	3	60.00	3	60.00
Expediting report	11	7	63.63	9	81.81	9	81.81
		Overall accuracy	60.19	**Overall accuracy**	73.31	**Overall accuracy**	61.16

6.8 Assessment/derivation examples of actionable results from the algorithm outcomes

The ultimate purpose of the three categories of outcomes for Early Warning and Weak signal prediction algorithms was to allow managers to derive actionable results from the predicted/identified classes. To achieve this, the classes were labelled under “common sense” linguistic representation such as “Clear Signals” identified with a “Lack of Manpower” observed via a KNN-clustering algorithm indicates the required action to be undertaken to not only consult the relevant document for manual assessment but also to enable the provision of required manpower to address or minimise the impact of staff shortage. As the algorithms still focus on categories, it is the responsibility of the project manager to then further analyse the document to undertake an appropriate action. Moreover, as the same document also indicates a “Clear Signal”, the manager should also analyse the document for additional aspects that may lead to other delays during the entire project lifecycle. Another example would be a document classified under an Early Warning category of “Commitment to project milestones” and a Weak Signal identification of “Exact Signal” category. Upon reviewing this classification, the project manager can review the relevant document first to see in which milestones there are deviations (e.g. delay in completion or lack of required resources to fulfil a milestone). Furthermore, as the document is also indicated to have an “Exact Signal”, the management can further check and see if this “signal” correlates with the identified Early Warning or something else. Two examples of “Exact Signals” are shown in

Table **23** of “Delay in delivery time” and “Daily manpower shortage”. Ultimately, a combination of Early Warnings and Weak Signals can identify multiple, hidden issues within the documents with regards to the overall management of the project.

6.9 Summary

The initial tests from the hybrid Naïve Bayes technique addresses an information mining approach from project management documents to extract early warnings. The underlying approach targeted two core problems in construction data mining namely the development of a term-extractor and early-warning mapping corpus and the development of a machine learning early warning identification classifier.

The approach is novel in a sense that it improves on discrete word-level early warning extraction based on a supervised, expert-labelled project management document database. The approach was evaluated against 12 unique aspects that indicate early warnings including lack of management support, requirement delays, skills & material shortage and planning weaknesses. The underlying principles of extracting associated data-pairs were modelled via keywords extracted through qualitative information from an expertly guided questionnaire. The proposed methodology outperformed most these methodologies including standalone Nearest-neighbour and KNN approaches by improving respective accuracies of 58.82% and 49.11% to 68.06%. According to the work done by Caldas, et al (2002), this approach is only outperformed by SVM, which currently delivers an accuracy of 91.12% that is primarily limited to a two-class problem whereas the proposed approach addresses a multi-class solution.

The hybrid Naïve Bayes methodology can further be extended to identify sentence-level warnings by training sentence chains in a time-series state-transition fashion via a Markovian approach. This approach is likely to improve the accuracy of certain document sections depending upon how well the database is trained against a labelled dataset. The future extension of this research is to use the labelled qualitative information to train class-specific Genetic algorithm. The approach is expected to improve the existing accuracy as it would consider a sentence-level meaning instead of word-pairs and discrete (single) words as well as state solution.

The later part of this chapter presents the outcomes from the TF-IDF approach which show a strong correlation between the distinctness of classes and the

prediction accuracy of “early warnings”. The initial 12-class approach showed a high level of cross-class inaccuracy which, on further analysis, was found to be due to excessive closeness of class types. One such example was the class “Stable Project Scope” and “Commitment to Project Scope” which were deemed similar within the context of project management activities. This cross-class similarly lead to an overall accuracy of 54.92%. Once the classes with closest cross-class inaccuracies were merged, retrained and tested, the prediction accuracy increased to 72.78% which showed a marked improvement merely by eliminating cross-class bias via class combination.

To conclude, it must be noted that the confusion matrices currently did not initially report on false positives as this will require input from a group of experts capable of judging the test files to belong to one of the five categories. These inputs will act as the “ground truth” for these test samples and will give a deeper understanding of the text mining results obtained to-date. Moreover, the first-level classification via Naïve Bayes also demonstrates its shortcoming in detecting early warnings present within sentences instead of mere words. It must be noted that posterior probabilities in the Naïve Bayes classifier only worked on word frequencies or the so-called bag-of-word (Brynielsson, et al., 2013). This shortcoming is extended and addressed as a time-series identification problem which forms the extended part of this research. In order to address this issue, the research aims to utilise the Hidden Markov Models which are well-known in time-series analysis of real-world problems such as text synthesis, speech analysis, gesture recognition, etc.

7 CHAPTER VII: Conclusions

The work presents several novel methodologies to address the problem of data/text mining in the information retrieval and processing domain with a focus on construction project management. Unexpected discontinuities and abrupt problems in project management lifecycles are commonly reported. One of the most prevalent reasons of delays or failures in these projects has been the inability of the board, management and/or the policymakers to understand and identify these problems in advance. The area has extensively been investigated. However, it was found to be lacking substantially in terms of utilising trained AI models to pre-emptively identify such failures before they do actual damage. The research presented in this thesis can hence be categorised into two broad categories of identifying such “early warnings” and measuring subtle signals hidden within management documents that are capable of provide advance warnings into a wide range of failures.

7.1 Achievement of Objectives

The stated research objectives have been achieved as follows:

Critical analysis of the existing state-of-knowledge to extract actionable information from construction project management documents

A critical analysis of existing data and text mining techniques to from management documents revealed several aspects where a proper assessment would lead to the improvement of the overall project lifecycle including staff experience, project control, risk management, experience and knowledge transfer. Everyday challenges were identified to contain lack of collaboration & information granularity, inadequate management engagement and poor planning. Previous research into the domain of identification of anomaly indicators in the scope of construction management activities revealed certain hidden “early warning” aspects and the signal levels (as severity measures) that indicated the fact that a robust identification mechanism of these warnings and signals would improve the overall quality of project management.

Development of a project progress assessment methodology encompassing factors affecting project performance

To empirically validate the role of these “early warnings” or “signals”, a set of empirical, AI-based methodologies were explored and extended. It was understood that, if labelled effectively, the techniques would successfully classify these warnings. A qualitative labelling methodology was hence proposed by means of a secondary-research-based questionnaire to pinpoint the association of various keywords used in construction project documents to impending project failures or delays. The questionnaire responses were then mapped and categorised into several failure classes which were then trained initially over a Naïve Bayes classifier.

The classifier probabilistically mapped a set of keyword groups to a total of 12 Early Warning classes. The resultant Naïve Bayes implementation assigned a probabilistic score to each word and targeted two core problems of:

- Identification of early warnings on individual word usage and eventually coming-up with a cumulative probabilistic distribution for each class where the class with highest probability was then selected as the identified category.
- Identification of hidden word-pair relationships to improve sentence-level risks. This is where two words were paired due to their intrinsically similar association.

The above technique had a limitation where it couldn’t accommodate sentence-level structuring and hence generated false positives in cases where ambiguous word combinations were used. To resolve this issue, an unsupervised clustering mechanism was proposed via a KNN classifier.

The technique further handled the effect (bias) of common words on various classes which was addressed via a word frequency-based weighing mechanism commonly known as TF-IDF. When the technique was evaluated, it presented a number of additional challenges including those that were generated due to close association of classes. Once the classes were merged, the accuracy

improved substantially. The underlying KNN methodology was used to label the measure of severity (multi-level signals) in construction documents.

Development of a methodology to use the actionable information to indicate potential future failures

The qualitative responses classified the “actionable information” into 12 unique categories initially. These potential failures contained several aspects such as commitment to project scope, management support, manpower resource, team skills, project scope commitment, knowledge levels, project responsibility, purchases aspects, project milestones, materials on site, and project requirements. However, these requirements eventually showed cross-class similarities and hence presented cross-class misclassifications. Regardless of the methodology used, the classifiers successfully indicated six core actionable categories in the construction management domain.

Development of an analysis tool identify project risks and empirically validate its usefulness against various case studies

The trained system was evaluated against a total of 28 different document types including memos, bid contracts, engineering sheets, and expediting reports. The cases were evaluated against three main predictions including the Hybrid Bayesian Analysis & the KNN EW prediction for the early warning and KNN weak signal identification. The overall accuracy of these three predictions types was observed to be 60.19%, 73.31% and 61.16% respectively.

The data was further validated for sensitivity against outliers via cross-validation and jack-knife testing. The overall accuracy for Naïve Bayes early warning prediction was found to be 68.5%, KNN-based early warning prediction was 71.5% and KNN-based weak signal prediction was still assessed to be 64% obtained based on 5 runs with each containing 9 unique training files being tested with the remaining 4 groups. The accuracies are still very close to the actual testing accuracy given in the earlier paragraph hence demonstrating the viability of the datasets used.

7.2 Research contributions and concrete outcomes

The research encompasses several concrete research outcomes including the prepared dataset, the Hybrid Naive Bayes classifier for EW classification, the KNN classifier for EW classification and Weak Signal identification.

7.2.1 Training of an AI predictive model based on expertly-driven input

The problem addressed in this research is focussed mainly on modelling and predicting hidden sequences and signals where the training data must be tuned with previously labelled information. However, construction management related activities extensively rely on experience and hence any such model cannot be trained unless supervised via expert input. Based on this notion, this research focusses on training based on expert feedback.

7.2.2 Establishment of the ground-truth (labelled data)

A novel, expertly-driven dataset was used to quantify the measurement of early warnings via a qualitative (questionnaire-based) study that queried several experts into revealing their knowledge on how to categorise such aspects. This led to the discovery of certain keyword groups that were mapped onto various early warning and weak signal categories. One such example was that of the keyword/word-pair “employee absence” leading to problems with “Team's required knowledge and skills”. Hence, any appearance of similar word pairs was expected to identify the document using those words to belong to this class. However, any such document would also have other word groups indicating other early warnings. Hence, a probabilistic modelling strategy was adopted that trained itself on the likelihood of various feature groups.

7.2.3 Naïve Bayes modelling of early warnings in management documents

The second concrete outcome of this research was the supervised Naïve Bayes classification methodology which was trained on word data extracted from many MoM files. The results presented promising capabilities of the system. However, it was further noted that the system intrinsically had a shortcoming where the

meanings were only to be communicated via complete sentences and not mere word combinations. One such example was that of ambiguous word groups such as “there appears to be no significant delay in the delivery”. Such a word group may confuse a Naïve Bayes classifier due to the usage of superlative words such as “significant” combined with other negative word “delay”. It was understood ambiguity in such sentence-level cases could improve either by using a time-series algorithm model or a technique capable of removing insignificant words such as “the”, “then”, “is” or “yes/no”.

7.2.4 Extending to “early warnings” and “weak signals” via KNN TF-IDF classification

On the third contribution, a TF-IDF classification methodology was used to rank various words via a frequency-based bag-of-words approach which were then trained via a KNN classifier. The technique utilised the frequency of occurrence of keywords to ensure that words redundant in each class group were to be eliminated and only the most relevant words to each class were left before the commencement of model training. The TF-IDF approach was initially evaluated against a total of 12 classes that were extracted from the expert questionnaire. The outcome showed a marked cross-class false positive rate where each class had a few miss-classifications originating from other classes. Furthermore, it was observed that a large fraction of this miss-classification originated from one or two other classes which indicated a repeated pattern. These classes were then individually evaluated and a substantial inter-class similarity was reported. This similarity indicated the fact that these classes were intrinsically similar in nature to each other. For instance, the classes “team required skills” and “team required knowledge” both reported the following cross-class miss-classification:

- Out of 35 cases of “Team Required Skills” class, 13 (37.14%) cases were miss-identified as “Team Required Knowledge”
- Out of 39 cases of “Team Required Knowledge”, 11 (28.2%) cases were miss-identified as “Team Required Skills”

The pattern indicated a substantial level of similarity in the input class features. Moreover, merely by judging from the meaning of each class, it could be

understood that the classes originated from a single basic concept indicating a team with a lack of either skills or the knowledge required to operate. Similarly, other classes with a keyword similarity match above 10% threshold were merged. This resulted in a final class count of 6 from 11. The system was then retrained and then re-evaluated against the same data. The files used in total in the reduced-class case were 60 out of which only 17 were miss-identified to other classes. Moreover, the overall accuracy improved from 52.86% to 71.66%.

7.3 Performance comparison of EW identification with the previous research

The proposed Modified Naïve Bayes algorithm was compared with the existing word-based document identification classifiers study including SVM, ROCCHIO, IBM Miner, standard Naïve Bayes and standard KNNs. The proposed technique outperformed all the algorithms apart from the SVM-based approach which has so far reported the highest accuracy (91.12%) compared to the Modified Naïve Bayes accuracy of 68.02% which was retained for various validations tests. This approach was primarily limited to a two-class problem whereas the proposed approach addresses a multi-class solution. The extended KNN approach to EW prediction resulted in further improvement to 73.311% with the proposed modified KNN compared to the standard KNN accuracy of 49.11%.

7.4 Future directions and research extension

The proposed work largely focusses on a text parsing and processing domain. The research addresses the problem as a construction management problem aimed at extracting document information to classify certain signals. The information extraction can further be improved via recently investigated time-series-based AI algorithms such as Recurrent Neural Networks that aim to resolve the word-sequencing-level aspect of documents. Moreover, since the documents involved have highly complex data that often combines with other formats such as images, formatted tables and document sections, extracting meaningful information from such documents is a research problem originating

from image processing and ontology domains. Hence, this research can also be extended as a computer vision problem.

During project management meetings, documents are often generated and stored in varied formats such as PDFs, chart files and AutoCAD drawings. Parsing all these on a single training algorithm would indeed require a novel ontology. Moreover, sentence-level modelling of such documents is also a largely unexplored area which is likely to improve the overall accuracy of early warning identification if a time-series-level database of such documents is maintained and established.

Finally, the questionnaire used in this research mainly focussed on the identification of keywords that categorised various class groups. As the questions were largely developed via a secondary research survey, it was difficult to automatically update the underlying corpus if any new information was made available e.g. a new respondent completing a survey. Hence, an approach to completely automate the data extraction, parsing and algorithm training part would also contribute significantly to this research.

8 References

Abanda, F. H., Tah, J. H. M. & Cheung, F. K. T., 2013. Mathematical modelling of embodied energy, greenhouse gases, waste, time–cost parameters of building projects: A review', Building and Environment. Volume 59, pp. 23-37.

Adam, A., Josephson, P. & Lindahl, G., 2017. Aggregation of factors causing cost overruns and time delays in large public construction projects: trends and implications. *Engineering, Construction and Architectural Management,* 24(3), pp. 393-406.

Adeniyi, D., Wei, Z. & Yongquan, Y., 2016. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics,* 12(1), pp. 90-108.

Ahern, T., Leavy, B. & Byrne, P., 2014. Complex project management as complex problem solving: A distributed knowledge management perspective. *International Journal of Project Management,* 32(8), pp. 1371-1381.

Ajayi, S. O. et al., 2016. Reducing waste to landfill: A need for cultural change in the UK construction industry. *Journal of Building Engineering,* Volume 5, pp. 185-193.

Akil, H., Bouille, J. & Robert-Demontrond, P., 2017. Visual representations of climate change and individual decarbonisation project: an exploratory study. *Revue de l'organisation responsable,* 12(1), pp. 66-80.

Al Qady, M. & Kandil, A., 2010. Concept Relation Extraction from Construction Documents Using Natural Language Processing. *Journal of Construction Engineering and Management-Asce,* 136(3), pp. 294-302.

Al Qady, M. & Kandil, A., 2013. Automatic classification of project documents on the basis of text content. *Journal of Computing in Civil Engineering,* 29(3), pp. 0401-4043..

Al Qady, M. & Kandil, A., 2014. Automatic clustering of construction project documents based on textual similarity. *Automation in Construction,* Volume 42, pp. 36-49.

Alhawari, S., Karadsheh, L., Talet, A. & Mansour, E., 2012. Knowledge-based risk management framework for information technology project. *International Journal of Information Management,* 32(1), pp. 50-65.

Ali, H. A. E. M., Al-Sulaihi, I. A. & Al-Gahtani, K. S., 2013. Indicators for measuring performance of building construction companies in Kingdom of Saudi Arabia. *Journal of King Saud University - Engineering Sciences,* 2(125-134), p. 25.

Almazán, J., Gordo, A., Fornés, A. & Valveny, E., 2014. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence,* 36(12), pp. 2552-2566.

Almeida, T. A., Yamakami, A. & Almeida, J., 2009. *Evaluation of Approaches for Dimensionality Reduction Applied with Naive Bayes Anti-Spam Filters.* s.l., Machine Learning and Applications, 2009. ICMLA '09. International Conference on, 13-15 Dec. 2009, 517-522.

Alp, N. & Stack, B., 2012 . *Scope management and change control process study for project-based companies in the construction and engineering industries.* s.l., Proceedings of PICMET '12: Technology Management for Emerging Technologies, July 29 2012-Aug. 2 2012.

AlRababah, A., 2017. A new model of information systems efficiency based on key performance indicator (KPI). *Management,* Volume 4, p. 8.

Al-Shameri, F., 2012. Automated generation of metadata for mining image and text data. *U.S. Patent ,* Volume 8, pp. 145-677.

Alsubaey, M., Asadi, A. & Makatsoris, H., 2015. A Naïve Bayes approach for EWS detection by text mining of unstructured data: A construction project case. *SAI Intelligent Systems Conference (IntelliSys) IEEE,* pp. 164-168.

Alsubaey, M., Asadi, A. & Makatsoris, H., 2016. An Unsupervised Text-Mining Approach and a Hybrid Methodology to Improve Early Warnings in Construction Project Management. In: *In Intelligent Systems and Applications .* London: Springer International Publishing, pp. 65-87 .

Amarasinghe, K., Manic, M. & Hruska, R., 2015. *Optimal stop word selection for text mining in critical infrastructure domain,* s.l.: Resilience Week (RWS), 2015, 18-20 Aug. 2015, 1-6.

Amarasiri, R., Ceddia, J. & Alahakoon, D., 2013. *Exploratory data mining lead by text mining using a novel high dimensional clustering algorithm.* s.l., Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on, 15-17 Dec. 2005, 6 pp.

Anantatmula, V. & Webb, J., 2016. Critical chain method in traditional project and portfolio management situations. *In Project Management: Concepts, Methodologies, Tools, and Applications,* pp. 1005-1022). IGI Global.

Andrey, B., 2015. Social Aspects of Specialist Training in the Construction Industry. *Procedia Engineering,* Volume 117, pp. 60-65.

Ansoff, H. I., 1975. Managing strategic surprise by response to weak signals.California Management Review. *California Management Review,* p. 21–33.

Ansoff, I., 1980. Strategic issues management. Strateg. Manag. J. 1 (2), 131–148.. *Strateg. Manag. J.,* 1(2), p. 131–148.

Anupriya, P. & Karpagavalli, S., 2015. *LDA based topic modeling of journal abstracts.* s.l., Advanced Computing and Communication Systems, 2015 International Conference on, 5-7 Jan. 2015, 1-5.

An, X., 2005. Evaluation of research project on integrated management and services of urban development records, archives, and information. *Tsinghua Science and Technology,* 10(1), pp. 852-858.

Arashpour, M., Wakefield, R., Blismas, N. & Minas, J., 2015. Optimization of process integration and multi-skilled resource utilization in off-site construction', Automation in Construction, 50, pp. 72-80. *Automation in Construction,* Volume 50, pp. 72-80.

Asadi, A., Alsubaey, M. & Makatsoris, C., 2015. A machine learning approach for predicting delays in construction logistics. *International Journal of Advanced Logistics,* 4(2), pp. 115-130.

Assaf, S. A. & Al-Hejji, S., 2006. Causes of delay in large construction projects. *International journal of project management,* 24(4), pp. 349-357.

Atkinson-Abutridy, J., Mellish, C. & Aitken, S., 2003. A semantically guided and domain-independent evolutionary model for knowledge discovery from texts. *IEEE Transactions on Evolutionary Computation,* 7(6), pp. 546-560.

Atkinson-Abutridy, J., Mellish, C. & Aitken, S., 2004. Combining information extraction with genetic algorithms for text mining. *IEEE Intelligent Systems,* 19(3), pp. 22-30.

Atkinson, R. 1., 1999. Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria.. *International journal of project management,* 17(6), pp. 337-342.

Awolusi, I., Marks, E. & Hallowell, M., 2018. , 2018. Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices. *Automation in Construction,* Volume 85, pp. 96-106.

Azhar, S., Khalfan, M. & Maqsood, T., 2015. Building information modelling (BIM): now and beyond. *Construction Economics and Building,* 12(4), pp. 15-28.

Baghdadi, A. & Kishk, M., 2015. Saudi Arabian Aviation Construction Projects: Identification of Risks and Their Consequences. *Procedia Engineering,* pp. 32-40.

Bai, L. et al., 2016. Integrated Natural Resources Modelling and Management (INRMM)-library 5231 articles.. 113(23).

Bakshi, T., Sarkar, B. & Sanyal, S., 2012. *A new soft-computing based framework for project management using game theory.* s.l., In Communications, Devices and Intelligent Systems (CODIS), International Conference on (pp. 282-285). IEEE.

Bao, C., Ji, H., Quan, Y. & Shen, Z., 2016. *Dictionary learning for sparse coding: Algorithms and convergence analysis.* s.l., IEEE transactions on pattern analysis and machine intelligence, 38(7), pp.1356-1369..

Barki, H., Rivard, S. & Talbot, J., 2011. An Integrative Contingency Model of Software Project Risk Management. *Journal of Management Information Systems,* 17(4), pp. 37-69.

Bassett, B. & Kraft, N., 2013. *Structural information based term weighting in text retrieval for feature location.* s.l., In Program Comprehension (ICPC), 1st International Conference on (pp. 133-141). IEEE..

Basu, T. & Murthy, C., 2012. *Effective text classification by a supervised feature selection approach.* s.l., . In Data Mining Workshops (ICDMW), IEEE 12th International Conference on (pp. 918-925). IEEE.

Bavan, A. S., 2009. *An artificial neural network that recognizes an ordered set of words in text mining task.* s.l., Current Trends in Information Technology (CTIT), 2009 International Conference on the, 15-16 Dec. 2009, 1-5.

Bechtel, (1994). *Bechtel On Line Reference Manual,* USA : MD .

Bellavita, C. & Gordon, E., 2006. Changing homeland security: teaching the core. *Homeland Security Affairs,* 2(1), pp. 1-19.

Berger, M. & Doban, V., 2014. Big data, advanced analytics and the future of comparative effectiveness research.. *Journal of Comparative Effectiveness Research,* 3(2), pp. 167-176.

Bertone, E. et al., 2006. State-of-the-art review revealing a roadmap for public building water and energy efficiency retrofit projects. *International Journal of Sustainable Built Environment,* 5(2), pp. 526-548.

Biggs, M., 1990. Information overload and information seekers: What we know about them, what to do about them. *The Reference Librarian,* 11(25-26), pp. 411-429.

Bijalwan, V., Kumar, V., Kumari, P. & Pascual, J., 2014. KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application,* 7(1), pp. 61-70.

Bilal, M. et al., 2016. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics,* 30(3), pp. 500-521.

Binder, J., 2016. *Global project management: communication, collaboration and management across borders.* s.l.:CRC Press.

Biro, I., Benczur, A., Szabo, J. & Maguitman, A., 2008. *A Comparative Analysis of Latent Variable Models for Web Page Classification.* s.l., Web Conference, 2008. LA-WEB '08., Latin American, 28-30 Oct. 2008, 23-28.

Boehm, B., 2009. Software Risk Management Principles and Practices. *IEEE Software,* 8(1), pp. 32-41.

Bosch-Sijtsema, P. & Henriksson, L., 2014. Managing projects with distributed and embedded knowledge through interactions. *International Journal of Project Management,* 32 (8), pp. 1432-1444..

Brady, T. & Davies, A., 2014. Managing structural and dynamic complexity: A tale of two projects. *Project Management Journal,* 45(4), pp. 21-38.

Braglia, M. & Frosolini, M., 2014. An integrated approach to implement project management information systems within the extended enterprise. *International Journal of Project Management,* 32(1), pp. 18-29.

Brandon Jr, D., 2010. *Project performance measurement.* 9 pp.75 ed. s.l.:The Wiley Guide to Project Control .

Bromiley, P., McShane, M., Nair, A. & Rustambekov, E., 2015. Enterprise risk management: Review, critique, and research directions. *Long range planning,* 48 (4), pp. 265-276.

Brynielsson, J. et al., 2013. Harvesting and analysis of weak signals for detecting lone wolf terrorists. Secur. Inf.. *Secur. Inf.,* 2(11), pp. 5-15.

C. Hendrickson, P. M. f. C. F. C. f., 2008. *Owners, Engineers, Architects and Builders.* 2.2 ed. s.l.:www.ce.cmu.edu/pmbook .

C.Tung-Tsan, 2010. Partnerships among different participants in construction industry of Taiwan: Critical success and failure factors. *Industrial Engineering and Engineering Management (IE&EM),* pp. 1912-19.

Caldas, C., Soibelman, L. & Gasser, L., 2005. Methodology for the integration of project documents in model-based information systems. *Journal of Computing in Civil Engineering,* 19(1), pp. 25-33.

Caldas, C., Soibelman, L. & Han, J., 2002. Automated classification of construction project documents. *Journal of Computing in Civil Engineering,* 16(4), pp. 234-243.

Candy, J., 2016. *Bayesian signal processing: classical, modern, and particle filtering methods.* s.l.:ohn Wiley & Sons.

Cao, Y., Chau, K., Anson, M. & Zhang, J., 2002. *An intelligent decision support system in construction management by data warehousing technique,* s.l.: Lecture Notes in Computer Science, 2480, pp. 360–369.

Cassidy, A., 2016. *A practical guide to information systems strategic planning.* s.l.:CRC press.

Cerulo, L., Ceccarelli, M., Di Penta, M. & Canfora, G., 2013. *A Hidden Markov Model to detect coded information islands in free text.* s.l., Source Code Analysis and Manipulation (SCAM), 2013 IEEE 13th International Working Conference on, 22-23 Sept. 2013, 157-166..

Chamatkar, A. J. & Butey, P. K., 2015. *Implementation of Different Data Mining Algorithms with Neural Network.* s.l., Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on, 26-27 Feb. 2015, 374-378..

Chapman, C. a. W. S., 2003. *Project risk management: processes, techniques and insights.* s.l.:s.n.

Chaturvedi, K. K. & Singh, V. B., 2012. *'Determining Bug severity using machine learning techniques.* s.l., Software Engineering (CONSEG), 2012 CSI Sixth International Conference on, 5-7 Sept. 2012, 1-6.

Cheng, C. et al., 2012. Applying HFMEA to prevent chemotherapy errors. *Journal of medical systems,* 36 (3), pp. 1543-1551.

Cheng, M. & Hoang, N., 2014. Interval estimation of construction cost at completion using least squares support vector machine. *Journal of Civil Engineering and Management,* 20(2), pp. 223-236.

Chen, L. & Luo, H., 2014. A BIM-based construction quality management model and its applications. *Automation in construction,* Volume 46, pp. 64-73.

Cheung, K. W., Kwok, J. T., Law, M. H. & Tsui, K. C., 2003. Mining customer product rating for personalized marketing. *Decision Support Systems,,* 35(2), pp. 231-243.

Chi, B. & Hsu, C., 2012. A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications,* 39(3), pp. 2650-2661.

Chih-Tien, F., Wei-Jen, W. & Yue-Shan, C., 2011. *Agent-Based Service Migration Framework in Hybrid Cloud.* s.l., High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on, 2-4 Sept. 2011, 887-892.

Chilipirea, C. et al., 2017. An integrated architecture for future studies in data processing for smart cities. *Microprocessors and Microsystems.*

Chi, S., Suk, S.-J., Kang, Y. & Mulva, S. P., 2012. Development of a data mining-based analysis framework for multi-attribute construction project information. *Advanced Engineering Informatics,* 26(3), pp. 574-581.

Choi, W., Pantofaru, C. & Savarese, S., 2013. A general framework for tracking multiple people from a moving camera. *IEEE transactions on pattern analysis and machine intelligence,* 35(7), pp. 1577-1591.

Christen, M., Adey, B. T. & Wallbaum, H., 2016. On the usefulness of a cost-performance indicator curve at the strategic level for consideration of energy efficiency measures for building portfolios. *Energy and Buildings,* Volume 119, pp. 267-282.

Cicmil, S., Cooke-Davies, T., Crawford, L. & Richardson, K., 2017. Exploring the complexity of projects: Implications of complexity theory for project management practice. *Project Management Institute.*

Cleland, D. I., 1994. Project Management, Strategic Design and Implementation. Issue 2.

Coombs, W., 2014. Ongoing crisis communication: Planning, managing, and responding. In: s.l.:Sage Publications.

Cooney, J., 2016. *Health and safety in the construction industry-a review of procurement, monitoring, cost effectiveness and strategy.* s.l.:Doctoral dissertation, University of Salford.

Cotelo, J. M., Cruz, F. L., Enríquez, F. & Troyano, J. A., 2016. Tweet categorization by combining content and structural knowledge. *Information Fusion,* Volume 31, pp. 54-64.

Cox, R., Issa, R. & Ahrens, D., 2003. Management's perception of key performance indicators for construction. *Journal of construction engineering and management,* 129(2), pp. 142-151.

Cserhati, G. & Szabo, L., 2014. The relationship between success criteria and success factors in organisational event projects. *International Journal of Project Management,* 32(4), pp. 613-624.

Cui, Z. & Loch, C., 2014. *A rational framework on the causes and cures of collaborative projects failure.* s.l., s.n.

Czarnigowska, A. & Sobotka, A., 2013. Time–cost relationship for predicting construction duration. *Archives of Civil and Mechanical Engineering,* 13 (4), pp. 518-526.

Dadachev, B., Balinsky, A., Balinsky, H. & Simske, S., 2012. *On the Helmholtz Principle for Data Mining.* s.l., Emerging Security Technologies (EST), 2012 Third International Conference on, 5-7 Sept. 2012, 99-102.

Dagan, D. & Isaac, S., 2015. 'Planning safe distances between workers on construction sites. *Automation in Construction,* Volume 50, pp. 64-71.

Dansoh, A., Frimpong, S. & Oteng, D., 2017. Industry environment features influencing construction innovation in a developing country: a case study of four projects in Ghana. *International Journal of Technological Learning, Innovation and Development, 9,* 9(1), pp. 65-95.

Dao, B. et al., 2016. Exploring and Assessing Project Complexity. *Journal of Construction Engineering and Management,* 143(5), p. 04016126.

Dau, H., Begum, N. & Keogh, E., 2016. *Semi-supervision dramatically improves time series clustering under dynamic time warping.* s.l., In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 999-1008).

Davison, H., Mishra, V., Bing, M. & Frink, D., 2014. How individual performance affects variability of peer evaluations in classroom teams: A distributive justice perspective. *Journal of Management Education,* 38(1), pp. 43-85.

de la Rosette, J. J. M. C. H. et al., 2012. Categorisation of Complications and Validation of the Clavien Score for Percutaneous Nephrolithotomy. *European Urology,* 62(2), pp. 246-255.

Decker, R., Wagner, R. & Scholz, S., 2005. An internet-based approach to environmental scanning in marketing planning. *Mark. Intell. Plan. ,* 23(2), p. 189–190.

Demirkesen, S. & Ozorhon, B., 2017. Impact of integration management on construction project management performance. *International Journal of Project Management,* 35(8), pp. 1639-1654.

Dey, L., Bharadwaja, H. S. M. G. & Shroff, G., 2013. *Email Analytics for Activity Management and Insight Discovery.* s.l., Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, 17-20 Nov. 2013, 55.

Diamantini, C., Potena, D. & Storti, E., 2016. SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Computer Systems,* Volume 54, pp. 352-365.

Ding, L. Y. et al., 2013. Real-time safety early warning system for cross passage construction in Yangtze Riverbed Metro Tunnel based on the internet of things. *Automation in Construction ,* Volume 36, pp. 25-37.

Ding, R., 2016. Communication Is the Key. *In Key Project Management Based on Effective Project Thinking,* Volume Springer Berlin Heidelberg, pp. 275-298.

Dokas, I., Karras, D. & Panagiotakopoulos, D., 2009. Fault tree analysis and fuzzy expert systems: Early warning and emergency response of landfill operations.. *Environmental Modelling & Software,* 24 (1), pp. 8-25.

Doloi, H. S. A. I. K. &. R. S., 2012. Analysing factors affecting delays in Indian construction projects. *International Journal of Project Management,* 30(4), pp. 479-489.

Don, 2011. *Use early warning indicators as a management tool.* [Online] Available at: http://www.donlowe.org/risk-management/use-early-warning-indicators-as-a-management-tool/
[Accessed 12 May 2016].

Dong, D., 2016. *Content-aware compression for big textual data analysis.* s.l.:s.n.

Donkers, S., Ledoux, H., Zhao, J. & Stoter, J., 2016. Automatic conversion of IFC datasets to geometrically and semantically correct CityGML LOD3 buildings. *Transactions in GIS,* 20(4), pp. 547-569.

Duffield, S. & Whitty, S., 2015. Developing a systemic lessons learned knowledge model for organisational learning through projects. *International journal of project management,* 33(2), pp. 311-324.

Dulewicz, V. a. H. M., 2000. Emotional intelligence: the key to future successful corporate leadership?. *Journal of general management,* 25(3), pp. 1-14.

Eastman, C., Teicholz, P., Sacks, R. & Liston, K., 2011. *BIM Handbook: a Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators.* 2nd ed ed. Hoboken, NJ, USA: John Wiley and Sons.

Ehsanifar, M., Hamta, N. & Hemesy, M., 2017. A Simulation Approach to Evaluate Performance Indices of Fuzzy Exponential Queuing System (An M/M/C Model in a Banking Case Study). *Journal of Industrial Engineering and Management Studies,* 4(2), pp. 35-51.

Elmualim, A. & Gilder, J., 2014. BIM: innovation in design management, influence and challenges of implementation. *Architectural Engineering and design management,* 10(3-4), pp. 183-199.

Emonet, R., Varadarajan, J. & Odobez, J., 2014. Temporal analysis of motif mixtures using Dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence,* 36(1), pp. 140-156.

Erohin, O., Kuhlang, P., Schallow, J. & Deuse, J., 2012. Intelligent utilisation of digital databases for assembly time determination in early phases of product emergence. *Procedia CIRP,* Volume 3, pp. 424-429.

Estruch, V., Ferri, C., Hernández-Orallo, J. & Ramírez-Quintana, M. J., 2006. Web Categorisation Using Distance-Based Decision Trees. *Electronic Notes in Theoretical Computer Science,* 157 (2), pp. 35-40.

Fang, L. & Qingyuan, B., 2010. *A refined weighted K-Nearest Neighbors algorithm for text categorization.* s.l., Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on, 15-16 Nov. 2010, 326-330.

Fan, H., Xue, F. & Li, H., 2014. Project-based as-needed information retrieval from unstructured AEC documents. *Journal of Management in Engineering,* 31(1), p. A4014012.

Feilmayr, C., 2011. *Text Mining-Supported Information Extraction: An Extended Methodology for Developing Information Extraction Systems.* s.l., Database and Expert Systems Applications (DEXA), 22nd International Workshop on, pp. 217-221.

Filippetto, A., Barbosa, J., Francisco, R. & Klein, A., 2016 . *A project management model based on an activity theory ontology.* s.l., In Computing Conference (CLEI), XLII Latin American (pp. 1-11). IEEE.

Fleming, Q. & Koppelman, J., 2016. Earned value project management. *Project Management Institute,* 12(4).

Floricel, S., Michela, J. & Piperca, S., 2016. Complexity, uncertainty-reduction strategies, and project performance. *International Journal of Project Management,* 34(7), pp. 1360-1383.

Flyvbjerg, B. B. N. a. R. W., 2003. Megaprojects and risk: An anatomy of ambition. *Cambridge University Press..*

Gajzler, M., 2010. Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry. *Technological and Economic Development of Economy,* 16(2), pp. 219-232.

Ganiz, M. C., Tutkan, M. & Akyoku, S., 2015. *A novel classifier based on meaning for text classification.* s.l., Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on, 2-4 Sept. 2015, 1-5.

Ganz, F., Barnaghi, P. & Carrez, F., 2016. Automated semantic knowledge acquisition from sensor data. *IEEE Systems Journal,* 10(3), pp. 1214-1225.

Gaurav, J., Ginwala, A. & Aslandogan, Y. A., 2004. *An approach to text classification using dimensionality reduction and combination of classifiers.* s.l., Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on, 8-10 Nov. 2004, 564-569..

Gee, D. & Greenberg, M., 1896–2000. *5. Asbestos: from 'magic'to malevolent mineral. Late lessons from early warnings: the precautionary principle,* s.l.: European Environment Agency, Environmental Issue Report No 22.

Getoor, L., 2007. *Introduction to statistical relational learning.* London: MIT press.

Ghaffari, M. & Emsley, M., 2015. Current status and future potential of the research on Critical Chain Project Management. *Surveys in Operations Research and Management Science,* 20(2), pp. 43-54.

Gilad, B., 2003. *Early warning: Using competitive intelligence to anticipate market shifts, control risk, and create powerful strategies.* New York: AMACOM – American Management.

Goldratt, E., 1990. Theory of constraints. *Croton-on-Hudson: North River.*

Golini, R. & Landoni, P., 2013. *International development projects: peculiarities and managerial approaches,* s.l.: Project Management Institute.

Gomez, J. C. & Moens, M.-F., 2014. Minimizer of the Reconstruction Error for multi-class document categorization. *Expert Systems with Applications,* 41(3), pp. 861-868.

González, P., González, V., Molenaar, K. & Orozco, F., 2013. Analysis of causes of delay and time performance in construction projects. *Journal of Construction Engineering and Management,* 140(1), pp. 2401-3027.

Gordon, C., Mulley, C., Stevens, N. & Daniels, R., 2013. How optimal was the Sydney Metro contract?: Comparison with international best practice. *Research in Transportation Economics,* 39(1), pp. 239-246.

Goswami, S. et al., 2016. A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal.*

Gowtham, S., Goswami, M., Balachandran, K. & Purkayastha, B., 2014. *An Approach for Document Pre-processing and K Means Algorithm Implementation.* s.l., In Advances in Computing and Communications (ICACC), 2014 Fourth International Conference on (pp. 162-166). IEEE.

Groen, B., Wouters, M. & Wilderom, C., 2017. Employee participation, performance metrics, and job performance: A survey study based on self-determination theory. *Management accounting research,* Volume 36, pp. 51-66.

Gunasinghe, U. & Alahakoon, D., 2010. *A biologically inspired neural clustering model for capturing patterns from incomplete data.* s.l., Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on, 17-19 Dec. 2010, 126-131.

Haidar, A., 2015. *Construction Program Management–Decision Making and Optimization Techniques.* s.l.:Springer.

Haji-Kazemi, S. A. B. & Krane, H., 2013. A review on possible approaches for detecting early warning signs in projects. *Project management journal,* 44(5), pp. , pp.55-69..

Haji-Kazemi, S., Andersen, B. & Klakegg, O., 2015. Barriers against effective responses to early warning signs in projects. *International Journal of Project Management,* 33(5), pp. 1068-1083.

Hajjar, D. & AbouRizk, S., 2000. Integrating document management with project and company data. *Journal Computing in Civil Engineering, ASCE,* 14(1), p. 70–77.

Halter, J., 2015. A strategic approach to corporate communication: an analytic creating strategic alignment and measuring results. *Doctoral dissertation, University of Southern Queensland.*

Hamzah, N. et al., 2011. Cause of Construction Delay-Theoretical Framework.. *Procedia Engineering,* Volume 20, pp. 490-495.

Handford, M. & Matous, P., 2015. Problem-solving discourse on an international construction site: Patterns and practices. *English for Specific Purposes,* Volume 38, pp. 85-98.

Harding, J., 2012. Avoiding Project Failures. *Chemical Engineering,* 119(13), pp. 51-54..

Harisinghaney, A., Dixit, A., Gupta, S. & Arora, A., 2014. Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. *In Optimization, Reliabilty, and Information Technology (ICROIT),International Conference on. IEEE.,* pp. 153-155 .

Harrison, F. & Lock, D., 2017. *Advanced project management: a structured approach.* s.l.:Routledge.

Herbst, A., 2017. Capturing knowledge from lessons learned at the work package level in project engineering teams. *Journal of Knowledge Management.*

Herbst, A., 2017. Capturing knowledge from lessons learned at the work package level in project engineering teams. *Journal of Knowledge Management,* 21(4), pp. 765-778.

Hirschman, A., 2014. *Development projects observed.* 2 ed. s.l.:Brookings Institution Press.

H, N., ez & Ramos, E., 2012. *Automatic classification of academic documents using text mining techniques.* s.l., Informatica (CLEI), 2012 XXXVIII Conferencia Latinoamericana En, 1-5 Oct. 2012, 1-7.

Hoegl, M., Weinkauf, K. & Gemuenden, H., 2004. Interteam coordination, project commitment, and teamwork in multiteam R&D projects: A longitudinal study.. *Organization science,* 15(1), pp. 38-55.

Hoła, B., 2015. Identification and evaluation of processes in a construction enterprise. *Archives of Civil and Mechanical Engineering,* 15(2), pp. 419-426.

Hoła, B. & Sawicki, M. 2., 2014. *Tacit knowledge contained in construction enterprise documents.* s.l., Procedia Engineering, 85, pp.231-239..

Holopainen, M. & Toivonen, M., 2012. Weak signals: Ansoff today. *Futures,* 44(3), p. 198–205.

Hovde, M., 2014. Factors that enable and challenge international engineering communication: A case study of a United States/British design team. *IEEE Transactions on Professional Communication,* 57(4), pp. 242-265.

Hsu & J.Y., 2013. Content-based text mining technique for retrieval of CAD documents. *Automation in Construction,* Volume 31, pp. 65-74.

Hsu, J., 2013. Content-based text mining technique for retrieval of CAD documents. *Automation in construction,* Volume 31, pp. 65-74.

Huang, Y. et al., 2014 . Training Effectiveness and Trainee Performance in a Voluntary Training Program Are Trainees Really Motivated?. *Nonprofit and Voluntary Sector Quarterly,* 43(6), p. 1095–1110.

Huang, Z., Zhou, Z. & He, T., 2013. Associative classification with kNN. *Journal of Theoretical and Applied Information Technology,* 49(3), pp. 1013-1019.

Hui, S. C. & Jha, G., 2000. Data mining for customer service support. *Information & Management,* 38(1), pp. 1-13.

Hu, W., 2008. *Framework of Knowledge Acquisition and Sharing in Multiple Projects for Contractors.* s.l., . KAM '08. International Symposium on, 21-22 Dec. 2008, 172-176.

Ika, L., 2009. Project success as a topic in project management. *Project Management Journal,* 40(4), pp. 6-19.

Ilmola, L. & Kuusi, O., 2006. Filters of weak signals hinder foresight: monitoring weak efficiently in corporate decision-making. *Futures ,* Volume 38, p. 908–924.

Indhuja, K. & Reghu, R. P. C., 2014. *Fuzzy logic based sentiment analysis of product review documents.* s.l., Computational Systems and Communications (ICCSC), 2014 First International Conference on, 17-18 Dec. 2014, 18-22.

Ingason, A. et al., 2015. Expression analysis in a rat psychosis model identifies novel candidate genes validated in alarge case–control sample of schizophrenia. *Translational psychiatry,* 5(10), p. 656.

Isa, D., Lee, L. H., Kallimani, V. P. & RajKumar, R., 2008. *Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine.* s.l., IEEE Transactions on Knowledge and Data Engineering, 20(9), pp. 1264-1272..

Ithra, T. K. A. C. f. K. a. C., 2009. *kingabdulazizcenter.* [Online] Available at: http://www.kingabdulazizcenter.com/home-en/ [Accessed 13 June 2013-2014].

Jaffali, S. & Jamoussi, S., 2012. *Principal component analysis neural network for textual document categorization and dimension reduction.* s.l., '. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, 21-24 March 2012, 835-839..

Jiang, S., Zhang, H. & Zhang, J., 2013. Research on BIM-based Construction Domain Text Information Management. *JNW,* 8(6), pp. 1455-1464.

Jiang, W. & Wong, J. K. W., 2016. Key activity areas of corporate social responsibility (CSR) in the construction industry: a study of China. *Journal of Cleaner Production,* Volume 113, pp. 850-860.

Jia, R., Yuan, J., Li, Q. & Chen, Y., 2014. *The application of Dempster-Shafer theory in soft information management of construction projects.* s.l., In Management Science & Engineering (ICMSE), International Conference on (pp. 1814-1819). IEEE.

Jindal, R., Malhotra, R. & Jain, A., 2015. *Mining defect reports for predicting software maintenance effort.* s.l., Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on, 10-13 Aug. 2015, 270-276..

Jindal, R., Malhotra, R. & Jain, A., 2015. *Mining defect reports for predicting software maintenance effort.* s.l., Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on, 10-13 Aug. 2015, 270-276.

Jin, X., Wah, B., Cheng, X. & Wang, Y., 2015. Significance and challenges of big data research. *Big Data Research,* 2(2), pp. 59-64.

Jones, C., 2004. Software Project Management Practices: Failure Versus Success CrossTalk. *The Journal of Defense Software Engineering,.*

Kabakchieva, D., 2013. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies,* 13(1), pp. 61-72.

Kacprzyk, J. & Pedrycz, W., 2015. *Springer handbook of computational intelligence.* s.l.:Springer.

Kadir, M. K. A. et al., 2011. *Grain Security Risk Level Prediction Using ANFIS.* s.l., Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third.

Kaivo-oja, J., 2012. knowledge management theory and systemic socio-cultural transitions. *Futures 44 (2012), 206–217.,* 44(3), p. 206–217.

Kajava, J., Savola, R. & Varonen, R., 2005. Weak signals in information securitymanagement.. *Lecture Notes in Computer Science,* Volume 3802 , pp. 508-517.

Kappelman, L. A., McKee man, R. & Zhang, L., 2006. Early warning signs of IT project failure: The dominant dozen. *Information Management Systems,* 23(4), p. 31–36.

Karaali, O. et al., 1998. *A high quality text-to-speech system composed of multiple neural networks.* s.l., '. Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, 12-15 May 1998, 1237-1240 vol.2..

karenu, 2015. *stackoverflow.* [Online] Available at: http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and [Accessed 12 June 2016].

Kartam, N., 1994. *Knowledge-intensive database system for making effective use of construction lesson learned.* New York, In Proceedings Computing in Civil Engineering.

Kaur, R. & Singh, S., 2016. A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal,* 17(2), pp. 199-216.

Kaya, M. & Alhajj, R., 2005. *A novel approach to multiagent reinforcement learning: utilizing OLAP mining in the learning process.* s.l., IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35(4), pp. 582-590.

Kazaz, A. U. S. &. T. N. A. (., 2012. Causes of Delays in Construction Projects in Turkey. *Journal of Civil Engineering and Management,* 18(3), pp. 426-435.

Kelly, J., Male, S. & Graham, D., 2014. Value management of construction projects. *John Wiley & Sons..*

Kennedy, D., 2013. *The utility of planning in large projects.* s.l., Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET), July 28 2013-Aug. 1 2013, 1753-1760..

Kersting, H. & Železný, S., 2013. *Machine Learning and Knowledge Discovery in Databases.* s.l.:Springer.

Kerzner, H., 2013. *Project management: a systems approach to planning, scheduling, and controlling.* Eleventh edn ed. New Jersey: John Wiley & Sons, Inc, Hoboken,.

Kerzner, H., 2017. Project management metrics, KPIs, and dashboards: a guide to measuring and monitoring project performance. *John Wiley & Sons.*

Khan, A. B. B. L. L. H. a. k. K., 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technolodgy,* 1(1).

Kim, H. & Chang, J., 2007. *"Integrating Incremental Feature Weighting into Naive Bayes Text Classifier".* s.l., in Machine Learning and Cybernetics, 2007 International Conference on, pp. 1137-1143..

Kim, H., Soibelman, L. & Grobler, F., 2008. Factor selection for delay analysis using knowledge discovery in databases. *Automation in Construction,* 17(5), pp. 550-560.

Kim, M. S., Whang, K. Y. & Moon, Y. S., 2012. *Horizontal Reduction: Instance-Level Dimensionality Reduction for Similarity Search in Large Document Databases.* s.l., Data Engineering (ICDE), 2012 IEEE 28th International Conference on, 1-5 April 2012, 1061-1072.

King, J., 1987. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of information science,* 13(5), pp. 254-255.

Klakegg, O. J. W. T. D. B. &. M. O. M., 2012. *Early warning signs in complex projects.* Newtown Square: Projects Management Institute.

Klakegg, O., Williams, T. & Magnussen, O., 2009. *Governance frameworks for public project development and estimation.* s.l.:Project Management Institute..

Ko, C.-H. a. C. M.-Y., 2007. Dynamic prediction of project success using artificial intelligence. *Journal of Construction Engineering and Management-Asce,* 133(4), pp. 316-324..

Ko, C.-H. & Cheng, M.-Y., 2007. Dynamic prediction of project success using artificial intelligence. *Journal of Construction Engineering and Management-Asce,* 133(4), pp. 316-324..

Kondor, D. et al., 2013. *Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages.* s.l., Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, 2-5 Dec. 2013, 393-398.

Köster, K., 2009. *International project management.* pp.15-20 ed. s.l.:Sage .

Krier, M. & Zaccà, F., 2002. Automatic categorisation applications at the European patent office. *World Patent Information,* 24(3), pp. 187-196.

Kruschke, J., Aguinis, H. & Joo, H., 2012. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods,* 15(4), pp. 722-752.

Kumaraswamy, M. et al., 2017. Developing a clients' charter and construction project KPIs to direct and drive industry improvements. *Built Environment Project and Asset Management,* 7(3), pp. 253-270.

Kumar, B. & Ravi, V., 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems,* Volume 114, pp. 128-147.

Kumar, B. S. & Ravi, V., 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems,* Volume 114, pp. 128-147.

Kummamuru, S., 2014. HR Management Challenges of Indian IT Sector: An Application of the Viable Systems Model. *ASCI Journal of Management,* 43(2), pp. 1-17.

Kunz, J., 2015. Objectivity and subjectivity in performance evaluation and autonomous motivation: An exploratory study. *Management Accounting Research,* Volume 27, pp. 27-46.

Kuosa, T., 2010. Futures signals sense-making framework (FSSF): A start-up tool to analyse and categorise weak signals, wild cards, drivers, trends and other types of information.. *Futures ,* Volume 42, p. 42–48.

L. Yuan, C. Y. X. G. a. Y. Y., 2010 . Text-Aided Image Classification: Using Labeled Text from Web to Help Image Classification. *Web Conference (APWEB)12th International Asia-Pacific ,* pp. 267-273.

Lam, H. W., Dillon, T. & Chang, E., 2010. *Towards the use of semi-structured annotators for Automated Essay Grading.* s.l., Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on, 13-16 April 2010, 228-233.

Lamkanfi, A., Demeyer, S., Soetens, Q. D. & Verdonck, T., 2011. *Comparing Mining Algorithms for Predicting the Severity of a Reported Bug.* s.l., Software Maintenance and Reengineering (CSMR), 2011 15th European Conference on, 1-4 March 2011, 249-258.

Larouk, O. & Batache, M., 1995. *An evaluation the uncertainty of textual data with logic and statistics.* s.l., Uncertainty Modeling and Analysis, 1995, and Annual Conference of the North American Fuzzy Information Processing Society. Proceedings of ISUMA - NAFIPS .

Larsen, K. R., Monarchi, D. E., Hovorka, D. S. & Bailey, C. N., 2008. Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems. *Decision Support Systems,* 45(4), pp. 884-896.

Leach, L., 1999. Critical chain project management improves project performance. Project Management Journal, 30, pp.39-51.. *Project Management Journal,* Volume 30 , pp. 39-51.

Lee, H., Oh, H., Kim, Y. & Choi, K., 2015. Quantitative analysis of warnings in building information modeling (BIM). *Automation in Construction,* Volume 51, pp. 23-31.

Lee, W., Tse, K. & Ma, W., 2016. Applied Technologies in Minimizing Accidents in Construction Industry. *Procedia Environmental Sciences,* Volume 36, pp. 54-56.

Legodi, I. a. B. M. L., 2010. 'The current challenges and status of risk management in enterprise data warehouse projects in South Africa. *Technology Management for Global Economic Growth (PICMET), Proceedings on PICMET '10.*

Lei, D. et al., 2007. *The Strategies of Initial Diversity and Dynamic Mutation Rate for Gene Expression Programming.* s.l., Natural Computation, 2007. ICNC 2007. Third International Conference on, 24-27 Aug..

Leung, R. W. K. L. H. C. W. a. K. C. K., 2003. On a responsive replenishment system: a fuzzy logic approach. *Expert Systems,* 20(1), pp. 20-32.

Lewis, J., 2006. *The Project Manager's Desk Reference.* 3 ed. s.l.:McGraw-Hill Pub. Co..

Lewis, J. P., 1993. The Project Manager's Desk Reference, Probus Publishing Com.

Liberatore, M. J. & Pollack-Johnson, B., 2009. *Quality, time, and cost trade offs in project management decision making.* s.l., PICMET '09 - 2009 Portland International Conference on Management of Engineering & Technology, 2-6 Aug. 2009, 1323-1329..

Lientz, B. & Rea, K., 1995. *Project management fort he 21st century.* San Diego, Academic .

Li, H., n.d. *Project integration method based on knowledge set theory in science and technology project management.* s.l., In Information Management, Innovation Management and Industrial Engineering (ICIII),International Conference on (Vol. 1, pp. 390-392). IEEE.

Li, H. & Sima, Q., 2015. Parallel mining of OWL 2 EL ontology from large linked datasets. *Knowledge-Based Systems,* Volume 84, pp. 10-17.

Lines, B. C., Sullivan, K. T., Smithwick, J. B. & Mischung, J., 2015. Overcoming resistance to change in engineering and construction: Change management factors for owner organizations. *International Journal of Project Management,* 33(5), pp. 1170-1179.

Liu, S. & Lee, I., 2015. *A Hybrid Sentiment Analysis Framework for Large Email Data.* s.l., Intelligent Systems and Knowledge Engineering (ISKE), 2015 10th International Conference on, 24-27 Nov. 2015, 324-330..

Liu, S. & Lee, I., 2015. *A Hybrid Sentiment Analysis Framework for Large Email Data.* s.l., Intelligent Systems and Knowledge Engineering (ISKE), International Conference on, 24-27 Nov. 2015, 324-330.

Li, Z. et al., 2014. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering,* 26(9), pp. 2138-2150.

Loss, J., 1987. AEPIC project: update. *Journal of Performance of Constructed Facilities, ASCE,* 1(1), p. 11–29.

Love, P., Teo, P., Morrison, J. & Grove, M., 2016. Quality and safety in construction: creating a no-harm environment. *Journal of Construction Engineering and Management,* 142(8), p. 05016006.

Lu, W., Chen, X., Peng, Y. & Shen, L., 2015. *Benchmarking construction waste management performance using big data.* s.l., Resources, Conservation and Recycling, 105, Part A, pp. 49-58. .

Lv, L. & Liu, Y., 2005. Research of English text classification methods based on semantic meaning. *International Conference on Information and Communication Technology,Enabling Technologies for the New Knowledge Society, IEEE,* pp. 689-700.

Mao, W., Zhu, Y. & Ahmad, I., 2007. Applying metadata models to unstructured content of construction documents: A view-based approach. *Automation in Construction,* 16 (2), pp. 242-252..

Mao, Y. h., Liu, Y. & Li, Q. c., 2009. *Investment Decision-Making of Construction Projects Based on Modified Risk-Adjusted Discount Rate.* s.l., Information Science and Engineering (ICISE), 2009 1st International Conference on, 26-28 Dec. 2009, 4355-4358.

Marjaba, G. & Chidiac, S., 2016. Sustainability and resiliency metrics for buildings–Critical review. *Building and Environment,* Volume 101, pp. 116-125.

Martínez-Rojas, M., Marín, N. & Vila, M., 2015. The role of information technologies to address data handling in construction project management. *Journal of Computing in Civil Engineering,* 30(4), p. 4015064.

MartíN-Valdivia, M., MartíNez-CáMara, E., Perea-Ortega, J. & UreñA-LóPez, L., 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications,* 40(10), pp. 3934-3942.

Matharage, S., Ganegedara, H. & Alahakoon, D., 2013. *A scalable and dynamic self-organizing map for clustering large volumes of text data.* s.l., Neural Networks (IJCNN), The 2013 International Joint Conference on, 4-9 Aug. 2013, 1-8.

Matta, A., Chahed, S., Sahin, E. & Dallery, Y., 2014. Modelling home care organisations from an operations management perspective. *Flexible Services and Manufacturing Journal,* 26(3), pp. 295-319.

Maynard, D., Bontcheva, K. & Rout, D., 2012. *Challenges in developing opinion mining tools for social media.* s.l., Proceedings of the@ NLP can u tag# usergeneratedcontent, pp.15-22..

McCombie, C. & Jefferson, M., 2016. Renewable and nuclear electricity: Comparison of environmental impacts. *Energy Policy,* Volume 96, pp. 758-769.

McFarlan, W., 1982 . Portfolio Approach to Information Systems. *Journal of Systems Management ,* pp. 12-19.

McKeeman, R., 2001. *Early Warning Signs of Project Failure,* s.l.: Report for University of North Texas Information Systems Research Center.

Mench, J. W., 2002. *Electrical education for construction engineers.* s.l., SoutheastCon. Proceedings IEEE, 2002, 147-151..

Mendonça, S., Cardoso, G. & Caraca, J., 2012. The strategic strength of weak signal analysis. *Elseveir ltd, Futures,* Volume 44, p. 218–228.

Mendonça, S., Cunha, M., Kaivo-oja, J. & Ruff, F., 2004. Wild cards, weak signals and organisational improvisation. *Futures,* 36(2), pp. 201-218.

Mendonça, S., Cunha, M., Ruff, F. & Kaivo-oja, J., 2009. Venturing into the wilderness:preparing for wild cards in the civil aircraft and asset-management industries. *Long Range Planning,* 42(1), pp. 23-41.

Merkl, D. & Schweighofer, E., 1997. *En route to data mining in legal text corpora: clustering, neural computation, and international treaties.* s.l., Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop on, 1-2 Sep 1997, 465-4.

Mintzberg, H., 1987. The strategy concept I: Five Ps for strategy. *California management review,* 30(1), pp. 11-24..

Mintzberg, H. L. J. a. A. B., 1998. Stratagy Safary: a guided tour through the wilds of strategic management..

Min, Y. & Shou-rong, L., 2013. *Influence of behavioral factors on project schedule management: A Monte Carlo method.* s.l., 2013 25th Chinese Control and Decision Conference (CCDC), 25-27 May 2013, 4831-4835.

Mir, F. & Pinnington, A., 2014. Exploring the value of project management: linking project management performance and project success. *International journal of project management,* 32(2), pp. 202-217.

Mohammadi, A. & Tavakolan, M., 2013. *Construction project risk assessment using combined fuzzy and FMEA.* s.l., IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint, 24-28 June 2013, 232-237.

Morphy, 2016. *Stakeholder Analysis.* [Online] Available at: http://www.stakeholdermap.com/stakeholder-analysis.html [Accessed 14 11 2016].

Moura, R., Beer, M., Patelli, E. L. J. & Knoll, F., 2016. Learning from major accidents to improve system design. *Safety Science,* Volume 84, pp. 37-45..

Mubarak, S., 2015. *Construction project scheduling and control.* s.l.:John Wiley & Sons.

Musa, H., Yacob, M., Abdullah, A. & Ishak, M., 2017. Enhancing subjective well-being through strategic urban planning: Development and application of community happiness index. *Sustainable Cities and Society.*

N. Kartam, I. F., 1997 . Constructability feedback systems: issues and illustrative prototype. *Journal of Performance of Constructed Facilities,* 11(4), p. 178–183.

Neto, R., Adeodato, P. & Salgado, A., 2016. A Framework for Data Transformation in Credit Behavioral Scoring Applications Based on Model Driven Development. *Expert Systems with Applications.*

Ngai, E., Xiu, L. & Chau, D., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications,* 36(2), pp. 2592-2602.

Nicolaescu, S., Palade, H., Dumitrascu, D. & Kifor, C., 2017. A new project management approach for R&D software projects in the automotive industry-continuous V-model. *International Journal of Web Engineering and Technology,* 12(2), pp. 120-142.

Nikander, I., 2002. *Early warnings: a phenomenon in project management.* Helsinki University of Technology: s.n.

Nikander, I. O. a. E. E., 2001. Project management by early warnings. *International journal of project management ,* 19(7), pp. 385-399 .

Njadat, M., Salo, F. & Nassif, A. B., 2016. Data mining techniques in social media: A survey. *Neurocomputing,* Volume 214, pp. 654-670.

Nokhbeh Zaeem, R., Manoharan, M., Yang, Y. & Barber, K. S., 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security,* Volume 65 , pp. 50-63.

Odeh, A. M. &. B. H. T., 2002. Causes of construction delay: traditional contracts. *International Journal of Project Management ,* 20(1), pp. 67-73.

Oehmen, J., Thuesen, C., Ruiz, P. & Geraldi, J., 2015. Complexity management for projects, programmes, and portfolios. *In Project management Institute ,* Volume 1, pp. 2-32.

Olaru, M., Şandru, M. & Pirnea, I. C., 2014. *Monte Carlo Method Application for Environmental Risks Impact Assessment in Investment Projects.* s.l., Procedia - Social and Behavioral Sciences, 109, pp. 940-943.

Olawale, Y. & Sun, M., 2015. Construction project control in the UK: Current practice, existing problems and recommendations for future improvement. *International Journal of Project Management,* 33 (3), pp. 623-637.

Oppong, G., Chan, A. & Dansoh, A., 2017. A review of stakeholder management performance attributes in construction projects. *International Journal of Project Management,* 35(6), pp. 1037-1051.

Othman, I., Idrus, A. & Napiah, M., 2011. *Effectiveness of Human Resource Management in Construction project.* s.l., National Postgraduate Conference (NPC), 2011, 19-20 Sept. 2011, 1-6. .

Otsuka, N. & Matsushita, M., 2014. *Constructing knowledge using exploratory text mining.* s.l., Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on, 3-6 D.

Pang, X.-L., Feng, Y.-Q. & Ieee, 2006. *An improved economic early warning based on rough set and support vector machine.* s.l., Proceedings of 2006 International Conference on Machine Learning and Cybernetics, Vols 1-7, pp. 2444-2449..

Patanakul, P., 2014. Managing large-scale IS/IT projects in the public sector: Problems and causes leading to poor performance. *The Journal of High Technology Management Research,* 25(1), pp. 21-35.

Pedrycz, W., 2014. Allocation of information granularity in optimization and decision-making models: towards building the foundations of granular computing. *European Journal of Operational Research,* 232(1), pp. 137-145.

Pérez-Godoy, M.D., R. A., Carmona, C. & del Jesús, M., 2014. Training algorithms for radial basis function networks to tackle learning processes with imbalanced data-sets. *Applied Soft Computing,* Volume 25, pp. 26-39.

Pinto, J., 2014. Project management, governance, and the normalization of deviance. *International Journal of Project Management,* 32(3), pp. 376-387.

Pitigala, S., Li, C. & Seo, S., 2011. *A comparative study of text classification approaches for personalized retrieval in PubMed.* s.l., Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, 12-15 Nov. 2011, 919-921.

Pi, T. L. X. & Zhang, Z., 2014. *Structural Bregman Distance Functions Learning to Rank with Self-Reinforcement.* s.l., Data Mining (ICDM), 2014 IEEE International Conference on, 14-17 Dec. 2014, 500-509.

Polig, R. et al., 2014. Giving Text Analytics a Boost. *EEE Micro,* 34(4), pp. 6-14.

Ponsteen, A. & Kusters, R., 2015. Classification of human-and automated resource allocation approaches in multi-project management. *Procedia-Social and Behavioral Sciences,* Volume 194, pp. 165-173.

Porwal, A. & Hewage, K., 2013. Building Information Modeling (BIM) partnering framework for public construction projects. *Automation in Construction,* Volume 31, pp. 204-214.

Proverbs, D. G. & Holt, G. D., 2000. 'Reducing construction costs: European best practice supply chain implications. *European Journal of Purchasing & Supply Management,* 6(3–4), pp. 149-158.

Puche, J. et al., 2016. Systemic approach to supply chain management through the viable system model and the theory of constraints. *Production Planning & Control,* 27(5), pp. 421-430.

Puntheeranurak, S. & Sanprasert, S., 2011. *Hybrid Naive Bayes Classifier Weighting and Singular Value Decomposition Technique for Recommender System.* s.l., Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on, 15-17 July 201.

Qazi, A., Quigley, J., Dickson, A. & Kirytopoulos, K., 2016. Project Complexity and Risk Management (ProCRiM): Towards modelling project complexity driven risk paths in construction projects. *International Journal of Project Management,* 34(7), pp. 1183-1.

Raghuram, S. et al., 2009. *Bridging Text Mining and Bayesian Networks.* s.l., Network-Based Information Systems, 2009. NBIS '09. International Conference on, 19-21 Aug. 2009, 298-303.

Ramanathan, C. T. & Narayanan, S. P., 2014. *A comparative study among stakeholders on causes of time delay in Malaysian multiple design and build projects.* s.l., 2014 IEEE International Conference on Industrial Engineering and Engineering Management, 9-12 Dec. 2011.

Ramisch, C., 2014. *Multiword expressions acquisition: A generic and open framework.* s.l.:Springer.

Rathore, D., Jain, R. & Ujjainiya, B., 2013. *A text mining method for research project selection using KNN.* s.l., In Green Computing, Communication and Conservation of Energy (ICGCE), International Conference on (pp. 900-904). IEEE.

Rathore, H., 2016. *Artificial Neural Network.* s.l.:In Mapping Biological Systems to Network Systems (pp. 79-96). Springer International Publishing.

Ratsiepe, K. B. & Yazdanifard, R., 2011. Poor Risk Management as One of the Major Reasons Causing Failure of Project Management. *Management and Service Science (MASS),International Conference on,* pp. 1-5.

Romero, C. V. S. a. G. E., 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education,* 51(1), pp. 368-384.

S. R. M. Nasir, I. C. M. M. I. a. J. M. J., 2012. Comparative studies on factors influencing success completion of a project. *Science and Engineering (CHUSER),* pp. 319-324.

Sambasivan, M. a. S. Y., 2007. Causes and effects of delays in Malaysian construction industry. *International Journal of project management,* 25 (5), pp. 517-526.

Saritas, O. & Smith, J., 2011. The big picture — trends, drivers, wild cards, discontinuities and weak signals. *Futures,* 43(3), p. 292–312.

Saritas, O. & Smith, J., 2011. The big picture — trends, drivers, wild cards, discontinuities and weak signals. *Futures,* Volume 43, p. 292–312.

Scanlin, J., 1998. The Internet as an enabler of the Bell Atlantic project office. *Project Management Institute.*

Schaufelberger, J. & Holm, L., 2017. Management of construction projects: a constructor's perspective. *Taylor & Francis.*

Schmidt, R. L. K. a. M. K. P., 2001. Identifying software project risks: An international Delphi study. *Journal of management information systems,* 17(4), pp. 5-36..

Schoemaker, P., Day, G. & Snyder, S., 2013. Integrating organizational networks,weak signals, strategic radars and scenario planning. Technol. Forecast. Soc. Chang.. *Technological Forecasting and Social Change,* 80(4), p. 815–824..

Schouten, K. & Frasincar, F., 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering,* 28(3), pp. 813-830.

Sears, S. et al., 2015. Construction project management. *John Wiley & Sons.*

Sedita, S. R. & Apa, R., 2015. 'The impact of inter-organizational relationships on contractors' success in winning public procurement projects: The case of the construction industry in the Veneto region'. *International Journal of Project Management,* Volume 33.

Senaratne, S. & Ruwanpura, M., 2016. Communication in construction: a management perspective through case studies in Sri Lanka. *Architectural Engineering and Design Management,* 12(1), pp. 3-18.

Shaikh, S. & Darade, M., 2017. Key Performance Indicator for Measuring and Improving Quality of Construction Projects.

Shariff, A. H. M. a. K. S., 2011. Leveraging unstructured data into intelligent information: analysis & evaluation. *In Proceedings of the 2011 International Conference on Information and Network Technology ,* pp. 153-157.

Sheffield, J. & Lemétayer, J., 2013. International Journal of Project Management. *Factors associated with the software development agility of successful projects,* 31(3), pp. 459-472.

Silva, T. J., M. & Chen, Y., 2015. Process analytics approach for R&D project selection. *ACM Transactions on Management Information Systems (TMIS),* 5(4), p. 21.

Sivarajah, U., Kamal, M., Irani, Z. & Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* Volume 70, pp. 263-286.

Sjödin, D., Frishammar, J. & Eriksson, P., 2016. Managing uncertainty and equivocality in joint process development projects. *Journal of Engineering and Technology Management,* Volume 39, pp. 13-25.

Slankas, J. & Williams, L., 2013 . *Automated extraction of non-functional requirements in available documentation.* s.l., In Natural Language Analysis in Software Engineering (NaturaLiSE), 1st International Workshop on (pp. 9-16). IEEE.

Smith, B., Damphousse, K. & Roberts, P., 2006. *Pre-incident indicators of terrorist incidents: the identification of behavioral, geographic, and temporal patterns of preparatory conduct,* FULBRI: ARKANSAS UNIV FAYETTEVILLE TERRORISM RESEARCH CENTER .

Söderholm, A., 2008. Project management of unexpected events. *International Journal of Project Management,* 26(1), pp. 80-86.

Sohrabi, M. & Barforoush, A., 2012. Efficient colossal pattern mining in high dimensional datasets. *Knowledge-Based Systems,* Volume 33 , pp. 41-52.

Soibelman, J., M.ASCE & H. Kim, 2002. Data preparation process for construction knowledge generation through Knowledge Discovery in Databases. *Journal of Computing in Civil Engineering,* 16(1), p. 39–48.

Soibelman, L. & Kim, H., 2002. Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering,* 16(1), pp. 39-48.

Soibelman, L. et al., 2008. Management and analysis of unstructured construction data types. *Advanced Engineering Informatics,* 22(1), pp. 15-27.

Soucy, P. & Mineau, G., 2001. *A simple KNN algorithm for text categorization.* s.l., In Data Mining, ICDM, Proceedings IEEE International Conference on (pp. 647-648). IEEE.

Speicher, K. et al., 2013. Estimation of plate temperatures in hot rolling based on an extended Kalman filter. *IFAC Proceedings Volumes,* 46(16), pp. 409-414.

Suliman, M. O., Kumar, V. S. S. & Abdulal, W., 2011. *Optimization of uncertain construction time-cost trade off problem using simulated annealing algorithm'. Information and Communication Technologies (WICT).* s.l., World Congress on, 11-14 Dec. 2011, 489-494.

Suli, Z. & Xin, P., 2011. *A novel text classification based on Mahalanobis distance.* s.l., Computer Research and Development (ICCRD), 2011 3rd International Conference on, 11-13 March 2011, 156-158.

Sun, S., Luo, C. & Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion,* Volume 36, pp. 10-25.

Sweis, G., Sweis, R., Abu Hammad, A. & Shboul, A., 2008. Delays in construction projects: The case of Jordan. *International Journal of Project Management,* 26(6), pp. 665-674.

Syamil, A., Doll, W. & Apigian, C., 2004. Process performance in product development: measures and impacts. European Journal of Innovation Management, 7(3), pp.205-217. *European Journal of Innovation Management,* 7(3), pp. 205-217.

Tabassi, A. A. et al., 2016. Leadership competences of sustainable construction project managers. *Journal of Cleaner Production,* Volume 124, pp. 339-349.

Tabatabaei, N., 2011. Detecting weak signals by internet-based environmental scanning. *Master thesis Waterloo University.*

Tahan, M., Tsoutsanis, E., Muhammad, M. & Karim, Z., 2017. Performance-based health monitoring, diagnostics and prognostics for condition-based

maintenance of gas turbines: A review. *Applied Energy,* Volume 198 , pp. 122-144.

Tascikaraoglu, A. & Uzunoglu, M., 2014. A review of combined approaches for prediction of short-term wind speed and power. *Renewable and Sustainable Energy Reviews,* Volume 34, pp. 243-254.

Thamhain, H., 2013. Managing risks in complex projects. *Project Management Journal,* 44(2), pp. 20-35.

Thorleuchter, D., Scheja, T. & Van den Poel, D., 2014 . Semantic weak signal tracing. *Expert Systems with Applications,* 41 (11), p. 5009–5016.

Tianyi, J. & Tuzhilin, A., 2006. 'Segmenting Customers from Population to Individuals: Does 1-to-1 Keep Your Customers Forever?. *Knowledge and Data Engineering, IEEE Transactions on,* 18(10), pp. 1297-1311.

Tixier, A., Hallowell, M., Rajagopalan, B. & Bowman, D., 2016. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction,* Volume 62, pp. 45-56.

Trautmann, N. & Baumann, P., 2009. *Resource-constrained scheduling of a real project from the construction industry: A comparison of software packages for project management.* s.l., IEEE International Conference on Industrial Engineering and Engineering Manage.

Tsai, W. H., Lin, S. J., Lin, W. R. & Liu, J. Y., 2009. The relationship between planning & control risk and ERP project success. *Industrial Engineering and Engineering Management (IEEM), IEEE International Conference on,* pp. 1835-1839.

Tseng, Y.-H., Lin, C.-J. & Lin, Y.-I., 2007. Text mining techniques for patent analysis. *Information Processing & Management,* 43(5), pp. 1216-1247.

Tsirakis, N., Poulopoulos, V., Tsantilas, P. & Varlamis, I., 2016. PaloPro: monitoring the public opinion in social media and news streams. *Journal of Systems and Software..*

Ulhaq, I., Khalfan, M., Maqsood, T. & Le, T., 2017. Development of a conceptual framework for knowledge management within construction project supply chain. *International Journal of Knowledge Management Studies,* 8(3-4), pp. 191-209.

Ur-Rahman, N. & Harding, J., 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications,* 39(5), pp. 4729-4739.

Uskali, T., 2005. Paying attention to weak signals: the key concept for innovation journalism. *Innovation journalism,* 11(2), pp. 1-20.

Verma, V. K., Ranjan, M. & Mishra, P., 2015. *Text mining and information professionals: Role, issues and challenges.* s.l., Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on, 6-8 Jan. 2015, 133-137.

Vieira, D. C. d. L., Adeodato, P. J. L., Gon, P. M. & alves, 2010. *Improving reinforcement learning algorithms by the use of data mining techniques for feature and action selection.* s.l., Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on, 10-13 Oct. 2010, 1863-1870.

Vijayan, V., Bindu, K. & Parameswaran, L., 2017. *A comprehensive study of text classification algorithms.* s.l., In Advances in Computing, Communications and Informatics (ICACCI), International Conference on (pp. 1109-1113). IEEE.

Viljamaa, E. & Peltomaa, I., 2014. Intensified construction process control using information integration. *Automation in Construction,* Volume 39, pp. 126-133.

Walker, A., 2015. *Project management in construction.* s.l.:John Wiley & Sons.

Wallace, W., 1971. The Logic of Science in Sociology [sound Recording] . In: s.l.:Transaction Publishers.

Wang, H., Zhang, J., Chau, K. & Anson, M., 2004. 4D dynamic management for construction planning and resource utilization. *Automation in Construction,* 13(5), pp. 575-589.

Wang, L., Rege, M., Dong, M. & Ding, Y., 2012. Low-Rank Kernel Matrix Factorization for Large-Scale Evolutionary Clustering. *IEEE Transactions on Knowledge and Data Engineering,* 24(6), pp. 1036-1050.

Wang, Z. & Chu, L., 2010. *The algorithm of text classification based on rough set and support vector machine.* s.l., Future Computer and Communication (ICFCC), 2010 2nd International Conference on, 21-24 May 2010, V1-365-V1-368..

Warmerdam, A. et al., 2017. Workplace road safety risk management: An investigation into Australian practices. *Accident Analysis & Prevention,* Volume 98, pp. 64-73.

W., Chen, X., Ho, D. & Wang, H., 2016. Analysis of the construction waste management performance in Hong Kong: the public and private sectors compared using big data. *Journal of Cleaner Production,* Volume 112, pp. 521-531.

Web, 2013. *Pmsolutions.* [Online] Available at: http://www.pmsolutions.com/images/uploads/PMMM_Graphic_2013.jpg [Accessed 14 November 2016].

Weick, K. & Sutcliffe, K., 2001. Managing the unexpected: Assuring high performance in an age of complexity. *JosseyBass, a John Wiley & Sons.*

Weihong, Z. & Mingming, Z., 2009. *The research on network systems of quality control in construction stage.* s.l., Industrial Engineering and Engineering Management, 2009. IE&EM '09. 16th International Conference on, 21-23 Oct. 2009, 1160-1164.

Weiss, S. M. et al., 1999. *Maximizing text-mining performance.* s.l., IEEE Intelligent Systems and their Applications, 14(4), pp. 63-69.

Wen-Tsao, P., 2008. *The Study of Impact from the GAANFIS Model on Business Performance.* s.l., Computer Science and Information Technology, 2008. ICCSIT '08. International Conference on, Aug. 29 2008-Sept. 2 2008, 233-237.

Wex, F., Widder, N., Liebmann, M. & Neumann, D., 2013. *Early Warning of Impending Oil Crises Using the Predictive Power of Online News Stories.* s.l., System Sciences (HICSS), 2013 46th Hawaii International Conference on, 7-10 Jan. 2013, 1512-1521..

Whyte, J., Stasis, A. & Lindkvist, C., 2016. *Managing change in the delivery of complex projects: Configuration management, asset information and 'big data.* s.l., International Journal of Project Management, 34(2), pp.339-351..

Williams, T., 2016. Identifying success factors in construction projects: A case study. *Project Management Journal,* 47(1), pp. 97-112.

Williams, T. & Gong, J., 2014. Predicting construction cost overruns using text mining numerical data and ensemble classifiers. *Automation in Construction,* Volume 43, pp. 23-29.

Williams, T. et al., 2012. Identifying and Acting on Early Warning Signs in Complex Projects", Project Management Journal, 43:2, April, pp. 37–53.. *Project Management Journal,* 43(2), p. 37–53.

Wiltshire, A., 2006. *Developing early warning systems: A checklist.* , . In Proc. 3rd Int. Conf. Early Warning (EWC)..

Wise, R. et al., 2014. Reconceptualising adaptation to climate change as part of pathways of change and response. *Global Environmental Change,* Volume 28, pp. 325-336.

Witten, I., Frank, E., Hall, M. & Pal, C., 2011. Data Mining: Practical machine learning tools and techniques . In: s.l.: Morgan Kaufmann, p. 389.

Witten, I., Frank, E., Hall, M. & Pal, C., 2016. Data Mining: Practical machine learning tools and techniques. In: s.l.:. Morgan Kaufmann., pp. 512-516.

Wondimu, P. et al., 2016. *Early Contractor Involvement in Public Infrastructure Projects.* s.l., 24th Ann. Conf. of the Int'l. Group for Lean Construction, Boston, MA, USA (pp. 13-22)..

Wu, Z., Ann, T. & Shen, L., 2017. Investigating the determinants of contractor's construction and demolition waste management behavior in Mainland China. *Waste Management,* Volume 60, pp. 290-300.

Xu, J. & Li, Z., 2012. Multi-objective dynamic construction site layout planning in fuzzy random environment. *Automation in Construction,* Volume 27, pp. 155-169.

Xu, Y., Wang, D. & Liu, C., 2013. *Contemporary service theories integrated into construction project management.* s.l., In Service Systems and Service Management (ICSSSM), 10th International Conference on (pp. 90-95). IEEE.

Yang, H. C. & Lee, C. H., 2010. *A novel self-organizing map algorithm for text mining.* s.l., System Science and Engineering (ICSSE), 2010 International Conference on, 1-3 July 2010, 417-420..

Yan, X., Ye, Y. & Lou, Z., 2015. Unsupervised video categorization based on multivariate information bottleneck method. *Knowledge-Based Systems,* Volume 84, pp. 34-45.

Yongpeng, Y., Manoharan, M. & Barber, K. S., 2014. *Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories.* s.l., Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, 24-26 Sept. 2014, 184-191.

Yun, S., Choi, J., de Oliveira, D. & Mulva, S., 2016. Development of performance metrics for phase-based capital project benchmarking. *International Journal of Project Management,* 34(3), pp. 389-402.

Yusof, N. A., Zainul Abidin, N., Zailani, S. H. M. G. K. & Iranmanesh, M., 2016. Linking the environmental practice of construction firms and the environmental behaviour of practitioners in construction projects. *Journal of Cleaner Production,,* Volume 121, pp. 64-71.

Yu, S. & Zhang, J., 2009. *A Class Core Extraction Method for Text Categorization.* s.l., Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on, 14-16 Aug. 2009, 3-7.

Zaghbib, A., Pernelle, P. & Carron, T., 2012. *Managing the accompaniment to change using ict and experiential training: R&D Context.* s.l., Computing Technology and Information Management (ICCM), 2012 8th International Conference on, 24-26 April 2012, 265-272.

Zarei, B., Sharifi, H. & Chaghouee, Y., 2018. Delay causes analysis in complex construction projects: a Semantic Network Analysis approach. *Production Planning & Control,* 29(1), pp. 29-40.

Zavadskas, E. K., Vilutienė, T., Turskis, Z. & Šaparauskas, J., 2014. Multi-criteria analysis of Projects' performance in construction. *Archives of Civil and Mechanical Engineering,* 14(1), pp. 114-121.

Zeng, D. et al., 2016. Sentiment prediction by text mining medical documents using optimized swarm search-based feature selection. *Computerized Medical Imaging and Graphics.*

Zeng, J., Cheung, W. & Liu, J., 2013. Learning topic models by belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 35(5), pp. 1121-1134.

Zhai, C. & Massung, S., 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* s.l.:Morgan & Claypool.

Zhang, L., 2014. *Automatic signature generation for malicious PDF files.* s.l., U.S. Patent 8,695,096..

Zhang, L. et al., 2014. Bayesian-network-based safety risk analysis in construction projects. *Reliability Engineering & System Safety,* Volume 131, pp. 29-39.

Zhang, Y., Tsai, F. S. & Kwee, A. T., 2011. Multilingual sentence categorization and novelty mining. *Information Processing & Management,* 47(5), pp. 667-675.

Zhao, Z. Y. & Xue, B. X., 2010. *Construction change factors of direct current transmission line project and their impact on schedule and costs.* s.l., Industrial and Information Systems (IIS), 2010 2nd International Conference on, 10-11 July 2010, 171-174.

Zidane, Y., Andersen, B., Johansen, A. & Ahmad, S., 2016. “Need for Speed”: framework for measuring construction project pace–case of road project. *Procedia-Social and Behavioral Sciences,* Volume 226, pp. 12-19.

Zubrinic, K., Kalpic, D. & Milicevic, M., 2012. The automatic creation of concept maps from documents written using morphologically rich languages. *Expert systems with applications,* 39(16), pp. 12709-12718.

Zurada, J. & Fife, S. T., 2006. *Data Mining, Neural Networks and Rule Extraction.* s.l., IEEE CI Distinguish Lecture'. Neural Network Applications in Electrical Engineering, 2006. NEUREL 2006. 8th Seminar on, Sept. 2006, 1-1.

APPENDICES

Appendix A

Project early warnings:

1. Lack of keen commitment to project top management

2. Lack of keen commitment to project milestone

3. Lack of keen commitment to project scope

4. Lack of stable project requirements,

5. Lack of stable project scopes

6. Lack of stable project responsibility

7. Lack of stable project milestone

8. Lack of team required knowledge

9. Lack of team required skills

10. Lack of making purchases

11. Lack of materials on site

12. Lack of manpower resource

The Distribution of the Elements by the Level of Exactitude:

Inexact Signal: This is an uncertain warning level which will draw from keywords subtly indicating problems where the source is detectable but the information is very inexact

Clear Signal: This category will identify warnings via textual indicators but it would not be possible to give a number to those warnings.

Exact Signal: This category will precisely indicate each warning with clear numbers.

No signal: In the case of a normal project running condition, ideally, there should not be a warning at all as represented by this class

In addition to responding to the following questions, please select from the above Early warnings number and factors to inform how significantly the EWS and the factor(s) identified in the question contribute to the success/failure of the project

A.1

Q: Please specify keywords in project documents that indicate to the clients lack of following up of standards and procedures (e.g. proper submission of documents prior to attending follow-up /kick-off meetings)

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: Please indicate the factors that contribute to red-tape issues that uselessly delay the project (e.g. delay in permit issuance resulting in project delays)

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: What keywords indicate uncertainty in decision making for certain project activities such as schedule date confirmation?

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: Is there an impact if a proportion of sub-contractor task force is rejected due to some reason? Please state any possible reasons and the associated impacts.

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: What impact do you think experienced staff absence (e.g. leave) has on the project's progress?

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: Do think a sudden change of assignment of key persons to a certain task is a reason of alarm? and how is identified by key words?

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

Q: Please specify how lack of supply transport or an improper supply chain can be identified and how/what alternatives be arranged by key words.

Early warnings Number ( )

Inexact Signal Clear Signal Exact Signal No Signal

The research questionnaire is available in Google forms(https://docs.google.com/forms/d/14zCO9we2ZPq2ArQ0XIuS_SInyGK9cJyCDrBYVTQNJ0w/viewform?c=0&w=1&usp=mail_form_link)

A.2 List of questions, class mappings and scores

Table 39: List of questions used in the interview to identify text trends indicating early warnings in project management documents

Selected Questions	**Sample answer from one of the respondents**	**Extracted terms for Naïve Bayes training** ***Term (Score)***	**Class identified by the expert** **(Class code)**
How do you think specific staff delay affects different aspects of a project? Please name various types of delays?	delay in finishing the tasks assigned, delay in response and not attending meetings	VB: Delay (3), VB: Finishing (1), VB: Assigned (1), VB: Response (1), AD: Not Attending (1) CN: Tasks (1),	Keen commitment to project scope (KS)

		CN: Meetings (1)	
How supply chain problems can be identified in management documents?	untracked long lead item	VB: Untracked (1), AD: Long (1) CN: Lead (1), CN: Item (1)	Key management support (MS)
Explain the impact of lack of scope of personnel representing the managements to have to the overall efficiency of the project?	idle workers whereas the shortage of useful staff generally lead to tasks not completed on time	VB: Idle (1) CN: Workers (1), VB: Shortage (1), AD: Useful (1) CN: Staff (1), CN: Tasks (1), VB: Not completed on time (1)	Manpower resource (MR)
Please specify keywords in meeting Minutes that indicate to the key-project lack of following up of standards and procedures?	Drawing not meeting the specification	CN: Drawing (1), VB: Meeting (1), CN: Specification (1)	Team required skills (RS)
Please specify specific terms used in the meeting Minutes that indicate direct impact due to pending issues on the overall technical performance of the	Failure, restrictions, unavailability, poor performance	VB: Failure (1), VB: Restrictions (1), VB: Unavailability (1), AD: Poor (1), VB: Performance (1)	Stable project responsibility (SP)

project?			
What impact do you think increase employee absence has on the overall project performance? Also, please name a set of factors in project meeting Minutes that indicate problems due to absenteeism	quality manager not come or absent, no acting manager	CN: Quality manager (1), VB: Absent, VB: No acting manager (1)	Keen commitment to project scope (KS)
Please indicate the factors that contribute to red-tape issues that uselessly delay the project?	delay in document issuance, lack of properly documented guidelines	VB: Delay (1), VB: Documented (1) CN: Issuance (1), VB: Lack (1), AD: Properly (1), CN: Document (1), CN: Guidelines (1)	Team's required knowledge (RK)
What keywords indicate uncertainty in decision making for certain project activities?	Lack of the impetus needed to get the project underway	VB: Lack (1), CN: Impetus (1), VB: Needed (1), CN: Project (1), VB: Underway (1)	Keen commitment to project milestone (KM)
What modes of communication and activities indicate serious problems in a project?	Generally reminder letters indicate a more serious situation due	VB: Unavailability (1), VB: Lack (1), VB: Serious (1),	Stable project requirements (SR)

	to their documented change	CN: Change (1)	
Please exemplify how shortages and progress delays are indicated in project meeting Minutes?	Unavailability of materials, lack of inventory and task completion delays	CN: Unavailability (1), CN: Lack (1), VB: Delays (1),	Key management support (MS)
What practices in project management indicate lack of prioritization in various parties	excessive over-drafts and finance management	AD: Excessive (1), CN: Overdraft (1), CN: Finance (1), CN: Management (1)	Making purchases (MP)
Please state how repeated requests of similar items/quantities/actions be identified from the project documents	Requests not addressed, delays in delivery, completion or management of certain tasks	VB: Delays (1), VB: Completion (1), VB: Management (1),	Materials on site (MO)
Please state what impact the lack of resources such as not hiring an experienced engineers have and how can this impact be identified?	Any suddenly increased requirement will likely delay the project in such a case	ADJ: Suddenly, VB: Increased (1), CN: Requirement (1), ADJ: Likely (1), VB: Delay (1), CN: Project (1)	manpower resource (MR)
Please state how	"item not	CN: Item (1),	Materials on site

shortages in certain items can be identified in project documents and how can this be compensated	supplied, item unavailable in inventory"	ADJ: Not Supplied (1), ADJ: Unavailable (1), CN: Inventory (1)	(MO)

A.3 Naïve Bayes outcome

A.3.1 The main class

```
Class is : no signal. Value is : -122.949530029261
Result for TDM meeting minutes-optimisé-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -643.707065487727
Result for Weekly Minutes of Meeting 178-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : inexact signal. Value is : -504.018190366127
Result for Weekly Minutes of Meeting 179-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is inexact singal

Class is : no signal. Value is : -737.846765570398
Result for Weekly Minutes of Meeting 180-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is inexact signal

Class is : no signal. Value is : -847.920488214194
Result for Weekly Minutes of Meeting 181-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -1146.641757203
Result for Weekly Minutes of Meeting 182-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal
```

```
Class is : no signal. Value is : -887.668819065624
Result for Weekly Minutes of Meeting 183-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -952.466202670352
Result for Weekly Minutes of Meeting 184-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is exact signal

Class is : no signal. Value is : -1262.51632403429
Result for Weekly Minutes of Meeting 185-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : clear signal. Value is : -1057.10722730631
Result for Weekly Minutes of Meeting 186-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is clear signal

Class is : no signal. Value is : -1169.06977113218
Result for Weekly Minutes of Meeting 187-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -837.189472839812
Result for Weekly Minutes of Meeting 188-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -759.079105498071
Result for Weekly Minutes of Meeting 189-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -905.491662808821
Result for Weekly Minutes of Meeting 190-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is inexact

Class is : inexact. Value is : -357.717427019082
Result for Weekly Minutes of Meeting 191-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
```

```
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -660.334235229413
Result for Weekly Minutes of Meeting 192-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -802.73945935493
Result for Weekly Minutes of Meeting 193-2014.pdf
Classification is 6->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\Team_required_knowledge_skills.csv
Classification is inexact

Class is : inexact. Value is : -1.05880603826769
Result for Weekly Minutes of Meeting 194-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -943.255862298376
Result for Weekly Minutes of Meeting 195-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -857.13082858617
Result for Weekly Minutes of Meeting 196-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -822.680815101288
Result for Weekly Minutes of Meeting 197-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -660.932072230168
Result for Weekly Minutes of Meeting 198-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -1067.14509550014
Result for Weekly Minutes of Meeting 199-2014.pdf
Classification is 6->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\Team_required_knowledge_skills.csv
Classification is inexact

Class is : inexact. Value is : -1.05880603826769
Result for Weekly Minutes of Meeting 200-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
```

```
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -939.846366113899
Result for Weekly Minutes of Meeting 201-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -626.482058745286
Result for Weekly Minutes of Meeting 204-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -961.078706041572
Result for Weekly Minutes of Meeting 205-2014.pdf
Classification is 6->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\Team_required_knowledge_skills.csv
Classification is inexact

Class is : inexact. Value is : -1.05880603826769
Result for Weekly Minutes of Meeting 206--201417 JULY- AT.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -802.73945935493
Result for Weekly Minutes of Meeting 207--2014- 24 JULY.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : exact signal. Value is : -651.721731858192
Result for Weekly Minutes of Meeting 208-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is exact signal

Class is : no signal. Value is : -144.181869956935
Result for Weekly Minutes of Meeting 209-2014.pdf
Classification is 6->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\Team_required_knowledge_skills.csv
Classification is inexact

Class is : inexact. Value is : -1.05880603826769
Result for Weekly Minutes of Meeting 210-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -1120.11109990877
Result for Weekly Minutes of Meeting 213-2014.pdf
```

```
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -1528.21294436081
Result for Weekly Minutes of Meeting 214-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : exact signal. Value is : -1420.25773372018
Result for Weekly Minutes of Meeting 215-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is exact signal

Class is : no signal. Value is : -1324.922359636
Result for Weekly Minutes of Meeting 216-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : no signal. Value is : -1433.57071745719
Result for Weekly Minutes of Meeting 217-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : clear signal. Value is : -1592.60311132439
Result for Weekly Minutes of Meeting 218-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is clear signal

Class is : no signal. Value is : -1890.03339613188
Result for Weekly Minutes of Meeting 219-2014.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : clear signal. Value is : -1644.27813155172
Result for Weekly Minutes of Meeting 220.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is clear signal

Class is : no signal. Value is : -1902.65323268833
Result for Weekly Minutes of Meeting 221.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal
```

```
Class is : no signal. Value is : -1773.46568212003
Result for Weekly Minutes of Meeting 222.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is no signal

Class is : clear signal. Value is : -1885.42822594589
Result for Weekly Minutes of Meeting 223.pdf
Classification is 5->C:\work\my_doc\PhD\work\BayesScanSystem\BayesScanSystem\bin
\Debug\Class_Files\lack_of_stable_project_requirements_scops_responsibility_mile
stone-org.csv
Classification is exact signal


Class is : no signal. Value is : -952.561512850156 CONT!!!!!!!
```

A.4 KNN project outcome

A.4.1 Algorithm execution outcome

```
delay reporting of progress figures ,keen commitment to project scope,,,,,,,
   Proponent expressed their opinion that the actual progress at site has slow
progress ,Lack of key management support,,,,,,,
 mechanical trades appear to be more than what is seen in the weekly reports,Lack of
team required skills,,,,,,,
  Incomplete Structural Framing,keen commitment to project milestone,,,,,,,
            Planned date   31 January 2014 (as per contract amendment 5),Lack of
stable project requirements,,,,,,,
            Tower            incomplete drawings,Lack of team required skills,,,,,,,
            Library           uncomplete approval,Lack of team required skills,,,,,,,
            Keystone       status remains incomplate,Lack of team required
knowledge,,,,,,,
  Complete weathertight of Auditorium   15 April 14 (forecast 30 June 14),keen
commitment to project milestone,,,,,,,
  Complete weathertight of Keystone   15 April 14 (forecast 15 July 14),keen
commitment to project milestone,,,,,,,
  Complete weathertight of Library   15 May 2014 (forecast 30 June 14),keen
commitment to project milestone,,,,,,,
  Complete weather tight tower top   15 July 2014 (forecast 30 June 14),stable
project milestone,,,,,,,
  Energise the EC and start of HVAC   30 Jun 2014 (without snubbers),Lack of team
required knowledge,,,,,,,
 Outdated Schedule ,Lack of stable project responsibility,,,,,,,
   Client reqested Contractor to submit the updated schedule including Change ,keen
commitment to project scope,,,,,,,
 Orders awarded, but with time impact,Lack of stable project milestone,,,,,,
  Client emphasised that the ,Lack of stable project responsibility,,,,,,,
 completion date should be Clientme as the contract amendment No,Lack of stable
project milestone,,,,,,,
    Contractor clarified that their position is outlined in the correspondances and
also ,Lack of keen commitment to project scope,,,,,,,
 will reply to Client next week,Lack of stable project responsibility,,,,,,,
```

```
  RCA, Corrective, preventive action plan has,Lack of team required skills,,,,,
    been submitted officially dated May 19,2014,Lack of keen commitment to project
milestone,,,,,,
   3 HVAC,2Piping and Welding QC inspectors had been approved just waiting the PO
,Lack of making purchases,,,,,,
 6 are on ,Lack of stable project responsibility,,,,,,,
      board and awaiting for the 2 personnel to join the team,lack of manpower,,,,,,,
  Ncr is sugject for  ,Lack of stable project responsibility,,,,,,,
     Project closure,Lack of key management support,,,,,,,
    MAIN CONTRACTOR   to make sure site is ready (i,Lack of team required
skills,,,,,,,
  Inspection made on May 19,2014,Lack of stable project requirements,,,,,,
 2   preliminary design and calculation note to be issued by MAIN CONTRACTOR   as
soon as possible,Lack of team required skills,,,,,,,
  199 21 June MAIN CONTRACTOR   ,Lack of stable project responsibility,,,,,,,
  Shop Drawings will be issued next Week,Lack of stable project milestone,,,,,,,
  201 26 June CLIENT ,Lack of stable project requirements,,,,,,,
 Tower Platform  ,Lack of stable project responsibility,,,,,,,
   MAIN CONTRACTOR   is working on an alternative Main Contractor  ution for
scoffolding to accelerate and ,Lack of team required skills,,,,,,,
 provide continuity of progress for Seele pipes installation,Lack of keen commitment
to project scope,,,,,,,
 3   Planning and Preliminary Method Statement to be finalized by MAIN CONTRACTOR
on ,Lack of team required knowledge ,,,,,,,
 1   Bid Submitted by MAIN CONTRACTOR   to Contracting  North Park on 11 June,Lack of
stable project responsibility,,,,,,,
 2   Meeting with PROPONENT scheduled every Thursday  10hOO for follow up 193 Note
SOUCLIENT ,Lack of stable project scope,,,,,,,
   See correspondinQ Mom ,Lack of stable project responsibility,,,,,,,
 Rammed Earth  ,Lack of stable project responsibility,,,,,,,
  Inspection of Mock Up done by CLIENT/Proponent,Lack of stable project
responsibility,,,,,,,
   Visit of M/R is Expected on 07,Lack of stable project requirements,,,,,,,
   Factory at site is ordered, to be ready mid July,Lack of stable project
requirements,,,,,,
  Work shop at M/R is expected on 25 26 June 2014,Lack of stable project
responsibility,,,,,,,
   PROPONENT requested confirmation that DC Battery room will be ready as for ,Lack
of keen commitment to project milestone,,,,,,,
 temporary HVAC & Generators to Start Up the system on time,Lack of stable project
responsibility,,,,,,,
 4   MAIN CONTRACTOR   confirmed that Generators are on site,Lack of keen commitment
to project scope,,,,,,,
   MAIN CONTRACTOR   is waiting for confirmation of HVAC design ,Lack of key
management support,,,,,,,
  PROPONENT requested an action plan to expedite progress,ack of keen commitment to
project scope,,,,,,,
   Daily meeting is being held between MAIN CONTRACTOR   & PROPONENT,Lack of stable
project responsibility,,,,,,,
 8   MAIN CONTRACTOR   advised that material will be in Bahrain on 21 June &
installation shall 200 Note SOUCLIENT ,Lack of team required skills,,,,,,,
 be done by MAIN CONTRACTOR   under AQECLIENT supervision,Lack of stable project
responsibility,,,,,,,
   MAIN CONTRACTOR   submitted Traffic plan for Road 09 & will submit remaining plan
in July,Lack of keen commitment to project milestone,,,,,,,
  200 Closed MAIN CONTRACTOR   ,Lack of stable project responsibility,,,,,,,
 PROPONENT requested a comprehensive Layout for security items Lack of team required
skills,,,,,,,
 1 PROPONENT requested the Schedule of MAIN CONTRACTOR   for working Hours during
RAMADAN 201 22 Jun 14 MAIN CONTRACTOR   ,Lack of key management support,,,,,,,
```

```
 Notedbf V ,Lack of stable project responsibility,,,,,,,
    (3) new QC for MEP joined the site,Lack of team required skills,,,,,,,
  500007715 {Fire PumRs Aurora}  ,Lack of making purchases,,,,,,,
   MAIN CONTRACTOR   advised that NCR has been refuted through MAIN CONTRACTOR
letter KACWC ,Lack of key management support,,,,,,,
 CLIENTPID requested to avoid issuance of short notice Quality notifications and
,Lack of keen commitment to project scope,,,,,,,
 requested to comply with the 5 working days notice for VID inspection for any ,Lack
of keen commitment to project scope,,,,,,,
 procurement slow process,Lack of team required skills,,,,,,,
  Complete weather tight of Auditorium  15 April 14 (forecast 30 June 14) ,Lack
stable project milestone,,,,,,,
  Complete weather tight of Keystone  15 April 14 (forecast 15 July 14) ,Lack of team
required skills,,,,,,,
  Complete weather tight of Library  15 May 2014 (forecast 30 June 14) ,Lack of team
required skills,,,,,,,
  Complete weather tight tower top   15 July 2014 (forecast 30 June 14) ,Lack of key
management support,,,,,,,
  Energise the EC and start of HVAC   30 Jun 2014 (without snubbers) ,Lack of
materials on site,,,,,,,
 Subject ,Lack of stable project responsibility,,,,,,,
 stated that an updated schedule that include all awarded change ,Lack Lack of stable
project requirements,,,,,,,
 shall be submitted as promised by end of June 2014,Lack of keen commitment to
project milestone,,,,,,,
  PROPONENT ,Lack of stable project responsibility,,,,,,,CONT!!!!...
```

A.4.2 Algorithm execution outcome for the 6-class case

```
" '''>tnl""""nF'n again that all awarded change orders where it clearly stated ",
lack of stable project scopes and requirements,clear signal,,,,,,
 impact as not applicable shall not show any schedule impact,Lack of team required
knowledge and skills,no  signal,,,,,,,
 schedule shall be in accordance to Amendment No,Lack of team required knowledge and
skills,clear signal,,,,,,
 5 Completion ,Lack of stable project scopes and requirements,no signal,,,,,,
 PROPONENT stated that Contractor shall use the early Planned Progress Curve as ,Lack
of key management support,inexact signal,,,,,,
 a reference for project progress as it was used from the start of the project, Lack
of team required knowledge and skills,clear signal,,,,,,
 3 MAIN CONTRACTOR   clarified that there is no contract obligation to use early
curve and it is ,Lack of keen commitment to project scope,no signal,,,,,,
 purely the discretion of the Contractor to use the most practical one,Lack of keen
commitment to project scope,no signal,,,,,,
  Second, ,Lack of stable project scopes and requirements,no signal,,,,,
 MAIN CONTRACTOR   explained during the special meeting held on 22 June, that averga
curve ,Lack of stable project scopes and requirements,no signal,,,,,
 is internationally recognized as better and practical to control project, Lack of
team required knowledge and skills, inexact signal,,,,,,
  This was ,Lack of stable project scopes and requirements,no signal,,,,,,
 outlined also in MAIN CONTRACTOR   letters MW/4002 & 4046, Lack of key management
support,inexact signal,,,,,,
 MAIN CONTRACTOR   worked on an alternative Main Contractor  ution for scoffolding to
accelerate and , Lack of key management support,clear signal,,,,,,
 provide continuity of progress for Seele pipes installation,Lack of keen commitment
to project scope and milestone,exact signal,,,,,,
   Preliminary Method Statement to be finalized by MAIN CONTRACTOR   on 23 Jun
```

2014,Lack of making purchases,clear signal,,,,,,
A meeting held with PROPONENT on 18 June and MOM dipatched ,Lack of stable project scopes and requirements ,no signal,,,,,,
Meeting with PROPONENT scheduled every Thursday 1 OhOO for follow up , scopes and requirements,no signal,,,,,,
(23) Rooms uncompleted,Lack of keen commitment to project scopes and milestone,exact signal,,,,,,
(13) will be completed next week , keen commitment to project scope and milestone,no signal,,,,,,
Workshop at L TE (MlR) conducted by MAIN CONTRACTOR ON 25 26 June 2014, keen commitment to project scope and milestone,no signal,,,,,,
A visit of M/R is Expected on 07 July 2014 ,Lack of stable project scopes and requirements,inexact signal,,,,,,
Factory at site to be ready by end of July,Lack of key management support,,,,,,,
Subject ,Lack of stable project responsibility,no signal,,,,,,
DC Battery room PROPONENT requested confirmation that DC Battery room will ,Lack of keen commitment to project scope and milestone,clear signal,,,,,,
be ready as for temporary HVAC & Generators to Start Up the system on time,Lack of team required knowledge and skills,clear signal,,,,,,
CLIENT requested to expedite the availability of all Battery room installations such , Lack of team required knowledge and skills,exact signal,,,,,,
as fire extinguisher, Clientfety signs, temporary eye wash and shower,Lack of stable project scopes and requirements,no signal,,,,
MAIN CONTRACTOR advised that HVAC will be completed on 28 June ,Lack of stable project scopes and requirements,no signal,,,,,,
Arcade Wall ,Lack of stable project responsibility,,,,,,,
MAIN CONTRACTOR advised that (5) elements have been cast; (4) installed, but is waiting enhanced team ,Lack of manpower resource,clear,,,,,,
Lifts IEscalator ,Lack of manpower resource,no signal,,,,,,
PROPONENT request Status of Shipment,Lack of keen commitment to project scopes and milestone,clear signal,,,,,,
MAIN CONTRACTOR advised that status has been disapproved, Lack of team required knowledge and skills,exact signal,,,,,,
shared with Roger/CLIENT ,Lack of stable project responsibility,,,,,,,
PROPONENT requested Clientmples for Lighting,Lack of keen commitment to project scope and milestone,no signal,,,,,,
MAIN CONTRACTOR expects arrival by end of ,Lack of stable project responsibility,,,,,,,
Daily meeting is being held between MAIN CONTRACTOR regarding the pending PO & PROPONENT,Lack of making purchases,clear signal,,,,,, CONT!!!!!!!