The use of text mining to improve knowledge discovery in a project environment

CT Mushonga

orcid.org/0000-0002-9727-992X

Thesis accepted in fulfillment of the requirements for the degree *Doctor of Philosophy in Information Technology* at the North-West University

Promoter: Prof PD Pretorius

Graduation: July/August 2022

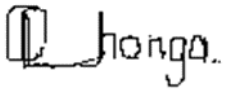
Student number: 27724255

Declaration

I, Cleopatra Tsungai Mushonga, declare that

The use of text mining to improve knowledge discovery in the project environment

is my own work and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.



Signature: __ _____ Date: 21 March 2022

Abstract

This thesis investigates how project managers can use knowledge discovery from text on the available project data to improve future projects using a petro-chemical company in South Africa as a case. This concept has been explored and developed in the analysis of scientific databases. Currently, project managers analyse project reports by manually reading through each report and attempting to decipher trends and patterns. The study employed an action research method. Empirical data was collected using interviews on project managers to establish the use of text data mining on the available data and implementing the action plan by analysing the data using the chosen text data mining technique. Other parts of the interviews entailed evaluating with the project managers whether the technique extracts the anticipated knowledge efficiently in a user-friendly format. The findings of the study confirm that the kind of information collected by different projects differ, but in general it includes the project scope, people as well as the terms of reference.

Information collected from projects should be used to learn and predict the future of future projects. Multiple participants concurred that information collected to learn and predict the future may certainly improve future projects. Project managers confirmed that lessons learned should be captured in a database where specific information can be retrieved as and when required. The data analysis process should be measured where lessons learned should show trends automatically and provide useful graphical reports. A well-structured database which can search or retrieve data from various projects to make sense of the lessons learned is ideal. The more predictable the database can be the better and the ability to convert learnings to easily extractable knowledge is important. A system that can show patterns or trends of critical information from the projects is needed.

The findings from this study confirm that there is significant amount of information gathered and created during the lifecycle of each project. In addition, most of the information is essential for the success of current projects and future projects. It is important to know the historic project performance for insights such as cost estimation, pitfalls from previous projects as well as scheduling. The project managers indicated that it is critical to have a centralized repository where all the projects' information is stored and accessible. A centralized repository would be instrumental in correlating data from similar projects.

Lessons learned reports are critical to the function of project managers in order to avoid re-learning a lesson that has been learnt in another project. It was also confirmed that project managers need a better way of analysis which summarizes the key aspects of the lessons learned as well as best practices. The key insights that were found to be important for project managers include summarization of the numerous reports into a digestible format. Furthermore, the study confirms that text mining could be used to improve knowledge discovery in the project environment. The study therefore recommends the use of Python text mining tool to discover knowledge from past projects. Finally, study proposed an 11th project management area called Project Knowledge Management. The application of Project Knowledge Management contributes the success of projects in project environments.

Keywords: Project Management, knowledge areas, Text Mining, lessons learned, issues, Project Knowledge Management

Acknowledgements

Firstly, I would like to thank God, from whom all wisdom and blessings flow. For advice and ideas, thanks are due to my study supervisor, Professor Philip Pretorius for his advice, encouragement, patience, trust, and guidance throughout my academic journey.

I would like to thank my family for being a constant source of inspiration. My dream of becoming a recognized scholar was first inspired by my late granduncle- Dr. Abraham Dumisani Maraire. There are many times when I wanted to give up, but my parents and my siblings always encouraged me to push on. A special thank you to my mother, Tsitsi Nyangari and my father, Zecks Nyangari for believing in me and supporting me since I started school as a child. My parents always believed that I would be a great scholar. To my siblings, Tendai, Simon, Salome, Craig and Zelda for being the best cheerleaders a sister could ask for.

To my husband Johnathan, and my children- Rutendo and Simba, I would like to express my deepest gratitude for their patience and continuous support throughout my studies. We encountered many challenges as a family during this journey including battling with depression, managing my studies through periods when we had to migrate to a new country and surviving the Covid-19 pandemic, among other things. It is truly the best feeling that this chapter is finally coming to an end as I look forward to new beginnings and giving you all my utmost attention.

I would also like to thank all participants who shared their time, knowledge, and experiences. Without you, this study was not going to be possible. I would like to further thank all those who contributed to this thesis, my associates Ved Bhardwaj and Sandeep Sharma and to all those who supported me but are not mentioned here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Introduction

The first chapter introduces the study by providing background information, a problem statement, and objectives of the study and the associated research questions. The chapter also explains the significance of the study as well as the organization of the chapters in this thesis. A brief overview of what is contained in each chapter is given in the chapter organization section. The chapter concludes with a chapter summary.

1.2 Research Background

Organizations are faced with large quantities of data coming from various information sources as a result of the advances in technology (Provost and Fawcett, 2013, Mizgier and Willis, 2014, da Silva Cezar and Maçada, 2021). One would expect an organization to be able to exploit this high availability of data and be able to learn from it in order to improve processes and profitability. However, the vast amounts of data have made the search and analysis of data more complex (Alamsyah and Peranginangin, 2013). More so, the analysis of unstructured textual data is intricate. Nevertheless, organizations are forced to exploit this data and derive potentially meaningful insights to learn from and consequently increase their productivity and much-needed competitive advantage (Davis and Bhattacharyya, 2016, Rajpathak and Narsingpurkar).

Knowledge discovery from databases (KDD) is a theory established by Piatetsky-Shapiro (1991) to help with the automatic discovery of knowledge implicit within the database. According to Piatetsky-Shapiro (2000), knowledge is the result of a data driven discovery process. KDD has been described in literature as “the overall process of discovering useful knowledge from data” (Fayyad et al., 1996). The aim of the KDD process is to discover new knowledge that can be utilised in a particular domain (Klösgen, 2021). There are nine steps to follow in the process of discovering useful knowledge from data follow, before arriving at the point of knowledge discovery. One of the steps in the KDD process is identified as data mining. Data mining is used to extract particular patterns from data by using algorithms (Data Mining Techniques, 2020). Knowledge discovery is therefore viewed as a process of finding useful insights from data and data mining is a step that is at the core of the KDD process.

The concept of data mining was developed to address the pertinent need for data analysis (Fayyad et al., 1996) with the goal of discovering hidden patterns and insights from available data. Numerous algorithmic tools of data mining exist that can be used to report on information. Each of these algorithmic tools is unique and suitable for extracting information

for a particular purpose. An extension of data mining is text data mining which is also known as knowledge discovery from textual databases (Tan, 1999)). Text mining is atypical to traditional data mining because data mining is performed on structured data while textual data mining is performed on unstructured textual databases.

This research focuses on knowledge discovery from text in the projects environment. Knowledge discovery in scientific databases using text mining and social network analysis has since been explored by Jalalimanesh (2012). Throughout the project lifecycle of each project, a significant amount of data is collected concerning that specific project (PMBOK, 2019). Cumulatively, organizations collect data from various projects which is usually stored in a database. This data can be structured or unstructured. According to PMBOK (2019), project data usually takes the form of reports such as project plans, lessons-learned reports, project review, project summary, and debriefing report among other useful documentation. Literature studies have shown that effective knowledge management can be attributed to project success (Emiliano de Souza et al., 2021). It is imperative to state that most of these documents are text intensive narratives. Retrieval of relevant information from such databases and identifying all the useful information is time consuming. The process requires careful selection of keywords as well as drafting of queries as highlighted by Themistocleous and Morabito (2017). This research explores whether the information collected here can be analysed for hidden patterns and insights that may not have been identified using conventional data analysis methods. The study assists project managers to analyse the textual data obtained from reports.

Since most of the documents are text intensive, this research implements knowledge discovery from textual databases (KDT) on project data. KDT was established by various scholars including Feldman and Dagan (1995), Kerzendorf (2019) as well as Davis and Bhattacharya (2016) as a means of handling unstructured data sets using text categorization. Feldman and Sanger (2007) later defined KDT as the process of finding information in comprehensive text collections and automatically identifying fascinating patterns and relationships in textual data. Other sources like Larose and Larose (2014) consider KDT as text mining or textual data mining.

In contrast to KDD which it evolved from, KDT applies TDM or simply text mining as its core process (Aggarwal and Zhai, 2012). Text mining has advanced tremendously to incorporate semantic search algorithms that can acquire a lexicon of synonyms and context for analysis (Rubin, 2012:1) rather than keyword/phrase search. As a result, more detailed and accurate insights may be discovered from the process of KDT. This study establishes the textual data

mining methods (TDM) that are suitable for analysing project data. The textual data mining method is deemed suitable based on consultation with the project managers about the applicability of the textual data mining method and relevance of the results to the projects environment. The following section gives us the purpose of the study.

1.3 Purpose of the study

The purpose of this study is to establish a way for project managers to improve the process of information retrieval, analysis, and reporting from the numerous project reports available to gain insights. Creswell (2016) affirms the importance for researchers to indicate the purpose of the study for the benefit of all stakeholders concerned. It is well known that research contributes to scholarly knowledge and improves processes in project management, especially for this study. Project management is a diverse field. It goes beyond simply tracking deadlines and allocating a budget. Effective project management involves monitoring a project from start to finish. It requires that the initiatives and goals are strategically aligned, the project has stakeholder support, and all parties are in accord. An effective project manager will ensure that the project goals and objectives are met (PMBOK, 2019). The results obtained from this study might help project managers to understand and apply text mining to improve knowledge discovery in the project environment.

Text mining will help project teams to transform unstructured text data into meaningful and actionable information with the aim to assist project management. The study explores the utilization of text mining to improve knowledge discovery in the project environment. Previous studies exclusively looked at text mining (Davis and Bhattacharyya, 2016) or knowledge discovery (Sokhanvar, Matthews and Yarlagadda, 2014) without a proper application of the text mining process in a real project. This gap in the literature justifies the need to thoroughly examine the use of text mining in project management and eventually offer a solution to the problems identified.

1.4 Problem statement

Several authors (Todorović et al., 2015, Polyaninova, 2011, Sokhanvar et al., 2014)) have written on knowledge management in project management environments. Knowledge management within project environment has been highlighted by Rex et al. (2018) as a key factor in the success of projects. According to Sokhanvar et al. (2014) the capturing and creation of knowledge are paramount to knowledge management. Studies by Mazorodze and Buckley (2019) confirmed that knowledge sharing is the key knowledge management process in knowledge intensive organizations. However, in most project management environments the focus is on knowledge capturing (Polyaninova, 2011).

The necessity of knowledge management in project management has been mentioned by scholars such as Polyaninova (2011), Pretorius and Steyn (2005), Bresnen et al. (2003) among others. Knowledge management is explained as a collection of activities, initiatives, and strategies that companies employ to create, store, transfer, and apply knowledge for the advancement of organizational performance (Donate and Sánchez de Pablo, 2015:362). Knowledge can be placed in two categories: tacit knowledge and explicit knowledge (Nonaka and Takeuchi, 2001). Tacit knowledge is the knowledge that is embedded in an individual's experiences, ideas and skills that cannot be easily articulated. On the other hand, explicit knowledge is knowledge that can easily be articulated to others in the form of documents or other tangible artefacts. The Project Management Institute (2008) states that project management requires the application of knowledge. Consequently, knowledge is one of the inputs required for the success of a project. As evident in literature (O'Dell and Hubert, 2011), knowledge may be acquired from various sources. This study focuses on the knowledge acquired from reports that are available from previous projects which is in essence, explicit knowledge.

The challenge for project managers is that the codified data is usually in formats that are poorly presented for analytical purposes (Serrat, 2012, Pretorius and Steyn, 2005). The codified knowledge in the project databases is unstructured and as such renders itself cumbersome and unwieldy for project managers. Moreover, the time constraints do not afford project managers substantial opportunity to examine the multitude of reports from the various projects. As a result, there is a lack of knowledge discovery and knowledge management from abundant project data that is collected. With the rise of data mining, the ease of data analysis has improved (Feldman and Sanger, 2017). However, the output from the data analysis needs to be in a format that is easily consumable by the project managers.

Camilleri (2010) claims that a general need exists to arrange and organise information for specific end-user problems. This research focuses on understanding the challenges that project managers face with the analysis of data from previous projects. Significant knowledge is generated throughout every project but the challenge that project managers face is retaining and utilizing this knowledge once each project has ceased (Rose et al., 2020). This research also investigates how project managers can benefit from using data mining techniques on the available project data to gain knowledge to improve present and future projects.

While there has been a number of studies of text mining in databases (Larose and Larose, 2014), there has been little or no research on mining documents in their existing structure. It

is almost certain that the data has to be pre-processed and cleaned. Yet, this potentially improves the insights that can be identified because some semantic meanings may be lost during the cleaning process. Therefore, it is necessary to optimize the process of knowledge discovery from the project data if an organisation is to become effective, efficient, and competitive in its operations. This research establishes how project managers can improve the process of information retrieval, analysis, and reporting from the massive project reports available to provide more insights. The focus will be on the lessons learned reports and the end result could be replicable on other types of reports.

The assumption for this study is that project managers can benefit from the knowledge that may be discovered from the data collected from various projects using text mining which analyses semantic meanings. Furthermore, Carrillo et al. (2011) established that there is need to improve the output of the data mining for the end user. This study establishes how the textual data mining methods (TDM) can improve the analysis of project data. The textual data mining method is deemed successful based on consultation with the project managers about the applicability and efficiency of the textual data mining method, and relevance of the results to the projects environment. The following section presents the research questions for the study.

1.5 Research questions

This study aimed to answer the following main research question.

> How can project managers improve analysis of project data using KDT to gain adequate knowledge to improve future projects?

The study builds on the following research sub-questions.

- What are the knowledge requirements and challenges for knowledge discovery in the projects environment?
- What is the value of KDT for the projects environment in improving project management processes?
- How can knowledge be discovered from projects data using KDT?
- How can project data management improve data driven decision making in project environments?

The research questions directed the researcher to find a solution to the problem. The questions are aligned with the study objectives described in the subsequent section. The questions were formulated from a comprehensive literature review.

1.6 Research objectives

The main objective of this study was to investigate whether KDT can improve the analysis of project data by yielding useful insights for the project environment that can become part of their body of knowledge.

The study was guided by the following sub-objectives which sought to:

- Determine the knowledge needs of the projects environment and identify. the current challenges for knowledge discovery from projects data.
- Determine and demonstrate the value of textual data mining for the projects environment for improving project management processes.
- Determine the methods of knowledge discovery from textual project data.
- Determine how project data management knowledge area improves data driven decision making in project environments.

1.7 Contribution to the field

Considering the growth of and high availability of data and information in organizations, it is increasingly crucial to be able to determine the knowledge that can be discovered. The knowledge era demands new insights that can improve effectiveness and efficiency in the economies. This research empirically determines the value of KDT for project managers and in turn organizations.

The study provides a comprehensive literature review in areas of knowledge management, project management and text mining, which contributes to the existing body of knowledge through the unification of approaches in these areas. The research provides a model of knowledge discovery from text that may be applicable to other organisations, and this enable organizations to gain better insights from their textual databases. The research adds to the existing body of knowledge on KDT, knowledge management and project management. The research proposes to add an eleventh project management area called project knowledge management.

This study closes the gaps to determine the applicability of using a text mining tool such as Python for text mining. The theoretical and methodological contribution is therefore the applicability of performing KDT using Python on lesson learned reports from projects. Below are the specifications of learning.

1. The knowledge environment is currently challenged with vast amounts of data that is not being analysed. As such, project managers are unable to discover useful patterns that can potentially improve the quality of their projects. The research provides a model knowledge discovery from text that may be applicable to other organisations, and this enable organisations to gain better insights from their textual databases.
2. Textual data mining can provide significant value to the project environment, particularly the analysis of lessons learned reports. Lessons learned reports are rich in project knowledge that informs future projects; hence it is crucial to find ways to analyse them efficiently and effectively.
3. Because of the vast amounts of data, a taxonomy should be created. A taxonomy can also improve search results by showing the levels immediately above, below, and adjacent to the search term in the hierarchy, providing both a meaningful context and ideas for further exploration.
4. Since knowledge management is a critical component to the success of a project, an eleventh area for project management body of knowledge is recommended called project knowledge management.

1.8 Limitations of the study

The study only focuses on how text mining can be utilized to improve knowledge discovery in the project environment. Knowledge discovery is the process of extracting useful knowledge from data. Thus, the solution developed is only applicable to project environments. The solution is developed to understand patterns and relationships within a dataset, consequently making the best decisions during projects.

1.9 Structure of thesis

This thesis has ten chapters. The chapters are organised as follows:

Chapter 1: Introduction and background

This chapter gives a synopsis of the study, the research problem, and the research objectives. More so, the chapter explains the importance and relevance of the study. It provides the background to the study on the use of text mining to improve knowledge discovery in the project environment. The problem statement is explained outlining the areas which are addressed by the study. To accompany the problem statement, a list of research questions and objectives have been clarified.

Chapter 2: Research strategy

This chapter presents the research strategy adopted to conduct this study in South Africa. The chapter discusses the data collection tools used to gather data together with full justification why they have been chosen. The chapter further explains the data analysis tools and concludes with a statement on ethics that the researcher abides by throughout the study. The chapter also provides the basis for data collection and analysis.

Chapter 3: Knowledge discovery and data mining

The third chapter reviews literature on knowledge discovery and data mining. The researcher also illustrates the significance of knowledge discovery and data mining in project environments. The literature review in this Chapter defines knowledge discovery which was the primary focus of this study. Knowledge discovery as well as its associated processes are discussed to lay a solid foundation to this specific study.

Chapter 4: Text mining

The fourth chapter reviews literature from accredited sources on text mining methods and processes. A variety of text mining methods and processes are explored. This literature review provides a segue into how text mining could be applied to projects.

Chapter 5: Knowledge discovery with text mining

This chapter discusses and proposes the actions to be taken to address the problem identified. Action research is employed with the aim to emancipate users from the bondage of their current project management processes by improving the same through text mining. The purpose of the action planning phase is to ensure that the outcome of the action solves the problems identified in the diagnosed area of application before commencing the actual development work. Further work may be required to improve the text mining methodology.

Chapter 6: Project Knowledge Management

The sixth chapter reviews literature on Project Management (PM) and the PM knowledge areas that are currently recognized in the industry. The chapter proposes the 11th project management knowledge area termed "*Project Knowledge Management*" and how this could be applied in project environments.

Chapter 7: Text data on data analysis

The seventh chapter interprets the qualitative data obtained from the research participants during interviews to determine the current problem that the Project Managers are currently faced with. The analysis of the data collected from project managers highlights that,

knowledge discovered should be utilized for the success of future projects. Project managers confirmed that they rely on expert discussions to determine which knowledge is useful to improve on the current projects and avoid relearning. The analysis also showed that knowledge extracted from lessons learned may certainly increase efficiency of the project. Participants, specifically project managers further submitted that there should be a proper database which can provide graphical reports that are easy on the eye. From the submissions we can infer that a well-structured database which has the ability to search or retrieve data from various projects to make sense of the lessons learned is ideal.

Chapter 8: Recommendations to analyse text data

The eighth chapter is the plan of action. It shows how the Project Knowledge Management area is formulated and crafted to fit into the project management domain. This chapter presents the plan of action which dealt with the needs for knowledge in the project environment through the problem-solving cycle. The chapter provides effective methods for textual data analysis in project environments. Moreover, the chapter demonstrates the value of textual data mining for the projects environment. The practical requirements for knowledge management during projects are also highlighted. The chapter concludes with emphasizing for analysis of the lessons learned from past projects.

Chapter 9: Results from Text Analysis using recommendations

This chapter makes recommendations based on the findings from text analysis. From the findings, the study recommends that, if one would like to use this algorithm seamlessly, it is best to use a standardized format for capturing the lessons learned. The study recommends project knowledge management area to improve data driven decision making in project environments.

Chapter 10: Summary, overall conclusion, and future work.

The chapter provides a comprehensive summary of how the research achieved the research objectives, provides a conclusion, and recommendations for future work.

Figure 1 below shows the structure of this thesis.

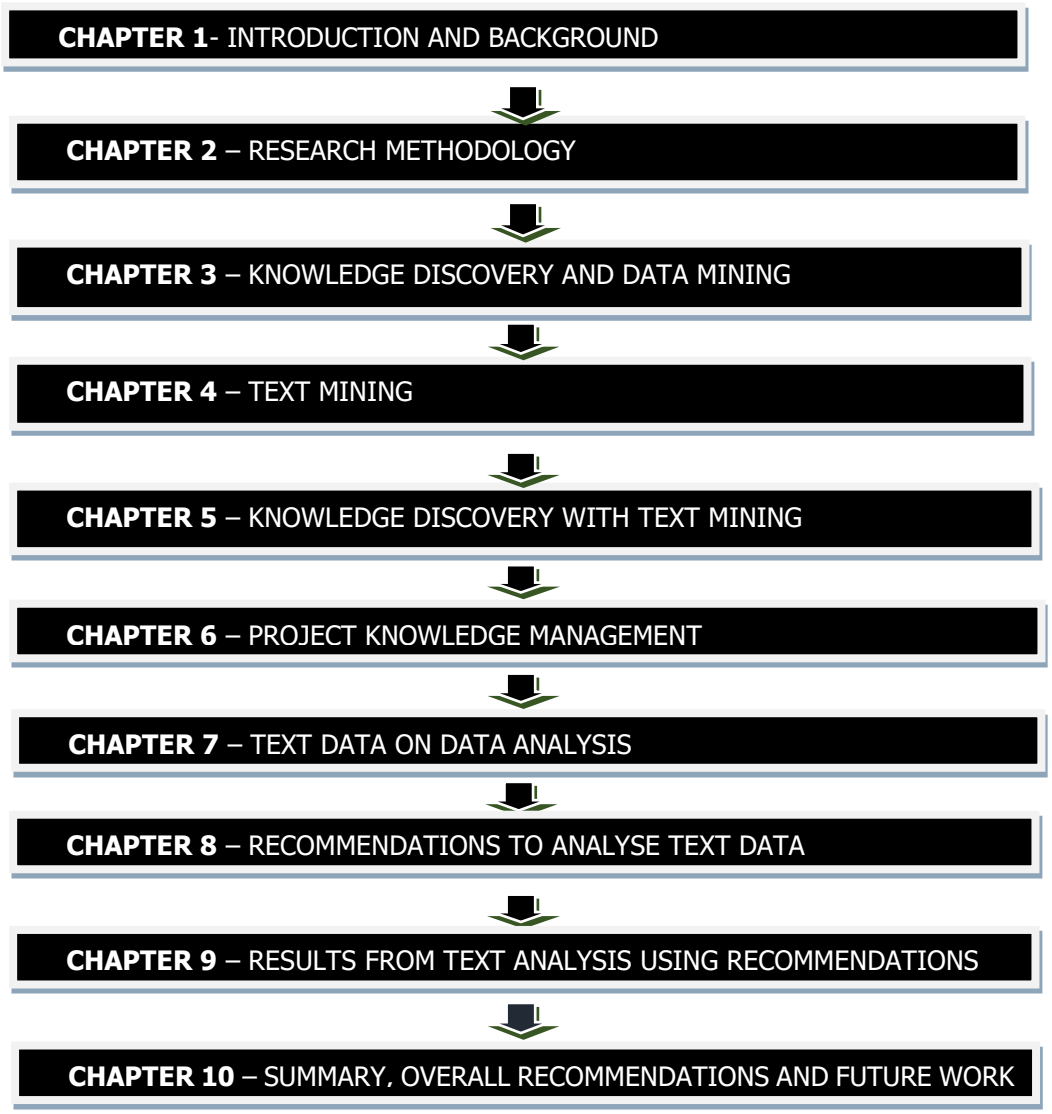


Figure 1 : Thesis structure

1.10 Chapter summary

The background to the study on the use of text mining to improve knowledge discovery in the project environment has been explained. An explanation of the problem statement given, outlining the areas which are addressed in the study. To accompany the problem statement, a list of research questions and objectives have been explained. The second chapter reviews literature from accredited journal articles, textbooks and other relevant online sources which are linked to both knowledge management and project management.

CHAPTER 2: RESEARCH STRATEGY

2.1 Introduction

In this chapter the methodology used during the study is explained. The chapter also discusses the research philosophy, approach, research paradigm, methodological choice, strategy, data collection tools, data analysis tools as well as necessary ethical considerations which the study abides by. The methodology adopted is of such importance that Ngulube (2005:12) asserts that the presentation of valid results is dependent on using the correct methodology. It is important to highlight that the research was conducted in the information systems domain to bridge the gap in project management and knowledge management. The following section describes the research philosophy.

2.2 Research philosophy and paradigm

According to Oates (2006) a research paradigm is a worldview about how things operate. Creswell (2016) and Saunders et al. (2012) compares paradigms to philosophies as ways in which the researcher views the world relative to the study. This kind of perception and understanding of the world in turn enlightens the approach, methodological choice and the strategy adopted in a specific study. A research paradigm is an accepted set of beliefs of how incidents are observed and analysed (Scotland, 2012) Thus, a paradigm is a way of looking at a set of beliefs and a way of thinking about the world. Different researchers have differing worldviews and as such operate in different research paradigms. Researchers align themselves to a world view that locates them in a particular frame of reference in line with their views of the world around them. A research paradigm is mainly defined by its ontological assumptions, i.e., what is believed to be reality, and its epistemological assumptions, i.e., how do we learn about our perceived reality (Scotland, 2012:9).

The four philosophies that are commonly used are positivism, interpretivism, realism and pragmatism. Philosophies (Saunders et al., 2012) and paradigms (Oates, 2006; Creswell, 2016) are defined by their ontology, epistemology, methodology and axiology. On one hand, ontology refers to what already exists and is a view about the nature of reality whereas epistemology is our perceived relationship with the knowledge we are realising. Saunders et al. (2012), define axiology as a branch of philosophy that studies judgements about a value. Ontology and epistemology make up the paradigmatic base of research in all disciplines, including information systems which is the domain of this study. Sarantakos (2013) suggests that methodological, epistemological, and ontological prescriptions of research are packaged in paradigms. In this chapter, the researcher explores the four research paradigms identified in terms of ontology, epistemology, methodology and axiology.

Ontological assumptions are beliefs about the essence of the reality of the phenomenon that is being studied. Thus, ontological questions strive to understand the nature of physical and social reality. According to De Figueiredo and Da Cunha (2007:9), the ontological question seeks to understand what can be known about reality. Ontology aims to understand what is real. A researcher's ontological beliefs influences his beliefs about the empirical world and determines what can be known about it. Sarantakos (2013) posits that ontology determines whether the researcher takes an objective stance or a subjective stance about reality. An objective reality assumes that the world exists independent of humans, whereas the subjective stance believes that the world's existence is dependent on humans Bhattacherjee (2012).

Epistemology is the convention that is applied for constructing and evaluating valid knowledge. The epistemological question seeks to understand what is knowledge (de Figueiredo and da Cunha, 2007:9); how we come to know what we know. The epistemology of a study is concerned about the creation, acquisition and expression of knowledge (Scotland, 2012:9). Epistemological assumptions are the beliefs of the acceptable knowledge. It is important to note that epistemology is dictated by a researcher's ontological beliefs. Methodology dictates how we discover knowledge in a systematic way and is driven by ontology and epistemology. Methodology defines suitable research methods for generating valid knowledge (Creswell, 2016) and relates to the process or procedures by which we create these knowledge claims.

Research paradigms are said to be incommensurable (Aldrich, 1992) in that it is difficult to compare competing paradigms. As such some paradigms are suitable for certain studies. The paradigms of Burrell & Morgan (2014) can be applied to certain studies but may not necessarily be applicable in behavioural information systems. In other words, the relationship between information and systems and human activity may not be purely understood in these paradigms. A number of scholars as cited by Myers and Klein (2011:19) have widely accepted a paradigm shift. Information systems research is complex and cannot exclusively be limited to the dominant perspectives of Burrell and Morgan (1979). The dominant perspectives limit information systems research particularly in "the aspects of phenomenon that can be studied and how it can be studied" Orlikowski and Baroudi (1991) .

Orlikowski and Baroudi (1991) identified three research paradigms which are positivist, interpretive and critical research that are suitable for the study of behavioural information systems. The majority of research in information systems have been conducted under two philosophical paradigms, which are positivism and interpretivism (Orlikowski and Baroudi, 1991). However, recent studies (Bhattacherjee, 2012) in information systems are being conducted in the critical research paradigm. Critical studies are designed to critique the

situation of affairs, eliminate contradictions and strive to reform organizations and societies for a better world (Asghar, 2013:3121). This study was conducted in the critical realism paradigm which is a philosophical paradigm popularised by Roy Bhaskar and it combines critical naturalism and transcendental realism (Archer et al., 2013). The following section describes positivism and interpretivism philosophies before the critical realism is described.

2.2.1 Positivism

The aim of the positivist paradigm is to statistically explore and grasp a phenomenon (Creswell, 2016). Positivism is deductive based on the scientific premise focusing on a hypotheses and verifying a priori hypotheses (Park et al., 2020). Consequently, positivist researchers hold that there is an absolute reality which can be measured scientifically. Absolute reality is a reality that is unaffected by the researchers beliefs or any limitations (Saunders et al., 2007). Scientific concepts are generally precise and have fixed meanings that cannot vary. Thus, scientific concepts are mainly focused on "testing, confirmation and falsification, and predictive abilities of generalizable theories about an objective, readily apprehended reality" (Wynn Jr and Williams, 2012a). Orlikowski and Baroudi (1991:6) mention the scientific nature of positivism by emphasizing the testing nature of positivist studies. It becomes clear that positivist researchers focus on testing to understand the phenomenon through the ability to predict.

Positivist research is often viewed synonymously with quantitative research (Creswell, 2016). We can therefore underscore that quantitative research is scientific in nature. The positivist inquiry in some cases is alluded to as the scientific method since it falls back on the thoughts of reductionism and determinism (Pather and Remenyi, 2004). In the real world, positivism is viewed as a paradigm which has provided the world with numerous logical and designing victories. When researchers started to direct their concentration toward how organizations and the people inside them worked, they looked towards the scientific method as their first methodology. This prompted another community known as social sciences. For this reason, positivism has been largely dominant in the studies of information systems (Orlikowski and Baroudi, 1991) The following section describes the ontological, epistemological, methodological and axiology aspects of the positivist philosophy.

2.2.1.1 Ontology of positivism

The positivist paradigm views the nature of reality as being objective (Oates, 2006). In most cases, positivists expect that the nature of reality is dispassionately given and can be depicted by quantifiable properties independent of the scientist and his or her instruments (Myers, 1997). Positivist researchers believe the social world exists independent of humans and that

the truth we perceive is autonomous from us. Therefore, the truth exists before we attempt to know it, it is conceivably understandable, reasonable by changeless laws, and it is multifaceted in nature (De Figueiredo and Da Cunha, 2007:10). Additionally, the social reality portrays human action as intentional and rational. In a study that uses the positivist approach, the researcher assumes a neutral stance and does not intervene. Furthermore, the results of the study are considered to be valid, reliable, and replicable.

2.2.1.2 Epistemology of positivism

Epistemologically, the positivist paradigm believes that knowledge is that which we can learn by investigating the reasons for the issues we encounter (Oates, 2006). The knowledge we obtain is subsequent to our investigation into the reasons for the issues. Faith in the deterministic theory communicates not just the likelihood of depicting a reality free from the subject, but also the likelihood of clarifying it in a perpetual way (de Figueiredo and da Cunha, 2007:11) The positivist paradigm is focuses on empirical verifiability and it applies deductive methods to predict patterns and behaviours across different situations. The positivist researcher believes that knowledge is what we learn by delving into the reasons for the problems we encounter and the solutions we obtain can be applicable in any context. Positivists are concerned about the hypothetic deductive testability of hypotheses. Scientific information ought to permit testability and generalization of the outcomes. In that capacity, a causal relationship is generally displayed and a close connection among clarification, forecast and control is normal (Orlikowski and Baroudi, 1991:7)

2.2.1.3 Methodology of positivism

According to Leedy and Ormrod (2010), methodology defines the research methods and techniques that are suitable for collecting evidence. Looking at methodology, positivists believe that research must have a purely objective position and utilize target estimation to gather investigate proof (Chen and Hirschheim, 2004). The positivist researcher makes use of quantitative methods that test theories or hypotheses (Gay, Mills and Airasian, 2012). Positivist research mostly uses the survey strategy.

2.2.1.4 Axiology of positivism

Positivists believe that all truth is obtained from methodical scientific processes. Objectivity and neutrality are central to the positivist researcher (Park et al., 2020). Therefore, any subjective inclinations such as experiences and values do not have any relevance to a positivist study. Neutrality means that the researcher should not interact with participants during the process of inquiry.

2.2.1.5 Limitations and weaknesses of positivism for this study

This study applies the data analytics technology in order to improve knowledge discovery in the project environment. The study involves human and computer interaction. It can therefore be considered as behavioural information systems or social science study. In social science, the researcher cannot be independent of the phenomenon they are studying (Reinecke and Bernstein, 2013). Positivism leads to studies that are primarily centred on observation and measurement, classification, experiment, and statistical analysis. The results of such investigations are intended to confirm or falsify hypotheses about an objectively observable, independent reality (Mingers, 2000). This view of science as a pure, objective perception. Hence, observation activity is not suitable for this study because it limits the brain to a blank slate on which the external world imposed itself. Yet, perception and conceptualization are evident construction of a phenomenon.

The key characteristic of positivism is its nomothetic epistemological position suggesting the existence of regularities or law-like speculations in material or social settings that give the premise to both clarification and forecast. This consistency enables positivists to believe that they can put forth causal expressions. On the off chance that two occasions take place sequentially and routinely, one is said to clarify the other. However, this basic and rich plan has various of issues making its utilization in any research study an issue. The central problem is that a steady combination of components or factors is not a clarification of any sort, because it doesn't answer the inquiry (Easton, 2010).

Mingers (2004) highlights falsification as one of the limitations of positivism in research. He further adds that theories often need to be developed regardless of initial failures, and not just be abandoned. In addition, Mingers (2004b) highlights that positivism disagrees, in many ways, with the actual practices of scientists and could therefore not adequately explain the de facto success of science.

Nevertheless, the statistical analysis approach of positivism can help in several other areas. Statistical analysis tends to be useful to identify specific patterns in data. Some statistical methods are more useful for uncovering hidden structures, particularly something like factor analysis which aims to reduce the dimensions of regular elements that includes a progression of interrelated conditions. Finally, the fundamental use of positivist statistical analysis may be in approving conceivable clarifications by substantiating or falsifying them. The following section discusses the interpretivist philosophy.

2.2.2 Interpretivism

According to Schaffer (2020), interpretivism aims to understand the social aspects of the phenomenon of interest in its natural setting. Reality and knowledge are social items unequipped for being understood free of the social actors. The world is viewed as a rising social procedure, an augmentation of human cognizance, and abstract understanding. The interpretivist intends to understand how individuals perceive the world in the social settings they find themselves in. To Orlikowski and Baroudi (1991), interpretivism endeavours to comprehend intersubjective importance implanted in social convictions and attempts to clarify business as usual. Interpretivists expect that scientific learning ought to be obtained by way of comprehending human and social interaction through which subjective meaning of reality is constructed (Chen and Hirschheim, 2004).

Interpretive researchers believe that access to reality is solely through social developments (Alharahsheh and Pius, 2020). Interpretivism aims to study the phenomenon through the perceptions of people. It centres around the full unpredictability of human sense-making as the conditions develop (Myers, 1997)Interpretivists repudiate the likelihood of understanding what is real and dismiss the likelihood of observing causality. Because interpretivists use their own interpretation it is difficult to determine which interpretation is superior to another (Easton, 2010). Interpretivists recognize that the issue which they are looking into exists in a social setting and that the method most suitable for grasping activities of social performing artists may actually not be through numbers and thorough measurable tests. The following section discusses the ontological and epistemological dimensions of the interpretivist philosophy.

2.2.2.1 Ontology of interpretivism

The interpretivist ontology believes that the nature of reality is socially constructed (Alharahsheh and Pius, 2020). Interpretivism believes that the subjects construct reality through their interactions with the world (De Figueirdo and Da Cunha, 2007:10)(de Figueiredo and da Cunha, 2007:10) As alluded by Schaffer (2020), in the interpretivist paradigm, subjective meanings are critical. Thus, interpretivists endeavours to see how and why people offer significance to the world. The interpretivist assumes that the social world is not given but rather created and fortified by human activity and association. Because interpretivist paradigm assumes that importance and request are shaped and arranged (Orlikowski and Baroudi, 1991), social reality must be translated.

2.2.2.2 Epistemology of interpretivism

The interpretivist paradigm assumes that understanding social procedure includes getting inside the universe of those creating it. The interpretivist worldview asserts that understanding

social practices requires seeing how practice and implications are framed. The interpretivist believes that knowledge brings an intended outcome (Peng, 2013). It is, along these lines, a characteristic result of the phenomenological theory and it plays a part in the development of information (de Figueiredo and da Cunha, 2007:11).

2.2.2.3 Methodology of interpretivism

Interpretivists (Schaffer, 2020; Alharahsheh and Pius, 2020) contend that to understand the significance of human and social intercommunication, researchers need to be part of the social setting researched to examine how the interaction takes place from the members' point of view. Researchers who form part of the genuine social setting being studied, will be able to present studies progressively suitable for producing interpretive information (Chen and Hirschheim, 2004). Interpretivism can be studied from a non-deterministic point of view. Interpretivist research can be conducted through the commitment of the researcher in the specific social setting explored. The study is dependent on members' perspectives.

2.2.2.4 Axiology of interpretivism

Since the interpretivist believes that the nature of reality is socially constructed, the interpretivist believes that there is value to be obtained in social inquiry. Interpretivists believe that the researchers' values will have an influence on the study as it pertains to human interaction. According to Saunders et al. (2007), the interpretivist is reflexive to the universe in which the study is being conducted.

2.2.2.5 Limitations and weaknesses of interpretivism for this study

The central focus of interpretivism is grasping the subjective meanings participants attach to a given phenomenon within a specific, unique context (Myers and Klein, 2011). From the submissions by Myers and Klein (2011), we can see that interpretivism is difficult to replicate, maybe unreliable and this makes it difficult to generalize.

2.2.3 Pragmatism

Pragmatism research is intended to make a difference in the practices of organisations. Pragmatic research is inductive as it works to provide or improve a solution from a complex problem and a theory. The pragmatist paradigm is concerned with a rational explanation of why a particular phenomenon is occurring and recommending a resolution. As a result, the nature of pragmatism is often referred to as regulatory because of this. The pragmatist researcher views reality as a practical result of ideas. The outcome is to develop a set or recommendation. Because pragmatists focus on reality, pragmatism considers all theories

and ideas as instruments of thought and in turn action. Pragmatists recognize the connection of experience, knowing and acting (Kelly and Cordeiro, 2020).

2.2.3.1 Ontology of pragmatism

The pragmatist views the world as possessing many different realities. The pragmatist researcher argues that a singular world view cannot portray the entire world. The nature of reality for the pragmatist is considered to be in constant flux with various processes, experiences and practices interacting (Saunders). The focus of pragmatism is a solution that works rather than what is objectively real (Frey, 2018).

2.2.3.2 Epistemology of pragmatism

The acceptable knowledge for a pragmatist researcher is that which provides actionable results (Kelly and Cordeiro, 2020). The pragmatist believes that knowledge exists independent from human activity. The epistemology of pragmatism considers acceptable knowledge to be that which enables actions (Saunders et al., 2007).

2.2.3.3 Methodology of pragmatism

Research problems and research questions are central to the pragmatist. This implies that pragmatist researchers are capable and flexible to manoeuvre in complex and dynamic situations. The methodology that is assumed by the pragmatist emphasizes practical solutions and outcomes (Kelly and Cordeiro, 2020).

2.2.3.4 Axiology of pragmatism

Pragmatist researchers aim to provide value through the research process conducted while remaining amenable to their personal doubts and beliefs. The core element of pragmatist is that they are open to questioning the foundations of philosophies (Elder-Vass, 2022).

2.2.3.5 Limitations and weaknesses of pragmatism

While the pragmatist approach offers, some advocates of pragmatism believe strongly in the scientific methods of research which are independent of individual opinions. Additionally, Frey 2018 highlights the limitation of pragmatism as that of being solution driven rather than objective. For these reasons the pragmatic approach was discounted from this study. The subsequent section describes critical realism with full justification why it has been adopted.

2.2.4 Critical realism

According to Patomaki, Heikki, and Colin (2000), post-positivism is pluralist in its function as it balances both positivist and interpretivist approaches. Post-positivism has many strengths,

one being that it is a flexible research perspective. This supports the researcher to utilise a number of methods to carry out the research in accordance with the research questions. It is necessary to underscore that in post positivism, arguments are the basic units of analysis, as alluded by Panhwar, Ansari and Shah (2017). Researchers like Fluck (2017), stress that positivists are also realists. A well-known form of post-positivism is a philosophy called critical realism.

Orlikowski and Baroudi (1991) asserts that the aim of critical studies is to critique the status quo by exposing structural contradictions resulting in transformation of the social conditions. Critical social theory therefore aims to identify an oppressing structure and promote liberation. Critical researchers expect that social reality is comprised of historical underpinnings and that it is created by individuals (Hoddy, 2018). Although individuals can wilfully act to transform their social and financial circumstances, critical researchers recognise that their ability to do so is constrained by different types of social and political regulation. The principle goal of critical research is seen as that of social investigation whereby the prohibitive and estranging states of the existing conditions are exposed emancipation is attempted (Myers and Klein, 2011).

Critical realism has been posited as an alternative to the traditional positivist and interpretivist paradigms (Wynn Jr and Williams, 2012a, p. 787:787). Studies conducted in the critical paradigm focus on assessment, depiction, and clarification. Information systems are considered the result of human activity, appearing, and being continued through human activity, and being established through its use. Information systems constrains and empowers social practices (Carlsson, 2003)). Critical social scientists accept the truths of both positivism and interpretivism. We can add that social scientists accept the need for both causal theories based on objective observation and interpretive descriptions based on inter-subjective understanding.

Critical realism can be viewed as an explicit form of realism and its goal is to recognize the reality of the innate structure and the happenings and dialogue of the social world. Bhaskar (2009) asserts that the social world can be comprehended by distinguishing the systems at work producing such happenings and dialogue. These structures are unexpectedly unclear in the detectable pattern of events. The structures can be recognized only through the applied and hypothetical work of sociology. Therefore, the world is comprised of events and experiences with underlying the structures and mechanisms (Dobson et al., 2007).

Bhaskar (2009) distinguishes three domains within the critical realism paradigm: the real, the actual, and the empirical as illustrated in Figure 2 below. The real domain comprises of basic structures and mechanisms and relations; events and conduct; and encounters that exist autonomously but are capable of creating patterns of events. Thus, behaviours are a result of relations that exist in the social world. The actual domain is comprised of these behaviours and events. The empirical domain comprises of what we experience, subsequently, it is the area of experienced events. The structures in the real domain are autonomous of the actual events which in turn are autonomous of experiences.

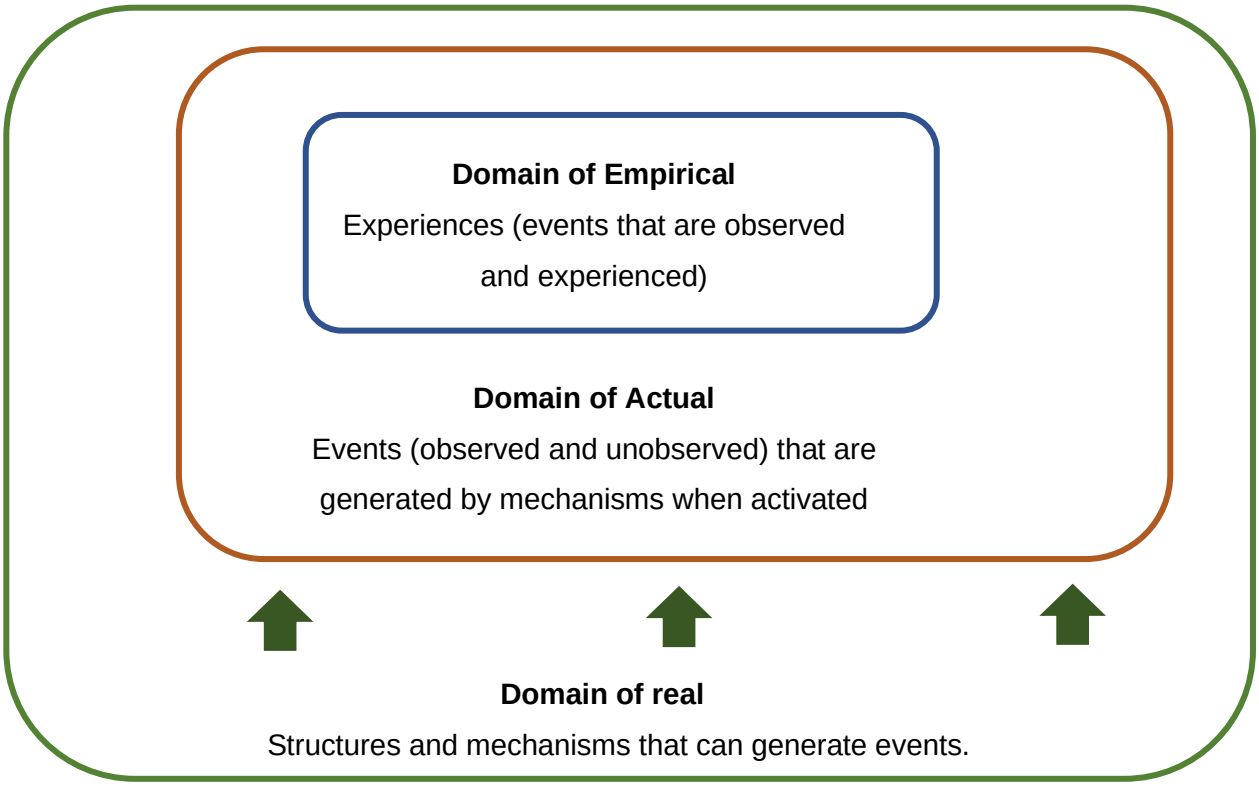


Figure 2 : Domains of Reality in Critical Realism

The subsequent sections describe the ontological dimension of critical realism.

2.2.3.1 Ontology of critical realism

Critical realism is based on a "stratified ontology which is comprised of structures, mechanisms, events, and experiences; emergent powers dependent upon but not reducible to lower-level powers; and an open systems perspective" (Wynn Jr and Williams, 2012a, p. 789:789). The nature of reality for critical realism comprises of a number of structures that

create the occasions. Critical realism preserves an emphatically pragmatist philosophy that there is a current world free of our insight that exerts causal influence. Critical realism defends against both established positivism that would lessen the world to that which can be exactly observed and quantified, and interpretivism that would diminish the world to our human explanations of it (Mingers et al., 2013). Ontologically, critical realism relies on participants' observation and experience, as well as other types of data. Additionally, Hoddy (2018) state that the nature of reality for critical research is historically constituted, determined by contextual conditions, and has no confinements to the current state and believes in the totality of relationships in the context rather than isolation. Thus, organisational issues are inseparable from the systems, people, and technology.

Critical realism recognizes that the world and elements that establish reality exist autonomous of human learning or our capacity to see them. Critical realism perceives that the world is not reducible to our recognitions and encounters. The autonomous nature of critical realism does not rely upon any immediate knowledge or emotional convictions with respect to the existence of entities (Wynn Jr and Williams, 2012b).

The reality of critical realism is stratified into three domains. The domain of the *real* incorporates the substances and structures of the real world and the random forces natural to them as they autonomously exist. The *actual* domain is a subset of the real and it incorporates the occasions that happen when the causal forces of structures and elements are ordered, despite being seen by people (Wikgren, 2005). The last domain, the *empirical,* is a subset of the actual domain and comprises of those occasions which we encounter through recognition or estimation. The three spaces are organized with the purpose that events in the space of the domain of the real which initiate a system, will not be seen as encounters in the area of the observational. Moreover, there are components – initiated and not yet initiated - existing in the domain of the real that are balanced by different systems. These components do not cause events in the space of the real (Wynn Jr and Williams, 2012b).

The critical research paradigm assumes that entities are autonomous from, and final to, the parts in which they are included. The attributes, capacities, and forces that can be credited to a given substance or structure rely upon those totalled from the parts. They also rely on the synergistic impacts starting to happen due to the example of their association. In this manner, the attributes of a specific structure emanate from the interrelationships among the parts and their random forces (Easton, 2010). This is very important in social structures.

2.2.3.2 Epistemology of critical realism

The epistemological assumption of critical realism is mediated knowledge (Wynn Jr and Williams, 2012a) and explanation of a phenomena. Critical realism uses various mechanisms to arrive at the knowledge. Pather and Remenyi (2005, p. 79:79) explain that knowledge in critical research should result in the emancipation and transformation of society. Orlikowski and Baroudi (1991:20) posit that in critical research, valid knowledge about a phenomenon can be constructed and evaluated from social and historical practices. This research discovers the truth using various mechanisms which include interviews with participants to understand the historical practices and propose data analysis using KDT as an emancipatory measure. The realist approach is non-positivistic which means that values and facts are intertwined and hard to disentangle (Mingers, 2013). Critical realism perceives that our approach to this world is constrained and constantly intervened by our perceptual and hypothetical focal points. The philosophy acknowledges epistemic relativity, yet not judgmental relativity.

Avison et al. (2018) confirm that critical realism aims to place depictions of reality dependent on an investigation of the encounters observed and translated by the members, alongside different kinds of information. The subsequent information claims indicate and depict the elements of reality without which the events and experiences under examination cannot have happened. The nature and type of these information claims are obtained from explicit epistemological assumptions connected to the ontological premises of critical realism. The epistemological assumptions comprise mediated knowledge, explanation rather than prediction or understanding, explanation by mechanisms, unobservability of mechanisms, and multiple possible mechanisms (Wynn Jr and Williams, 2012a, Sayer, 1999).

Knowledge is described by the critical realism paradigm as having both transitive and intransitive estimations (Mingers, 2013). The intransitive paradigm consists of the fragments of the world we endeavour to illuminate, commonly independent of our resources and experiences. The transitive paradigm constitutes experts' observations, similar to speculations about the free world that developed as the result of intelligent demand. The intransitive components generally, do not change in the typical world, yet learning objects of the transitive estimation will change. Critical realism acknowledges that our understanding of the intransitive elements containing autonomous truth is shaped in the transitive measurement, interceded by the social structures we are part of.

The ability to observe a mechanism makes a researcher more confident that the mechanism exists. However, the existence of the mechanism is not dependent on its observability (Sayer, 1999:12) and in turn knowledge about a mechanism is not dependent on observability. Where

mechanisms are not observable knowledge can be obtained from the researcher's capability to infer their existence based on experiences that would have potentially caused them. In critical realist studies, there may be various mechanisms that produce the outcome (Avison, 2018). This is because critical realist studies are positioned in an open system which can be subject to internal and external influences. As a result, there could be various explanations to an event that require evaluation and comparative analysis of the explanation which is most probable. The methodology of action research is summarized in the following section.

2.2.3.3 Methodology of critical realism using action research

Action research is a philosophy and methodology of research generally applied in the social sciences (Johnson, 2008). This study adopted the action research philosophy and methodology. Although action research used to be viewed as a post-positivist approach, Carr and Kemmis (2003:209) ascertain that action research is a suitable method for critical research based on Habermas' emphasis on enlightenment and emancipation in critical social sciences. According to Avison (1999), action research is an iterative process asking of researchers and practitioners to work together to diagnose a problem, plan an intervention, intervene, and reflect on the actions. As such, the philosophy and methodology combine the elements of theory and practice. Action research is set apart from another research approaches due to the fact that researchers are an integral part of the phenomenal under study. The researcher's input often influences the outcomes of the phenomenon. Thus, the researcher's role could change from researcher to subject (Chen and Hirschheim, 2004). Figure 3 below shows the action study adopted.

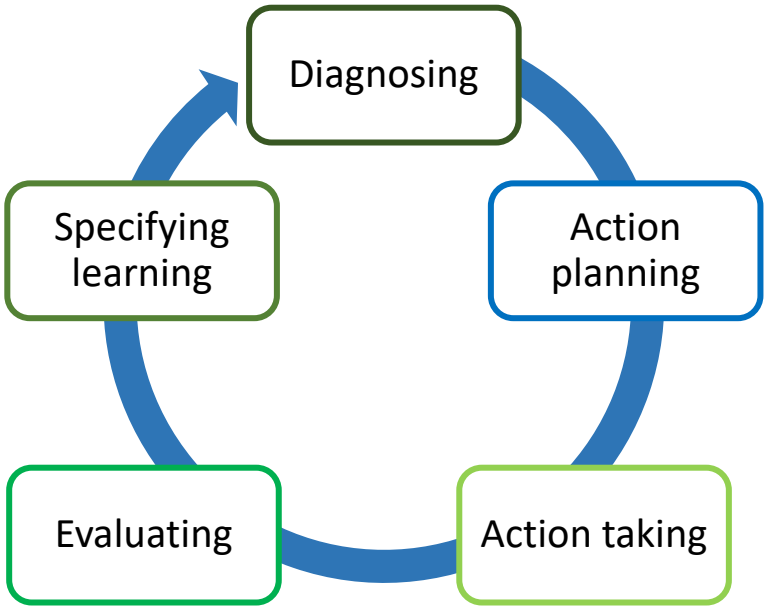


Figure 3: Action study framework: Adapted from Park (2017: 27)

This research involved interviewing the project managers, establishing how to use text data mining on the available data, implementing the action plan by analysing the data using the chosen text data mining technique, evaluating whether the technique extracts the anticipated knowledge efficiently in a user-friendly format and finally specifying the learning from the study. A critical approach infers an obligation to contribute to the improvement of the world (Wilson and Greenhill, 2004). Critical realism pursues transformation of the world through actions that prevent disintegration of impartiality and ultimately improves the status quo.

Several issues were taken into consideration before adopting the action research philosophy and methodology, including rigour and evaluation. To ensure rigour and validity of action research, Heikkinen et al. (2012:8) proposed a set of principles. The same principles by Heikkinen et al. (2012) were applied in this study as follows:

a) The principle of historical continuity analyses: the history of action, i.e., how the action progressed through history; and employment, i.e., the logic and coherence of the narrative. The researcher strived to portray the narrative of the study logically and coherently while documenting it in the thesis.
b) The principle of reflexivity determines subjective adequacy, i.e., what is the essence of the researcher's connection with the research object; ontological and epistemological presumptions, i.e., what are the researcher's beliefs about knowledge and reality; and transparency, i.e., how does the researcher describe material and methods used. The researcher ensured that she remained aware of the presumed knowledge and reality; and provided a transparent account of research study.
c) The principle of dialectics evaluates dialogue, i.e., has the researcher's comprehension develop in conversation with others; polyphony, i.e., the method the researcher used to report and present different voices and interpretations; and authenticity, i.e., how true, and real are the protagonists of the narrative. The researcher provided a platform to hear different voices and perspectives and provided an authentic report.
d) The principle of workability and ethics evaluates pragmatic quality, i.e., how well does the research reach the goal of constructing workable practices; criticalness, i.e., what is the character of discussions encouraged by the researcher; ethics, i.e., what methods are used to deal with ethical problems; and empowerment, i.e., does the research succeed to make people believe in their own capabilities to act, thus encouraging new conventions and ways of doing. The researcher ensured that the evaluators of the outcome of research pay attention to an outcome of change in social action.

e) Lastly, the principle of evocativeness determines how well the research narrative bring forth mental images, memories or emotions related to the theme.

The methodologies adopted for critical realism offer researchers an opportunity to conduct a holistic study of organizations (Delanty, 2011). The critical realist paradigm blends different ontological and epistemological characteristics and therefore mandates different approaches in terms of methodologies (Mingers et al., 2013). According to Easton (2010), critical realism assumes a transcendental realist ontology, an eclectic realist/interpretivist epistemology and a generally emancipatory axiology. The goal of critical realists is to critique the power and promote transformation in the social world (Delanty, 2011:75, Wilson and Greenhill, 2004). The critical realist paradigm is therefore capable of addressing the need for improved processes of developing theories in information systems studies. This enables specialists to create and bolster comprehensive causal explanations for the sociotechnical phenomena that is related to data innovation, social, organization, and natural elements. This is necessary to understand the role of information technology in the organizational context (Wynn Jr and Williams, 2012a).

A major goal of research studies directed under critical realism is to create clarifications for the way things operate and how they can do so (Vincent and O'Mahoney, 2018). Various strategies for critical realist-based research have been identified and this study employed the emancipatory nature of critical theory studies in order to leverage the capability and liberate human injustice. In critical realism, realism and naturalism are combined; and the concept of explanatory critique is conjoined with a radical emancipatory concentration (Bhaskar, 2009).

Basically, people's activities are compelled by society's social structures, systems, and mechanisms (Mingers, 2013). 'It is only if social phenomena are genuinely emergent that realist explanations in the human sciences are justified; and it is only if these conditions are satisfied that there is any possibility of human self-emancipation worthy of the name' (Bhaskar, 2009:103-104). Critical realism can be applied in the study of any situation if the process involves thoughtful, all-embracing research with the aim to comprehend why things are the way they are. Critical realism also has an emancipatory dimension and the researcher is capable of critiquing his/her role and influence in social activity as a result of the explanatory critique (Wikgren, 2005).

2.2.3.4 Axiology of Critical Realism

Similar to interpretivist axiology, the critical realist is value laden (Saunders et al., 2007). However, the researcher endeavours to remain objective as is the case with positivist

researcher. The critical realist researcher aims to remain objective by evaluating their own value position in relation to the study.

2.2.3.5 Suitability of action research for this study

Although there are various debates about the different paradigms, scholars generally accept that there is no one superior paradigm. Critical realism was found to be suitable for this study because it combines the concepts from positivism and interpretivism in answering research questions (Mingers, 2013). This therefore provides a dependable and understandable philosophy. The purpose of this study was to investigate whether KDT can improve the analysis of project data by yielding useful insights for the project environment that can become part of the body of knowledge. The critical realist paradigm is open to experimental and observational analysis while at the same time it acknowledges the explanatory powers and assumes that the social world is influenced by the actors within it (Sayer, 1999:10-12) . Therefore, it addresses both natural and social science which are important dimensions of this study.

Critical realism embraces specific frames of mind towards the positivist and interpretivist approaches. The critical realist is at no time content with mere qualitative or quantitative descriptions but regards these as steps in the research process. Critical researchers need to uncover the state of affairs in order to comprehend and clarify why things are as they are. This leads to an understanding of the structures and components that shape detectable occasions. Further, critical realism perceives the existence of different objects of knowledge material, applied socially and mentally and each requiring distinctive research techniques to be understood (Mingers 2013). Critical researchers perceive the unavoidable uncertainty of perception, particularly in the social world, and requires the analyst to be mindful of the presumptions and confinement of their examination (Pather and Remenyi, 2004)

Sayer (1999:5) outlines the eight key assumptions of critical realism as follows:

a) “The world exists independently of our knowledge of it.
b) Our knowledge of the world is fallible and theory laden. Concepts of truth and falsity fail to provide a coherent view of the relationship between knowledge and its object.
c) Knowledge develops neither wholly continuously, as the steady accumulation of facts within a stable conceptual framework, nor discontinuously, through simultaneous and universal changes in concepts.
d) There is necessity in the world; objects—whether natural or social— necessarily have particular powers or ways of acting and particular susceptibilities.

e) The world is differentiated and stratified, consisting not only of events, but objects, including structures, which have powers and liabilities capable of generating events.
f) Social phenomena such as actions, texts and institutions are concept dependent. We not only have to explain their production and material effects but to understand, read or interpret what they mean.
g) Science or the production of any kind of knowledge is a social practice. For better or worse (not just worse) the conditions and social relations of the production of knowledge influence its content.
h) Social science must be critical of its object. In order to be able to explain and understand social phenomena we have to evaluate them critically".

The following section describes the data collection methods used in detail.

2.3 Data collection instruments

Research instruments include questionnaires, interviews and focus group discussions (Yin, 2016:40). The study used interviews to gather data form the research participants in South Africa. Interviews allow quality information to be gathered (Creswell, 2016) as questions can be clarified to participants in the study. The questions asked are all well-structured and grouped into two categories, namely open-ended and closed questions trying to answer the research questions defined in chapter one. Interviews were conducted with project managers at selected organisations, specifically including a petrochemical company in South Africa. These interviews were conducted to get more information on projects using the knowledge discovery techniques. The interviews gave the researcher a better understanding of the projects and project management in general. The researcher established rapport with participants to make them feel more comfortable and at ease. This rapport generates more insightful responses especially regarding sensitive issues surrounding the topic (Bhattacherjee, 2012).

2.3.1 Trustworthiness of data collection instruments

All interview questions were open-ended to gather adequate and relevant information from the identified project managers. The way the data is collected and evaluated significantly impacts on the trustworthiness of the findings (Leedy and Ormrod, 2016). To ensure that valid and sound responses were obtained, the research participants were given time to ask questions for further clarity on areas which were unclear to them. Participants were allowed to make sure the notes written down by the interviewer were exactly what have been said. This ensured conformability of the findings. The structured interview guide **(Appendix B)** was used in this

research to ensure the qualitative findings were reliable and trustworthy. The questioning style was consistent throughout all interviews. The following section describes how the data collected was analysed and presented to the audience.

2.4 Data analysis

The data analysis is a pivotal part of the study since it is crucial to drawing conclusions and making recommendations. The purpose of qualitative data analysis is to understand and interpret social interactions (Gay, Mills and Airasian, 2012). For this study, qualitative data analysis was done using descriptions of participants' views and opinions. A thematic approach was used during data interpretation. Thus, the research objectives acted as the themes according to which data was presented. Leedy and Ormrod (2016) asserts once a reader truly understands the problem and the method of investigation, the next question pursues the evidence. Authors like (Orlikowski, 1991; Reinecke and Abraham, 2013) consider empirical evidence as valid knowledge. The collected data is presented in a consistent and logical sequence within this thesis. Each sub-problem and its associated data are discussed, and this keeps the reader informed of the progress of the research.

2.5 Study population and sample

A single petrochemical company in South Africa was used for this research during the period 2017 - 2019. The organisation runs a number of projects regularly. Therefore, the organisation possesses vast amounts of historical project data that was analysed to identify useful insights such as trends. A sample is a subset of data values drawn from a population (Wegner, 2012). Purposive sampling, with ten participants were used to get valid results in this study. According to Leedy and Ormrod (2016), purposive sampling is determined by the researcher's judgement of the best source of information to achieve the objectives of the study. The researcher identified ten project managers with suitable characteristics for providing the required information on knowledge and project management in a project environment. The project managers were involved in the overall project management processes in the organisation, specifically the analysis of the project reports.

2.6 Ethical considerations

Research ethics set the standards of conduct for scientific researchers (Manti and Licari, 2018). Ethical codes ensure that the risk of harm is minimized, and the anonymity and confidentiality of the research participants are protected. Besides, ethics gives participants the right to withdraw from the study at any time. It also avoids deceptive practices. The researcher ensured that the participants gave informed consent. According to Creswell (2016), informed consent is a voluntary agreement to participate in research. To be valid, such consent must

be in a straightforward written language, effortlessly understood by all subjects. Furthermore, written consent must keep the possibility of intimidation or inappropriate influence a minimum. Also, participants must be given enough time to contemplate participation. For this study, the participants received a letter setting out the details and purpose of the study and requesting their participation for purposes of data collection. **Appendix C** contains a copy of the informed consent form for this study.

Anonymity means there is no way of identifying participants in the study (Manti and Licari, 2018). Anonymity and confidentiality were given due consideration in this study. The researcher made sure to follow the ethical principles in research. The personal information of the participants was kept confidential and will be kept on a password protected computer for a period of five years. The researcher gained permission for data collection from the organisation where the data was gathered. In addition, an ethical clearance certificate from the North-West University was obtained and is attached in **Appendix A** of this thesis.

The North-West University (NWU) has a clear code of ethics which all researchers must follow when conducting research study. This code of ethics is also available online. Some of the principles in this ethics code include, but is not limited, to honesty, integrity, respect for intellectual property and confidentiality. Respect for intellectual property ensures that the researcher gives proper acknowledgements for all contributions to the research. All information sources used have been cited and a complete list of references used is provided in the bibliography of this thesis.

Figure 4 on the following page provides a summary of the methodology employed.

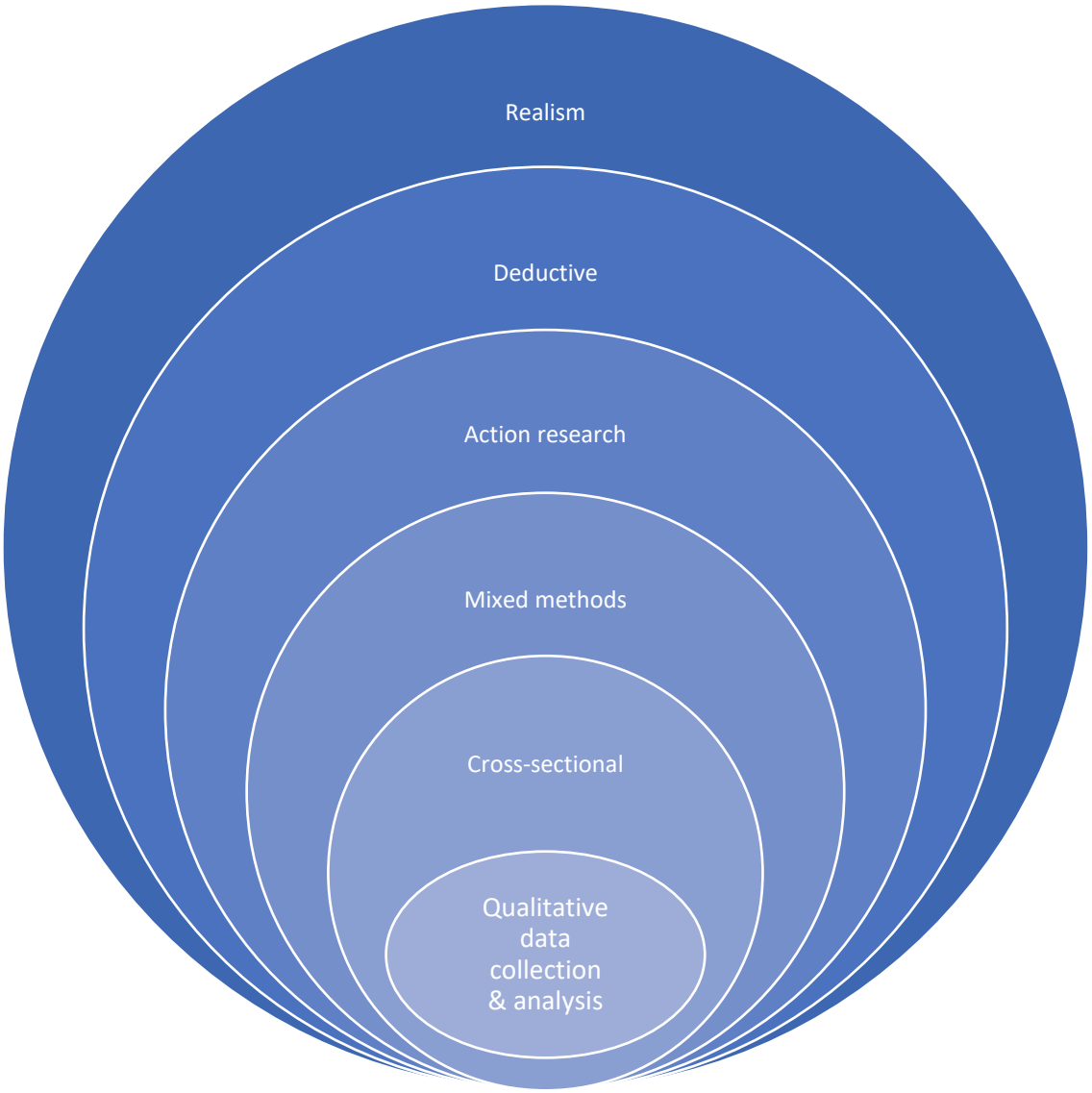


Figure 4 : Research methodology

2.7 Chapter summary

This chapter presented the methodology adopted to conduct this study in South Africa. The chapter discussed the data collection tools used to gather data together with full justification why they have been chosen. The chapter further explained the data analysis tools and concluded with a statement on ethics the researcher abides by throughout the study. It can therefore be stated that the chapter provides the basis for data collection and analysis.

Since the study investigates the use of text mining to improve knowledge discovery in the project environment, the next chapter therefore reviews literature on knowledge discovery and data mining in project environments.

CHAPTER 3: KNOWLEDGE DISCOVERY AND DATA MINING

3.1 Introduction

This chapter reviews literature on knowledge discovery and data mining. This literature review was conducted to describe and understand knowledge discovery and data mining in the project environment. Literature review helps researchers to discover research studies on similar problems (Webster and Watson, 2002; Chris, 2018). To have a greater understanding of the research topic, multiple sources of information were consulted, including accredited journal articles, textbooks and other online sources related to the topic. Webster and Watson (2002) as well as Babbie and Mouton (2010) state that it is important to review literature as it analyses the past in preparation for the future.

3.2 Knowledge Discovery

With the spread of digitization and globalisation, data is constantly being created and made accessible to individuals and organisations (Fayyad et al., 1996:37, Bramer, 2007, Omidipoor et al., 2020). Hand (2006:1) supports this assertion and attributes the increase in the amount of data to both digitization and the widespread internet. According to Hand (2006:2), the internet and the World Wide Web have made gathering of information less demanding, adding to the colossal amount of information accessible to organisations. Numerous organisations (Shaw et al., 2001:127, Silwattananusarn and Tuamsuk, 2012:13) appreciate the availability of knowledge in these tremendous databases. Organisations acknowledge that this information may be vital to supporting the organisational decisions. There is progressive acknowledgement that such information contains hidden knowledge that can impact an organisation's success (Bramer, 2007:3-4). This is due to the advances in innovation, generally minimising the effort to store such huge measures of information.

Bramer (2007) continues to explain that the vast amounts of information available today could empower humanity to foresee the climate and catastrophic events, distinguish the reasons for and conceivable remedies for deadly diseases, learning that could mean the contrast among life and demise. However, the current dilemma is that a large portion of the tremendous volumes of information is being stored, never to be analysed in the most basic way (Bramer, 2007:3-4). While the availability of data is continually expanding, our capacity to assimilate and process this data stays steady. Innovative technologies such as web crawlers, intensify the issue by making more and more records accessible in a matter of a couple of keystrokes. This leads to data overload, and additionally miss essential patterns and connections (Maimon and Rokach, 2005:809). It is apparent that there is a critical need to expand human capabilities of analysing the large amounts of information available in order to understand the underlying

patterns and relationships hidden in the information. The available literature for application of text mining during projects is minimal Camilleri (2010:338).

3.2.1 Knowledge Discovery in Databases

According to Fayyad et al. (1996:39) the phrase knowledge discovery in databases (KDD*)* was coined by Piatetsky-Shapiro (1989) to emphasize that knowledge is the end product of a data-driven discovery (Scheidler and Rabe, 2021). Further, Bramer (2007:2) identifies knowledge discovery as a "non-trivial extraction of implicit, previously unknown and potentially useful information from data." Maimon and Rokach (2005:1) define KDD as "the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets." Taking clue from Fayyad et al. (1996), Azevedo (2019) describes KDD as a process of using data mining methods to extract what is considered knowledge according to the specification of measures, using a database along with any required pre-processing and transformation of the database.

Azevedo (2019) underscores that KDD is different from machine learning and Artificial Intelligence (AI) in that KDD only uses the data mining methods in one of its steps. The data mining part of KDD currently depends on known strategies from AI, design acknowledgment, and insights to discover designs from data in the data mining step of the KDD procedure (Fayyad et al., 1996:39, Ur-Rahman, 2010:10). KDD centres around the general discovery of knowledge, including how the information are stored; how calculations can be scaled to enormous informational indexes and still run effectively; how results can be translated and imagined; and how the human-computer interaction can conveniently be demonstrated and upheld. The usefulness of the information is a critical product of the KDD process. As a result, KDD can be seen as a multidisciplinary action that envelops strategies past the extent of any one specific discipline. KDD creates knowledge from structured and unstructured sources. The key is that knowledge has to be in machine-readable and interpretable format. It must also represent knowledge in such a way that enables inferencing. The subsequent section reviews the knowledge discovery process in detail.

3.2.2 The KDD process

The KDD process is comprised of a set of steps to achieve the KDD goals. The Cross Industry Standard Process for Data Mining (CRISP DM) model is the de facto model for data mining (Schröer et al., 2021). The CRISP DM model identifies the main steps of data mining as:

a) business understanding,
b) data understanding,
c) data preparation,

d) modelling,
e) evaluation and
f) deployment.

According to Maimon and Rokach (2005:2), the knowledge discovery process comprises of nine steps which are interactive and iterative and many of these steps require the user's assessment and judgment. Silwattananusarn and Tuamsuk (2012:15) derives five sequential methods from the nine steps as follows: selection, pre-processing, transformation, data mining and interpretation/evaluation. Being iterative at each progression implies that moving back to modify past advances might be required. It is vital to stress that the process has no single set formula which means that it can be done from many angles. This point has been emphasized in literature to an extent that Fayyd et al. states that it is not an elementary calculation of predetermined quantities. It is therefore necessary to profoundly understand the process and the distinctive needs as well as potential outcomes in each stage.

The KDD process begins with deciding the KDD objectives, and ends with the execution of the discovered knowledge (Fayyad et al., 1996:42). Subsequently, the knowledge would need to be applied in the domain area. The implementation of changes closes the circle, and the impacts are then measured on the new information stores, and the KDD process is restarted. The following is a brief portrayal of the nine advance KDD process, beginning with an administrative step (Maimon and Rokach, 2005:3). The detailed review of the processes follows in Figure 5.

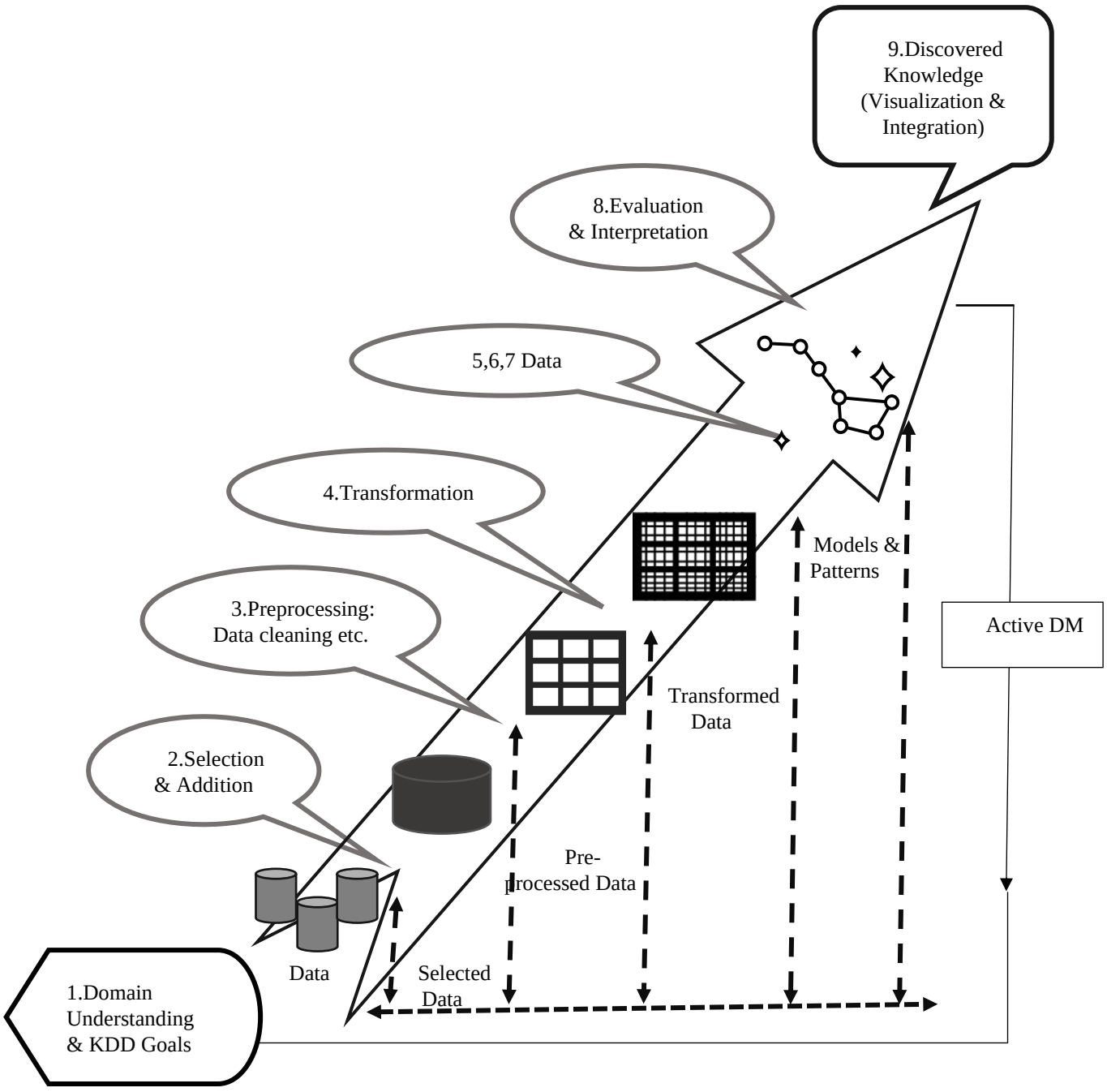


Figure 5: An Overview of the steps that compose the KDD Process (Maimon and Rokach, 2005)**.**

3.2.2.1 Step 1 - Domain understanding and Knowledge discovery goals

According to Fayyad et al. (1996:39) as well as Azevedo (2019), the preliminary step of the KDD process leads to a better understanding of the domain, understanding the required foregoing knowledge and establishing the goals of the KDD process. Maimon and Rokach (2005) describe this step as a preparatory step that sets the stage for understanding the knowledge objectives (what knowledge the user is looking for) and the prior knowledge that should exist. The step also involves understanding of the environment where the knowledge will be applied. As the KDD process continues, a revision and update of this progression might occur (Azevedo, 2019). Having understood the KDD objectives, the pre-processing of the information commences, and it involves the application domain, the relevant prior knowledge and finally the goals for the end-user.

3.2.2.2 Step 2 - Creating a target data set

The second step entails creating a target data set which will be utilized for knowledge discovery (Maimon and Rokach, 2005). A data set comprises a selection of related, discrete data items that may be accessed individually or in combination or managed as a whole entity. Azevedo (2019) adds that a data set can be organized into some form of data structure. The data set is created from the accessible data along with any additional data and attributes that may be relevant. This is a crucial step, because data mining is typically performed on the data that is made available. On top of this data models emanate from the data.

3.2.2.3 Step 3 - Data cleaning and pre-processing.

Data cleaning and pre-processing is the third step identified by (Maimon and Rokach, 2005:2-5). This is an essential step where the reliability of the data is improved. Basic activities such as removing noise from the data, gathering the fundamental data to model, settling on techniques for dealing with missing information fields and representing time sequence information take place at this stage. According to Bhattacharjee and Petzold (2016:1), noise may include grammatical errors, misuse of punctuations, spelling errors and so forth. Therefore, the goal of this step is to eliminate noise, errors, and also handling missing data.

3.2.2.4 Step 4 - Data reduction and projection or data transformation

The fourth step is concerned with determining the characteristics that are essential for data representation based on the goal of the data mining task. In other words, “better data” is generated (Maimon and Rokach, 2005:4). Various methods are used in this step to achieve data quality that will be suitable for a particular project. Maimon and Rokach (2005:4) identifies data reduction and transformation methods to include feature selection, extraction, discretization of numerical attributes and functional transformation. The authors further

highlight that this step is critical for the success of a KDD project as it influences succeeding steps of the KDD process.

3.2.2.5 Step 5 - Choosing the data mining task

After data reduction or transformation, the next step is to choose the type of data mining method to use. There are a variety of data mining methods such as classification, regression, clustering, summarization, dependency modelling, change and deviation detection (Fayyad et al., 1996:45). Each of these methods fall under one of the two major goals of data mining which are prediction and description (Maimon and Rokach, 2005:4). The choice of the data mining task is therefore dependent on the project's goal or data mining as well as the data reduction or transformation step.

3.2.2.6 Step 6 - Choosing the data mining algorithm(s)

As indicated in the previous step, there are various data mining methods available. The sixth step of the KDD process involves selecting a data mining algorithm to use as well as appropriate models and parameters. According to Microsoft (2020), an algorithm is a specific procedure for solving a well-defined computational problem. It is important to mention here that an algorithm should be complete, infinite, and unambiguous.

3.2.2.7 Step 7 - Data mining

This is the actual step of the KDD process in which data mining is performed (Azevedo, 2019). The data mining step involves the application of the selected data mining algorithm defined above to search for interesting patterns and this data mining algorithm may require fine tuning to achieve the goals of the project.

3.3.2.8 Step 8 - Interpreting mined patterns

The eighth step involves interpreting the results that have been mined by the data mining process. Azevedo (2019) adds to the initial findings by Maimon and Rokach (2005) that the eighth step centres around the functionality of the initiated model. In this progression the information found is likewise reported for further utilization. The last advance is the utilization and feedback on the revealed results acquired and as stated by Maimon and Rokach (2005), there can be iterations to previous step in a bid to gain the required insights.

3.3.2.9 Step 9 - Consolidating discovered knowledge

The final step is the utilization of the discovered knowledge. Such knowledge may be applied directly to a particular situation. Alternatively, the knowledge may be documented in a knowledge base for future applications. Azevedo (2019) states that any existing related

knowledge will be checked against the discovered knowledge to resolve conflicts, if any. It is imperative to highlight that this step is also iterative. The knowledge becomes active. This means we can make changes to the system and measure the effects. The subsequent section of this review looks at the actual data mining and the associated data mining models.

3.4.3 Data mining

According to Fayyad et al. (1996:41), “data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.” It is essentially the application of specific algorithms for extracting patterns from data (Bramer, 2007:2). Shaw et al. (2001:128) define data mining as a process of seeking and studying information to uncover conceivably valuable data. However, the definition of data mining, as per Larose and Larose (2014:2), is the discovery of useful patterns and trends in large datasets. Aggarwal (2015) defines data mining as a process by which usable data is extracted from a larger set of raw data. This involves analysing data patterns in large clusters of data using one or more software programmes.

Data mining includes choosing, investigating, displaying large amounts of data to reveal obscure patterns, and intelligible data, from expansive databases (Shaw et al., 2001:128). The authors go on to explain that data mining utilizes an expansive group of computational strategies that incorporate statistical analysis, decision trees, neural networks, rule induction and refinement, and graphic visualization. Despite the fact that data mining instruments have existed for quite a while, the advances in computerization have made data mining progressively appealing. (Shaw et al., 2001:128).

Data comes from numerous sources and in numerous forms. It is coordinated and set in a basic data store. A portion of the data is then taken and pre-prepared into a standard format. This processed data is passed to a data mining algorithm which delivers output as tenets or patterns which are translated resulting in new and conceivably valuable learning (Bramer, 2007:2). Hence, data mining is fundamental to the KDD process, and includes the deduction of algorithms that explore the data, model development and the discovery of previously unknown patterns (Maimon and Rokach, 2005:2). By application of the model, phenomena discovered through data analysis can be understood. Data mining and knowledge discovery is a significant matter due to the availability and plenitude of information (Atluri et al, 2018). The next section reviews the data mining methods.

3.4.4 Data mining methods

The main objective of data mining is prediction and description (Themistocleous and Morabito, 2017:86). Prediction includes utilizing some fields in a database to anticipate obscure or future values of different areas of interest, whereas description centres around discovering human-interpretable examples portraying the information. According to Maimon and Rokach (2005:5), descriptive goals are focused on understanding the data and centres around comprehending the manner in which the fundamental data identifies with its parts. Maimon and Rokach (2005:5) go on to stipulate that prediction goals intend to naturally figure a behavioural model, which gets new and concealed illustrations and can foresee estimations of at least one variable identified with the illustration. Prediction additionally creates designs, which structure the knowledge discovered into a way which is reasonable and simple to work upon (Themistocleous and Morabito, 2017:86).

The boundaries between prediction and description in data mining are not clear However, the distinction is valuable for comprehending the general discovery objective (Themistocleous and Morabito, 2017:86). The relative significance of prediction and description for certain data mining applications can shift significantly. Chen et al. (2021) study focuses on the following data mining methods “multivariate adaptive regression splines (MARS), k-nearest neighbours (KNN), extreme learning machine (ELM), eXtreme gradient boosting (XGBoost), and stochastic gradient boosting (SGB)”, which also have the same objectives of prediction and description. The objectives of prediction and description can be accomplished utilizing an assortment of specific data mining methods (Fayyad et al., 1996:42). Authors like Fayyad et al. (1996) assert that there are six main methods of data mining as defined as discussed individually below. Other authors have also recognized the same methods (Silwattananusarn and Tuamsuk, 2012:16) . This study explores the data mining methods below.

3.4.4 .1 Classification

Classification involves the partitioning and arrangement of items with the aim that each item is allocated to one of a number of comprehensive and limited groupings known as classes (Bramer, 2007:23). Fayyad et al. (1996:44) define classification as “learning a function that maps (classifies) a data item into one of several predefined classes”. Classification therefore finds models that analyse data into several predefined classes (Gan, Lin, Chao, and Zhan, 2017).

3.4.4.2 Regression

Regression is a familiar procedure in statistics for determining weighting of data from the attribute values (Witten et al., 2016:17). The basic regression method is incapable of locating

nonlinear relationships but different representations can be used for predicting numeric quantities (Witten et al., 2016:17). In the data mining context, regression involves grouping data into one of a few predefined classes. Thus regression inputs space into a real-valued domain as alluded by (Maimon and Rokach, 2005:6). There are numerous instances of regression applications in projects. A typical example given by Maimon and Rokach (2005) is that “a regressor can predict the demand for a certain product given its characteristics”.

3.4.4.3 Clustering

Clustering is a way to recognize a limited arrangement of categories to depict data (Fayyad et al., 1996:47). The classes can be commonly restrictive and comprehensive or comprise of a richer representative, for example, hierarchical or overlapping classes. Clustering algorithms look at information to discover categories of data that are comparative (Bramer, 2007:8). Clustering therefore involves gathering objects that are alike and different to the other items that have been placed in categories.

3.4.4.4 Dependency modelling

Dependency modelling comprises of finding a model that depicts dependencies between variables (Gan, Lin, Chao, and Zhan, 2017). Dependency models exist at two levels: structural and quantitative level. The structural level of the dependency model of the display indicates which factors are locally reliant on one another and the quantitative level of dependency model indicates the qualities of the conditions utilizing some numeric scale.

3.4.4.5 Deviation and detection

Change and deviation detection centres around finding the most critical changes in the data from pre-existing values (Gan et. al, 2017). These types of rules represent an association of values between attributes and the others. Such rules are known as association rules. The procedure by which such rules are removed from a given dataset is called association rule mining (ARM) (Bramer, 2007:188). Association rules are useful for analysing and predicting behaviour.

3.4.4.6 Summarization

Summarization includes strategies for finding a minimized depiction for a subset of data (Al-Hashemi, 2010:164). For example, organizing the mean and standard deviations for all fields in the project data. Progressively modern techniques include the inference of summary rules, multivariate representation systems, and the discovery of relationships between factors. Summarization techniques are frequently connected to intelligent exploratory information investigation and automated reporting. In a nutshell we can state the summarization is finding

a compact description for a subset of data. Since the study investigates the use of text mining to improve knowledge discovery in projects, the subsequent chapter reviews literature on text mining in the context of knowledge discovery.

3.5 Chapter summary

The literature review defined knowledge discovery which is the primary focus of this study. Other sections reviewed include project reviews and critical success factors for projects. Knowledge discovery as well as its associated processes were discussed to lay a solid foundation to this study.

CHAPTER 4: TEXT MINING

4.1 Introduction

This chapter reviews literature on text mining. The objective of the study is the discovery of knowledge from unstructured data using text mining tools. It is therefore imperative to review the text mining process and the tools associated with it. The chapter reviews different text mining techniques and models.

4.2 Text mining

Maimon and Rokach (2005:810) as well as Sigle and Robinson (2017), assert that text mining can be defined as the spontaneous discovery of new information that happens through the analysis of various textual resources. The textual sources are used to extract facts and events that will be used for text mining (Alfattni et al., 2021). Thereafter, traditional data mining methods are used to further explore the data and analytical purposes (Maimon and Rokach, 2005:810). While data mining has the capability of extracting insight most of the data that resides in structured databases, most of the information that exists in organisations is in unstructured textual format (Dörre et al., 1999:398, Feldman and Dagan, 1995:112, Ur-Rahman, 2010:1).

Typical data mining tasks use data that is presented in tabular format with particular attributes which are selected before the data collection ((Bramer, 2007:241). In contrast, text mining uses data sets comprising of the actual documents. Xu et al. (2014) points out that the attributes of these documents are extracted automatically. In addition to the use of data mining methods, text mining (TM) uses a combination of natural language techniques as well as information retrieval techniques in order to analyse the textual data for hidden insights (Dörre et al., 1999:398, Hotho et al., 2005:22). Text mining is also referred to as Knowledge Discovery from Text (KDT). The term was coined by Feldman and Dagan (1995:112).

Text mining has as its goal to find or get new knowledge from data, discovering patterns across datasets, or potentially identifying insights from noise (Sigle and Robinson, 2017). Situations where a user obtains information from an information retrieval system are not considered knowledge discovery as the user already had the knowledge about the information (Maimon and Rokach, 2005:811). Text mining is increasingly becoming an important practical problem as the volume of text-based data in many fields keeps increasing (Bramer, 2007:240). The ability to identify key issues that exist within textual data and the classification could aid in the functions of decision makers or knowledge workers for the improvement of their future activities (Ur-Rahman, 2010:1).

Text mining aligns with the CRISP model (Schröer et al., 2021) and applies a procedure similar to the KDD process. Text mining usually requires an adaptation of the known data mining algorithms to extract the data. Furthermore, text mining uses a combination of natural language techniques as well as information retrieval techniques (Uramoto et al., 2004:520). These areas rely on the application of data mining methods and statistics to handle their specific tasks. Figure 6 that follows shows the phases of the CRISP model in text mining.

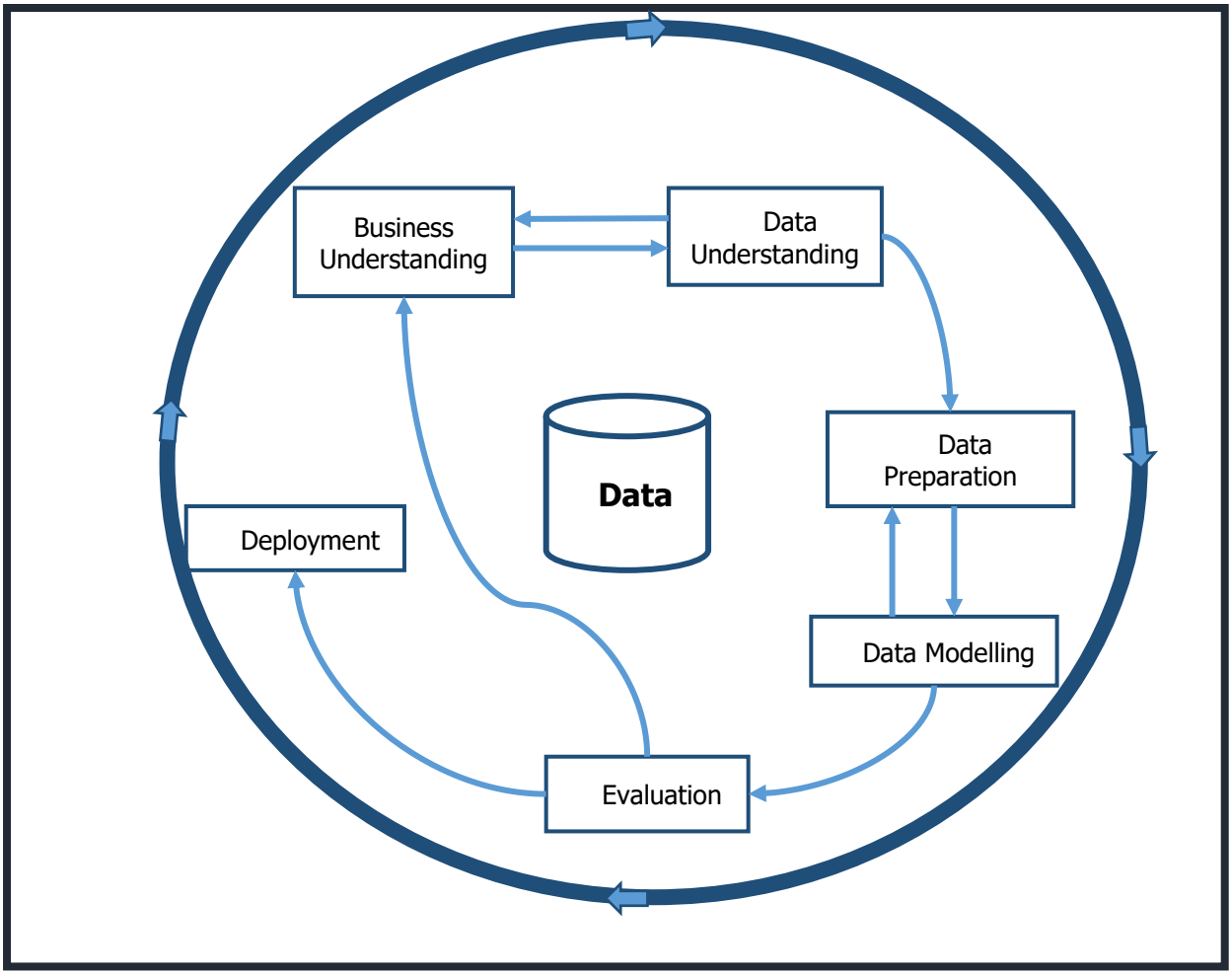


Figure 6: Phases of CRISP DM. Adapted from (Hotho et al., 2005:21).

The text mining process involves multiple phases as illustrated in Figure 8. in the following section extracted from (Vijayarani and Janani, 2016). The phases of text mining are explained in the next section.

4.2.1 Preparatory processing

For textual data to be mined, the text needs to be converted to instances with a fixed number of attributes (Sigle and Robinson, 2017). Preparatory processing also known as pre-processing converts the raw textual data into a structure suitable for further linguistic processing (Bramer, 2007:240). The processing step is an important step text mining technique (Merten et al., 2014:10). Sigle and Robinson (2017) underline that in text mining, a document is usually regarded to be a collection of words occurring in the document at least once. Such documents are commonly referred to as a 'bag-of-words' (BoW). The meaning, order of words, punctuation and sentence structures are ignored and the focus is on the measure of importance of each word. Bramer (2007:241) considers the count of how many times a word occurs as an example of measure of importance of the word.

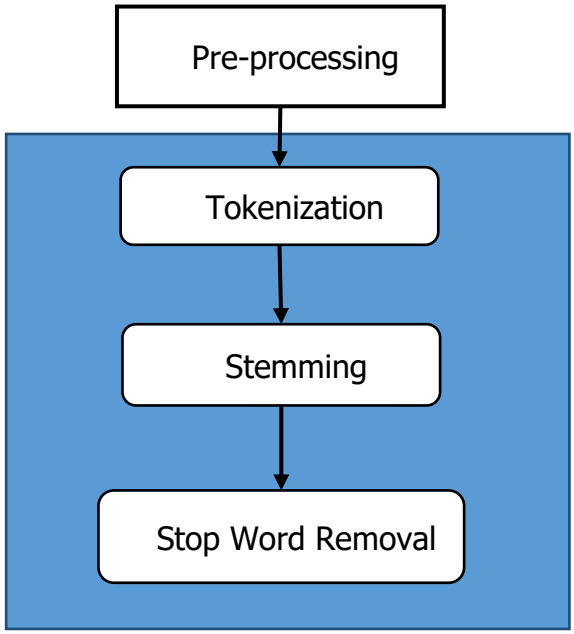


Figure 7: Pre-processing operations (Vijayarani and Janani, 2016:38)

Figure 7 shows above the pre-processing operations on textual data. The first step in preparatory processing is known as tokenization. In this step, a paragraph of text is broken down into smaller parts (Perkins, 2010:8). This task of breaking a sequence of characters into pieces is known as tokenization (Allahyari et al., 2017:3). The smaller parts of the paragraph which are words or phrases are used for further processing are called tokens. In addition to breaking down the paragraph into tokens, tokenization also eliminates certain characters such as punctuation marks from the original text (Allahyari et al., 2017:3). The second step in preparatory processing is called stemming. Stemming is a technique that is used to decrease

the number of words in a body of text by treating words with the same root structure as equivalent (Bramer, 2007:242), where the root structure of the word is the affix of the word.

According to Allahyari et al. (2017:4), stemming is aimed at obtaining the stem or root of a word in order to group the words that have the same linguistic stem together. Stemming is performed on words that have the same semantic meaning and morphological form (Vijayarani et al., 2015:10). An alternative to stemming is lemmatization. However, rather than using a root stem, an actual root word is used (Perkins, 2010:28). The last step of preparatory processing is the removal of stop words. In this step, words that are considered unusable for classification are removed (Salloum, et al, 2017). The words that are removed are words that do not add value to such as articles, determiners, and quantifiers. Figure 8 below shows the text mining steps according to Vijayarani and Janani. This will be discussed in relation to KDT in Chapter 5.

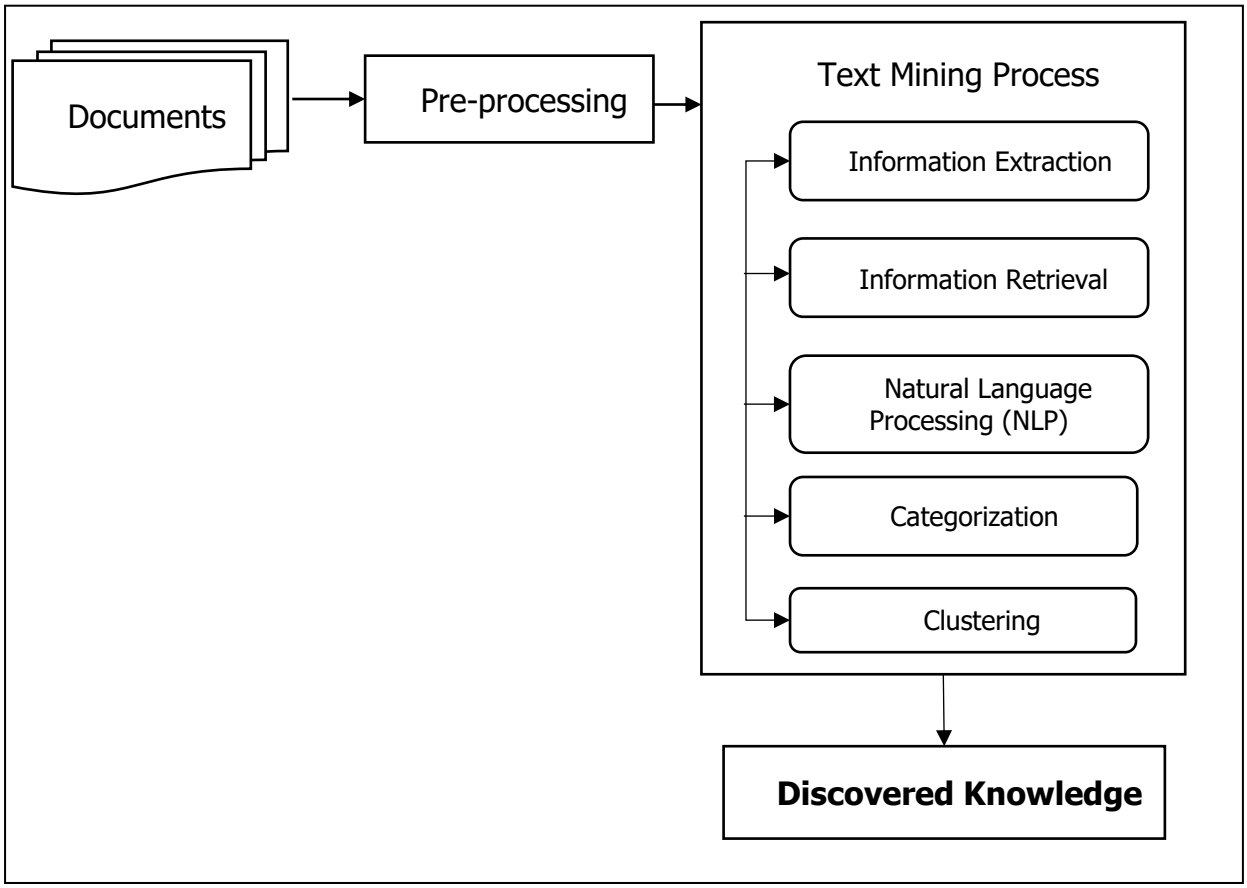


Figure 8: Text mining steps (Vijayarani and Janani, 2016:37)

4.2.2 Information extraction

Traditional data mining is performed on structured data which resides in a relational database. Today vast amounts of data are generated mainly in the form of natural language documents

rather than structured databases (Vijayarani et al., 2015:8). It is vital to sift through this data and identify the useful information and knowledge. The task of extracting useful information in unstructured and semi-structured forms of data is referred to as information extraction (Hashimi et al., 2015:3). Vijayarani et al. (2015:8) describe the method of information extraction as being performed by identifying key words and relationships within the text and looking for predefined sequences in the text. This process is called pattern matching. According to Allahyari et al. (2017:2), information extraction is usually the starting point for other text mining algorithms.

4.2.3 Information Retrieval

Krallinger et al. (2017:7680) considers information retrieval as a key concept in text mining. First introduced by Mooers in 1950, it was only in the 1970s that the concept of full text analysis and document indexing became common. The task of information retrieval involves the association of and finding information resources from a large number of unstructured data (Vijayarani et al., 2015:8). The extracted features from text documents are used to create a vocabulary of unique words that appear in the documents. This vocabulary is known as a Bag of Words (BoW). In text mining, information retrieval facilitates accesses the required information from a large text corpus (Allahyari et al., 2017:2). A corpus (plural corpora) is a large and structured set of texts.

4.2.4 Natural Language Processing

Natural language processing (NLP) is a significant step in text analytics since textual data uses natural language. Natural language processing is useful for defining ways to use computers to understand and manipulate the natural language text (Nayak et al., 2016:16877). Vijayarani et al. (2015:8)claim NLP collects knowledge on human beings' understanding and utilization of language. This knowledge is used to develop tools and techniques to make computer systems understand and manipulate natural languages for the accomplishment of specific tasks. There are various NLP techniques, such as part of speech tagging (POS), syntactic parsing and other types of linguistic analysis.

4.2.5 Categorization

Another important task in text mining is categorization which is about recognizing the key aspects of a document by comparing the document to a pre-determined set of topics (Nayak et al., 2016:16876). The document is treated as a "bag of words" rather than attempting to process the information that is in the document. The main topics can be identified by counting all the words that appear in that document. The predefined topics and relationships are identified by looking for large terms, narrower terms, synonyms, and related terms.

4.2.6 Clustering

Text mining is mainly derived from data mining (Sigle and Robinson, 2017). As such text mining also utilizes clustering to augment retrieval of information and support effective searching. In text mining, clustering involves the automatic organisation of documents, information filtering, rapid retrieval, as well as topic extraction (Hashimi et al., 2015:3). Similar documents are grouped together into clusters from a collection of documents. The clusters can be documents, paragraphs, sentences, or terms.

4.2.7 Summarization

One of the end goals of text mining is summarization of textual data. Text summarization is the technique used to produce a concise and precise summary of voluminous texts. During the process the focus is on the sections that convey useful information and without losing the overall meaning (Roul et al., 2019:64). According to Al-Hashemi (2010:164) automatic summarization involves reducing a text document or a larger corpus of multiple documents into a paragraph or short set of words that conveys the core meaning of the text. Summarization is achieved by determining the similarity of sentences on two levels: word space based level and semantic space based level (Yin and Pei, 2012:832).

Two methods can be employed for automatic text summarization, namely extractive and abstractive. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation. It uses natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original text. However, abstractive methods are considered to be weaker in comparison to the extractive counterparts (Al-Hashemi, 2010). The next section discusses the text mining models used in text mining which is the focus of this study.

4.3 Text Mining Models

According to Lee et al. (2010:2), text mining models have evolved from discriminative models to generative models. Ng and Jordan (2002) differentiated the two types of models based on the classifiers. “Generative classifiers learn a model of the joint probability, *p(x, y),* of the inputs *x* and the label *y,* and make their predictions by using Bayes rules to calculate *p(ylx),* and then picking the most likely label *y.* On the other hand, discriminative classifiers model the posterior *p(ylx)* directly or learn a direct map from inputs *x* to the class labels”. The difference between the models stems from methods of modelling documents and estimating variables.

Generative models differ from discriminative models because they start by establishing a model and approximating unknown variables utilizing a word document matrix (Lee et al., 2010). Different types of text mining models are therefore suitable for different scenarios. Various models were proposed to estimate continuous representations of words, including the well-known Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM). These models have been highly advocated for as means of estimating continuous representations of words (Mikolov et al., 2013).

4.3.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a discriminative model used in text mining. LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer et al., 1998:2). LSA analyses relationships between a corpus of documents and the terms within the documents and produces a set of concepts that pertains to the document and the terms. According to Lee et al. (2010:2), LSA creates a small factor space by reducing the original vector space or term-document matrix. A singular decomposition value is used to achieve the dimensional reduction of a matrix. This is accomplished by decomposing the original matrix into three matrices, namely a document eigenvector matrix, an eigenvalue matrix, and a term eigenvector. Multiplying these three matrices with a high eigen value can therefore yield an estimation of the original matrix.

4.3.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) is a generative model also used in text mining. According to Oneata (1999:1), pLSA is a technique that was initially developed by Thomas Hoffmann in 1999 for the purpose of indexing, retrieval and clustering in text-based applications. The goal of pLSA is to model co-occurrence information under a probabilistic framework to discover the underlying semantic structure of the data. Lee et al. (2010:3) assert that the probabilistic theory can be used to model the occurrence of words based on the assumption of “exchangeability”. Exchangeability is also known as the bag of words concept. It is a common concept in text mining, as the word sequence can be changed without an influence on analysis outcome. Lee et al. (2010:3) identified the three-step assumption of documents generation from a pLSA assumption. They assume that under pLSA, there is a probability of a document to be generated or selected; there is a probability of a topic to be selected and there is a probability for each word in a topic to be selected and from these probabilities. pLSA is commonly used for identifying themes and trends in documents.

4.3.3 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. Therefore, a custom corpus is really just a bunch of text files in a directory, often alongside many other directories of text files. (Perkins, 2010:46). The fundamental idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei et al., 2003). In LDA, document generation follows three steps as identified by Lee et al. (2010:3) Step one involves using Poisson distribution to determine the number of words used in a document. Step two uses Dirichlet distribution to prompt the distribution of document over topics. The final step involves topic generation and subsequently words for each topic are generated according to the document-specific distribution.

These steps by Blei (2013) are summarized as follows:

a) Choose N _ Poisson(x).
b) Choose q _ Dir (a).
c) For each of the N words w_n:
    (i) Choose a topic z_n _ Multinomial (q).
    (ii) Choose a word w_n from p (w_n j z_n;b), a multinomial probability conditioned on the topic z_n"

LDA is suitable for long documents with multiple topics. Additionally, LDA incorporates an adjective and noun structure which empowers researchers with the ability to mark the topics. The topics provided by LDA consist of words that have probability values (Lee et al., 2010).

4.3.4 Correlated topic model

According to Blei and Lafferty (2007:18), the correlated topic model (CTM) models the correlation between the latent topics in the collection. This enables the construction of topic graphs and document browsers allowing the user to navigate the collection in a topic-guided manner. Topic models can extract interpretable and useful structure without any clear "understanding" of the computer language. CTM uses logistic normal distribution and follows a generative process to address topic relations and considers the covariance structure (Lee et al., 2010:5). Table 1 on the following page shows the characteristics and limitations of four text mining models discussed here (Lee et al., 2010:5).

Table 1: The characteristics and limitations of four text mining models (Lee et al., 2010:5)

Models	Characteristics/Limitations
Latent Semantic Analysis	**Characteristics** • Reduces dimensionality of *tf-idf* using Singular Value Decomposition. • Captures synonyms of words. • Not robust statistical background.
	Limitations • Difficult to determine the number of topics. • Difficult to interpret loading values with probability meaning. • Difficult to label a topic in some cases using words in the topic.
Probabilistic Latent Semantic Analysis (PLSA)	**Characteristics** • Mixture components are multinomial random variables that can be viewed as representations of "topics." • Each word is generated from a single topic; different words in a document may be generated from different topics. • PLSA partially handles polysemy.
	Limitations • No probabilistic model at the level of documents
Latent Dirichlet Allocation (LDA)	**Characteristics** • Provides full generative model with multinomial distribution for words in topics and Dirichlet distribution over topics. • Handles long-length documents. • Shows adjectives and nouns in topics.
	Limitations • Incapable to model relations among topics
Correlated Topic Model (CTM)	**Characteristics** • Considers relations among topics using logistic normal distribution • Allows the occurrences of words in other topics. • Allows topic graphs.
	Limitations • Requires complex computations. • Contains too general words in topics

4.4 Text mining tools

There are several text mining tools that can be used to achieve different goals. Kaur and Chopra (2016:191) identified text mining tools and their uses as follows:

> *Text Analytics:* involves extracting useful information and patterns from text.
>
> *Text Processing:* involves transforming and manipulating unstructured text so that analysis methods can be applied to it.
>
> *Classification/Categorization:* Many tools are used for classification and categorization of text/documents.
>
> *Sentiment Analysis*: is used to identify subjective information from text. Many tools provide for sentiment analysis. Other sources refer to sentiment analysis as Opinion Mining.
>
> *Knowledge Discovery*: deals with identification of useful information from huge amount of text.
>
> *Semantic Analysis*: involves checking the syntactic structures with the meaning of the text as a whole. Many tools are available that not only provide syntactic analysis but also semantic analysis of the text.

In addition, (Maimon and Rokach, 2005:811) identify the following Text Mining (TM) tools:

> *Trend analysis*: The process finds trends or relations between people/places/organisations etc. by aggregating and comparing information extracted from the documents.
>
> *Information retrieval:* Retrieving documents is based on the various sorts of information about the document content.
>
> *Clustering:* The process of clustering organizes the documents based on their similarities and present the groups or clusters of the documents in certain graphical representations.

The following section reviews literature on the text mining applications which are very important to achieve the study objectives.

4.5 Text Mining Applications

In the technology dependent 21st century there are a number of text mining and text analysis applications in use. Kaur and Chopra (2016:186) classified these applications into three categories which are: *proprietary text mining tools, open-source text mining tools and online*

text mining tools. Proprietary text mining tools are commercial products owned by a company and therefore need to be purchased. One of the popular proprietary text mining tools is SAS. Open-source text mining tools are available to the public for free including the source code for the software to promote improvements in developing the software. R programming, Rapidminer Text mining and Python are among the popular open-source text mining tools. Finally, there are a few online text mining tools which provide limited functionality that can be executed from the web using only a web browser. An example of an online text mining tool is Textalyser (Kaur and Chopra, 2016:187).

4.6 Text Mining with Python

Python is an object oriented interpreted programming language consisting of a large number of libraries (Madnani, 2007:3). Python is a straightforward, yet powerful programming language with outstanding functionality for processing linguistic data (Bird et al., 2009). It can be added that Python is a dynamic language which facilitates rapid development facilitates and interactive exploration and has transparent syntax and semantics. This study employs the open-source text mining tool with Python. Python is one of the major tools that is currently being adopted for text mining (Vel, 2021).The rationale for choosing Python is because of its shallow learning curve and excellent online learning resources. Furthermore, Python is heavily used in industry, scientific research, and education around the world, hence the study adopted the tool. Two text mining applications with Python are discussed below.

4.6.1 Natural Language Tool Kit

Natural Language Tool Kit (NLTK) is the Python Natural Language Processing Toolkit which is a suite of libraries and programs used for developing Python programs that work with human language data and statistics (Vijayarani and Janani, 2016:41). NLTK is a comprehensive Python library for natural language processing and text analytics (Perkins, 2010:7). The tool offers the basic interfaces for NLP techniques such as part-of-speech tagging, syntactic parsing, and text classification among others. The tool also provides text processing libraries to perform tasks such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning (Perkins, 2010:10). NLTK also includes a WordNet corpus reader, which is a dictionary designed for programmatic access and exploration (Perkins, 2010:9). From the submissions above, we can see that the tool is very important for text processing which this study tries to achieve.

4.6.2 Word2vec

While traditional NLP use the simplistic bag of words (BoW) representations, more recent approaches use distributed word representations which is achieved by constructing word

embeddings (Weston, Bordes, Yakhnenko and Usunier, 2017). Word2Vec is a popular NLP technique used to learn embeddings from text corpora (Treviso et al, 2017) with disregard of the word order in a window. Treviso et al. (2017:4) further explains that Word2Vec is divided into two types of modelling namely:

a) Continuous Bag of Words (CBOW), which if given a window of words as input, the network tries to predict the word in the middle as output.
b) Skip-gram model, which tries to predict the window given the centre word as input.

The two processes work hand in hand to mine text and make sense out of the extracted text to improve decision making (Mikovol, 2013). The next section reviews literature related to the actual usage of text mining.

4.7 Usage of text mining

Text mining is gradually being adopted in many disciplines where there is need to collect and condense textual data for decision support purposes (Sigle and Robinson, 2017). Some of the advantages of text mining include the capacity to rapidly process substantial volumes of textual data that cannot be performed effectively by human readers. In addition to processing large amounts of textual information, Roul (2019) adds the objectivity facet of the process. This means that the results entirely depend on the end-product of the linguistic processing algorithms and statistical calculations provided by the text mining technology. It can be added that, with text mining, it is possible to automate labour-intensive routine tasks and entrust human readers with the more demanding tasks.

In short, text mining is used to

- Extract relevant information from a document e.g., summarization, feature extraction (Chen and Hirschheim, 2004, Wang and Lo, 2021).
- To better understand trends, associations between people, places, organisations, etc. by automatically aggregating and comparing information extracted from documents of a certain type e.g., incoming mail, customer letters, news-wires. (Bose, 2009).
- Classify and organize documents based on their content; i.e., automatically pre-selecting groups of documents with a specific topic and assign them to the relevant person (Phan et al., 2008).
- Organize repositories of document-related meta-information for search and retrieval (Patel and Soni, 2012).

4.8 Chapter summary

From the submissions by various authors, it can be deduced that text mining helps to retrieve documents based on a variety of information about document content. The following chapter discovers knowledge using text mining tools.

CHAPTER 5: KNOWLEDGE DISCOVERY WITH TEXT MINING

5.1 Introduction

The chapter further provides action plans and guidelines for the use of knowledge discovery from text for project managers. It is vital to state that the study's aim was to understand the use of knowledge discovery from text to assist project managers with data analysis. The action planning phase of the study is synthesised in this chapter, proposing an action plan to improve the area of concern. The chapter further demonstrates how the recommended text mining method fits into the organisation. As part of the study contribution, the chapter reflects on the necessary improvements for the organisation.

5.2 Action planning

An action research approach was used in this study as described in the methodology section. As defined in Chapter 2, action research is an iterative process during which researchers and the practitioners collaborate to diagnose a problem, plan an intervention, intervene and reflect on the actions (Avison et al., 1999).Although action research was largely viewed as a post-positivist approach, Carr and Kemmis (2003:209) ascertain that action research is a suitable method for critical research based on Habermas' emphasis on enlightenment and emancipation in critical social sciences. Action research is known for three main features: 1) its usefulness managers in context; 2) its ability to integrate academia and corporate industry; and 3), contribution to advancement of theory (Alfaro-Tanco et al., 2021). Consequently, there are five iterative phases of action research which are diagnosis, action planning, action taking, evaluating, and specification of learning (Baskerville and Wood-Harper, 1996: 238). In this section of the chapter the focus is on the action planning phase of the study.

Action research focuses on invoking change within the organisation (Park et al., 2017:28) . Therefore, the action researcher was involved in creating organisational change firstly by identifying the problem that necessitates the change. Once the diagnosis is complete the second phase is the action planning phase. The researcher identified certain actions to resolve identified problems during the action planning phase (Baskerville and Wood-Harper, 1996:238). The illustration below shows how the researcher was led by the conceptual framework of ideas (F) using key concepts from critical social theorists. Using this process, the researcher could discover and guide the planned actions of an action research intervention. The conceptual framework of ideas describes the desired state and provide guidelines on what needs to be done to achieve the desired future state. The conceptual framework of ideas aids in identifying methodology/(ies) (M) which is the change that is required to be applied to the (diagnosed) area of application (A) to achieve the desired state (Baskerville and Wood-Harper, 1996:240).

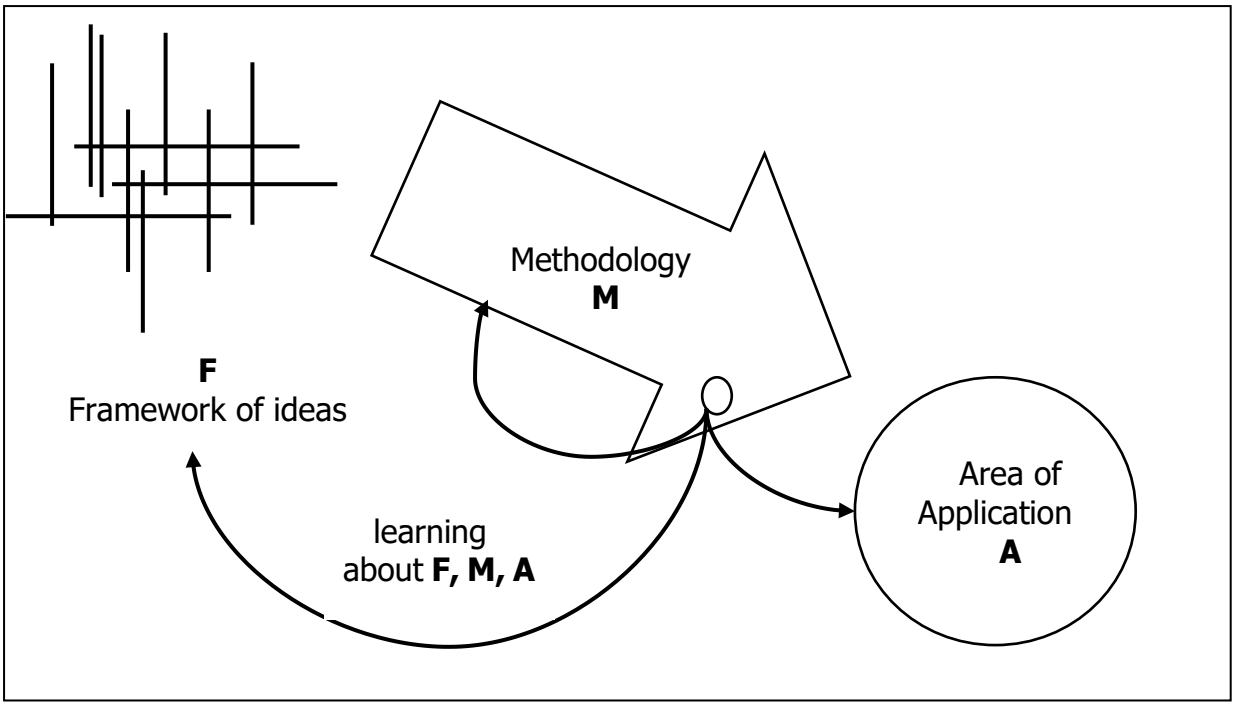


Figure 9 : A conceptual framework for Action Research adapted from (Baskerville and Wood-Harper, 1996)

As indicated in Figure 9 above, the action planning phase endeavours to develop an action plan for improving the method of knowledge retrieval and knowledge presentation from project data in order to emancipate the project managers. In the action planning phase, the researcher established how to use textual data mining on the available data.

5.3 Guidelines for the use of Text Mining in Project Management

Cope III et al. (2011:56) emphasized the need for knowledge capturing and discovery as the critical component of knowledge management in projects. Although project knowledge management theoretically exists, in practice the knowledge is simply collected but not being retrieved and therefore not being utilized effectively. As a result of the organisational memory loss (Polyaninova, 2011:1), the knowledge in the organisation dwindles and in turn affects the success of the organisation. Novins (2002) as cited by Cope III et al. (2011:60) stated that "the solution is not creating the world's greatest database repository of all wisdom with the world's fanciest search engine. Instead, we need to give people specific tools designed to help them do their job and solve specific business problems".

Traditional databases store data in formats that can easily be queried to get information because the data is organized in a structural manner that can be presented using tables.

However, the type of data collected in projects is usually narrative textual reports that are highly unstructured. Furthermore, the conventional method of turning data into knowledge depends on manual analysis and interpretation (Fayyad et al., 1996: 37-38). This research contributes to the knowledge workers' (the project managers) endeavour by providing an efficient methodology of gaining the information from project data that may be necessary to perform their jobs.

According to Loh et al. (2000:2) the main goals of KDT are meant to

a) Investigate concepts instead of words or attribute values, enabling the user to discover ideas, ideologies, trends, and intentions in texts.
b) Minimize the effort required to identify concepts in texts and discover knowledge.
c) Find interesting patterns in textual collections using straightforward statistical techniques.
d) Allow users to perform ad hoc discovery (with ill-defined goals) without spending time and effort to create formal models.

Textual data can be analysed for useful insights that have a high commercial potential using knowledge discovery from textual databases also known as text mining (Tan, 1999). Text mining has been defined as the means to discover information in large unstructured text bundles, and automatically identifying interesting patterns and relationships in textual data (Feldman and Sanger, 2007). Text mining utilizes methods from various disciplines such as machine learning, data mining, and natural language processing as well as knowledge management among others. Text mining is synonymous to knowledge discovery from textual databases which Tan (1999) defines as extraction of knowledge from unstructured text documents.

This thesis is related to work done by Carrillo et al. (2011) which is an application of data mining on post project reports. The gap that this study closes is to determine the applicability using specifically a text mining tool such as Python text mining. The theoretical and methodological contribution is therefore the applicability of performing KDT using Python on lesson learned reports from projects.

5.4 Application of Text Mining for project environments

The automatic analysis of different textual resources to discover new information, is known as text mining (TM) (Silge and Robinson, 2016). TM aims to discover or get new knowledge from data, unlocking patterns across datasets, or potentially identifying insights from noise. The ability to identify key issues that exist within textual data and the classification aids in the

functions of decision makers or knowledge workers for the improvement of future activities (Ur-Rahman, 2010:1). The TM process consists of many phases as per the illustration by Vijayarani and Janani (2016) in Chapter 4, Figure 8.

5.4.1 Pre-processing- Data cleaning

Because textual data is highly unstructured and may not be in the format of structured relational databases, pre-processing the data is an essential step of text mining (Vijayarani et al., 2015:9). During pre-processing each text file was read and segregated into its separate lines. Each of the separate lines was separated further into words. For this study, a set of lessons learned document files was provided by the project managers who partook in the study. The files were in different formats including PDFs, spreadsheets, and pure text files. Each of the files underwent the pre-processing step in order to prepare the files for analytics and machine learning. The researcher recommends that these files be in one format in order to make them relational during the analysis. The specific format used is the spreadsheet format .xls.

5.4.2 Information Extraction - Tokenization

Tokenization is a critical step of text mining. According to Vijayarani (2015), tokenization is a method of breaking up a piece of text into many pieces, such as sentences and words. In Python, the word *tokenize* function is used to separate words from punctuation and stop words (Perkins, 2014:10). Tokenization helps in identifying the data within the dataset that is deemed as not useful. In order to utilize the processing resources optimally, it was necessary to drop the unwanted data and only retain the data that met the study goals. After separating punctuation and stop words was completed, these unwanted data were removed from the dataset.

5.4.3 Information retrieval – Bag of Words

Information retrieval is a crucial task that feeds into the NLP step of the text mining process. After completion of the data cleaning, the most frequent data from the documents was retrieved using queries on the project data provided by the project managers forming a BoW. The BoW was used as a source for information retrieval to create vectors from the data. Python supports two ways of converting sentences to vectors. One method that can be used to convert words into vectors is by using Word2Vec whilst the other method involves the use of TF**-**IDF (term frequency- inverse document frequency). Below is comparison of Word2Vec and TF-IDF.

Table 2: Comparison of TF/TF-IDF and Word2Vec

TF/TF-IDF	Word2Vec
Creates one number per word	Creates one vector per word.
Good for classification documents as a whole	Good for identifying contextual content.

The TF**-**IDF (term frequency- inverse document frequency) was utilized in this study to convert sentences to vectors using the bag of words concept and was later (in the clustering step) used to determine the weighting of unique words in the dataset. The Term frequency was calculated using the bag of words. Usually, the counts from Term Frequency focus on longer sentences. Therefore, the term frequencies required undergoing through a process of normalization. TF-IDF was found to be suitable for this study because it also provides search query relevance.

5.4.4 Natural Language Processing- Stemming / lemmatization

As described earlier on, stemming is a process used to reduce inflected or derived words to their stem or root form (Ali and Ibrahim, 2012). The most popular stemmer in text mining applications and information retrieval is the Porter stemmer. However, stemming has been denoted of having the limitations of over-stemming and under-stemming (Karaa and Gribâa, 2013:245). Two words that have different meanings can have the same stem, for instance “university” and “universal” are both stemmed to “univers”. Over-stemming is when a word is reduced to a stem that can have a different meaning. Conversely, two words that are supposed to be stemmed to the same root may not necessarily be stemmed that way, for instance “ability” and “able” are stemmed differently while they have a similar meaning.

Lemmatization, on the other hand, involves identifying different parts of speech, removal of inflections, and reducing words to their simplest form. According to Karimov et al. (2017:2) lemmatization takes into account the identified part of speech when associating the root forms and the respective lemmas. To address the problems of over-stemming and under-stemming. Gupta et al. (2012) proposed a combination of stemming with partial lemmatization. In Python, NLTK comes with an implementation of the Porter stemming algorithm. The Porter stemmer uses its knowledge base of word forms and suffixes to stem the words in the data set to their root form (Perkins, 2014:30). When combined with lemmatizer, the resulting words which are derived are root words that preserve the original meaning of the words that have been stemmed.

5.4.5 Categorization- Abstractive summarization

Automatic text summarization converts lengthy documents into shorter meaningful versions. This process could be demanding and costly to undertake manually. This study employed the abstractive summarization technique. While Al-Hashemi (2010) considers abstractive summarization as a weaker technique, Yin and Pei (2012) argue that extractive summarization is strict and rigid as it depends on hard matching of terms. Furthermore, (Khatri et al., 2018), assert that extractive summarization is not scalable to larger datasets. Machine learning algorithms can be trained to understand documents and recognize the segments conveying important facts and information.

This study employed abstractive summarization using the Gensim Word2vec. Word2Vec is a word embedding algorithm used for this study in Python to train a set of fixed-length dense and continuous-valued vectors based on the corpus of text from the lessons learned reports from the projects (Brownlee, 2017). Brownlee (2017) defines word-embedding as a method to provide a dense vector representation of words that capture something about their meaning. Word embeddings to represent text in natural language processing. Gensim Word2vec which is used for this study is a uses a shallow neural network to learn the representations of words/phrases in a corpus.

5.4.6 Clustering- Generating the similarity matrix

As described in Chapter 4, clustering is instrumental to separate different groups based on the similar attributes or properties. According to Sieg (2018), text similarity is used to establish how ‘close’ two pieces of text are both in lexical similarity (surface closeness) and semantic similarity (meaning). Similarity matrices have been identified by researchers as effective techniques of retrieving the most relevant that meet a certain given criteria (Jain et al., 2017). Applying a similarity matrix has proved particularly useful for information retrieval given the vast amount of data that is continuously being created today(Mardiana et al., 2015:2).

There are a number of models that have been suggested for similarity measures to retrieve the desired information such as the ones of Jaccard, Dice, or Cosine, (Egghe, 2010). However, the similarity measure that provides a precise calculation is not yet available (Thada and Jaglan, 2013). This study considered three similarity coefficients which are Jaccard, Dice and Cosine coefficients for retrieving the most relevant information from the lessons learned documents that were provided by the Project Manager. The similarity coefficients are illustrated in Table 3 on the following page.

Table 3: Three similarity coefficient (Thada and Jaglan, 2013:203)

Similarity Coefficient (X, Y)	Actual Formula
Dice Coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Cosine Coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2}.\lvert Y \rvert^{1/2}}$
Jaccard Coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$

According to (Thada and Jaglan, 2013:202), a similarity coefficient determines the similarity between two objects and categorizes them according to the presumed relevance based on a given set of keywords. This study used the Cosine similarity coefficients which outperforms Jaccard and Dice Coefficient approaches (Hadi et al., 2007:5) and also provides the optimum fitness values(Thada and Jaglan, 2013:204).

Cosine Similarity was deemed as suitable and used for this study in Python. Cosine similarity calculates similarity by measuring the cosine of angle between two vectors (Wei, 2017). The focus when using the Cosine similarity matrix is the size of the angle rather than the Euclidean distance. Mathematically, the Euclidean distance is the straight-line distance or shortest distance between two points. When the angle between two vectors is smaller, then the cosine similarity is higher (Sieg, 2018). This means that regardless of the Euclidean distance is not of consequence when using Cosine similarity matrix because the orientation of the documents may still be closer together. Table 4 below shows how Cosine Similarity is calculated.

Table 4: Cosine Similarity calculation for two vectors (Gupta, 2018)

$$similarity = \cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

As described in section 5.4.3, Python supports two methods of converting sentences to vectors. The TF/IDF methods was employed. During this phase, Inverse Data Frequency (IDF)

was used to determine the weight of unique words from the project data at hand. The most relevant words and phrases identified during this stage where then used for the next step to extract meaningful insights.

5.4.7 Summarization- Identification of critical data from the summarized text

Summarizing techniques in Python are used to compile and understand the data from the text mining. Summarization further breaks the clustered data into subsets that can help in identifying patterns and providing insights (Sarkar, 2020). Based on certain selected keywords from the project data received from the project managers, insights were captured using regular expressions. The package used for this study was Regular (reg-ex) expression builder. Regular expressions (regex) are text patterns used to automate searching through text and eliminating the need for manual search by project managers.

For the purpose of this study, the author applied the high-level text mining functional architecture by (M K and K, 2016) which classifies the steps in the TM process under three categories: Pre-processing tasks, Processed Document Collection and Core Mining Operations was adopted. Figure illustrates the high-level text mining functional architecture.

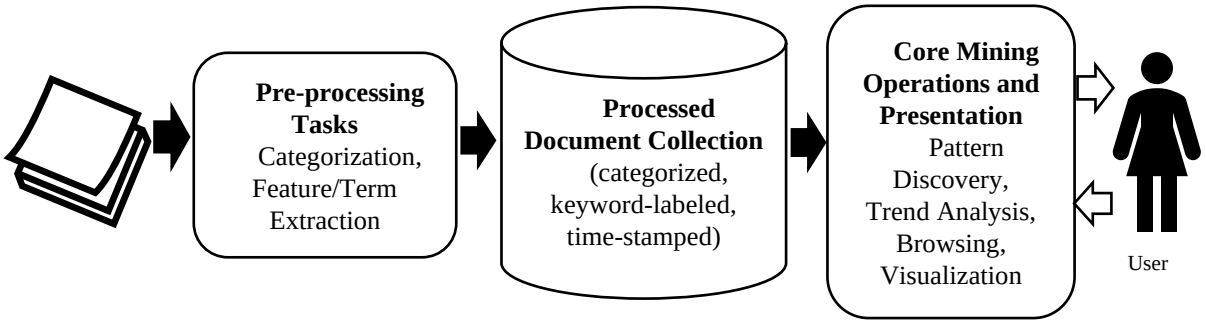


Figure 10 High-level text mining functional architecture (M K and K, 2016)

The author has used these categories to align the TM process with the practical application of TM that was employed in this study. Table 5 which follows shows the classification of the three categories.

Table 5 Classification of Text Mining processes (Author's work)

Text mining process	Categorization for this study
Pre-processing Tasks	• Pre-processing/Data Cleaning • Information Extraction • Information Retrieval
Processed Document Collection	• Natural Language Processing • Categorization
Core Mining Operations	• Clustering • Summarization

5.5 Chapter summary

This chapter discussed and proposed the actions to be taken to address the problem identified. Action research is aimed at emancipating users from the bondage of their current process by improving their processes. The purpose of the action planning phase is to ensure that the outcome of the action will solve the problems identified in the diagnosed area of application before commencing the actual development work. The text mining methodology that is proposed here has the required improvement that will be established by text mining for project managers.

CHAPTER 6: PROJECT KNOWLEDGE MANAGEMENT

6.1 Introduction

This chapter further reviews literature on Project Knowledge Management which is part of the study contribution. Project Knowledge Management (PKM) involves people, process, and tools. In this regard, people understand the gravity of knowledge and information to project success. On the same domain, the process requires that the organization has a structure for knowledge management in place and ingraining the structure in project management processes and methodology. A couple of the most effective tools to guide the management and sharing of knowledge in projects include communities of practice and repositories for storing and retrieving lessons learned. Thus, Knowledge Management and Project Management compliments each other in working together to improve performance in project environments.

6.2 Project Management

Project management is a knowledge intensive activity. It is used by to employ their abilities and experience to make informed decisions during the project (Barros et al., 2002)). Currently, project-based organisations (PBOs) is described as a growing trend in organisational terms (Pemsel et al., 2014). Projects constitute a large part of their activities (Institute, 2008). The generation of knowledge takes place during projects; however, project managers do not sufficiently use existing organisational knowledge or transfer project knowledge (Fernie et al., 2003, Newell et al., 2006, Goffin et al., 2010, Pemsel et al., 2014) . For many organisations in both developed and developing countries, transferral of knowledge and experience to new projects is challenging (Newell et al., 2006, Scarbrough et al., 2004). Due to ineffective knowledge transfer in PBOs, resource wastage and consequent increase in costs, takes place (Ren, Deng, and Liang, 2018). The complexity of project-based environments imposes difficulties in their operations, however, this complexity also provide opportunities for continuous improvement of the organisation's performance through knowledge (Kotnour, 2000).

Knowledge management is a challenge for every organisation (Prusak and Davenport, 1998, Davenport and Prusak, 2000). Managers in project-based organisations agree that notwithstanding their enterprise or industry, correct knowledge asset management leads to the successful completion of projects (Lierni and Ribière, 2008). In this regard, knowledge governance mechanisms are key to derive the maximum advantages from the knowledge created through project activities (Peltokorpi and Tsuyuki, 2006, Pemsel et al., 2014). The successful transfer of knowledge among members of the organisation enables the organisation to use knowledge effectively, fostering its development and prosperity (Argote,

2011). Most enterprises agree that knowledge acquisition during projects creates a wealth of information that positively influence project management (Abu Bakar et al., 2016). However, under normal circumstances team members are disbanded after completion of projects. Work on new projects start immediately after completion of the previous projects, and no effective knowledge transfer takes place. This leads to a loss of valuable project knowledge (Sokhanvar et al., 2014). The following section discusses the general nature of projects.

6.2.1 The nature of projects

Major challenges are created for project managers and project-based organisations owing to the intricate and unpredictable character of projects (PMI, 2018). Knowledge represents one of the core project capabilities that ensures that projects and project-based organisations can cope with these challenges. Knowledge transfer across projects moves knowledge from project sources of knowledge to project recipients with the goal of improving performance and capabilities (Landaeta, 2008). In PBOs, projects are the fundamental structure of those organisations. Project managers recreate the process at the start of a new project, instead of learning from previous projects and reusing valuable knowledge (Ren et al., 2018). This is related to the project's own nature. The relationship between project nature and knowledge transfer has been discussed by various authors in different contexts. Uniqueness (Ren et al., 2018), temporality (Ren et al., 2018, Zhao et al., 2015, Cheng, 2009), geographical distance (Ho and Liu, 2013, Ren et al., 2018, Wiewiora et al., 2009), urgency and similarity have been mainly researched.

To understand a project-based organisation, it is necessary to first recognize the meaning of a project. Kerzner (2017:2) define a project as a series of activities and tasks that:

- "have a specific objective to be completed within certain performance specifications (e.g., cost, quality, schedule),
- have limited resources (e.g., time, personnel),
- have defined start and end dates,
- have a project manager and a project team with the authority and responsibility over the accomplishment of the project objectives, and
- have knowledge needs".

In addition to the factors identified by Kerzner (2017 above, Yang (2015) extrapolates that uncertainty and ambiguity (i.e., lack of knowledge) is part of every project from beginning to end. This finding highlights the need for managing knowledge in project related work.

The nature of projects plays a fundamental role in knowledge discovery and exchange. Projects are normally topographically scattered, while individuals prefer to retrieve knowledge from, or diffuse knowledge to, the place where it can easily be reached (Ho and Liu, 2013, Ren et al., 2018). Consequently, the circulation of knowledge only takes place between projects within close proximity. Furthermore, because of projects' temporality, timely storage of knowledge after project completion proves to be a challenge. Findings by Newell et al. (2016) revealed that time constraints restrict information exchange among team members from different projects.

Geographical distance is another critical factor affecting knowledge transfer since team members are usually situated in various geographical areas. Geographical distance can be classified as shorter-distance and longer-distance. Knowledge is more likely to be shared over shorter-distances, as individuals are more inclined to access and share information on projects to places within the immediate area (Haldin-Herrgard, 2000). Similarly, due to geographical distances, the topographical separation and lack of formal links across projects, face-to-face communication and social networks are greatly reduced (Ren et al., 2018). From this finding, we can theorize that geographical distance has an influence on knowledge transfer in project-based organisations.

In both principle and practice, a project is a temporary exercise and has a beginning and an end (Turner and Müller, 2003). It is therefore imperative to state here that a project is short-term-oriented, such that the knowledge management cycle is rarely completed. This is unfavourable to knowledge transfer among project teams and specifically, project managers. Several studies about knowledge learning proved that the loss of knowledge is increased because of the temporality of projects (Zhao et al., 2015, Hanisch et al., 2009, Connelly et al., 2009). Hence, temporality leads to communication challenges between project teams in PBOs.

Pressure to finish on time and deliver projects, is a challenge project teams must deal with regularly. This is known as the urgency feature of projects (Haldin-Herrgard, 2000). At this juncture, contention exists about the effect of urgency on knowledge transfer. A number of researchers have demonstrated that project teams have insufficient time for communication and knowledge sharing under confined due dates and tight schedules (Loo, 2002, Zhao et al., 2015). Interestingly, Newell et al. (2006) found that time urgency propels a project team to seek knowledge from other project teams when they face difficulties. Contrary to the results of studies above, Connelly et al. (2009) conducted two experimental studies proving time actually does not impact knowledge sharing.

The paradox in project-based organisations is that project teams simultaneously face serious knowledge needs and yet have the opportunity to use the existing knowledge of projects to increase performance and capabilities (Kotnour, 2000). The necessary project knowledge can potentially be found, fully or partially, in concurrent or closed projects within the same project-based organisation (Dixon, 2000).

Kotnour (2000) proposed a general project learning framework to aid project-based organisations in benefiting from their knowledge. Kotnour (2000) also clarified how a project, and subsequently a project-based organisation, can expand its capabilities (e.g., knowledge) and performance (e.g., cost, schedule, quality) from both the understanding developed within the project and the understanding developed by other projects. To Kotnour (2010), knowledge frameworks in projects help project managers, leaders, and senior managers of project-based organisations to articulate how a project would benefit from their knowledge, as well as from the knowledge of other projects.

The project lifecycle is comprised of five process groups. Each process that takes place during a project's lifecycle falls into one of these process groups. Abyad (2018) has described the process groups as follows:

> *"**Initiation** - Setting up the project for success by identifying the right team and scope, as well as determining the relationship between the project and its alignment with the organization's overall charter.*
>
> ***Planning** – Developing the relevant resources, timelines and milestones, and mapping project delivery to business priorities (i.e., risk management, communications, quality, cost/budgeting, duration and sequencing, external dependencies).*
>
> ***Execution** – Assigning the project team and distributing information to ensure the proper activities are undertaken. This process also includes ensuring quality assurance methods are in place to address change management, organizational updates, possible changes to the plan, etc. the main elements are:*
>
> ***Controlling and Monitoring** – Ensuring the resulting product maps back to the original plan, and risk from uncontrolled external actions is mitigated.*
>
> ***Closing** – Making sure you have delivered everything expected of the project. Once you close, you need to review the project vis-à-vis the plan and likewise ensure contract closure"*

The subsequent section reviews literature on the project management knowledge areas which these processes fall under.

6.2.2 Project Management Knowledge Areas (PMKA)

According to Liu and Javed (2017), a knowledge area is a specific area of project management defined by its knowledge requirements and described in terms of its component processes, practices, input, outputs, tools, and techniques. The Project Management Institute (USA) distributes the project management knowledge into ten interlinked areas (PMI, 2018). Each knowledge area belongs to one of the process groups identified in the previous section. The completeness or incompleteness of the number of these project management knowledge areas (PMKAs) can be debatable. Nevertheless, the importance of these knowledge areas for project success is established across the board. The PMI ten project management knowledge areas are listed below.

i. Project Integration Management (PIM)
   -Deals with the end-to-end execution and delivery of project work.
ii. Project Scope Management (PSM)
   -Deals with ensuring that the project remains within the defined scope.
iii. Project Time Management (PTM)
   -Deals with defining schedule and ensuring that the project remains on schedule.
iv. Project Cost Management (PCoM)
   -Deals with defining and managing costs to ensure that projects remain on track with the defined budget.
v. Project Quality Management (PQM)
   -Deals with quality assurance for the project.
vi. Project Human Resource Management (PHRM)
   -Deals with ensuring that resources are available, adequate, and allocated to relevant tasks accordingly.
vii. Project Communications Management (PCmM)
   -Deals with ensuring effective information dissemination for the project
viii. Project Risk Management (PRM)
   -Deals with identifying, monitoring, and controlling risks that may affect the project and minimizing the impact of such risk.
ix. Project Procurement Management (PPM)
   -Deals with ensuring that all purchasing activities are seamless.
x. Project Stakeholder Management (PSHM)
   -Deals with ensuring that stakeholders are engaged for all necessary elements of the project

Since all projects require skills and knowledge (Hayajneh and Hamada, 2020), it is clear that the knowledge areas lack the knowledge management facet. The PMBOK has classified

knowledge management under the Project Integration Management area. However, knowledge management during projects are critical to warrant a unique knowledge area. This study therefore proposes a new knowledge area for project management called Project Knowledge Management (refer to *Figure 11: Project knowledge Management- Author's own work, page 68*). The motivation for introducing this new knowledge area was born from the potential benefits for project management practise. One such benefit can be derived from data driven decision making given the proliferation of data in this era. The introduction of a new knowledge area may have a positive impact on project success as knowledge from previous projects can be useful in new projects as depicted in Figure 11 on the following page.

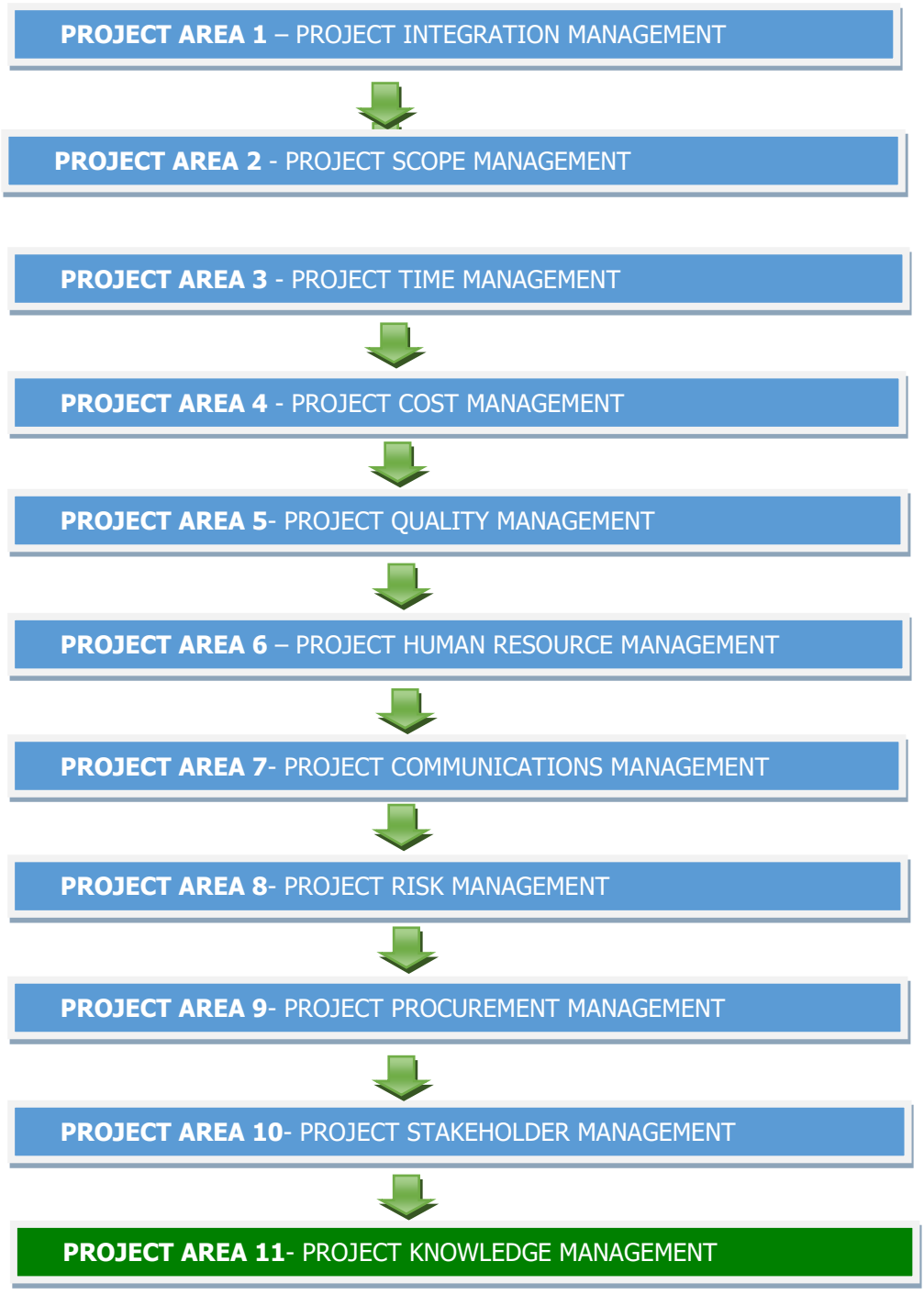


Figure 11: Project Knowledge Management (Author's own work) has eleven knowledge area vs the existing ten knowledge areas.

6.2.3 Project Performance Factors

An organised project management work environment makes it simpler to deploy promising project management practices that lead to predictable project outcomes, enhance

management performance, and enable knowledge management (Ozdemir Gungor and Gozlu, 2016). Structured project management advances the training of project personnel, establishes project management policies and procedures, priorities for projects, and consistent project management processes. Nevertheless, as projects are unique, the adjustable nature of such projects is attractive. Subsequently, organisations are faced with the responsibility to establish a balance between flexibility and stringent structure (Voss, 2012). When effort is made to enhance project team members' performance, flexibility is accomplished, and project team performance is achieved because of people-related factors. Project teams, in turn, are in charge of project performance.

Accordingly, interpersonal skills and communication are crucial for a project manager's success in managing projects. However, Kerzner (2011, 2017) suggests that critical success factors (CSF) of projects include corporate understanding of project management, executive commitment, organisational adaptability, project manager selection criteria, project leadership style, and ultimately commitment to planning and control. As evident in literature, effective communication also plays a critical role in project team development, conflict management, negotiations, decision-making, and project performance (Anantatmula, 2016).

Early investigations by Pinto and Slevin (1987) identified clearly defined goals, top management support, project plan and execution processes, endeavours to recognize expectations of clients and stakeholders, project monitoring and feedback, adequate communication with key stakeholders, and capacity to deal with unexpected problems as several factors in project achievement. Other studies have considered clearly defined project mission, detailed plans, communication, and top management support as predictors of project success (Hartman and Ashrafi, 2012). Likewise, Fedor et al. (2013) considers top management support as an essential contributor to project success. A later report identified project size, project type, plans and procedures, and project organisation as basic achievement factors (Park, 2019). Another recent study by Ozdemir et al. (2016) proves that strategic top management support enables more effective operational help, enhancing project performance.

6.2.4 Project Competence

Competence may be defined as a set of qualities an individual possesses, making him an effective or superior performer when given a task (PMI, 2018). Project management competence, in this manner, is the powerful or unrivalled execution of management tasks by a project manager or adding value to the successful achievement of project goals. There is an impressive assertion in literature (Alam, Gale, Brown, and Kidd, 2007) that competence is a

multifaceted construct. When acquired knowledge is put into practice, skill development takes place with time. This is called experience, particularly if deployed in a conducive working environment. Such progression should boost competence and is feasible only if there is continuous professional development. Consequently, knowledge acquisition and sustenance can be viewed as cyclic process.

Knowledge acquisition is, therefore, an essential building block for nurturing a competent project manager. There are solid indications to propose that quality, depth, and breadth of primary knowledge acquisition impact a person's ability to acquire appropriate professional skills and experience. For that reason, insufficient primary knowledge in most situations negatively affects the quality of skills and experience acquired. It has been proposed, for instance, by Baldwin and Ford (2008) that acquisition of knowledge is a combination of an appropriate body of knowledge (BoK), pedagogy, a conducive and well-endowed training environment as well as a correct disposition of attitude by the trainee. As the beginning stage of developing PM competence, the concern is about BoK, tools, and techniques (PMI, 2018) that an individual acquires. The knowledge can be acquired through training on the job, short courses, and professional courses. Hands-on training is a decent method of conducting apprenticeship and obtaining experience.

6.2.5 Understanding project successes and failures

Enhancing project performance and guaranteeing project success are challenges that project-based organisations face daily. Deficiencies to put in use the fundamental knowledge, skills, tools, and techniques to project activities creates the danger of project failure (PMI, 2018). In 1994, the Standish Group started to share statistics from global research projects. A study among directors found that only 16% of projects were viewed as successful. According to the Standish Group International (2001, 2013), the projects were divided into three classifications, namely project success (timely completion of projects and in budget, with all features and functions as specified), project challenged (projects finished, but over budget and timeframe, as well as coming short of the highlights and functions originally specified), and project impaired/failed (projects deserted or dropped at some point, thus becoming total losses).

Literature on project management describe project success and failure. However, the core of what project success really means and the methods by which success is evaluated, is debatable. debated what represents project success and the methods by which project success is evaluated. Hughes, Tippett, and Thomas (2014) and PMI (2018) differentiate between project success measures and project success factors. Project success measures are the standards by which achievement or disappointment is judged, whereas project

achievement factors are seen as inputs leading to project success. Generally, a fruitful project is characterised by the ability to yield the ideal results within a timeframe that all agreed on and using agreed-upon timeframe by using the selected resources (Hughes et al., 2014; PMI, 2018). Researchers like Aladwani (2012), Parsons (2006), and Rosenfeld (2013) portray success or failure regarding the established objective outcome measures, such as project cost (below, on, or over budget), project duration (early, on time, or late), and outcome quality (with less or superior to the required features and capacities).

Langston (2013) affirms that the connection linking the three main constraints that boost successful project delivery was originally defined by Martin Barnes (2014) as the iron triangle: time, cost, and output as shown in Figure 12 below.

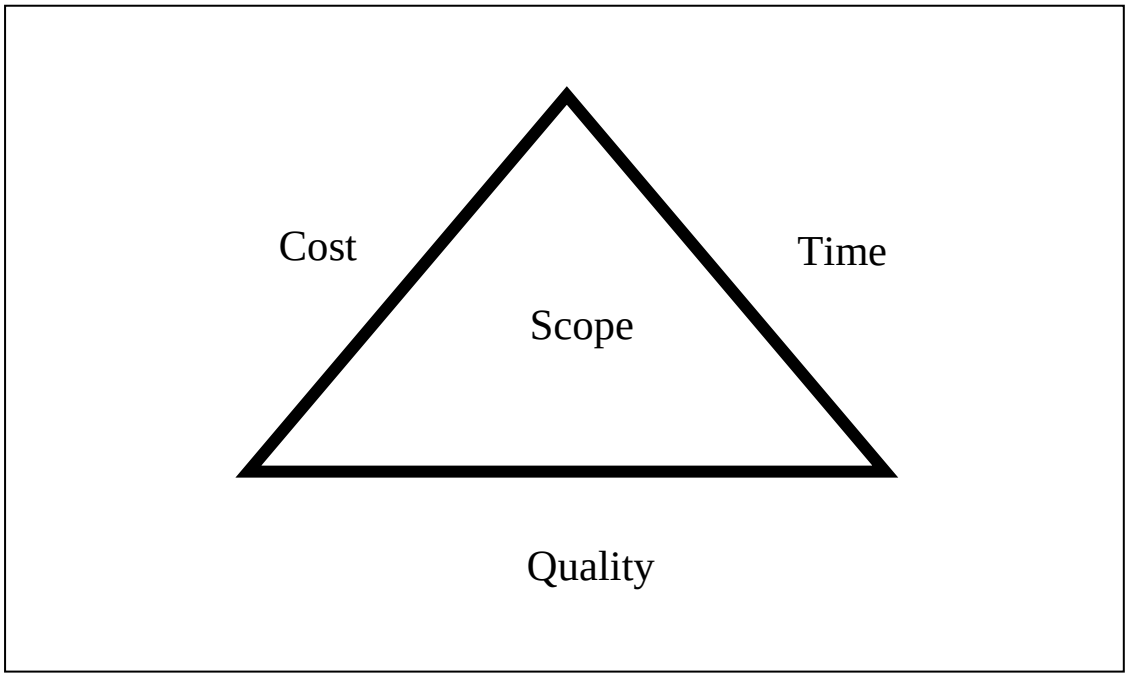


Figure 12: The iron triangle: Extracted from Martin Barnes (2014)

The three dominant project limitations continued are illustrated throughout the literature in different sets of terms - “time, cost, and output” (Langston, 2013), “time, cost, and quality” (Ika, 2009), “spending plan, schedule, and scope” (Agarwal and Rathod, 2006), and “cheap, fast, and good” (Langston, 2013). However, progress in the PM profession has led to the triangle becoming increasingly unpopular. This is due to the great number of project limitations that have been pointed out in PM literature (Langston, 2013).

Early researchers, such as Baker, Murphy, and Fisher (1974), described projects as successful if it fulfilled objective as well as subjective factors. The American Heritage

Dictionary defines objective factors as those showing a connection to actual events and verifiable data or information as opposed to thoughts. Subjective factors, in contrast, relate to personal feelings, interpretation, perception, attitudes, beliefs, or opinions, instead of reliance on actual events. Baker, Murphy, and Fisher (1974) studied 650 projects and concluded that subjective factors described by the distinctive nature of perception have a significant influence on project success. Hughes et al. (2014) studied subjectively measured versus objectively measured factors in evaluating project performance. Although the focus was on metrics other than the traditional objective metrics of cost, time, and specifications, they acknowledged that more subjective factors exist. These factors, while not easy to quantify, can have a remarkable impact on projects.

Pinto and Slevin (2007) identifies three subjective measures of project success, namely project perceived value, project implementation process, and customer satisfaction with the project outcome. Kirsch (2010) puts forward that measurement of project success be inclusive of project team member and stakeholder satisfaction with the project team. Hughes et al. (2014) and PMI (2018) distinguishes between measuring project management performance and project performance. The evaluation of project management performance takes place against objective factors (cost, time, quality, etc.), whereas objective as well as subjective factors characterized by perception (customer satisfaction, project team satisfaction, etc.) is used to judge project performance.

To begin with, the criteria for successful projects typically entail various dimensions. The number of dimensions has increased since the discipline of PM as a management discipline was initiated in the mid-twentieth century. PM success was at first measured by time, cost and quality which are criteria related to the iron triangle (Atkinson 1999). To manage the iron triangle, KM in PM focused on these three dimensions. For example, KM was done by generating and sharing information related to project performance by the means of schedules (time criterion), budgets (cost) and specifications/standards (quality).

It still is critically important to manage these dimensions in terms of effective PM. Yet, over time there has been a growing acceptance that other important performance dimensions exist and need to be considered. Health and safety, client, user and stakeholder satisfaction, sustainability, and quality assurance through adherence to defined processes and procedures are some of the dimensions (Mir and Pinnington 2014). The implication of this development is that amount of knowledge needed to be generated, shared and, ultimately, managed has dramatically increased to include all these dimensions. The volume and diversity of data which

need to be converted into useable knowledge, to both measure and manage performance against the various success criteria, poses a great challenge.

Alongside the broadening scope of KM in terms of the dimensions of success, a second issue is the acknowledgement that project success varies depending on stakeholders' perspectives (Davis 2014). This is further complicated by the fact that competing values may exist among stakeholders (Walton and Dawson 2011). Some stakeholders may desire contrasting outcomes from the same project. In some cases, a particular stakeholder may deem a project successful if it either fails or is called off (Bryde, 2015).

Considering that knowledge is power, project stakeholders may use their knowledge to advance their own goals and agendas in relation to a specific project. This might not align with the stated project goals and may clash with stated project objectives. This issue must be acknowledged and mitigated. If not, the power of KM could be undermined. For example, withholding information from people both inside and outside the project team; feeding misleading data about the actual progress of the project into the PM system (for instance, in order to hide perceived bad news from clients and other interested parties).

Thirdly, statistics on project failure and poor PM performance, indicate that participants often do not absorb lessons learned from previous projects (Von Zedtwitz 2012). The failure to learn results in underutilization of PM skills for the future delivery of projects. The importance of the PM capacity and capability agenda is on the increase. Project-focused industries and PM-related professional bodies want to further improve levels of performance to reach better outcomes for stakeholders. KM has a crucial role in achieving better outcomes. By sharing knowledge from present-day projects and from past projects to current projects, managers are provided with a wealth of knowledge to use in future projects. This will lead to delivery of the desired outcomes. Furthermore, building PM capacity and capability is subjected to time constraints.

A further drop in PM skills is forecasted with the rise of the age profile of the typical project manager (Hoxha and McMahan, 2019). However, this can be curbed by effective mechanisms to share the wealth of knowledge individuals gained during the many years of their PM careers. Project knowledge will be lost if experienced managers leave the industry without sharing their knowledge of PM best practices. According to evidence this can be countered if KM processes positively relate to PM competence and retention. Such processes can include formal lesson sessions, incorporating activities such as the establishment of opportunities for career development and growth (Ekrot et al. 2016).

Lastly, the structures formed to guarantee successful delivery of projects, proved to pose some challenges. Projects have specific start and finish dates. Therefore, it is necessary to form a temporary organisation for the duration of the project. This temporary organisation exists only for the duration of the project and is dismantled once the project is complete. The temporary nature of the organisational structure brings some specific difficulties for KM as team members cooperate temporarily and each go their own way at the end of the project. This means the knowledge developed and gained during the project needs to be managed while the project is running. The next section reviews the post project activities.

6.2.6 Post project reviews

A post project review (PPR) is the procedure by which an organisation reflects on the project process with the goal of learning from project activities. This is done to avoid future mistakes and to learn from successes and failures (PMI, 2018). A PPR is also defined as “a formal audit of the project which looks at the lessons which may be learnt and used to the advantage of future projects” (Media, 2012). Whilst many authors acknowledge the benefits of conducting PPRs, several recurring problem areas are identified. These problems provide project teams with opportunities to share, discuss and explain their experiences through face-to-face, facilitated interactions before a project is closed and the team is dissolved. Thus, PPRs allow multi-disciplinary teams to critique a project and determine the positive and negative aspects, potentially capturing tacit knowledge as learning points to improve the planning and execution of future projects.

Debates between team members during PPRs may lead to greater innovation and creativity by individuals. This communication is important as each individual contributor has his own perspective of the project story (Kerth, 2010). In the United Kingdom (UK), major companies such as BP Amoco, BAA plc, National Grid Transco, and construction companies such as Bovis Lend Lease, IPSL, Simons Design and Buro Happold, have adopted PPRs to learn from experience. Conducting PPRs takes time, is manpower intensive and expensive in terms of company overheads. Disterer (2010) mentioned that after finishing the project, team members do not have close contact anymore and project documentation is stored in folders without retaining the essentials for later use. In his assessment of PPRs, Disterer (2010) highlights that “PPRs are important learning mechanisms, and their value seems to be underestimated by individuals who do not appreciate the need to disseminate insights throughout the organisation”.

Companies conduct PPRs as a part of their quality systems. However, the fact that companies have insufficient resources to act on the outcome of PPRs proves to be a problem (Carrillo, 2015). Companies appoint individuals to create PPRs, but do not have individuals or teams responsible for analysing the PPRs to identify the good or bad practice and areas of improvement across a range of projects. In addition, Kamara (2013) identified that PPRs have huge potential for much more thorough exploitation. Effective extraction of information and knowledge from PPRs leads to identification of good and bad practices. This will point out lessons learned from past projects. Knowledge will be reused to improve the quality and levels of success in future projects. PPRs have several benefits as explained in the subsequent section.

6.2.7 Benefits of Post Project Reviews

Benefits gained by organisations from conducting PPRs have been highlighted in (Tan, 2016) and Carrillo (2014). The benefits stressed by the two authors are listed below.

- **Facilitating collective learning:** PPRs provide an opportunity to get involved in a project and examine what went right or wrong during execution of the project. There should be space for knowledge sharing, exchange of ideas, brainstorming, identifying good and bad practices and contributions that enhance learning.
- **Provide usable knowledge:** Knowledge that can be used for future projects should be the outcome of a successful PPR. However, the result is often tacit knowledge and is therefore difficult to reuse.
- **Benefit client organisations:** The aim of review processes should be to provide better insight into how assets are managed. This should help the project organisation to improve its processes and asset management.
- **Better project phase management:** Reviewing each phase of a project provides opportunities for better project management at the phase level, rather than carrying out a single review at the end. Hence, mistakes will be detected early. This is beneficial to the remaining phases of the project.
- **Prevent knowledge loss:** Knowledge may be lost at the end of a project when team members disband. The related knowledge should be captured by a PPR process and made available for others to use.

6.2.8 Difficulties with Post Project Reviews

The resulting benefits of PPRs have been applauded by various. However, several issues have also been highlighted. Challenges associated with PPRs include the following:

- **Ad hoc processes** – although organisations may have existing procedures to conduct project reviews, it is not to say these are systematic and followed throughout the organisation.
- **Availability of key staff** - key staff may not always be available to conduct reviews as some of them may be transferred to other projects.
- **Timing** - this is a key issue regarding PPRs. Key staff have to return to pressing duties during the period all project expenses are being accounted. This means key staff may not be available for reviews (Carrillo, 2014).
- **Content** – If the PPR does not have a useable format, length, and content, it could be mistrusted (Fairclough, 2012).
- **Dissemination** – although review reports are available in organisations, the dissemination of key knowledge and lessons might be ineffective.

The above challenges can be overcome. However, they call for resources not necessarily available. Rather than leaving these potentially useful reports untapped, text mining is explored to help discover knowledge from PPRs. Since the literature is organised into three sections as stated in the introduction section of this chapter, the next section reviews literature on knowledge management and how the process can be applied in project management.

6.3 Knowledge Management

Knowledge has been accepted as a powerful asset for almost all organisations in the knowledge-based economy (Liebowitz, 2008; Liu, 2016). Sharing knowledge is regarded as truly important for firms in their journey to business achievement (Mazorodze and Buckley, 2019). An organisation with a wealthy source of knowledge is more likely to keep up its competitive advantage and achieve commercial success (Gonza´lez-Loureiro, 2015). As such, knowledge management is paramount to projects as knowledge transfer across projects impacts project performance in terms of costs, schedule, and quality. Consequently, Berteaux and Javernick-Will (2015) suggest knowledge is one of the strongest competitive advantages in project-based organisations. Organizations' effective management of knowledge has become a critical organisational capability (Davenport and Prusak, 1998; O'Dell and Hubert, 2011; Liu, 2016; Girard and Girard, 2015). The next section reviews literature on KM in project environments.

6.3.1 Knowledge Management in Project Environments

A project environment is regarded as any organisation that handles various projects as part of their core business operation. The definition of a project underscores the temporary nature of the activities that are conducted by various individuals with different set of skills, working

towards the same goal for a defined period of time (Nicholas 2011). Koskinen, Pihlanto and Vanharanta (2013) have concurred that a project involves a group of multi-disciplinary individuals working towards the realisation of a specific objective. These individuals being experts, bring together significant knowledge. Studies have shown that the knowledge gained from project environments can be a powerful source of wealth, and add to economic improvements (Fernie, Green, Weller and Newcombe, 2013; Bresnen, Edelman, Newell, Scarbrough and Swan, 2013). Such knowledge can further add to existing bodies of knowledge and advances in the industry. The primary contribution of knowledge from project environment is the improvement in process and systems, which lead to optimum project performance and in turn leads to the organisation's competitive advantage (Weller and Newcombe, 2013). Through the sharing of best practices and lessons learned, projects have a higher chance of success.

Repetition of errors may be prevented by applying knowledge gained through lessons learned during the life cycle of previous projects (Carrillo et al, 2014). This can therefore diminish the expenses and time needed for rework. The systematic documentation of mistakes or potential predicaments decreases project risk (Schindler and Eppler, 2013). From the submissions by these and other authors it can be concluded that it is beneficial for the knowledge that is discovered and created during the project to be conveyed to future projects. Similar to any KM process, the knowledge management process in projects involves the generation, administration, dissemination, and utilisation of knowledge within and outside the project (Bresnen et al., 2016). Projects always generate project knowledge in the form of technical, procedural, and organisational knowledge. The challenge of KM in project environments is the dissemination and sharing of a newly generated architecture of knowledge because of the nature of projects.

By nature, projects' time and resources are limited, processes are multi-disciplinary, and teams keep evolving, yielding clear challenges regarding KM. The individuals who form the project teams are often busy with their specific tasks that contribute to the project tasks and this results in little or no time for review writing or knowledge sharing activities (Purvis and McCray, 2013) (Carrillo et al., 2014). Consequently, project knowledge is captured, but not shared and will dissipate along with participating members once the project is terminated. Explicit knowledge can be captured in project documentation such as project schedules and project technical reports. However, it is challenging to capture tacit knowledge in documentation. This is a point consistent in literature and emphasised by Schindler et al. (2013). Tacit knowledge is best transferred through the social interactions that occur among individuals (Nonaka and Takeuchi, 1995). Nonetheless, the effectiveness of tacit knowledge

transfer depends on strong relationships, built on trust and through time (Fernie et al., 2013). Time constraints within projects therefore present a challenge to tacit knowledge sharing as it is difficult for people to build trust over short periods of time (Nicholas, 2011). The result is that the trust conducive for effective knowledge transfer by the end of a project might still not exist.

In some cases, project team members could be scattered both organisationally and spatially, (Kasvi et al., 2013), which demands non-personal communication like electronic mail. Remote working which resulted from the Covid-19 pandemic demonstrated how project members can be spatially separated but continue to successfully complete a project16 (Bushuyev et al., 2020). However, effective communication may be hindered by detail, confidentiality and idiosyncrasies embedded in messages. Team members might shy away from would likewise personal interaction with other members, hindering the transfer of tacit knowledge (Koskinen et al., 2013). As can be observed from these submissions, knowledge transfer in multi-disciplinary teams is a matter of concern. Adding to the challenge, teams may comprise a scope of experts and artisans (Carrillo et al., 2014), using distinctive processes and practices (Nicholas, 2011) and who might speak different languages.

Some research studies postulate that the core of KM in project environments is the management of explicit knowledge (Fernie et al., 2013). The PMBOK Guide (Project Management Institute, 2018) highlights that the distribution of data and recovery of data is essential for project management. As such it is crucial to ensure that the captured explicit data is disseminated rather than archived. Most projects today undertake the process of codifying tacit knowledge through review packages (Liebowitz and Megbolugbe, 2013), project debriefings (Schindler et al., 2013), administrative closure procedures (Project Management Institute, 2018) and other knowledge repositories. It is, however, generally acknowledged that the effective way of sharing and exchanging tacit knowledge is through social interaction. Such social interactions can occur in workshops where individas are required to share their knowledge, and discussions (Fernie et al., 2013). At times knowledge transfer occurs during informal conversations (Koskinen et al., 2015), coaching and mentoring sessions (Mazorodze and Buckley, 2020) and brainstorming sessions (Carrillo et al., 2014). The culture of a company should bolster knowledge sharing (Liebowitz et al., 2013), it is therefore crucial that top management is involved in cultivating a culture that promotes knowledge sharing ad maintaining an environment that is conducive to the same.

Carrillo et al. (2014) assert the importance to assume ownership and accountability for the processes of knowledge management in project environments. Schindler et al. (2013)

mentions the need for a project debriefer and defines the role as preparing and facilitating review workshops. The debriefer should also take responsibility for documenting the results from the review workshop. Scholars have commented that managing knowledge is advantageous, for improvements within a single project (intra-project), but it is more beneficial for improvements across different projects (inter-project). According to Carrillo et al. (2014), knowledge management across projects has the potential to increase both competitiveness and profitability. Additionally, the documentation of project knowledge systematically will help organisations to conduct a comparison between various projects methodically and concentrate on the most effective problem-solving mechanisms (Schindler et al., 2013). Kasvi et al. (2013) proposes a conceivable way to deal with inter-project knowledge management, while Bresnen et al. (2013) underlines the obstacles experienced in integrating cross-functional contributions and perspectives.

6.3.2 Managing knowledge throughout the project life cycle

Because projects are predominantly temporary in nature, project management focuses on managing the activities crucial to accomplish project objectives throughout its life cycle (Patanakul et al., 2016). Documentation occurs at each phase of the project and used from start to finish. As a result, processes and activities that pertain to knowledge management need to be embedded in all the project stages to ensure that these processes and activities are carried out effectively and timely (PMI, 2018). There are five broad process groups that are identified in the Project Management Institute's PMBOK guide as part of the project life cycle. The process as identified in in section 6.2.1 are initiation, planning, execution, monitoring and controlling, and closing (Alwaly and Alawi, 2020). Initiation process involves the discovery, needs analysis and initial scoping of the project. Planning process group defines the constraints that the project operates under such as refined scope, time, and cost. The execution process group focuses on the carrying out of the project as defined to realize project completion. Monitoring and controlling, ensures that the project is on track and meeting the set milestones, where issues arise, they need to be addressed. Lastly closing, which is conclusion of each phase or the receiving of approval for the completion of work undertaken for the phase or project (PMI 2018).

A typical project lifecycle has various phases, namely concept, where the original idea is crafted; definition, where the preferred solution is recognized and refined; development, meaning the execution of the plan; handover and closure, which sees conveyance of the product or service and formal conclusion (APM 2018). While a few procedures are predominating in certain phases, they are not confined thusly. This way the focus will be on

codifying lessons learned as well as promoting the measurement of benefits, empowering the enhancement of both intra-project and inter-project practice (Fuller, 2017).

As far as the five processes identified by the PMI and the phases of the project life cycle as characterised by the UK-based Association of Project Management (APM, 2018), the knowledge management (KM) movement must not concentrate on a single process or single phase, but KM should be applied to all the processes and associated phases. Consequently, a change of mindset is essential to move away from the PM focus of conducting KM on the closure process and the closure and handover phase alone. For example, Patanakul and Shenhar (2016) describe the traditional mentality during the execution process, which has an operational spotlight on getting the job done. The shift proposed would include a focus on the wider business and strategic issues associated with executing the project; this way execution will be efficiently and effectively.

To achieve both efficiency and effectiveness, Patanakul et al. (2016) argue that PM should focus on continuous team learning. Post project reviews synonymous to lessons learned, is an example of a KM type activity. As such lessons learned sessions should be conducted throughout the life cycle of the project. Broadly, the APM outlines the types of KM-based activities that must take place as the processes are carried out in each phase. There are two broad categories of KM-based activities identified by the APM : knowledge capture in projects and knowledge use (APM 2018). Knowledge capture during projects must happen during meetings on project status and within the project teams themselves. Knowledge capture usually take places during project documentation and exercises such as project reviews (Newell et al., 2016). The argument posited by Patanakul et al. however requires that this activity should not be restricted to the occasions. Project team members should capture and utilize the knowledge from previous projects during and up to the end of the project. After capturing the information, it must be recorded in repositories that can be accessed to inform future projects.

The rationale behind such information recording is that different project teams have access to the reports, enrich their knowledge and learn. Knowledge capture might well incorporate diverse organisations and third parties that make up a transformation management office (TMO). Knowledge capturing can take place during internal reviews, but also amid audits and health checks. Some of these activities might be carried out by outside bodies given the responsibility to undertake KM activities in PM and facilitate knowledge sharing through means of a project management office (PMO) (Pemsel and Wiewiora, 2013). Capture should

be attempted at key points throughout the project life cycle dependent on the particular PM methodology used.

In project management, typical uses of knowledge at different points in time comprise of developing a strong business case to define the governance approach; to find solutions to problems; and to improve performance, both personally and in the team (APM, 2018). Research has differentiated between tools that capture and use knowledge through sharing enveloping endeavours to cover both explicit and tacit knowledge. Tacit knowledge is captured through post-project reviews, project meetings, consulting individuals, communities of practice, technical forums, brainstorming sessions, and conferences/training. Explicit knowledge is captured through project review files, intranets, skills/expertise databases, lessons learned documents, best practice sheets and audit documents (Carillo et al. 2014).

Knowledge structures in projects should promote foundational approaches to manage both tacit and explicit knowledge. For instance, Kasvi et al. (2013) proposes a Learning Project Model (LPM) and suggest the use of project workshops to dynamically updat: the project plan document and the team contract document. These two documents are the repositories in which knowledge is captured. Regarding the project plan, 'hard' project knowledge, including project definitions, activities and results are captured. In the case of the team contract, organisational knowledge is captured, including experience and capitalisation of lessons learned. The repositories are updated regularly, typically at milestone reviews at key stages. The key to success is knowing that that KM should be done orderly, which will only happen if the project is managed in a methodical manner. However, there are some challenges as explained below.

6.3.3 Challenges with capturing and using knowledge in PM

The goal of project management is to capture and use knowledge across projects mostly by project documentation and reviews (Newell et al., 2016). Knowledge from previous projects must be captured by project participants, for example as lessons learned at the end of the project. The captured knowledge must be written down as reviews and uploaded onto databases together with other project documentation. This to grant future project teams the opportunity to access the reports, read, enrich their knowledge, and learn. Unfortunately, it has been proven that PM processes to capture knowledge, for instance, through formal lessons learned and project reviews, often fail (Keegan and Turner, 2011).

To obtain and use project knowledge effectively proves to be challenging (APM, 2018). Typical challenges are a lack of time to engage in KM activities, a lack of resources and clear

guidelines and a lack of executive support (Shokri-Ghasabeth and Chileshe, 2014). Other difficulties include the view that transferability of lessons learned from one project to another is limited; in some cases, it is transfer of knowledge is not seen as a priority by the project manager who might consider it a diversion from delivering a successful project. One manner in overcoming these barriers is by formally making KM-based activities and specific tasks part of the PM process and putting aside time for such activities. By doing so, guidelines are set in the form of documented processes and procedures, time, resources, and formal roles allocated to knowledge capture and use.

If a formal PM methodology (PRINCE2 or an organisation's own propriety method) is mandated, and such activities are part of it, knowledge can be captured and used in each framework. The final, and very important, step is that top management support the process by guaranteeing the method is applicable to all projects and that the time, resources, and roles are available and sorted as planned. However, even with such formal PM strategies in place, shortfalls in KM remains a challenge (APM, 2018). Formal reviews, either during or post-project, is the most widely used way to capture and use knowledge in projects. This method creates four major barriers, three of which are: psychological, managerial, team-based, and epistemological. The psychological barrier means an inability to reflect while the managerial barrier relates to time constraints. Team-based barrier entails the blame shifting and a lack of internal communication structures (Von Zedtwitz, 2012). The fourth barrier classed as 'epistemological' highlights possible impediments with a methodology that assumes project knowledge is explicit and ready to be codified and generalised to other contexts.

To understand the restrictions of the codification-based approach to knowledge transfer across projects, one should explore the underlying assumptions about knowledge and how they fit in the PM context. One assumption is the 'knowledge as possession' view (Newell et al. 2016), asserting that knowledge is something that can be articulated and transferred from one entity to another. Knowledge is possessed by individuals (Nonaka and Takeuchi, 1995), project teams and organisations and such knowledge can easily be managed. Critics of the knowledge as possession view underlined that knowledge is a part of social and organisational practices and relationships (Tsoukas and Vladimirou 2011). In this perspective, knowledge, or rather knowing, is embedded in practice, and cannot easily be detached. Tacit knowledge is considered as highly personal. In this view, knowledge presupposes values and convictions and is closely related to human action (Tsoukas and Vladimirou 2011). More so, knowledge is rooted in personal judgements and tacit commitments. Since knowledge is embedded in practice, it is impossible to directly transfer knowledge across projects where there exists no

link between practices. The two knowledge management strategies, namely codification versus personalisation, are reflected in these perspectives, (Hansen et al. 2009).

Codification is essential to making knowledge unequivocal and transferring it across projects, and thus, reflecting the possession view. Despite what might be expected, the emphasis of personalisation is the communication that takes place to share knowledge and encourage learning. According to the personalisation strategy knowledge is firmly connected to the activities of participants and needs to be shared, thus, reflecting the practice perspective. Relatively recent work to develop systemic lessons learned knowledge models, acknowledges that KM in projects is part of a wider complex adaptive system. In this wider system, the organisation undertaking a project comprises people-related elements such as learning, culture and social, as well as systems-related elements such as innovation, procedures, and infrastructure (Duffield and Whitty, 2015). KM models which align these diverse components, for instance by utilizing stories from past project encounters, can teach lessons for present use and future undertakings (Duffield and Whitty 2015). In order to comprehend the knowledge as practice perspective it is imperative to know more about knowledge and its underlying tacit dimension.

Wiewiora (2009) carried out an empirical study on barriers to effective knowledge transfer in project-based organisations. The need for a study on that topic was recognized in works by Desouza and Evaristo (2006) and Landaeta (2008). The scholars state that project-based organisations face serious knowledge needs in their projects. They also assert that PBOs tend to repeat the same mistakes too often due to the absence of effective knowledge transfer from previous projects in the same organisation. Carrillo (2014) states that despite a project's uniqueness, project experiences can be reused in other projects, providing valuable lessons. Consequently, it is important to share knowledge across projects in order to avoid unnecessary reinventions of what has already been done and decrease chances for failure in that way. Wiewiora (2009) grouped barriers to effective knowledge transfer into three categories as explained underneath.

- Barriers related to inter-project transfer of lessons learned, where collection of lessons learned almost never occurs, or if it does, it occurs periodically rather than throughout the performance, which causes important information to be missed or forgotten.
- Barriers related to social communication, where a lack of links between project teams results in a lack of knowledge sharing between them. A major barrier in this aspect can also be a negative atmosphere created in project-based organisations which makes employees feel unwelcome to share bad experiences. However, most interviewees agreed

that social communication is the most effective way to share valuable knowledge and information.

- Barriers related to the project manager, which mainly include situations when project managers hoard their knowledge, as they view it as a potential future threat for them.

In all three categories the biggest recognized barrier was a lack of time, in terms of keeping focus on the final delivery rather than on knowledge transfer activities which has the potential to speed up the process. When discussing the issue of knowledge transfer in project-based organisations, Ajmal and Koskinen (2018) emphasize that organisational culture is very often an obstacle to such activities. The authors confirm that while knowledge management is of crucial importance for project management, it fails due to cultural factors, rather than technological oversights. The subsequent section discusses some of the benefits of knowledge discovery in projects.

6.3.4 Benefits of Knowledge Discovery in projects

Popularity of data mining and knowledge discovery is the result of increasing demand for tools to reveal and comprehend information hidden in vast amounts of data (Cios, 2015). Knowledge discovery in databases is a nontrivial procedure of recognizing substantial, novel, potentially useful, and ultimately understandable patterns in data. Knowledge discovery (KD) is focused on developing methods and techniques to understand data. The essential issue addressed by KD is mapping low-level data (which is typically too voluminous to understand and digest easily) into more compact forms (e.g., a short report), more abstract (e.g., a descriptive approximation or model of the procedure that created the data), or more useful (e.g., a predictive model for estimating the value of future cases) (Fayyad, et, el.1996).

The financial benefits of discovering and sharing knowledge, are widely recognised. The powerful usage of organisational knowledge is perceived to be essential to “improve bottom line results’ (King, 2008) and to “create wealth” (Stewart, 2001). According to another view the new knowledge economy has the capacity for “dramatically increasing economic and social prosperity” (Allee, 2003). Knowledge discovery is imperative in relation to projects as knowledge transfer across projects affects project performance in terms of costs, schedule, and quality (Berteaux and Javernick-Will, 2015). Knowledge discovery is additionally essential in connection to projects because some projects depend on the generation of new knowledge that needs to be integrated and included in organisational practice, or else be lost (Javernick-Will and Levitt, 2009).

The discovery, transfer, and integration of knowledge between projects is imperative to avoid repetition of past mistakes at both the project and organisational levels (Javernick- Will and Hartmann, 2011). Despite the general advances in the PM profession, project-based organisations are continually faced with challenges to maintain their competitive edge. Moreover, additional challenges organisations face is improving project performance and addressing the risks of project failure, highlighting the significance of the role of knowledge discovery and its undisputable impact on unequivocal project success (Anantatmula, 2016; Lierni and Ribiere, 2018). Surely, it is important to share knowledge across projects in order to avoid unnecessary reinventions of what has been already done and decrease chances for failure. Consistent with this, it is accepted that the "discovery of knowledge is a necessary prerequisite for project success" (Love et al., 2013).

6.4 Chapter summary

The chapter reviewed literature on Project Knowledge Management and presented part of the contribution by proposing the 11th Project Management area termed Project Knowledge Management. Knowledge Management is concerned with how to systematically enhance and share knowledge throughout the project life cycle. This involves having a framework for Knowledge Management in the organization and embedding that framework into project management processes and methodology. Project Knowledge Management in projects therefore involves people, processes, and tools.

CHAPTER 7: TEXT DATA ON DATA ANALYSIS

7.1 Introduction

This chapter presents the results of the data collected from participants at a petrochemical company in the Republic of South Africa. Results are presented in chronological order according to the study objectives. This chapter is divided into two sections, namely the analysis of the qualitative data collected and the diagnosis of the problem. The diagnosis phase entails the diagnosis of an area of application in accordance with the action research fundamentals, and it focuses on how project data is currently being analysed in the organisation.

The following section presents the responses elicited from the interviews with the project managers to understand the knowledge requirements and diagnose the challenges currently faced in extracting data from the databases.

7.2 Analysis of qualitative data

Analysis of the transcripts from the semi structured interviews was executed using NVivo software a CAQDAS which was developed by QSR International (QSR, 1995). This software assisted the researcher in analysing the qualitative data from the interviews by using the capability of the query tool which analyses data and detects trends (QSR-International, 2018). The intention of the following section is to provide the reader with a step-by-step narrative description of the data analysis process. Figure 13 on the next page provides a visual illustration of the process that was utilized by the researcher to analyse and code the data. Data analysis was broken down into four major steps:

- Step 1- Explore source (transcript)
- Step 2- Explore broad themes
- Step 3- Review a theme node
- Step 4- Code on

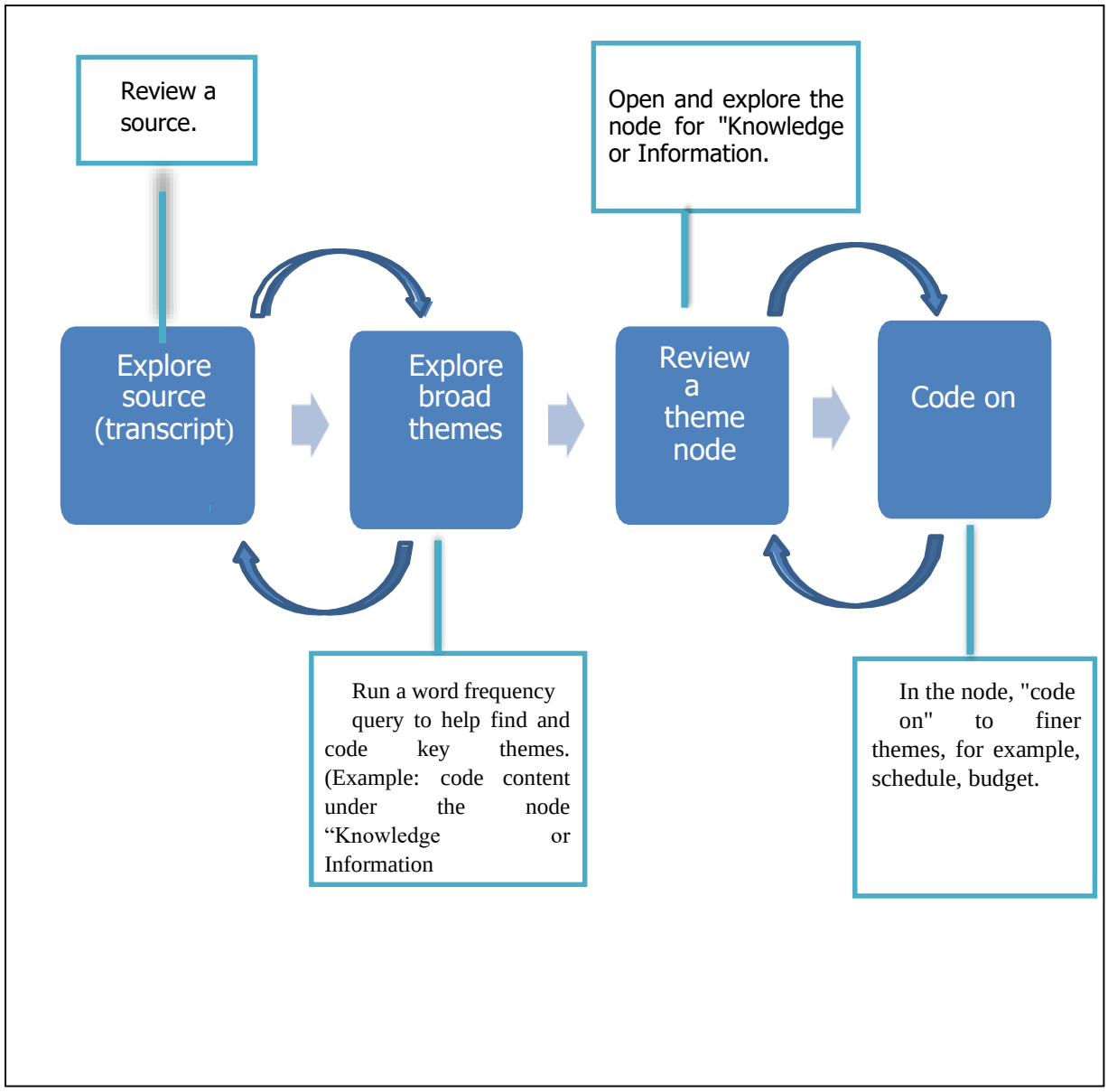


Figure 13: A visual illustration of textual data mining

7.2.1 Step 1-Exploring source (transcript).

Coding is a method of identifying all occurrences of a specific topic, theme, person or other entity and categorizing it accordingly (QSR-International, 2018). According to Buchanan and Jones (2010:15), the most critical step of the qualitative data analysis process is coding. Elliot (2018) adds that coding is a way of "indexing or mapping data, to provide an overview of disparate data that allows the researcher to make sense of them in relation to their research questions". Consequently, it can be deduced that coding can be viewed as a way of tagging data that is relevant to a theme or topic.

Interviews were recorded and then manually transcribed into a text file in MS Word which can be easily imported to Nvivo, a qualitative data analysis software. During the manual transcription of the audio data, the researcher utilized the time to establish initial impressions and detect trends. The initial information gathered revealed that the organization deals with a diverse range of projects. All interviewees suggested that there are many opportunities for improvement in the way in which information or data relating to projects is currently being extracted and converted into useful information. These findings were acquired through coding of the data that was gathered from the project managers at a petrochemical company in South Africa.

7.2.2 Step 2 - Exploring broad themes.

An approach called broad-brushing coding was used as a secondary step to understand categories within the data obtained from the participants. Broad brushing automatically codes sources based on words or phrases it contains (Elliot, 2018). The broad brushing technique also enabled the selection of relevant topics within the data. The researcher used the broad-brush approach, running multiple keyword and word frequency searches. For example, keyword searches for "difficult or inform" were ran. Based on these types of searches, NVivo 11 software produced 236 references throughout the transcript texts for the researcher to further review.

7.2.3 Step 3 - Reviewing a theme node.

The 236 references were broken down into 17 different nodes. According to (QSR-International, 2018), "a node is a collection of references about a specific theme, case, or relationship". Nodes are important to working with NVivo because they allow the researcher to deposit similar data in one place so that one can look for emerging patterns and ideas. Nodes can be used to classify a set of data under a particular theme (Wong, 2008). The nodes were then explored by the researcher to identify patterns in the data. Next, a process called "coding on" was used to establish a framework of thematic ideas. It is important to state that

this process was done manually. According to (Wong, 2008), this stage involves the reconciliation of the nodes where various nodes were explored for similarities and differences and grouped accordingly. The examination of the nodes can bring forth new discoveries that may require re-coding on to other nodes. For example, within the theme node “knowledge or data” the researcher coded on to a finer theme of “information”.

7.2.4 Step 4 - Coding on

From the data obtained from the project managers and analysed in a project environment, the findings have an implication on decision making in a project environment. In an effort to address the third objective which sought to determine the methods of knowledge discovery from textual project data, the following section presents and analyses the practical requirements for knowledge management in projects.

7.2.5 Empirical findings

The findings presented below were obtained from the project managers who partook in this study with an objective to determine the knowledge needs of the projects environment and also identifying the current challenges for knowledge discovery from project data. A total of ten questions were asked, the first question sought to identify the role of the participants in the organisation as explained below.

Question 1: What is your role in the organisation and this department? The responses from the participants are presented verbatim below.

Participant **A** said:
“I am a Project CEO for the Centre of Expertise. I am Senior Lead Specialist and therefore work closely with capital projects.”

Participant **B** stated that:
“I am the portfolio manager.”

Participant **C** pointed out that:
“Project manager for Engineering & Service project based in the Project Centre of Expertise. I am responsible for large projects from initiation to completion.”

Participant **D** stated that:
“I am a Senior Manager in the Capital Project or group technology. I oversee Major Tasks Portfolio for Tier 1(larger projects) up to Tier 3 projects (smaller projects)”

Participant **E** explained that:
" I am a Project manager who oversees the project feasibility for different phases of the project. I deal mainly with Tier 1 (larger projects) and Tier 2 projects (Medium projects)"

Participant **F** stated that:
"Project Director, working as an Elite Construction specialist in the organisation. "

Participant **G** stated that:
"Senior group technology Project manager, responsible for capital projects."

Participant **H** emphasised that:
" Manager of Project Portfolio. I have 5 managers reporting to me."

Participant **I** stated:
"I am a senior project officer responsible for Environment Project Controls."

Participant **J** confirmed that:
"Group Technology project manager, responsible for project controls and general project management."

From the submissions above, we can see that all participants were subject matter experts in the area of project management. The participants have practical working experience and knowledge within the project environment. The participants were working with projects; hence they could give relevant information for this study. According to PMI (2018) project managers are responsible for planning and implementing projects including managing their budgets. As such the project managers were best suited to provide information regarding gaps in the project lifecycle as it pertains to knowledge management. The following section presents responses from participants regarding the core functions of the project department.

Question 2: What is the core function of the project department? The responses from the participants are presented word for word below.

Participant **A** said:
"The name of the department is the project centre of expertise. The core function is to review projects and their quality and allocate the people and resources. It is a knowledge base we

gather information and insights. We work closely with the Capital projects to advise seniors and review and give them recommendations"

Participant **B** stated that:
"Group Technology (Capital Projects). Under the Capital Projects there are portfolios. The department that I am in is a small project. There are different portfolios. There are 36 small projects in the region and 9 project managers."

Participant **C** pointed out that:
"We are project reviewers for small projects. We act as gate keepers making sure/guaranteeing that project execution is maintained"

Participant **D** was of the view that:
"Execute project on behalf the business units within the mandates that have been agreed upon in terms of cost, quality, and schedule. "

Participant **E** explained that:
"Capital Projects in the Project Management department our core function is developing the feasibility phases for projects. The function we execute project we take a project from feasibility through all of the projects right through to beneficial operation. We take a concept, and we develop that to a solution and implement then, Capital Projects has two main departments which are Front Engineering and Execution. Some departments will have Small Projects which have a different scope. My portion is T1 and T2 project these are larger projects. These are projects above 500 million rand"

Participant **F** stated that:
"Currently I work for group technology, and we work for the function, which is called the Project Centre of Expertise. It is the department where we collect data from other departments as well as external sources to provide data. We are responsible for data collection, dissemination and providing recommendations that includes governance."

Participant **G** stated that:
"Currently in the front end of projects (feasibility thus defining the solution to the problem) so as to put it into execution. I am in 2 Tier projects. I have knowledge in Shut down projects, which are projects which are carried out when the plant is shut down. Core function is we take a business plan to generate an income stream or optimise or to restore integrity or an asset. We make Capital commitment."

Participant **H** emphasised that:
"We deal with all types of projects Tier 1, 2 & 3 mainly accountable for project execution."

Participant **I** stated:
" Managing different types of projects Small, Medium, and Large known as Tier 3, Tier 2, Tier 1 respectively."

Participant **J** confirmed that:
"We perform the supporting function for projects that includes creating a feedback loop for projects. Review projects quality and advice management on cost, resources, and implementation."

It is again very clear from the participants' submissions that their core function in the project department is to review projects and allocate resources accordingly. This finding is quite consistent in literature where the PMI (2018) states that some of the core functions of the project department is to review project budgets and allocate the resources to different project sections. Project reviewers act as gate keepers making sure that project execution is maintained at all times. The next question sought to identify the kind of information collected by the project management and what it is used for.

Question 3: What kind of information is collected by the project department and what is it used for?

Participant **A** said:
"Terms of reference, people, cost, scope. We collect data then review and analyse documents and form opinions based on the documents. We compare this data to set norms etc."

Participant **B** stated that:
"Terms of references, people, cost schedule, scope. You also need to utilise your expertise."

Participant **C** pointed out that:
"We have to review the scope, methodology and project charter just post initiation stage. It is mostly textual and to some extent numeric. Past procedural, safety performance, lessons learnt, risk management, engineering."

Participant **D** was of the view that:

“We deal with various internal reports that re receive on a period basis. There is the monthly management report, weekly progress reports and other reports about project status. The structure of the reports is mainly word documents/ textual format “

Participant **E** explained that:
” It depends on where you are in the cycle of the project. It depends how you want to execute the project to a point where you procure material and execute the projects. Engineering information and role it into estimates and execution plan into a schedule. We use information from previous projects. In the estimate department they will have an estimate of previous projects in their database. In the planning department they will have template for schedules on how long certain activities will take. Lessons learnt is used for our execution project.”

Participant **F** stated that:
“The project depends on the triple constraint and the factors that are considered knowledge are cost, quality, schedule, scope, safety, engineering, and quality. This is collected to determine the scope of your project. The data is mainly textual data in discipline specific function. The data is not connected to a single database that is different projects have independent databases.”

Participant **G** stated that:
“Engineering documentation such as drawings and plans to execute the required work.”

Participant **H** emphasised that:
“We look at overall project performance – scheduling performance, cost performance, safety performance. We also collect safety lessons learned”

Participant **I** stated:
” There is a lot of documentation which include drawings, sketches, SharePoint documentation- all together over 36 000 documents.”

Participant **J** confirmed that:
“We look for the requirements of the project. This is the first stage, and we call it a charter. The information will be distributed to the technical team and the project manager. Obviously, we will go through processes of scope clarification and then we will sign off that stage one review is complete. At this stage we know the amount of the hours needed (scheduling), we understand the scope (Scoping) and we have all the (technical information) which is required.

Stage 2 we look for the basic engineering package, thus all the information from the engineers. We look for the project execution."

The kind of information collected by different projects differ, but in general it includes the project scope, people as well as the terms of reference. The data is then reviewed and analysed from the documents. Hereafter, the project managers form opinions based on the documents. It was also confirmed that projects deal with various internal reports that are received on a period basis. The reports collected contain detailed information about the project. These reports include, but are not limited to, monthly management reports, weekly progress reports and other reports about project status.

It was also confirmed that the reports collated were in various formats. As such it would be complex to analyse these reports. While the structure of the reports is mainly in word or textual format, other formats such as spreadsheets and diagrams were also used for the reports.

Having identified the information collected by the project department, the next question sought to find the goal for capturing such information in the project management life cycle.

Question 4: What is the goal of gathering or capturing information at the various stages of the PM Lifecycle?

Participant **A** said:
"*The information gathered is shared with the other project teams to learn from"*

Participant **B** stated that:
"The information will be gathered to inform the project execution plan. The lessons learnt is not used as much as it should be."

Participant **C** pointed out that:
"The goal is to determine whether the project scope, schedule and cost are justified and if so, certify the project."

Participant **D** was of the view that:
"We rely on information sharing from other projects the goal is to inform other projects on how to achieve the project Key performance indicators. "

Participant **E** explained that:

" The information will be gathered is shared with the other project teams to learn from. It will also be used to inform future projects."

Participant **F** stated that:

"We collect the information to learn and predict the future and improve on future projects. Determine Primary and secondary factors that affect the success of the project."

Participant **G** stated that:

"We collect information in order to check if there are any deviations so that we can feed the information into current projects."

Participant **H** emphasised that:

"To improve on project performance from one project to the next. To avoid relearning or repeating the same mistakes."

Participant **I** stated:

" To inform during project initiation. When preparing for new projects we can use information captured throughout the lifecycles of previous projects."

Participant **J** confirmed that:

"Learn and predict the future and improve on future projects."

Information is gathered at different stages of project management (PMI, 2018).Participants confirmed that the information gathered is shared with the other project teams to learn from. Others further confirmed that the goal is to determine whether the project scope, schedule and cost are justified and if so, the project will be certified. Information is also collected to learn and predict the future and improve on future projects. Multiple participants concurred that information is collected to learn and predict the future and certainly improve on future projects. According to Verzuh (2015), capturing information in projects helps to achieve the goals and meet success criteria at a specified time. The next question presents responses as to what is considered knowledge by project managers.

Question 5: What is as considered knowledge in the department?

Participant **A** said:

"We look at trends such as the trends for scheduling and try to analyse these trends to predict what is to come like potential issues and try to come up with solutions. Scheduling, Budget, Technical Information"

Participant **B** stated that:

"Technical or engineering, lessons from similar projects."

Participant **C** pointed out that:

"For us the knowledge that is critical to our success is an understanding of the business process which helps us to provide the necessary guidance to project teams if something doesn't look right."

Participant **D** was of the view that:

"Whenever we have a big project, at the end of the project we normally have a Lesson lessons document that is fed into our Project Enablement team. They populate this document into the Lessons learned database which is then available for other projects to use and learn from for successful execution of their projects "

Participant **E** explained that:

" Pre-feasibility stage we are more concerned with engineering deliverables, estimates. However different knowledge is important at different times based on the PMBOK knowledge areas. Information from risk assessment department- risk identification is very important. Resources management and allocation which is with the cost department. Project support functions include scheduling.

Participant **F** stated that:

"Information about the triple constraints and safety. Safety issues are very important to consider. Schedule, cost, quality statistics are important to us."

Participant **G** stated that:

"We rely heavily on engineering documents from The Engineering Data Centre (EDC). We focus on permits processing, schedule, costs, safety risks and potential hazards. The data collected is drawings or plans for design."

Participant **H** emphasised that:
"This depends on the nature of the project – execution or operations. For execution the focus is on…. For operations the focus is on safety"

Participant **I** stated:
" We are interested in methods, techniques, procedures, system information and roles. This information helps us to make informed decisions for the project."

Participant **J** confirmed that:
"Quality, costing, resourcing information or knowledge is useful in order to assist with the implementation phase."

From the submissions by the various project managers, we can see that knowledge is perceived differently. In this domain of project management, knowledge depends on the project currently being executed. Because most projects rely on engineering documents, the drawings themselves could be considered as knowledge. According to KM practitioners (Nonaka and Takeuchi, 1995; Dalkir, 2011; Lee and Corney, 2018), knowledge refers to something which is new and novel. The next section looks at the extent to which lessons learnt from previous projects are used.

Question 6: Lessons learnt are the combined experiences and learning gained from the process of performing the project that should be considered for future projects. To what extent do you use lessons learnt from previous projects?

Participant **A** said:
"Review projects quality and advice management on cost, resources and implementation."

Participant **B** stated that:
"Textual data. Mostly we have to read through it and sometimes it's through verbal discussions."

Participant **C** pointed out that:
"This is sadly lacking or needs to be done better. As part of the feedback process, we receive lessons learned reports, but we do not have a database to store and no mechanisms to analyse the lessons. Unfortunately, the only lessons learned report that I currently use are from projects that they have reviewed."

Participant **D** was of the view that:

"We should really be doing this better or much more than we currently do. Project teams are expected to review these lessons learned before starting projects, but the honest truth is that there are time constraints. We have to look at each lessons learned document individually i.e., separate from each other. We have to read each document and do a comparison. There is not tool that allows us to do a common data search like scope, budgeting information. "

Participant **E** explained that:

"We look at the Project Execution plans from previous projects to use as input into our current projects. We are interested in the changes throughout the lifecycle which inform our Project Execution plan."

Participant **F** stated that:

"Lessons learnt reports are captured after every project therefore we have vast databases of these. They are simply filed, and no one reads them because no one has the time. The effort to read though all reports is too much. If there was a system to retrieve the relevant information, then that would be helpful."

Participant **G** stated that:

"We use these extensively because we need to understand the outcome of past project. What went well what were the shortcomings so that we can integrate these learnings into the projects. We want to avoid defects in material or process deficiency therefore we can pick that up from the lessons learned."

Participant **H** emphasised that:

"The reports are there but they are hardly used because when a project is starting, we are focused on getting it underway."

Participant **I** stated:

"This is an active document, lessons learned are taken throughout the process and incorporated as the project advances."

Participant **J** confirmed that:

"Not really. There is really no time to manually go through these."

Findings from the project managers revealed that lessons learnt from previous projects are crucial to the success of future projects. As stressed by Hiatt and Creasey (2013), lessons

learned are documented information that reflects both the positive and negative experiences of a project. Unfortunately, lessons learned reports were not stored in a database and no mechanisms were in place to analyse the lessons. It was therefore found very important to know how data was extracted and analysed in the organisation.

Question 7: How do you currently extract and analyse the available data?

Participant **A** said:
"*The date is mostly text based. We collect data and work with internal and external institutes such as the Centre of Expertise who will try to read through the schedule and analyse the data and provide estimates*."

Participant **B** stated that:
"The lessons learnt is currently analysed through discussions."

Participant **C** pointed out that:
"We read through text and also do face to face interviews. There is a checklist that we follow so it's a bit structured."

Participant **D** was of the view that:
"When we look at data, we look at it from a specific term of reference - a baseline plan. We read through or manually extract the data and validate against the current construction Key performance indicators. "

Participant **E** explained that:
" The Estimates Department has a database that has all the information. They have templates for scheduling, so the design input documents are given to the Estimates Department. It is loaded into an estimating program which pulls the figures or numbers out of the database."

Participant **F** stated that:
"The lessons learned are captured in a database therefore we retrieve information about the projects if you know the particular project that you want information about."

Participant **G** stated that:
"We request specific Lessons learned reports from the Centre of Expertise. The Centre of Expertise have access to all projects documentation so they can tell you which projects where similar to your current project therefore we get access to those projects and read manually."

Participant **H** emphasised that:
"We look at last similar projects and we meet with key persons and have verbal discussions.

Participant **I** stated:
" We use SharePoint Information Management analysis. Each team member is responsible for his own discipline to analyse relevant data."

Participant **J** confirmed that:
"We search for the data that is accessible to our team and read through."

According to the project managers, the lessons learnt data is currently analysed through discussions. Other project managers confirmed that lessons learned are captured in a database. Specific information can be retrieved about the projects if you know the particular project that you want information about. Other managers stated that there is an Estimates Department has a database that has all the information. Other participants submitted that they use SharePoint Information Management analysis where each team member is responsible for his own discipline to analyse relevant data. It can therefore be to a concluded that there is no standard for extracting and analysing project data.

Question 8: How is the knowledge that is discovered or identified from analysis utilised?

Participant **A** said:
"The lessons learnt that we are able to identify are used build into other projects.

Participant **B** stated that:
"It is not well structured and indigestible. There isn't a maintained and correctly captured database we don't really use it because of the challenges and also, we are not sure about the validity since you can only go through a few reports."

Participant **C** pointed out that:
"We check against the checklist of required documents and then analyse these documents. The devil is in the details- the documents provided have to meet a certain quality."

Participant **D** was of the view that:

“We validate current statuses against the baseline plan and identify factors contributing to the project success or failure “

Participant **E** explained that:
“We use this information to develop the model for input into our Project Execution plan.”

Participant **F** stated that:
“We look at the information and provide recommendations along with disseminated data.”

Participant **G** stated that:
“We rely on expertise discussions to determine which information is useful knowledge- during meetings with other project managers.”

Participant **H** emphasised that:
“The knowledge helps us to improve on the current projects and avoid relearning.”

Participant **I** stated:
” Throughout the process we rely on knowledge extracted from lessons learnt to increase the efficiency of the project.”

Participant **J** confirmed that:
“There is limited use because the knowledge extracted is also limited as a result of time constraints and also the systems in place. We rarely utilize knowledge but rely on the expertise of the project manager.”

Knowledge discovered should be utilized for the success of future projects. It was confirmed that collected information was ill structured and indigestible. More so, information collected was used to develop the model for input into the project execution plans. On the same spectrum, project managers confirmed that they rely on expert discussions to determine which information is useful knowledge. Hiatt and Creasey (2013) confirm that knowledge helps us to improve on the current projects and avoid relearning. From the submissions by the participants and literature we can conclude that knowledge extracted from lessons learnt may increase efficiency of the project. The next section presents data on the challenges faced with data analysis.

Question 9: Are there any challenges you are currently facing with data analysis?

Participant **A** said:
"It is not well structured and indigestible. There isn't a maintained and correctly captured database."

Participant **B** stated that:
"Data is not well structured, and the data isn't maintained"

Participant **C** pointed out that:
"There isn't a proper database and a manner to compare different projects. The accuracy of the information in the database depends on the person feeding the database. The resources to capture the data isn't in place for example time etc. It's not easy to navigate, not user friendly and too complex."

Participant **D** was of the view that:
"I can say we have system challenges in collecting and mining data and the categorization of data. Another challenge is accuracy and integrity of the data. The data needs to be properly validated for a project to be successful. "

Participant **E** explained that:
"Data is not well structured to allow direct comparison of different projects. There is an inability to recognize relevant searches"

Participant **F** stated that:
"Data collected isn't used as information, there isn't a single used integrated database. It is difficult. The data base is exceptionally complex and is not user friendly. There isn't a universal basis of comparison. The interpretation of data and understanding of the relationship within the data. Extracting relevant data/ insights is difficult. It has to be easy."

Participant **G** stated that:
"The process for data analysis is manual. There is limited access to the data because it is not properly secured that is why we have to go through Centre of Expertise to obtain relevant project documentation. Also, the quality of the Lessons learned is questionable."

Participant **H** emphasised that:
"Time constraints to manually read through various projects to search for patterns."

Participant **I** stated:
"Many hours for data analysis because they have to read through all the information."

Participant **J** confirmed that:
"The lack of database and mechanism to analyse the data. The data is not easily searchable. Also, the data is from dissimilar environments therefore different structure."

It was evident that data is not well structured and as a consequence the data isn't maintained. There is no proper database and a manner to compare different projects. The accuracy of the information in the database depends on the person feeding the database. From this finding, we can see that it's not easy to navigate, hence the complexity of the system. Other challenges highlighted by the project managers are collecting and mining data. This certainly poses another challenge in accuracy and integrity of the data.

Question 10: How would you measure an improvement in the data analysis process?

Participant **A** said the following:
"If we can take the lessons learned and show trends automatically and provide reports. This is where improvement is required."

Participant **B** stated that:
"If we can be able to extract and analyse data with different structures or formats/ from dissimilar environments."

Participant **C** pointed out that:
"*Have a proper database which can also provide graphical reports that are easy on the eye"*

Participant **D** was of the view that:
"We need a Lessons learned portal where a group of people can look at these Lessons learned collected across projects and look at the similarities and differences, categorize the lessons into distinct areas such as planning, cost control etc. "

Participant **E** explained that:
"A well-structured database and the ability to search or retrieve data from various projects to make sense of the lessons learned. Also, the ability to recognize relevant searches."

Participant **F** stated that:
"The more predictable the database can be the better. The ability to convert learnings to easily extractable knowledge is important."

Participant **G** stated that:
"A system that make it easy to extract information across different projects. A system that can show for example Top 5 issues, top 5 best practices. While projects are unique there are some components that are similar that can benefit other project."

Participant **H** emphasised that:
"A system that can show patterns or trends of critical information from the projects- something that is not manual."

Participant **I** stated:
" To be able to perform an analysis of similar projects, to specify a criterion and the relevant information is produced would save time. To have a database that is user friendly and quick to return results."

Participant **J** confirmed that:
"The ability to review dissimilar types or formats of data and get an idea of how to make your own project successful."

The data analysis process should be measured to an extent that the lessons learned should automatically show trends and provide useful reports. Project managers submitted that there should be a proper database which can provide graphical reports that are easy on the eye. From the submissions we can infer that a well-structured database which has the ability to search or retrieve data from various projects to make sense of the lessons learned is ideal. Therefore, the more predictable the database is, the better. The ability to convert learnings to easily extractable knowledge is important. A system that can show patterns or trends of critical information from the projects is needed.

7.3 Diagnosis of the problem

Knowledge discovery is critical to the project management process as it forms a core component of the lessons learned process. Knowledge discovered should be utilized for the success of future projects. The lessons learned process takes place throughout the project life cycle where project managers are required to document the project successes and failure (Jugdev, 2012:13) . According to Rowe and Sikes (2006), the lessons learned documents are

used to capture the positive and the negative experiences of each project. As such, lessons learned documents are critical in order to improve future projects. Knowledge extracted from lessons learnt may certainly increase efficiency of the project.

In order to gain an understanding of the problem context and diagnose the problem, the researcher engaged with the users of the system (Baskerville and Wood-Harper, 1996:237). Understanding the problem aids the researcher in formulating a practical solution that can be implemented to alleviate the problem. Since the researcher was not part of the system, the researcher conducted interviews to understand the problem and applied qualitative methods to analyse and interpret the data. Qualitative methods involve the use and study of a variety of empirical material that describe routine and problematic moments and meaning in individuals' lives (Denzin and Lincoln, 2007: 2). Qualitative data analysis is a "process of bringing order, structure and meaning to the mass of collected data" (Marshal et al, 1990).

Qualitative data analysis is, in fact, seeking a connection between classes and themes of data and looking to build comprehension of the phenomenon. Thus, rather than being systematic and structured, the researcher was required to be alert, adaptable and decidedly interface with information gathered. Traditionally, qualitative data was analysed manually through a laborious task (Wong, 2008:14). With the introduction of Computer-Assisted Qualitative Data Analysis Software (CAQDAS), the process of data analysis has been dramatically eased. The researcher was guided by an interview guide to ensure that the conversation remained in context and also guaranteeing that the interviewee had the opportunity to freely express their thoughts about difficulties and opportunities of improvement in the manner in which data or information was extracted within their departments. The semi-structured method enabled the researcher to identify other useful information which was used in analysing the data.

Literature shows that project-based organisations can find value from the lessons learned reports of past projects. One of the main causes of project failures according to (Desouza and Evaristo, 2004) is the lack of project knowledge management. Cope III et al. (2011) asserts that knowledge capturing, and knowledge discovery are critical components of knowledge management in projects. The importance of sharing knowledge across projects is also recognized by Love et al. (2005:155). In order to be able to share knowledge, the knowledge first needs to be discovered from past projects. Because project data is largely textual, the application of textual data mining can potentially improve the knowledge management process in projects through knowledge discovery.

The analysis of the semi- structure interviews highlights problems with knowledge discovery from the lessons learned reports. The critical problems are highlighted below:

Problem I

There is no centralized repository for capturing projects across the organization. Participants submitted that there should be a proper database which can provide graphical reports that are easy to decipher. From the submissions we can infer that a well-structured database which has the ability to search or retrieve data from various projects to make sense of the lessons learned is ideal. Therefore, the more predictable the database can be the better and the ability to convert learnings to easily extractable knowledge is important. Furthermore, there is a lack of process integration and no formalized standard for capturing knowledge for lessons learned purposes. The formats for capturing the lessons learned are inconsistent.

Problem II

There is no prevalent consistent utilization of lessons learned by project managers. This is primarily because of time constraints when commencing a new project. Furthermore, the lessons learned documents are cumbersome and would require extensive time commitment to effectively analyse the documents individually. There is no categorization of lessons which would make it easier to filter the required knowledge area for analysis. There is a lack of knowledge taxonomy that would improve search results by providing both a meaningful context and ideas for further exploration. Taxonomies are the basis of classification schemes and indexing systems in information management such as the Dewey Decimal System

Problem III

Because of the aforementioned challenges, some of the lessons learned are not captured at all. Some project managers rely on storytelling for knowledge transfer and knowledge discovery. From the interviews, project managers confirmed that they rely on expert discussions to determine which knowledge is useful to improve on the current projects and avoid relearning. While storytelling is an effective method for transferring tacit knowledge (Wijetunge, 2012), the risk remains of experts exiting the organization without completely sharing the knowledge that they possess (Alam et al., 2020). As such it remains critical to have explicit knowledge that is codified.

7.4 Chapter Summary

This chapter presented the diagnosis of the problem through the analysis of interviews conducted with the project managers. The problem identified was that project managers are frustrated by the current process of analysing and interpreting project data collected throughout the projects. This confirms the hypotheses that informed this study. The chapter confirms the challenge for project managers in analysing codified data in unstructured formats that are poorly presented for analytical purposes. The chapter also confirms the need for a centralized repository for managing the data if proper analysis is to be conducted. Finally, the chapter confirms that with the typical time constraints of projects, project managers need a more efficient method of analysing explicit knowledge that may be useful for future projects. From the expert interviews in this chapter, the conclusion can be made that there is a lack of knowledge discovery and knowledge management from collected project data.

Chapter 7 formed the diagnostic phase of the study which focused on understanding and expanding the knowledge area of data analysis in the project management environment. In Chapter 8, a framework of ideas is developed where knowledge management can be applied in project management.

CHAPTER 8: RECOMMENDATIONS TO ANALYSE TEXT DATA

8.1 Introduction to Action Planning

This chapter presents the plan of action. It deals with knowledge needs in the project environment through the problem-solving cycle. In addition, the chapter provides effective methods for textual data analysis in project environments. Moreover, the chapter demonstrates the value of textual data mining for the projects environment.

8.2 Applying Knowledge Management in the project environment.

The diagnostic phase focused on understanding and expanding the knowledge area of data analysis in the project management environment. This chapter focuses on the intervention phase of action research. The intervention phase is comprised of action planning and action taking (Susman, 1983, pp. 106-107).

According to Mathiassen et al. (2009), action researchers are faced with two goals when conducting their studies. The first goal is to develop theoretical knowledge of value to a research community. The second goal is to produce practical knowledge that is applicable in the context. Alfaro-Tanco et al. (2021) mentions a third goal, that of helping the integration between academia and industry. The action research process involves investigating existing theories around the problem and these theories were then applied in solving the practical problem. In line with the action requirements specifications, the researcher was involved in a research cycle. Secondly, action researchers have the goal of developing a way to solve a practical problem which is of value to the people in the relevant system setting. The researcher used the information gained from the interviews in order to get an understanding of the problem context. Such understanding aids the researcher in formulating a practical solution that can be implemented to alleviate the problem. Figure 14 on the following page depicts the action research cycle according to Susman. In this chapter, the action planning phase is discussed.

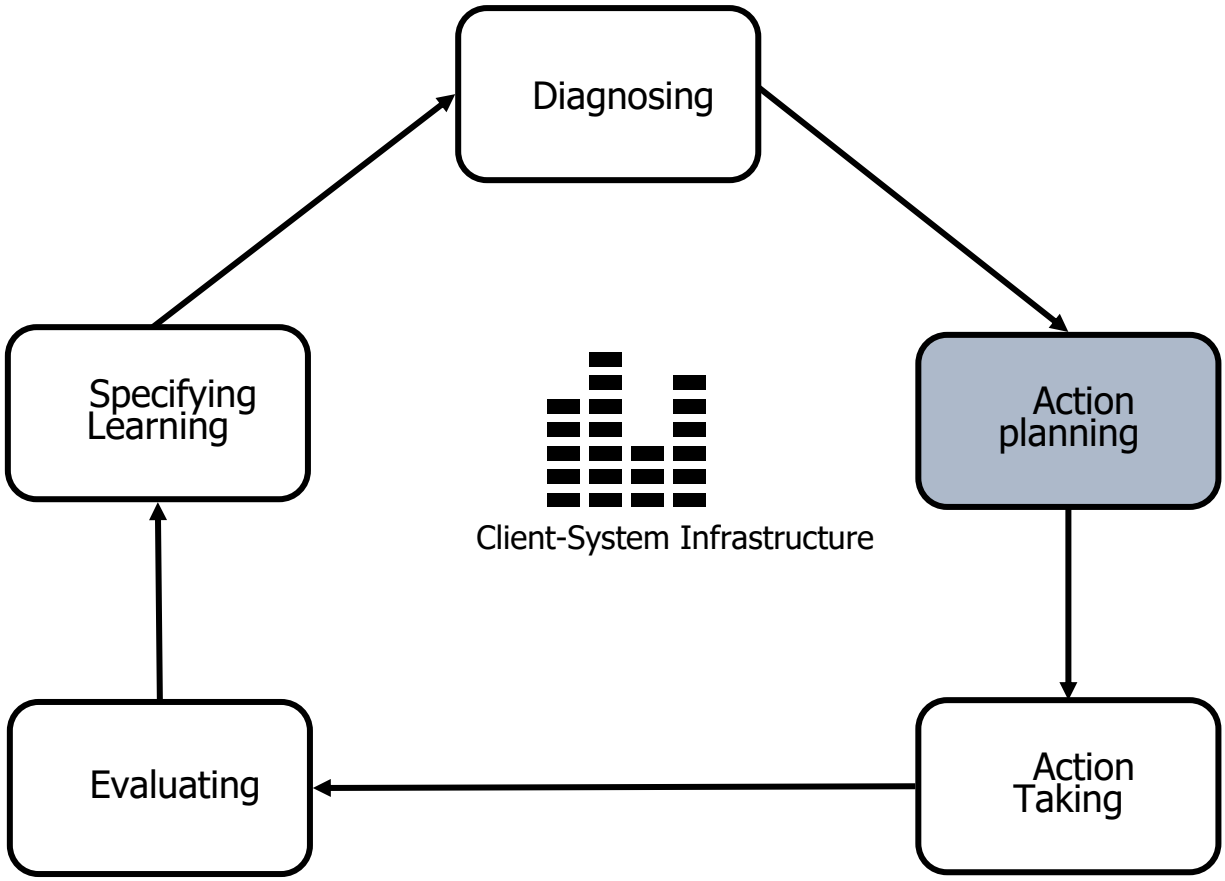


Figure 14 : The Action Research Cycle (adapted from Susman, 1983)

8.3 The problem-solving cycle

Action research entails solving a real-world practical problem (Eden and Ackermann, 2018). In order to diagnose the problem, the researcher was involved in the problem context and engaged with project managers within the problem context. The researcher acted as an agent of change to be consulted in order to improve the situation (Baskerville and Wood-Harper, 1996:237). The researcher identified a petro-chemical organisation and approached them to understand their project knowledge management process. It was found that the organisation faced challenges in performing knowledge discovery because of the numerous time sensitive projects that had to be performed. This finding is consistent to literature where Usai et al. (2018) confirmed that text mining to discover knowledge is a challenge for information and knowledge communities. The petrochemical company which was the focus of this study is an international integrated chemicals and energy company established in South Africa that leverages technologies and the expertise across 23 countries.

The organisation develops and commercializes technologies to build and operate world-scale facilities to produce a range of high-value product stream, including liquid fuels, chemicals, and low-carbon electricity. The company is one of the largest investors in capital projects, skills development and technological research and development. The annual report of the petro-chemical organisation shows that several projects were conducted in 2017. The organisation has a project management department which is accountable for managing capital projects and these capital projects expend several billions of Rands per year. These projects sustain the organisation's growth initiatives, which entails a wide variety of projects. As a consequence, the organisation possesses vast amounts of historical project data that was analysed to identify useful insights that could be valuable in improving future projects. The department comprises of project leaders or managers who oversee a portfolio of projects within the entire project department from the idea generation phase of the projects right up until the project close out phase.

From the information available, projects are a critical component of this organisation as projects form a key basis for the organisation to create and sustain their capital assets. For this organisation, projects form the core of the organization's growth strategy. The participants in this study were all project managers involved in overall project management processes in the organisation, specifically the analysis of the project reports. A total of ten managers were interviewed after obtaining their consent. Text mining provided a possible solution and could help project teams to gain knowledge from previous projects. Text mining (TM) could be a tool to "discover" useful knowledge from these reports. During text mining there is no need for manual searches through vast numbers of reports, likely including a variety of formats and foci, seeking trends useful for current and future projects. To avoid relearning and improve on current and future projects, text mining is imperative.

8.4 The practical requirements for project knowledge management

From the interviews with the project managers at a petrochemical company in South Africa, the researcher gathered information that is sufficient to confirm that there is a significant amount of information gathered and created during the lifecycle of each project. In addition, most of the information is essential for the success of current and future projects. As alluded by the PMI (2018), it is important to know the historic project performance for insights such as cost estimation; pitfalls from previous projects; scheduling. The following sub-section analyses data on the practical concerns of the project managers.

8.4.1 The practical concerns of the project managers

From the interviews conducted, the project managers indicated that it is critical to have a centralized repository where all the projects' information is stored and accessible. Currently the information gathered in each project is stored in various databases that can be accessed independently. It is essential to maintain a knowledge base to facilitate the knowledge discovery process. A knowledge base is a machine-readable centralized repository of information that is used as a resource for dissemination of problem-solving expertise (Ibrahim, 2003:113). Such a centralized repository would be instrumental in correlating data from similar projects. Furthermore, a centralized repository would make the searchability of information easier. This finding is consistent in literature of Doskocil and Lacko (2018) who also state that information can be searched easily from a centralized database.

According to (Rao and Dey, 2011) , "In text mining one of the key elements is to discover unknown information by linking together existing text data to form new facts or hypotheses." Currently, the data is presented in various formats. Traditional databases would have limited capabilities to analyse textual data in the various formats. With a refined data repository, text mining would scale-ably analyse the data for patterns and new insights that can be useful to project managers. Figure 15 shows how a central data repository can be used to store project data.

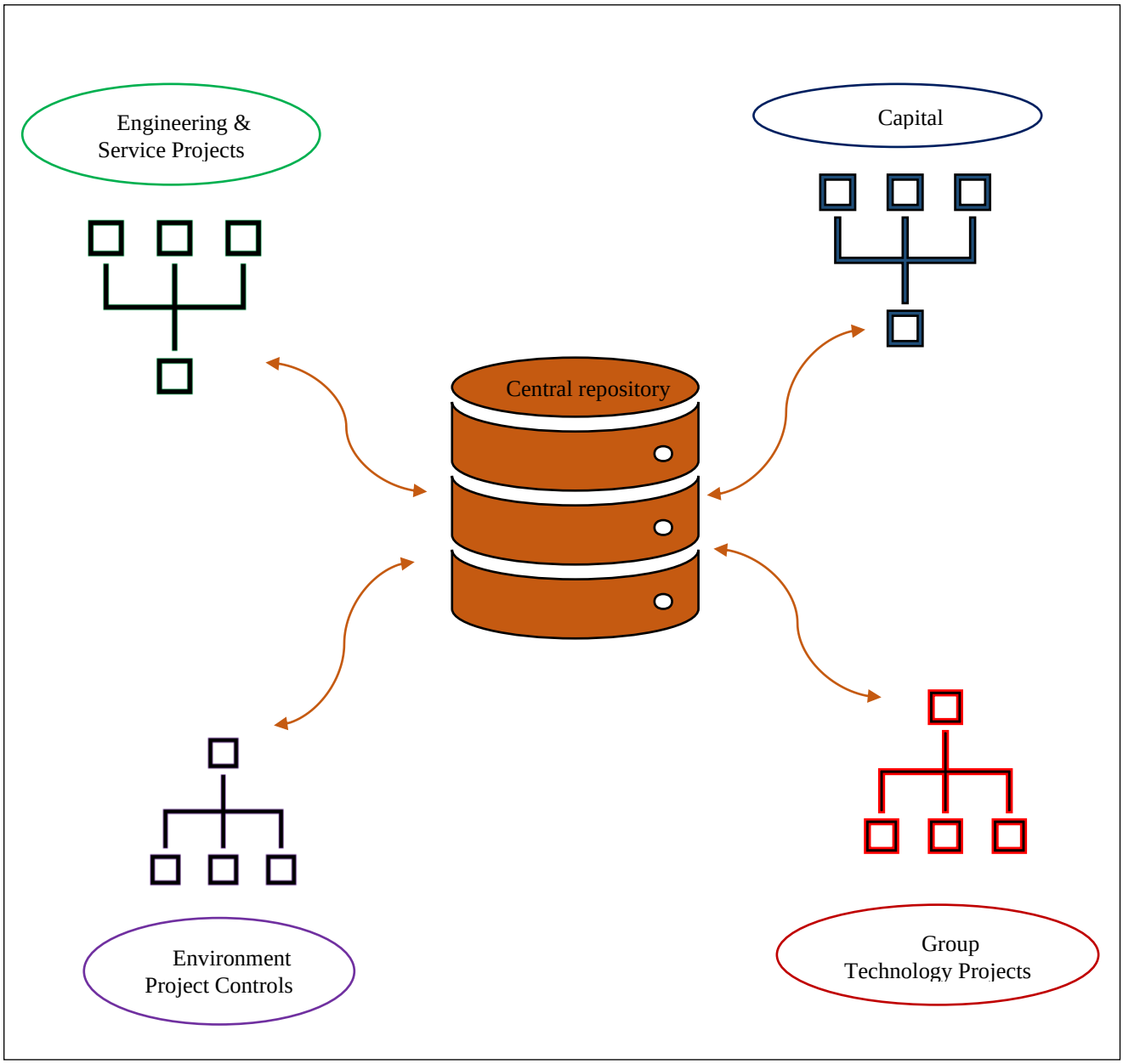


Figure 15: Centralized repository for project data

Another challenge identified by the project managers is the manual process of analysing the lessons learned. There are massive amounts of data that is collected as lessons learned reports and because the organisation yearly runs numerous, it becomes cumbersome to read through a number of project reports in an effort to gather the insights. Since projects usually have time constraints (PMI, 2018), the time resources may not be available to manually read through the relevant lessons learned reports. As a result, some projects commence without comprehension of the similar past projects and project managers will be left with no option but to rely on their own expertise.

At the petro-chemical company studied, the structure of the information collected during the projects is mainly textual documents including mandates, changes that occur during the project lifecycle, engineering information, format pdfs, spreadsheets, and word documents. As a result, when the project managers search for information, the output is in the same format of textual documents. Project managers also indicated that it would be valuable to be able to convert lessons learned into extractable knowledge having a graphical output in the form of a report. This would be useful for the project managers to easily consume the vast amount of information and increase efficiency in their day-to-day operations. The next section analyses data on the need for learning from past projects.

8.4.2 The need for analysis of the lessons learned from past projects

Lessons learned reports are critical to the function of project managers in order to avoid re-learning a lesson that has been learnt in another project. The project managers need a better way of analysis which summarizes the key aspects of the lessons learned such as top five best practices, or top five concerns/pitfalls. Summarization can assist to convert learnings to more condensed information which is better for consumption (Foote and Halawi, 2018). Summarization also reduces the time resources currently being used for the analysis of the lessons learned reports. The following were the key findings from the interviews with the project managers. The insights that are important for project managers include:

- Summarization of the numerous reports into a digestible format.
- Understanding deviations from scope, scheduling, cost, risks, safety issues, planning, and quality.
- Trend analysis in order to come up with mitigation strategies and improvement.

It can therefore be deduced that lessons learnt is particularly important in all project management activities.

Wang and Wang (2020) developed a model depicted in Figure 16 on the next page which portrays the synergistic relationship between data analysis and knowledge management. The model shows that in order for projects to have effective knowledge management, there needs to be strategic use of data. The knowledge that goes into other projects originates primarily from the lessons learned reports. As such, if text mining methods are employed that facilitate critical reading, coding, and interpretation; this may result in the gaining of insightful patterns.

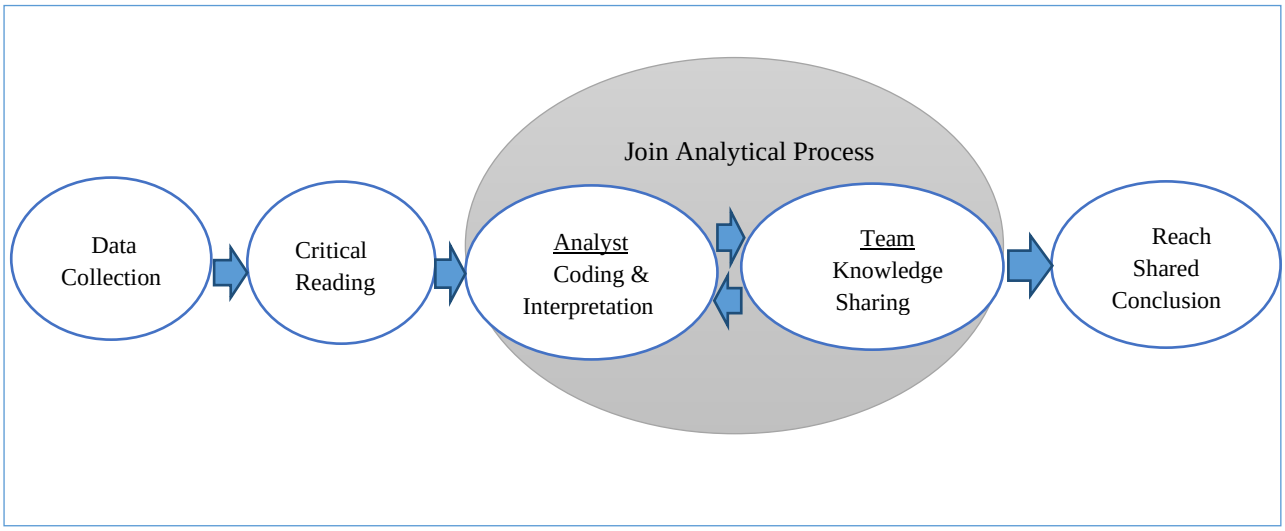


Figure 16. Data analysis and knowledge management (Wang & Wang, 2020)

8.4 Chapter summary

This chapter presented the plan of action. It dealt with the needs for knowledge in the project environment through the problem-solving cycle. The chapter provided effective methods for textual data analysis in project environments. Moreover, the chapter demonstrated the value of textual data mining for the project's environment. The practical requirements for knowledge management for projects were also highlighted. The chapter concluded with the needs for analysis of the lessons learned from past projects. The next chapter presents results from text analysis using recommendations.

CHAPTER 9: RESULTS FROM TEXT ANALYSIS USING RECOMMENDATIONS

9.1 Introduction to Action Taking

This chapter makes recommendations based on the findings from text analysis. The study objectives sought to determine the methods of knowledge discovery from textual project data and also demonstrate the value of textual data mining for the projects environment for improving project management processes. The Project Knowledge Management area improves data driven decision making in project environments.

9.2 Applying corresponding text mining methodologies

For this study, Python was used to perform text mining. Python is an “object oriented interpreted programming language which consists of a large number of libraries that provide excellent functionality for processing linguistic data” (Madnani, 2007:3). Python has notable linguistic processing capabilities which enables the processing of natural language. It should be made clear from the onset that natural language refers to the communication used in day-to-day human interactions. Computer programming languages that are capable of manipulation natural language refer to this capability as natural language processing (NLP). There are numerous technologies today which focus on natural language processing such as language translators, predictive text and hand writing recognition (Bird et al., 2009:2). NLP is sometimes referred to as computational linguistics. Over the years, there have been significant advances in NLP offering a potential for effective text mining (Venugopal et al., 2021).

For this study, the text mining process in this study followed the workflow depicted in Figure 17 on the next page:

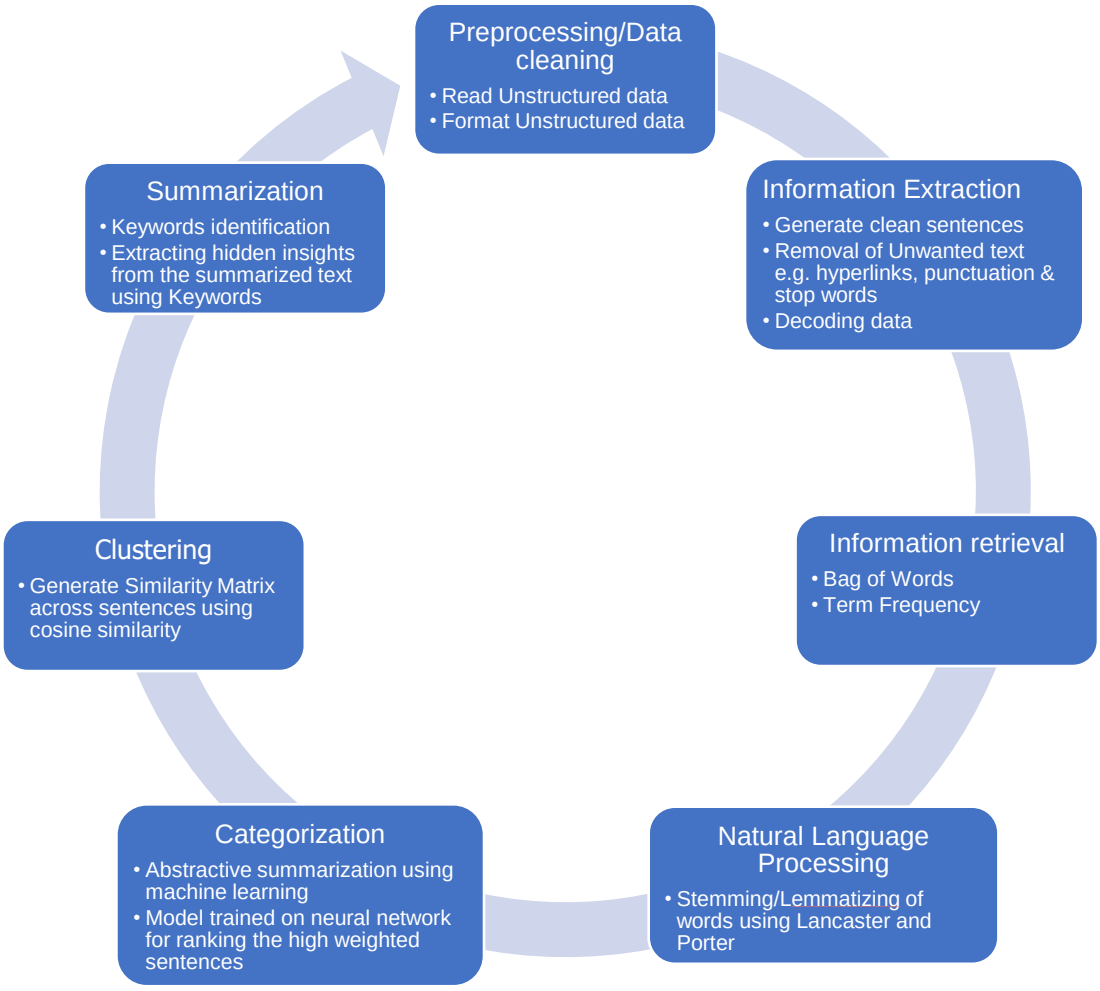


Figure 17: Text mining process workflow

All different file formats have been standardized into a single format type (.xls). In addition to standardization, the extracted columns with desired information and performed data cleaning were also identified. On data cleaning, it was confirmed that not all the categories of data in a dataset are useful to the user. Retaining these unnecessary categories will take up unnecessary space and potentially also bog down runtime. Pandas provides a handy way of removing unwanted columns or rows from a DataFrame with the drop () function.

The next step was to perform NLP which involved stemming and lemmatization. Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words. This process is normally aimed at removing inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. In the areas of natural language processing the researcher came across a situation where two or more words have a common root. For example, the three words - agreed, agreeing and agreeable have

the same root word agree. A search involving any of these words should treat them as the same word which is the root word. As such it became essential to link all the words into their root word. The NLTK library has methods to do this linking and give the output showing the root word. The next section summarises the steps followed in processing unstructured data.

9.3 Application of KDT intervention guidelines in a project management data analysis

KDT intervention involves the action taking and specification of learning. In this section, the process that was followed in the action taking phase to analyse data is described in detail. The process adhered to the defined practices of text mining as stated in Chapter 5.

Steps in processing unstructured data.

Unstructured data analysis is the process of using data analytics tools to automatically organize, structure and get value from unstructured data (Mahajan, 2018). For this specific study, Python was used to process unstructured data and the following steps were followed.

9.3.1 Pre-processing tasks

The files obtained from the project managers were in different formats including PDFs, spreadsheets, and pure text files. Each of the files underwent the pre-processing step in order to prepare the files for analytics and machine learning. Since the unstructured information from multiple sources is in different formats (pdf, doc, docs, xml, jpg, html etc.) a parsing system is used to transform the documents into the format, which has the capability to handle unstructured/semi-structured data (Rao and Dey, 2011:77) First of all, a “HTML” Parser is imported and “re” packaged into the python environment. Because PHP programs often interact with HTML pages, web addresses (URLs), and databases, there are functions to help work with those types of data. Thereafter, removal of all the HTML elements and hyperlinks from the paragraphs was done.

During apostrophe lookup, it is important to import regex libraries into the Python environment. Thereafter, it is imperative to create or load language dictionary to lookup for unbalanced quotation marks and convert abbreviated words into full words e.g., req’d = required. This step is followed by removal of stop words as explained in the next section.

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore when indexing entries for searching and when retrieving them as the result of a search query. One of the major forms of pre-processing is to filter out non-meaningful data. In natural language processing, non-meaningful words (data), are referred to as stop-words e.g., articles, prepositions etc. In the process of removal of stop-words, one

can either create an extensive list of stop-words or one can use predefined language specific libraries like NLTK.

All the punctuation marks according to the priorities were dealt with. For example: “.”, ”, ” ”? ” are important punctuations that should be retained while other needs to be removed. Punctuation is a pre-initialized string used as a string constant.

The decoding phase involves interpreting the meaning of the message. Unicode “Utf-8” was used for reading and writing files in Python. The method was used to convert from one encoding scheme, in which argument string was encoded to the desired encoding scheme. The package used was Pandas.

9.3.2 Processed Document Collection

The stemmer packages, Lancaster and Porter were used to split the attached words in Python. A split () method was used. Following successful data cleaning, the most relevant information from the documents was retrieved using a similarity matrix. As mentioned in Chapter 5, similarity matrices are effective techniques of retrieving the most relevant that meet a certain given criterion.

Categorization was used to convert the lengthy data files received from the project managers into shorter meaningful versions which is a process that could be difficult and costly to undertake if done manually. Text summarization is a subdomain of natural language processing (NLP) that deals with extracting summaries from huge chunks of texts (Harwani, 2012). This study employed the abstractive summarization technique and Gensim Word2vec was used. Gensim Word2Vec is a semantic learning framework that uses a shallow neural network to learn the representations of words/phrases. The text mining model for this study was trained on a neural network for ranking the high weighted sentences on the data set from the project managers to recognize the project management specific terms and language. His was manually done by tagging. However, once trained the model for the project management can be transferable and applicable to other project data sets.

9.3.3 Core Mining Operations and Presentation

Following the training of the model used for this study, a similarity matrix was generated. The sentence similarity was followed by weighting the sentences and finally ranking the sentences according to the weight. Based on certain selected keywords, insights were captured using regular expressions. The package used was regex expression builder. Having discussed the steps in processing unstructured data, the following section reflects on the application of KDT

in a project management data analysis. Visualization was performed which resulted in the word cloud in Figure 19, section 9.4. A detailed account of the application of the core mining operations and presentation is provided in the next section.

9.4 Application of text mining to project data

Ur-Rahman (2017) proposed a new method of discovering knowledge, Multiple Key Term Phrasal Knowledge Sequences (MKTPKS). The implementation of MKTPKS is designed to help decision makers in textual data. The benefit of using MTKPS is that it can identify some key issues discussed in the PPRs and for classifying these as good or bad information documents defined in the free formatted textual data. The results obtained in the form of MKTPKS are compared with the domain experts' key term phrases to determine the effectiveness of the knowledge processing units of the proposed system. The main study objective was to identify hidden key insights from unstructured and unorganized data. This would create a meaningful summarization report from the data available in multiple formats.

The main objective of this study was to investigate whether KDT can improve the analysis of project data by yielding useful insights for the project environment. The action taking phase of this study helped the researcher to evaluate the success of the application of KDT on project data. It is evident that text mining helps to develop an action plan for improving the method of knowledge retrieval and knowledge presentation. The results of the text mining exercise align with the areas that were identified as critical areas of project management. These are the areas that project managers need to gain more insights on.

During the text mining process, some key trends emerged and are presented in the Table 5 and the Word cloud in Figure 18. A total of twenty reports were analysed from the lessons learned reports provided by the project managers. Reports where numbered 1- 20 as they were analysed.

Table 6 : Key trends identified during data analysis

Key trends	Total count
Schedule	59
Risk	55
Impact	53
Requirement	53
Time	47
Scope	43
Delay	43
Safety	42
Cost	41
Issues	41
Estimate	36
Savings	33
Start	33
Contract	29
Manage	25
Change	24
Order	24
Communication	24
Design	23
Document	23

Table 6 is a concordance of the occurrence of each of these words together with an example of a statement in the report where the word occurs for the twenty key areas identified in table 4. Appendix F is a spreadsheet that shows the level of summarization for the lessons learned that were analysed. The percentage values show the level of summarization. The percentage shows the n-grams (a contiguous sequence of **n** items from a given sample of text) in the reference summary. Seventy percent shows the initial data analysis output which represents 70% of all the data that was analysed. After further iterations of analysis, the summarization was concluded at 40%. This is the highest level of summarization that could be achieved which retained meaningful data.

Table 7: Concordance of the occurrence of key trends

Key trends	Report of occurrence	Total count	Example of statement
Schedule	Report 1; Report 5; Report 9; Report 10; Report 12; Report 13; Report 14; Report 16; Report 17; Report 18; Report 19; Report 20	59	Utilise the job assignments according to the WBS and ensure organisation, schedule and execution are aligned. Ensure contractor assigns work scope responsibilities from top to bottom (craft level) and maintains these assignments. **Report 18.**
Risk	Report 1; Report 4; Report 5; Report 6; Report 8; Report 14; Report 15; Report 16; Report 18; Report 19;	55	Ensure the risk register is carefully managed. Have facilitated risk reviews with all stakeholders and assign responsibilities to mitigate risks. Projects that use similar resources during a shutdown, must make use of a combined activity schedule. **Report 14.**
Impact	Report 1; Report 5; Report 11; Report 16; Report 17; Report 19;	53	External problems like quality of electricity to the site also have a major impact on the production rates of coated pipes. **Report 17.**
Requirement	Report 1; Report 4; Report 5; Report 8; Report 9; Report 14; Report 16; Report 18; Report 19	53	Set the requirements upfront and make sure the client and EC teams review and approve the high-level content of the subcontractors' contracts prior to start of next TAR. **Report 18.**
Time	Report 1; Report 2; Report 5; Report 6; Report 8; Report 9**;** Report 10; Report 13; Report 14; Report 16; Report 18; Report 19;	47	Spend more time on details at project kick-off meeting and give enough time for new team members to prepare. **Report 9.**

Table 8: Concordance of the occurrence of key trends (continued…)

Key trends	Report of occurrence	Total count	Example of statement
Scope	Report 1; Report 4; Report 5; Report 6; Report 8; Report 9; Report 10; Report 14; Report 15; Report 17; Report 18; Report 19; Report 20	43	Need to set up correct control systems to separate productivity issues from deviation from scope issues, to properly manage the contract for productivity. **Report 5.**
Delay	Report 1; Report 2; Report 4; Report 6; Report 8; Report 16; Report 17; Report 18; Report 19	43	The liquid epoxy primer did work very well, but the premium and the delays and difficulties in applying the coating solution didn't justify the advantages and almost seriously delayed the project. **Report 19.**
Safety	Report 1; Report 4; Report 5**;** Report 8; Report 16; Report 18; Report 19; Report 20	42	Make equipment (vessels) more safely before work commences (acid still in equipment) (candles). **Report 20.**
Cost	Report 1; Report 2; Report 5; Report 7; Report 16; Report 18; Report 19	41	Prior to change in this key resource there was good alignment between cost controllers which ensured high accuracy on cost flow. **Report 1.** Poor change management of TIC costs resulted in the PBA not being updated in time to facilitate payment to service providers.
Issues	Report 1; Report 4; Report 5; Report 8; Report 18; Report 19; Report 20	41	Need to set up correct control systems to separate productivity issues from deviation from scope issues to properly manage the contract for productivity. **Report 5.**

Table 9: Concordance of the occurrence of key trends (continued…)

Key trends	Report of occurrence	Total count	Example of statement
Estimate	Report 5; Report 10; Report 16;	36	The agreement at SVP level was that risks be highlighted as they were realised and managed in such a way that cost implications were minimised through specialised commercial resources, independent quantity surveyors and project managers to assure capital cost result, well within 10% of ITC that was approved and well within estimate provided by Group Technology. **Report 16**
Savings	Report 14;	33	Consider the use of non-organisational approved vendors as there are substantial savings to be made. If this approach is followed contract on an EPC basis to ensure the risk associated with the unknown vendors reside with the contractor proposing the vendors. **Report 14**
Start	Report 5; Report 8; Report 15; Report 20	33	Work hours should be determined prior to shut down by management to avoid misunderstanding. Report 20
Contract	Report 3; Report 5; report 8; Report 10; Report 14; Report 15; Report 19	29	The proposal is to align the contracting strategy between the main contractor and the EC, thus contract both on lump sum, instead of having one on reimbursable contracting while the other is on lump sum contracting. Site communication is a serious problem and needs to be in place almost from the start of the project. Report 19.

Table 10: Concordance of the occurrence of key trends (continued…)

Key trends	Report of occurrence	Total count	Example of statement
Manage	Report 1; Report 3; Report 6; Report 14; report 18; Report 19; Report 20	25	The discipline to set up, maintain and utilise systems must be enforced. Develop workflow for controlling a master punch list. Develop a formal system to manage the packs and sign off on systems. Ensure the QC inspectors have field experience, formal work processes and inspection reporting mechanisms in place. Report 18.
Change	Report 1; Report 5; Report 7; Report 9; Report 11; Report 14; Report 18; Report 10	24	When a major change or innovation is done, consult with senior experienced people and get their buy-in and builds in order to make sure all angles are covered. Report 14.
Order	Report 1; Report 5, Report 14; Report 16	24	Extended engineering development is required to produce schedules for good contracting design freezes before order placement – although not always practically possible, there should be a complete design freeze prior to sending RFQ's into the market. Report 5.
Communication	Report 1; Report 5; Report 9; Report 18; Report 19; Report 20;	24	Ensure good communication and integration prior to the project starting regarding different company policies. Report 18.

Table 11: Concordance of the occurrence of key trends (continued…end)

Key trends	Report of occurrence	Total count	Example of statement
Design	Report 5; Report 6; Report 7; Report 8; Report 11; Report 12; Report 18	23	Engagement with the client early in the project, understanding their requirements, having discussions surrounding constraints and giving factual explanations when all design requirements "wants" could not be met, ensured alignment between the C&I team. Report 18.
Document	Report 7	23	Various systems (such as cost control and reporting, drawings, and document management) need to be designed and implemented before the work starts. Report 7

The information presented in the Table 7 on the previous pages is presented in the Word cloud in Figure 19 which shows the frequencies of the words used in a project environment. The words with the highest frequency were mined using Gensim Word2Vec. Gensim Word2Vec is a semantic learning framework that uses a shallow neural network to learn the representations of words/phrases in a particular text. The algorithm for this study used Gensim Word2vec and utilised the Lessons learned reports as a text corpus and produced an output of a set of word vectors. The model was then trained to compute the similarity of the vectors. These vectors were then weighted and the sentences with higher rank identified. This is represented by the 40% summarization level. The following Figure 18. depicts the summarization process followed.

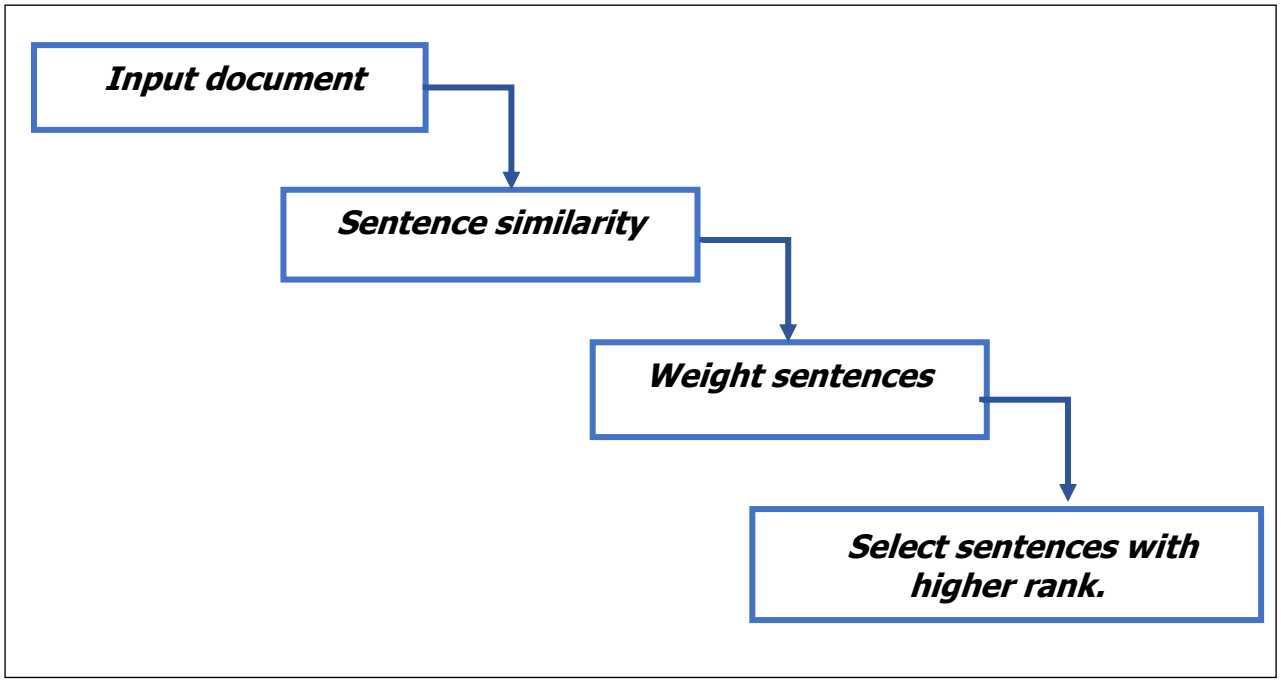


Figure 18 The Summarization process followed in this study

From the sentences with higher rank the word count function was used to identify the frequently occurring words. A total of 479 frequently occurring words was initially identified, stemming and lemmatization were then applied to remove the repetitive words based on the morphology and semantic meaning. The process was iterated until the twenty words identified in Table 6 (section 9.2) remained. Using these twenty words, the word cloud below was created using the Pro Word Cloud add-on in Microsoft Word to visualize the findings.

Figure 19: Word cloud of keywords

The keywords identified in the Word cloud in Figure 19 above were occurring across the project in most cases. However, there were some projects that had more of these keywords occurring. This type of analysis can assist the project managers with focusing on the project reports that have the most significant overall effect based on the number of keywords that are appearing in those reports. The graph in Figure 20 on the next page is a Pareto chart that represents the projects that were most affected by these keywords identified in the lessons learned report that were analyzed for this study.

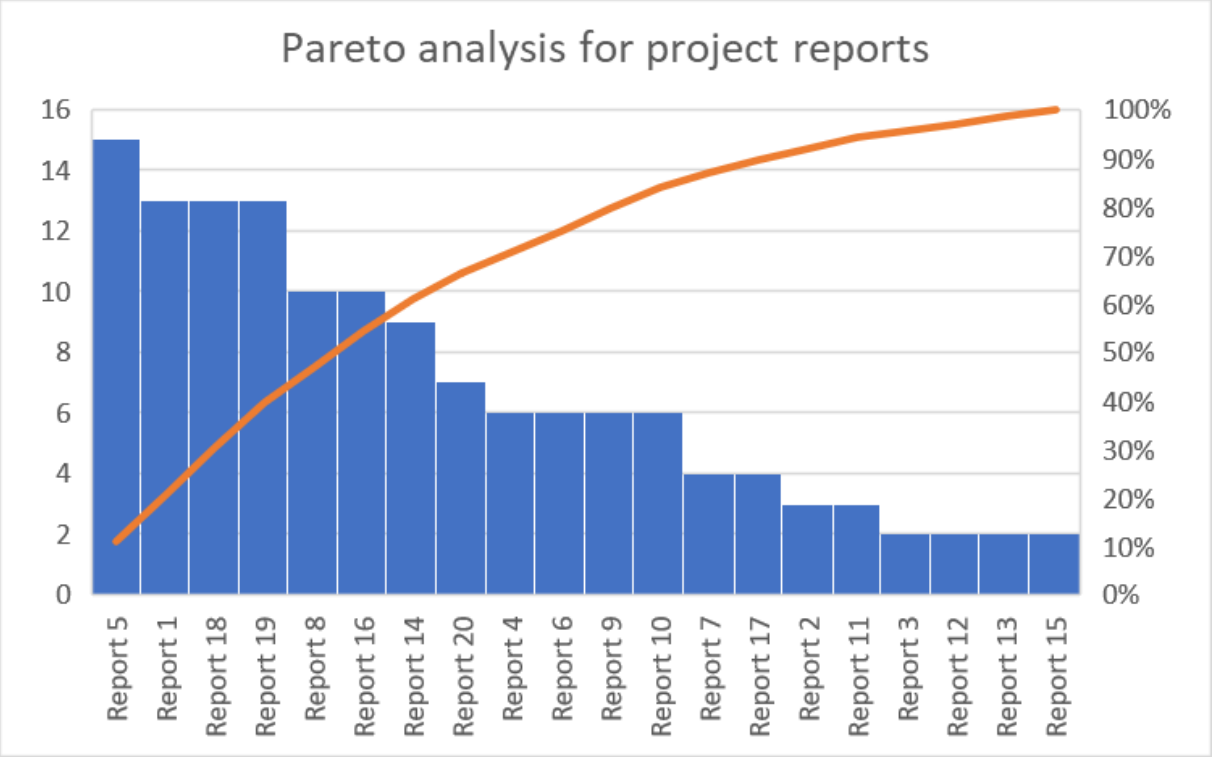


Figure 20. Pareto analysis for project reports

9.5 Evaluation of learning

This section evaluates the study by examining the principles described for action research. This section also assesses the degree to which the research problem was solved and proposes potential areas for future research. Susman (1983) suggests that during the evaluation stage of action research, the following questions must be asked:

1. *Did the actions taken bring about the conditions that the sociotechnical model led us to hypothesize will produce the outcome desired?*
2. *If the hypothesized conditions were brought about, were the desired outcomes produced?*
3. *If the desired outcomes were produced, how confident are we that it was the hypothesized conditions that produced them?*
4. *If the desired outcomes were not produced, what aspects of the sociotechnical model should be re-examined?*

This section of the study will attempt to address these questions.

9.5.1 Did the actions taken bring about the conditions that the sociotechnical model led us to hypothesize will produce the outcome desired?

One of the main challenges identified by the project managers is the manual process of analysing the lessons learned. The actions taken by the use of text mining in Python helped to address the problem with massive amounts of data being collected but not analysed. The various formats that were provided all managed to be included in the text mining process. The desired outcome of simplifying the data analysis process for various formats of data was achieved. Since projects usually have time constraints (PMI, 2018), the use of text mining would alleviate the time resources required available to manually read through the relevant lessons learned reports. As a result, projects will commence with the necessary insights and hence improve the project performance.

Project managers also indicated that it would be valuable to be able to convert lessons learned into extractable knowledge having a graphical output in the form of a report. This would be useful for the project managers to easily consume the vast information and increase efficiency in their day-to-day operations. The actions taken can also provide various graphical outputs like the Word cloud presented. Other graphical outputs can be for example charts or graphs.

9.5.2 If the hypothesized conditions were brought about, were the desired outcomes produced?

From the interviews with the project managers at a petrochemical company in South Africa, the researcher gathered information that is sufficient to confirm that there is a significant amount of information gathered and created during the lifecycle of each project that is difficult to analyse. The desired outcome was to develop an easier way of analysing the vast amounts of lessons learned reports collected. This was achieved in this study.

9.5.3 If the desired outcomes were produced, how confident are we that it was the hypothesized conditions that produced them?

Prior to the use of text mining, project managers expressed concerns that they were unable to analyse the vast amounts of lessons learned reports. It can be concluded with a high degree of confidence, that text mining enables the analysis of the lessons learned reports.

9.5.4 If the desired outcomes were not produced; what aspects of the sociotechnical model should be re-examined?

While the essential aspects of the desired outcomes were realized, there is always room for improvement. A standardized format of capturing lessons learned reports would yield seamless results. Furthermore, creating a taxonomy would contribute to the knowledge management efforts. Taxonomies are divisions into smaller subsets. This can make content, information, and knowledge resource retrieval and access faster and more accurate for text mining purposes. Instead of having to know the exact textual labels, i.e., keywords that describe relevant content, information or knowledge resources, project managers can interactively browse and search for documents by selecting the taxonomy categories that are relevant for their information need (Vu et al., 2018:1).

Each of the keywords identified matches a particular PMBOK knowledge area. Table 8 on the next page adopted from PMBOK shows how the different findings are classified for each knowledge area.

Table 12: The 10 Knowledge Areas & 49 Processes (PMBOK®, 6th ed.)

Knowledge areas	Processes	Keywords
Project Integration Management	Develop Project Charter; Manage Project Knowledge; Develop Project Management Plan; Direct and Manage Project Work; Monitor and Control Project Work; Perform Integrated Change Control; Close Project or Phase.	Start Manage Change
Project Scope Management	Plan Scope Management; Collect Requirements; Define Scope; Create WBS; Validate Scope; Control Scope.	Scope Requirements
Project Schedule Management	Plan Schedule Management; Define Activities; Sequence Activities; Estimate Activity Durations; Develop Schedule; Control Schedule.	Delay Time Schedule
Project Cost Management	Plan Cost Management; Estimate Costs; Determine Budget; Control Costs.	Costs Savings
Project Quality Management	Plan Quality Management; Manage Quality; Control Quality.	Document Design
Project Resources Management	Plan Resource Management; Acquire Resources; Estimate Activity Resources; Develop Team; Manage Team; Control Resources.	Estimates
Project Communication Management	Plan Communications Management; Manage Communications; Monitor Communications.	Communication
Project Risk Management	Plan Risk Management; Identify Risks; Perform Qualitative Risk Analysis; Perform Quantitative Risk Analysis; Plan Risk Responses; Implement Risk Responses; Monitor Risks.	Safety Risk
Project Procurement Management	Plan Procurement Management; Conduct Procurements; Control Procurements.	Order Contract
Project Stakeholder Management	Identify Stakeholders; Plan Stakeholder Engagement; Manage Stakeholder Engagement; Monitor Stakeholder Engagement.	Impact
Project Knowledge Management	Identify and review lessons learned	Issues

It can be argued that one word “Issues” does not fit concisely in the ten knowledge areas. Issues can span across a conglomerate of the existing knowledge areas and phases of a project. When the issues are drilled down in the lessons learned reports that were used in this study, it is evident that the issues range from scope creep, budgeting issues, poor communication, inadequate resources, poor risk management to mention but a few. These challenges continue to be faced by the project managers because there is no project management area that focuses on highlighting and addressing these issues before a new project starts. While lessons learned reports highlight both successes and areas of improvement for projects, further reflection on the statements highlighted from the text mining results demonstrates the numerous issues in each of the existing knowledge areas.

One primary reason why issues from past projects remain unaddressed is because of the typical time constraints for projects. With knowledge discovery from text mining, the time constraint is no longer a limiting factor. Therefore, project managers can discover issues that were pertinent during previous projects and improve on them to secure the success of current and future projects. When a new knowledge area Project Knowledge Management is added, the finding on “Issues” can be classified in this knowledge area. This knowledge area would focus on the lessons learned throughout the project and from past projects. The issues identified in the lessons learned reports would help improve future projects. This knowledge area can thus leverage from text analysis which is the core of this study.

9.6 New Project Management Knowledge Area - Project Knowledge Management

Based on the findings of this study, a new project management knowledge area is proposed. Project knowledge management areas. Project management knowledge areas are knowledge areas that are identified by the project management body of knowledge within project management philosophy. The knowledge areas define the fundamentals that are required to be understood by a project manager in order to execute successful projects. Currently there are ten project management knowledge areas. This study proposes an eleventh project knowledge management area with the proposed title of project knowledge management.

The primary focus of the project management knowledge area would be to identify and review lessons learned with the goal of improving future projects. The aim is to discover useful patterns from past projects that could help to improve current and future projects. The project knowledge management area will fall under the initiation phase and the planning phase of the project lifecycle. These are critical phases that could benefit from a good understanding of the lessons learned from past projects to avoid the same pitfalls.

Lessons learned affect multiple other knowledge areas if not all. For instance, it has been identified in this study that the lessons learned that were analysed from the data provided by the project managers highlighted safety and risk which falls under the project risk management. Such information is crucial to projects because these risks need to be avoided. Another example can be found in the project cost management knowledge area where costs and savings are highlighted as lessons learned. Therefore, while the new knowledge area will primarily focus on managing knowledge discovered from lessons learnt, the project knowledge management area will be relevant across the entire spectrum of knowledge areas and as such will improve the entire project lifecycle.

The input for this knowledge area would be the lessons learned reports from related projects. Various lessons learned report will be gathered from past related projects in order to be analysed for critical factors that may influence the success of the current project. The lessons learned reports will be subjected to a text mining process to identify key lessons that project managers can focus on. This process will be more efficient than reading through all the lessons learned reports. Figure 21 below is a simple illustration of the Project Knowledge Management knowledge area:

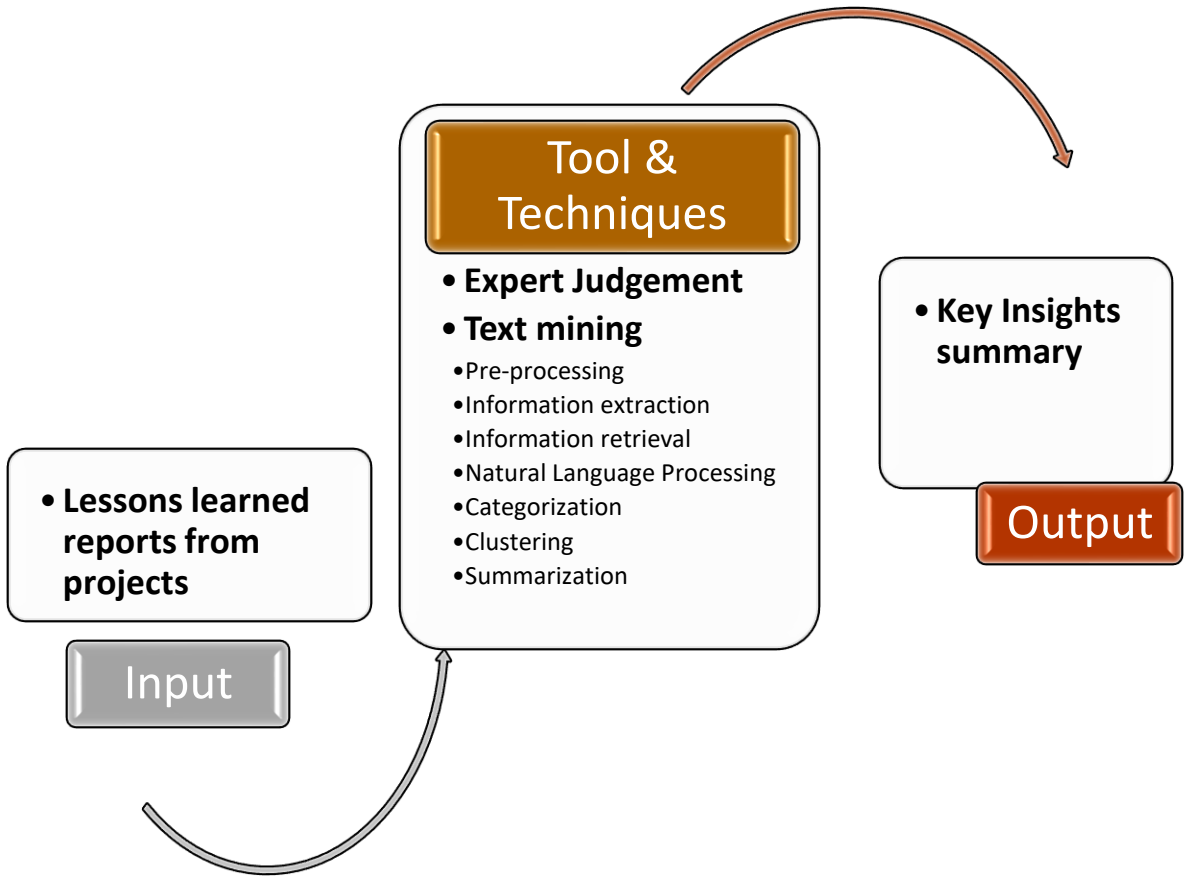


Figure 21. Input, Processes, Output for Project Knowledge Management

9.7 Chapter summary

From the findings, the study recommends that, if one would like to use this algorithm seamlessly, it is best to use a standardized format for capturing the lessons learnt. The study provides a model knowledge discovery from text that may be applicable to other organizations, and this enable organizations to get better insights from their textual databases. The study recommends project knowledge management area to improve data driven decision making in project environments. The next chapter provides a summary and the overall conclusion.

CHAPTER 10: SUMMARY, OVERALL CONCLUSION AND FUTURE WORK

10.1 Introduction

This chapter provides a summary for each chapter of the thesis and an overall conclusion based on the research findings. The aim of the study was to investigate the use of text mining to improve knowledge discovery in the project environment. The other main study objective sought to identify hidden key insights from unstructured and unorganized data in project environments.

10.2 Specification of learning

In this section, the learnings from the KDT process implemented are specified. While learning has been taking place throughout the phases of action research, the last phase is where the capability to achieve transformation is specified. This study closes the gaps to determine the applicability of using a text mining tool such as Python for text mining. The theoretical and methodological contribution is therefore the applicability of performing KDT using Python on lesson learned reports from projects. Below are the specifications of learning.

5. The knowledge environment is currently challenged with vast amounts of data that is not being analysed. As such, project managers are unable to discover useful patterns that can potentially improve the quality of their projects. The research provides a model knowledge discovery from text that may be applicable to other organisations, and this enable organisations to gain better insights from their textual databases.
6. Textual data mining can provide significant value to the project environment, particularly the analysis of lessons learned reports. Lessons learned reports are rich in project knowledge that informs future projects; hence it is crucial to find ways to analyse them efficiently and effectively.
7. Because of the vast amounts of data, a taxonomy should be created. A taxonomy can also improve search results by showing the levels immediately above, below, and adjacent to the search term in the hierarchy, providing both a meaningful context and ideas for further exploration.
8. Since knowledge management is a critical component to the success of a project, an eleventh area for project management body of knowledge is recommended called project knowledge management.

The application of project knowledge management is imperative to guarantee the success of all projects. The project manager serves as a mentor or change agent to establish knowledge management activities as part of the project work. In addition to that, once team members experience the benefit of knowledge sharing in projects, they are more inclined to participate in the whole process. The project artifacts like the schedule, communication plans, risk issues and change control documents can be used as blueprints or templates for the future projects, as depicted. Furthermore, project reviews provide an effective means to capture information and knowledge into the knowledge repositories.

Beneficial knowledge therefore provides long-term benefits regarding the improvement of the overall project performance and cultivating a learning organization. Knowledge is continuously more being valued as a strategic asset vital to sustaining a competitive advantage. Knowledge management provides a timely manner to capture project knowledge and apply the knowledge and learnings to future projects. When knowledge management techniques are applied to project management practices consequent results could include enhanced communication and better project integration, improved decision-making, reduced risks, and continuous improvement in project performance.

10.3 Summary of Chapters

The prime objective was to establish how the textual data mining methods could improve the analysis of project data. An action research approach within the critical research paradigm has been used to meet the objective. According to Baskerville and Wood-Haper (1996: 237), action research is comprised of five phases, namely diagnosis, action planning, action taking, evaluating, and specification of learning. The diagnosis sought to understand the knowledge requirements of project managers and diagnose the challenges currently being faced by project managers in extracting and retrieving informative data from the codified data that is stored in the databases. Action planning sought to develop an action plan for improving the method of knowledge retrieval and knowledge presentation from project data in order to emancipate the project managers.

Action taking sought to act by implementing the proposed action plan in the form of the text mining tools to perform analytics on project data. The evaluation phase sought to evaluate the success of the application of KDT on the project data in terms of improving the process of project data review for project managers. Specification of learning sought to specify the learning from the KDT process implemented and if necessary, iterate the previous stages. The following section presents the responses elicited from the interviews with the project managers

to understand the knowledge requirements and diagnose the challenges currently faced in extracting data from the databases.

10.3.1 Summary of chapter 1 - Introduction and background

Background to the study on the use of text mining to improve knowledge discovery in the project environment has been explained. The problem statement was well explained outlining the areas which are addressed by the study. To accompany the problem statement, a list of research questions and objectives have been clarified.

10.3.2 Summary of chapter 2 - Research methodology

This chapter presented the methodology adopted to conduct this study in South Africa. The chapter discussed the data collection tools used to gather data together with full justification why they have been chosen. The chapter further explained the data analysis tools and concluded with a statement on ethics that the researcher abides by throughout the study. It is therefore safe to state that the chapter provided the basis for data collection and analysis.

10.3.3 Summary of chapter 3 - Knowledge discovery and data mining

The literature review defined knowledge discovery which was the primary focus of this study. Knowledge discovery as well as its associated processes were discussed to lay a solid foundation to this specific study.

10.3.4 Summary of chapter 4 - Text mining

From the submissions by various authors, we can deduce that text mining helps to retrieve documents based on various sorts of information about the document content.

10.3.5 Summary of chapter 5 - Knowledge discovery with text mining

This chapter discussed and proposed the actions to be taken to address the problem identified. Action research was aimed at emancipating users from the bondage of their current process by improving their processes. The purpose of the action planning phase was to ensure that the outcome of the action solves the problems identified in the diagnosed area of aplication before commencing the actual development work. The text mining methodology proposed here required improvement that was established by text mining for project managers.

10.3.6 Summary of chapter 6 - Project knowledge management

The chapter reviewed literature on project knowledge management and presented part of the contribution by proposing the eleventh project management area termed project knowledge

management. The focus of knowledge management is to systematically develop and share knowledge throughout the project timeline. This entails an organisational knowledge management framework and ingraining that framework into processes and methodology of project management. Project knowledge management in projects therefore involves people, processes, and tools.

10.3.7 Summary of chapter 7 – Data analysis of expert interviews

Knowledge discovered should be utilized for the success of future projects. Project managers confirmed that they rely on expert discussions to determine which knowledge is useful to improve on the current projects and avoid relearning. Knowledge extracted from lessons learned may certainly increase efficiency of the project. Participants, specifically project managers further submitted that there should be a proper database which can provide graphical reports that are easy on the eye. From the submissions we can infer that a well-structured database which has the ability to search or retrieve data from various projects to make sense of the lessons learned is ideal.

10.3.8 Summary of chapter 8 - Recommendations to analyse text data

This chapter presented the plan of action which dealt with the needs for knowledge in the project environment through the problem-solving cycle. The chapter provided effective methods for textual data analysis in project environments. Moreover, the chapter demonstrated the value of textual data mining for the projects environment. The practical requirements for knowledge management during projects were also highlighted. The chapter concluded with emphasizing needs for analysis of the lessons learned from past projects.

10.3.9 Summary of chapter 9 - Results from text analysis using recommendations

From the findings, the study recommends that, if one would like to use this algorithm seamlessly, it is best to use a standardized format for capturing the lessons learned. The study recommends project knowledge management area to improve data driven decision making in project environments.

10.4 Overall conclusion

Unstructured data in project environments can be analysed using text mining tools to organize, structure and get value from it. Python is one of the tools which can be used to process unstructured data in project environments. It was confirmed that textual data mining for the projects environment improves project management processes. Furthermore, it was confirmed and proved that project knowledge management improves data driven decision making in project environments. The core of KM is systematic development and sharing

of knowledge throughout the organization. This includes having a knowledge management framework in the organization and ingraining that framework into processes and methodology of project management.

Project knowledge management in projects therefore involves people, processes, and tools. In this regard, people understand the importance of knowledge and information to project success. On the same domain, processes require a framework for knowledge management and embedding that framework into project management processes and methodology. Some of the most effective tools to promote managing and sharing project knowledge involve communities of practice and repositories for storing and retrieving lessons learned. Knowledge management and project management are therefore complimentary practices that can work together to enhance performance in project environments. At first the value of KM practices is demonstrated, and then KM is introduced into the project management process and methodology.

This study therefore recommends the use and adoption **Project Knowledge Management** in all project environments. The introduction of a new knowledge area may have a positive impact on the project success as knowledge from previous projects can be used in new projects.

10.5 Future work

For capturing the lessons learned in the projects environment, the study recommends the utilization of project knowledge management. The proposed project management area can be extended to other organizations which run different projects.

BIBLIOGRAPHY

Abu Bakar, A. H., Yusof, M. N., Tufail, M. A. and Virgiyanti, W. (2016) 'Effect of knowledge management on growth performance in construction industry', *Management Decision*, 54(3), pp. 735-749. doi:

Al-Hashemi, R. (2010) 'Text Summarization Extraction System (TSES) Using Extracted Keywords', *Int. Arab J. e-Technol.*, 1(4), pp. 164-168. doi:

Alam, M. Z., Kousar, S., Shafqat, N. and Shabbir, A. (2020) 'Drivers and challenges of tacit knowledge sharing in automotive workshop employees', *VINE Journal of Information and Knowledge Management Systems*. doi:

Alamsyah, A. and Peranginangin, Y. (2013) 'Effective Knowledge Management Using Big Data and Social Network Analysis', *Management and Business International Journal*, 1(1), pp. 10. doi:

Aldrich, H. E. (1992) 'Incommensurable paradigms? Vital signs from three perspectives', *Rethinking organization: New directions in organization theory and analysis*, pp. 17-45. doi:

Alfaro-Tanco, J. A., Avella, L., Moscoso, P. and Näslund, D. (2021) 'An Evaluation Framework for the Dual Contribution of Action Research: Opportunities and Challenges in the Field of Operations Management', *International Journal of Qualitative Methods*, 20, pp. 16094069211017636. doi:

Alfattni, G., Belousov, M., Peek, N. and Nenadic, G. (2021) 'Extracting drug names and associated attributes from discharge summaries: Text mining study', *JMIR medical informatics*, 9(5), pp. e24678. doi:

Ali, N. H. and Ibrahim, N. S. 'Porter stemming algorithm for semantic checking'. *Proceedings of 16th international conference on computer and information technology*, 253-258.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. and Kochut, K. (2017) 'A brief survey of text mining: Classification, clustering and extraction techniques', *arXiv preprint arXiv:1707.02919*. doi:

Alwaly, K. A. and Alawi, N. A. (2020) 'Factors affecting the application of project management knowledge guide (PMBOK® GUIDE) in construction projects in Yemen', *Journal of Construction Engineering and Management*, 9(3), pp. 81-91. doi:

Archer, M., Bhaskar, R., Collier, A., Lawson, T. and Norrie, A. (2013) *Critical realism: Essential readings.* Routledge.

Argote, L. (2011) 'Organizational learning research: Past, present and future', *Management learning*, 42(4), pp. 439-446. doi:

Asghar, J. (2013) 'Critical Paradigm: A preamble for novice researchers', *Life Science Journal*, 10(4), pp. 3121-3127. doi:

Avison, D. E., Lau, F., Myers, M. D. and Nielsen, P. A. (1999) 'Action research', *Commun. ACM*, 42(1), pp. 94-97. doi: 10.1145/291469.291479

Barros, M. d. O., Werner, C. M. and Travassos, G. H. 'Project management knowledge reuse through scenario models'. *International Conference on Software Reuse*: Springer, 227-239.

Baskerville, R. L. and Wood-Harper, A. T. (1996) 'A critical perspective on action research as a method for information systems research', *Journal of Information Technology (Routledge, Ltd.)*, 11(3), pp. 235-246. doi: 10.1080/026839696345289

Bhaskar, R. (2009) *Scientific realism and human emancipation.* Routledge.

Bhattacharjee, K. and Petzold, L. 'What Drives Consumer Choices? Mining Aspects and Opinions on Large Scale Review Data Using Distributed Representation of Words'. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*: IEEE, 908-915.

Bird, S., Klein, E. and Loper, E. (2009) *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blei, D. M. and Lafferty, J. D. (2007) 'A correlated topic model of science', *The Annals of Applied Statistics*, 1(1), pp. 17-35. doi:

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent dirichlet allocation', *Journal of machine Learning research*, 3(Jan), pp. 993-1022. doi:

Bose, R. (2009) 'Advanced analytics: opportunities and challenges', *Industrial Management & Data Systems*, 109(2), pp. 155-172. doi:

Bramer, M. (2007) *Principles of data mining.* Springer.

Bresnen, M., Edelman, L., Newell, S., Scarbrough, H. and Swan, J. (2003) 'Social practices and the management of knowledge in project environments', *International Journal of Project Management*, 21(3), pp. 157-166. doi: http://dx.doi.org/10.1016/S0263-7863(02)00090-X

Brownlee, J. (2017) *How to Develop Word Embeddings in Python with Gensim*. Available at: https://machinelearningmastery.com/develop-word-embeddings-python-gensim/ (Accessed: 26 April 2020).

Buchanan, J. and Jones, M. L. (2010) 'The efficacy of utilising Nvivo for interview data from the electronic gaming industry in two jurisdictions'. doi:

Camilleri, E. (2010) 'Data Mining and the Project Management Environment', *Data Mining in Public and Private Sectors: Organizational and Government Applications*: IGI Global, pp. 337-357.

Carlsson, S. A. 'Critical realism: a way forward in IS research'. *ECIS*, 348-362.

Carr, W. and Kemmis, S. (2003) *Becoming critical: education knowledge and action research.* Routledge.

Carrillo, P., Harding, J. and Choudhary, A. (2011) 'Knowledge discovery from post-project reviews', *Construction Management and Economics*, 29(7), pp. 713-723. doi: 10.1080/01446193.2011.588953

Chen, W.-J., Jhou, M.-J., Lee, T.-S. and Lu, C.-J. (2021) 'Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association', *Entropy*, 23(4), pp. 477. doi:

Chen, W. and Hirschheim, R. (2004) 'A paradigmatic and methodological examination of information systems research from 1991 to 2001', *Information systems journal*, 14(3), pp. 197-235. doi:

Cheng, M. 'Research on the knowledge transfer in construction projects'. *2009 16th International Conference on Industrial Engineering and Engineering Management*: IEEE, 2035-2039.

Connelly, C. E., Ford, D. P., Gallupe, B., Turel, O. and Zweig, D. 'The effects of competition and time constraints on knowledge transfer: Exploratory findings from two experiments'. *2009 42nd Hawaii International Conference on System Sciences*: IEEE, 1-10.

da Silva Cezar, B. G. and Maçada, A. C. G. (2021) 'Data literacy and the cognitive challenges of a data-rich business environment: an analysis of perceived data overload, technostress and their relationship to individual performance', *Aslib Journal of Information Management*. doi:

Davenport, T. H. and Prusak, L. (2000) *Working Knowledge: How Organizations Manage What They Know.* Harvard Business Review Press.

Davis, S. and Bhattacharyya, A. (2016) 'Using data analytics to improve process plant operations', (no. 8), pp. 38.

de Figueiredo, A. D. and da Cunha, P. R. (2007) 'Action research and design in information systems', *Information systems action research*: Springer, pp. 61-96.

Delanty, G. (2011) 'Varieties of Critique in Sociological Theory and Their Methodological Implications for Social Research', *Irish Journal of Sociology*, 19(1), pp. 68-92. doi: 10.7227/IJS.19.1.4

Desouza, K. C. and Evaristo, J. R. (2004) 'Managing knowledge in distributed projects', *Communications of the ACM*, 47(4), pp. 87-91. doi:

Dixon, R. S. (2000) 'Internet videoconferencing: coming to your campus soon!', *Educause Quarterly*, 23(4), pp. 22-27. doi:

Dobson, P., Myles, J. and Jackson, P. (2007) 'Making the case for critical realism: Examining the implementation of automated performance management systems', *Information Resources Management Journal (IRMJ)*, 20(2), pp. 138-152. doi:

Donate, M. J. and Sánchez de Pablo, J. D. (2015) 'The role of knowledge-oriented leadership in knowledge management practices and innovation', *Journal of Business Research*, 68(2), pp. 360-370. doi: https://doi.org/10.1016/j.jbusres.2014.06.022

Dörre, J., Gerstl, P. and Seiffert, R. 'Text mining: finding nuggets in mountains of textual data'. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, 398-401.

Easton, G. (2010) 'Critical realism in case study research', *Industrial marketing management*, 39(1), pp. 118-128. doi:

Egghe, L. (2010) 'Good properties of similarity measures and their complementarity', *Journal of the American Society for Information Science and Technology*, 61(10), pp. 2151-2160. doi:

Elder-Vass, D. (2022) 'Pragmatism, critical realism and the study of value', *Journal of Critical Realism*, pp. 1-27. doi:

Emiliano de Souza, D., Favoretto, C. and Carvalho, M. M. (2021) 'Knowledge Management, Absorptive and Dynamic Capacities and Project Success: A Review and Framework', *Engineering Management Journal*, pp. 1-20. doi:

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases', *AI magazine*, 17(3), pp. 37. doi:

Feldman, R. and Dagan, I. 'Knowledge Discovery in Textual Databases (KDT)'. *KDD*, 112-117.

Fernie, S., Green, S. D., Weller, S. J. and Newcombe, R. (2003) 'Knowledge sharing: context, confusion and controversy', *International journal of project management,* 21(3), pp. 177-187. doi:
Frey, B. B. (2018) 'The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation'. doi: 10.4135/9781506326139
Goffin, K., Koners, U., Baxter, D. and Van der Hoven, C. (2010) 'Managing lessons learned and tacit knowledge in new product development', *Research-Technology Management,* 53(4), pp. 39-51. doi:
Gupta, D., Kumar, Y. and Sajan, N. (2012) 'Improving unsupervised stemming by using partial lemmatization coupled with data-based heuristics for Hindi', *International Journal of Computer Applications,* 38(8), pp. 1-8. doi:
Hadi, W. e. M., Thabtah, F. A. and Abdel-Jaber, H. 'A Comparative Study using Vector Space Model with K-Nearest Neighbor on Text Categorization Data'. *World Congress on Engineering*, 296-300.
Haldin-Herrgard, T. (2000) 'Difficulties in diffusion of tacit knowledge in organizations', *Journal of Intellectual capital,* 1(4), pp. 357-365. doi:
Hand, D. J. J. E. o. E. (2006) 'Data Mining', 2. doi:
Hanisch, B., Lindner, F., Mueller, A. and Wald, A. (2009) 'Knowledge management in project environments', *Journal of knowledge management,* 13(4), pp. 148-160. doi:
Hashimi, H., Hafez, A. and Mathkour, H. (2015) 'Selection criteria for text mining approaches', *Computers in Human Behavior,* 51, pp. 729-733. doi:
Heikkinen, H. L., Huttunen, R., Syrjälä, L. and Pesonen, J. (2012) 'Action research and narrative inquiry: five principles for validation revisited', *Educational action research,* 20(1), pp. 5-21. doi:
Ho, M. H.-C. and Liu, J. S. (2013) 'The motivations for knowledge transfer across borders: the diffusion of data envelopment analysis (DEA) methodology', *Scientometrics,* 94(1), pp. 397-421. doi:
Hotho, A., Nürnberger, A. and Paaß, G. 'A brief survey of text mining'. *Ldv Forum*, 19-62.
Hoxha, L. and McMahan, C. (2019) 'The Influence of Project Manager's Age on Project Success', *Journal of Engineering, Project, and Production Management,* 9, pp. 12-19. doi: 10.2478/jeppm-2019-0003
Ibrahim, A. (2003) 'Expertise location: can text mining help?', *WIT Transactions on Information and Communication Technologies,* 29. doi:
Institute, P. M. (2008) *A Guide to the Project Management Body of Knowledge: PMBOK Guide*. Project Management Inst.
Jain, A., Jain, A., Chauhan, N., Singh, V. and Thakur, N. (2017) 'Information retrieval using cosine and jaccard similarity measures in vector space model', *International Journal of Computer Applications,* 164(6), pp. 28-30. doi:
Jalalimanesh, A. 'Knowledge discovery in scientific databases using text mining and social network analysis'. *2012 IEEE Conference on Control, Systems & Industrial Informatics*, 23-26 Sept. 2012, 46-49.
Jugdev, K. (2012) 'Learning from lessons learned: Project management research program', *American Journal of Economics and Business Administration,* 4(1), pp. 13. doi:
Karaa, W. B. A. and Gribâa, N. 'Information Retrieval with Porter Stemmer: A New Version for English'. *Advances in Computational Science, Engineering and Information Technology*. Heidelberg: Springer International Publishing, 243-254.
Karimov, R., Samkova, M., Nikitina, S. and Akinin, A. (2017) 'Using a Hybrid Algorithm for Lemmatization of a Diachronic Corpus'. doi:
Kaur, A. and Chopra, D. 'Comparison of text mining tools'. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*: IEEE, 186-192.
Kelly, L. M. and Cordeiro, M. (2020) 'Three principles of pragmatism for research on organizational processes', *Methodological innovations,* 13(2), pp. 2059799120937242. doi:
Kerzner, H. (2017) *Project management metrics, KPIs, and dashboards: a guide to measuring and monitoring project performance.* John Wiley & Sons.
Khatri, C., Singh, G. and Parikh, N. (2018) 'Abstractive and extractive text summarization using document context vector and recurrent neural networks', *arXiv preprint arXiv:1807.08000*. doi:
Klösgen, W. (2021) 'Assistant for knowledge discovery in data', *Computers as Assistants*: CRC Press, pp. 96-111.
Kotnour, T. (2000) 'Organizational learning practices in the project management environment', *International Journal of Quality & Reliability Management,* 17(4/5), pp. 393-406. doi:
Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J. and Valencia, A. (2017) 'Information retrieval and text mining technologies for chemistry', *Chemical reviews,* 117(12), pp. 7673-7761. doi:
Landaeta, R. E. (2008) 'Evaluating benefits and challenges of knowledge transfer across projects', *Engineering Management Journal,* 20(1), pp. 29-38. doi:

Landauer, T. K., Foltz, P. W. and Laham, D. (1998) 'An introduction to latent semantic analysis', *Discourse processes*, 25(2-3), pp. 259-284. doi:
Larose, D. T. and Larose, C. D. (2014) *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.
Lee, S., Song, J. and Kim, Y. (2010) 'An empirical comparison of four text mining methods', *Journal of Computer Information Systems*, 51(1), pp. 1-10. doi:
Lierni, P. C. and Ribière, V. M. (2008) 'The relationship between improving the management of projects and the use of KM', *Vine*, 38(1), pp. 133-146. doi:
Loo, R. (2002) 'Working towards best practices in project management: a Canadian study', *International Journal of Project Management*, 20(2), pp. 93-98. doi:
Love, P. E. D., Huang, J., Edwards, D. J. and Irani, Z. (2005) 'Chapter 7 - Building a learning organization in a project-based environment', *Management of Knowledge in Project Environments*. Oxford: Butterworth-Heinemann, pp. 133-154.
M K, V. and K, K. (2016) 'A Survey on Similarity Measures in Text Mining', *Machine Learning and Applications: An International Journal*, 3, pp. 19-28. doi: 10.5121/mlaij.2016.3103
Madnani, N. (2007) 'Getting started on natural language processing with Python', *ACM Crossroads*, 13(4), pp. 5. doi:
Maimon, O. and Rokach, L. (2005) 'Data mining and knowledge discovery handbook'. doi:
Mardiana, T., Adji, T. B. and Hidayah, I. 'The Comparation of distance-based similarity measure to detection of plagiarism in Indonesian text'. *International Conference on Soft Computing, Intelligence Systems, and Information Technology*: Springer, 155-164.
Media, B. P. P. L. (2012) *CIM Professional Diploma: 4 Project management in marketing 2012*. BPP Learning Media.
Merten, T., Mager, B., Bürsner, S. and Paech, B. 'Classifying unstructured data into natural language text and technical information'. *MSR*, 300-303.
Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*. doi:
Mingers, J. (2000) 'The Contribution of Critical Realism as an Underpinning Philosophy for OR/MS and Systems', *The Journal of the Operational Research Society*, 51(11), pp. 1256-1270. doi: 10.2307/254211
Mingers, J. (2004) 'Real-izing information systems: critical realism as an underpinning philosophy for information systems', *Information and organization*, 14(2), pp. 87-103. doi:
Mingers, J., Mutch, A. and Willcocks, L. (2013) 'CRITICAL REALISM IN INFORMATION SYSTEMS RESEARCH', *MIS Quarterly*, 37(3), pp. 795-802. doi. Available at: http://nwulib.nwu.ac.za/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=89477777&site=eds-live
Mizgier, K. and Willis, N. (2014) 'Systems disparity: The implications of data proliferation on business decisions', *Ivey Business Journal*, (July/August 2014). doi. Available at: https://iveybusinessjournal.com/publication/systems-disparity-the-implications-of-data-proliferation-on-business-decisions/
Myers, M. D. (1997) 'Qualitative research in information systems', *Management Information Systems Quarterly*, 21(2), pp. 241-242. doi:
Myers, M. D. and Klein, H. K. (2011) 'A Set of Principles for Conducting Critical Research in Information Systems', *MIS Quarterly*, 35(1), pp. 17-36. doi. Available at: http://www.jstor.org/stable/23043487
Nayak, A. S., Kanive, A., Chandavekar, N. and Balasubramani, R. (2016) 'Survey on pre-processing techniques for text Mining', *International Journal Of Engineering And Computer Science*, *ISSN*, pp. 2319-7242. doi:
Newell, S., Bresnen, M., Edelman, L., Scarbrough, H. and Swan, J. (2006) 'Sharing knowledge across projects: limits to ICT-led project review practices', *Management learning*, 37(2), pp. 167-185. doi:
Ng, A. Y. and Jordan, M. I. 'On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes'. *Advances in neural information processing systems*, 841-848.
Nonaka, I. and Takeuchi, H. (2001) 'Organizational knowledge creation', *Creative Management*, *SAGE*, *London*, pp. 64-82. doi:
Omidipoor, M., Toomanian, A., Neysani Samany, N. and Mansourian, A. (2020) 'Knowledge discovery web service for spatial data infrastructures', *ISPRS International Journal of Geo-Information*, 10(1), pp. 12. doi:
Oneata, D. 'Probabilistic latent semantic analysis'. *Proceedings of the Fifteenth conference on Uncertainty*, 1-7.
Orlikowski, W. J. and Baroudi, J. J. (1991) 'Studying information technology in organizations: Research approaches and assumptions', *Information systems research*, 2(1), pp. 1-28. doi:

Park, A., Chang, H. and Lee, K. J. (2017) 'Action research on development and application of Internet of Things services in hospital', *Healthcare informatics research*, 23(1), pp. 25-34. doi:
Park, Y. S., Konge, L. and Artino, A. R. (2020) 'The positivism paradigm of research', *Academic Medicine*, 95(5), pp. 690-694. doi:
Patel, F. N. and Soni, N. R. (2012) 'Text mining: A Brief survey', *International Journal of Advanced Computer Research*, 2(4), pp. 243. doi:
Pather, S. and Remenyi, D. 'Some of the philosophical issues underpinning research in information systems: from positivism to critical realism'. *Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*: South African Institute for Computer Scientists and Information Technologists, 141-146.
Pather, S. and Remenyi, D. (2005) 'Some of the philosophical issues underpinning research in information systems - from positivism to critical realism : reviewed article', *South African Computer Journal*, 2005(35), pp. 76-83. doi. Available at: http://journals.co.za/content/comp/2005/35/EJC27994
Peltokorpi, V. and Tsuyuki, E. (2006) 'Knowledge governance in a Japanese project-based organization', *Knowledge Management Research & Practice*, 4(1), pp. 36-45. doi:
Pemsel, S., Wiewiora, A., Müller, R., Aubry, M. and Brown, K. (2014) 'A conceptualization of knowledge governance in project-based organizations', *International Journal of Project Management*, 32(8), pp. 1411-1422. doi:
Perkins, J. (2010) *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd.
Perkins, J. (2014) *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.
Phan, X.-H., Nguyen, L.-M. and Horiguchi, S. 'Learning to classify short and sparse text & web with hidden topics from large-scale data collections'. *Proceedings of the 17th international conference on World Wide Web*: ACM, 91-100.
Piatetsky-Shapiro, G. (1991) 'Knowledge discovery in real databases: A report on the IJCAI-89 Workshop', *AI Mag.*, 11(5), pp. 68-70. doi:
Polyaninova, T. (2011) 'Knowledge management in a project environment: organisational CT and project influences'. doi:
Pretorius, C. and Steyn, H. (2005) 'Knowledge management in project environments', *South African Journal of Business Management*, 36(3), pp. 41-50. doi:
Provost, F. and Fawcett, T. (2013) 'DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING', *MARY ANN LIEBERT, INC*, 1(1), pp. 9. doi:
Prusak, L. and Davenport, T. (1998) 'Working knowledge: how organizations manage what they know'. doi:
QSR-International (2018) *Coding*. Available at: https://help-nv.qsrinternational.com/12/win/v12.1.90-d3ea61/Content/coding/coding.htm (Accessed: November 25 2019).
Rajpathak, T. and Narsingpurkar, A., Manufacturing *Managing knowledge from Big Data analytics in product Development.*: Tata Consultancy Services.
Rao, G. K. and Dey, S. (2011) 'Decision support for e-governance: a text mining approach', *arXiv preprint arXiv:1108.6198*. doi:
Ren, X., Deng, X. and Liang, L. (2018) 'Knowledge transfer between projects within project-based organizations: the project nature perspective', *Journal of Knowledge Management*, 22(5), pp. 1082-1103. doi:
Rose, A.-L., Dee, J. and Leisyte, L. (2020) 'Organizational learning through projects: a case of a German university', *The Learning Organization*. doi:
Roul, R. K., Joshi, P. M. and Sahoo, J. K. 'Abstractive Text Summarization Using Enhanced Attention Model'. *International Conference on Intelligent Human Computer Interaction*: Springer, 63-76.
Rowe, S. F. and Sikes, S. 'Lessons learned: taking it to the next level. '. *PMI® Global Congress 2006*, North America, Seattle, WA. Newtown Square, PA: Project Management Institute.
Rubin, S. 2012. Knowledge discovery and dissemination of text by mining with words. Google Patents.
Sarkar, T. (2020) 'A simple intro to Regex with Python'. Available at: https://towardsdatascience.com/a-simple-intro-to-regex-with-python-14d23a34d170.
Saunders, M., Lewis, P. and Thornhill, A. (2007) 'Research methods', *Business Students 4th edition Pearson Education Limited, England*. doi:
Sayer, A. (1999) *Realism and social science*. Sage.
Scarbrough, H., Swan, J., Laurent, S., Bresnen, M., Edelman, L. and Newell, S. (2004) 'Project-based learning and the role of learning boundaries', *Organization studies*, 25(9), pp. 1579-1600. doi:
Scheidler, A. A. and Rabe, M. (2021) 'Integral verification and validation for knowledge discovery procedure models', *International Journal of Business Intelligence and Data Mining*, 18(1), pp. 73-87. doi:
Schröer, C., Kruse, F. and Gómez, J. M. (2021) 'A systematic literature review on applying CRISP-DM process model', *Procedia Computer Science*, 181, pp. 526-534. doi:

Scotland, J. (2012) 'Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms', *English Language Teaching*, 5(9), pp. 9. doi:
Serrat, O. (2012) 'Managing Knowledge in Project Environments'. doi:
Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. J. D. s. s. (2001) 'Knowledge management and data mining for marketing', 31(1), pp. 127-137. doi:
Sieg, A. (2018) *Text Similarities : Estimate the degree of similarity between two texts*. Available at: https://medium.com/@adriensieg/text-similarities-da019229c894 (Accessed: February 2020).
Silwattananusarn, T. and Tuamsuk, K. J. a. p. a. (2012) 'Data mining and its applications for knowledge management: a literature review from 2007 to 2012'. doi:
Sokhanvar, S., Matthews, J. and Yarlagadda, P. (2014) 'Importance of Knowledge Management Processes in a Project-based organization: A Case Study of Research Enterprise', *Procedia Engineering*, 97, pp. 1825-1830. doi: http://dx.doi.org/10.1016/j.proeng.2014.12.336
Susman, G. (1983) 'Action Research: A Sociotechnical Systems Perspective', *Beyond Method: Strategies for Social Research*. Newbury Park: Sage, pp. 95-113.
Tan, A.-H. 'Text mining: The state of the art and the challenges'. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 65-70.
Thada, V. and Jaglan, V. (2013) 'Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm', *International Journal of Innovations in Engineering and Technology*, 2(4), pp. 202-205. doi:
Themistocleous, M. and Morabito, V. (2017) *Information Systems: 14th European, Mediterranean, and Middle Eastern Conference, EMCIS 2017, Coimbra, Portugal, September 7-8, 2017, Proceedings.* Springer International Publishing.
Todorović, M. L., Petrović, D. Č., Mihić, M. M., Obradović, V. L. and Bushuyev, S. D. (2015) 'Project success analysis framework: A knowledge-based approach in project management', *International Journal of Project Management*, 33, pp. 772-783. doi: 10.1016/j.ijproman.2014.10.009
Treviso, M. V., Shulby, C. D. and Aluisio, S. M. (2017) 'Evaluating word embeddings for sentence boundary detection in speech transcripts', *arXiv preprint arXiv:1708.04704*. doi:
Turner, J. R. and Müller, R. (2003) 'On the nature of the project as a temporary organization', *International journal of project management*, 21(1), pp. 1-8. doi:
Ur-Rahman, N. (2010) *Textual Data Mining Applications for Industrial Knowledge Management Solutions.* © Nadeem Ur-Rahman.
Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. and Takeda, K. (2004) 'A text-mining system for knowledge discovery from biomedical documents', *IBM Systems Journal*, 43(3), pp. 516-533. doi: 10.1147/sj.433.0516
Vel, S. S. 'Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries'. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*: IEEE, 879-884.
Venugopal, V., Sahoo, S., Zaki, M., Agarwal, M., Gosvami, N. N. and Krishnan, N. A. (2021) 'Looking through glass: Knowledge discovery from materials science literature using natural language processing', *Patterns*, 2(7), pp. 100290. doi:
Vijayarani, S., Ilamathi, M. J. and Nithya, M. (2015) 'Preprocessing techniques for text mining-an overview', *International Journal of Computer Science & Communication Networks*, 5(1), pp. 7-16. doi:
Vijayarani, S. and Janani, R. (2016) 'Text mining: open source tokenization tools-an analysis', *Advanced Computational Intelligence: An International Journal (ACII)*, 3(1), pp. 37-47. doi:
Vu, B., Mertens, J., Gaisbachgrabner, K., Fuchs, M. and Hemmje, M. 'Supporting taxonomy management and evolution in a web-based knowledge management system'. *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, 1-11.
Wang, L. L. and Lo, K. (2021) 'Text mining approaches for dealing with the rapidly expanding literature on COVID-19', *Briefings in Bioinformatics*, 22(2), pp. 781-799. doi:
Wang, S. and Wang, H. (2020) 'Big data for small and medium-sized enterprises (SME): a knowledge management model', *Journal of Knowledge Management*. doi:
Wiewiora, A., Trigunarsyah, B., Murphy, G. D. and Liang, C. 'Barriers to effective knowledge transfer in project-based organisations'. *Proceedings of the 2009 International Conference on global innovation in construction proceedings*: Loughborough University UK, 220-230.
Wijetunge, P. (2012) 'Organizational storytelling as a method of tacit-knowledge transfer: Case study from a Sri Lankan university', *The International Information & Library Review*, 44(4), pp. 212-223. doi:
Wikgren, M. (2005) 'Critical realism as a philosophy and social theory in information science?', *Journal of documentation*, 61(1), pp. 11-22. doi:

Wilson, M. and Greenhill, A. (2004) 'Theory and action for emancipation: Elements of a critical realist approach', *Information Systems Research*: Springer, pp. 667-674.
Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016) *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.
Wong, L. (2008) 'Data analysis in qualitative research: A brief guide to using NVivo', *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia,* 3(1), pp. 14. doi:
Wynn Jr, D. and Williams, C. K. (2012a) 'Principles for conducting critical realist case study research in information systems', *Mis Quarterly,* 36(3), pp. 787-810. doi:
Wynn Jr, D. and Williams, C. K. (2012b) 'Principles for conducting critical realist case study research in information systems', *MIS quarterly*, pp. 787-810. doi:
Yin, W. and Pei, Y. 'Automatic multi-document summarization based on new sentence similarity measures'. *Pacific Rim International Conference on Artificial Intelligence*: Springer, 832-837.
Zhao, D., Zuo, M. and Deng, X. N. (2015) 'Examining the factors influencing cross-project knowledge transfer: An empirical study of IT services firms in China', *International Journal of Project Management,* 33(2), pp. 325-340. doi:

APPENDICES

Appendix A: Ethical Clearance Certificate

FNASREC-NWU-013
14-21-A9-2021-2772

Appendix B: Interview guide

INTERVIEW QUESTIONS

1. What is your role in the organisation and this department?_____

   _____

2. What is the core function of the project department?_____

   _____

3. What kind of information is collected by the projects department and what is it used for?_____

   _____

4. What is the goal of gathering/capturing information at the various stages of the PM Lifecycle?_____

   _____

5. What is considered knowledge in the department?_____

   _____

6. To what extent do you use lessons learnt from previous projects?

   _____

   _____

7. How do you currently extract and analyse the available data?

   _____

   _____

8. How is the knowledge discovered or identified from the analysis utilized?

   _____

   _____

9. Are there any challenges you are currently facing with data analysis?

   _____

   _____

10. How would you measure an improvement in the data analysis process?

    _____

    _____

Appendix C: Participant information sheet or consent form

Dear Participant

This letter is an invitation to consider participating in a study I am conducting as part of my Doctoral degree in the Department of Information Technology at the North West University under the supervision of Professor Philip Pretorius, Dr Carin Venter, and Dr Daan De Villiers**.** I would like to provide you with more information about this project and what your involvement would entail if you decided to take part.

This research is focused on knowledge discovery from databases in the projects environment. Throughout the project lifecycle of each project, a significant amount of data is collected concerning the project. Retrieval of relevant information from such data and identifying all the useful information may be time consuming and requires careful selection of keywords and drafting of queries. This research aims to determine the value that knowledge discovery from text may offer to emancipate project managers.

Participation in this study is voluntary. It will involve an interview of approximately 20 minutes in length to take place in a mutually agreed upon location. You may decline to answer any of the interview questions if you so wish. Further, you may decide to withdraw from this study at any time without any negative consequences by advising the researcher. With your permission, the interview will be tape-recorded to facilitate collection of information, and later transcribed for analysis. Shortly after the interview has been completed, I will send you a copy of the transcript to give you an opportunity to confirm the accuracy of our conversation and to add or clarify any points that you wish.

All information you provide is considered completely confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission anonymous quotations may be used. Data collected during this study will be retained for a period of approximately 5 years in an encrypted folder. Only researchers associated with this project will have access. There are no known or anticipated risks to you as a participant in this study. Should you require any further information feel free to email the researcher on the email below:

nyangceet@yahoo.com

Kind regards,

Cleopatra Tsungai Mushonga **(student no. 27724255)**

Appendix D: Source code for text mining software

Below is the source code that was used for text mining in this study. The code was developed using Python version 2.0 but can be compatible with other versions of Python.

Appendix E: Lessons Learnt Summarization Spreadsheet

LIST OF LESSONS LEARNED FILES

S.No	Files	Status
1	Copy of 2016-09-23 - DMR - Project Lessons Learnt - Dhashan Naidoo.xls .xlsx	Summarised
2	Demin Lessons learned.xlsx	Summarised
3	Copy of SNI Lessons Learnt Session 2 29 Jul 2011 - Final - 2011-07-29.xlsx	Summarised
4	VOCA Lessons Learnt - U13- Workshop 1 - 29 Mar 2017 - Rev 1.xlsx	Summarised
5	FTWEP I - Lessons Learnt form.xlsx	Summarised
6	Copy of Lessons learned.xlsx	Summarised
7	20160623 CTFE Project - Broad Lessons Learnt.xlsx	Summarised
8	Copy of GNP Lessons Learnt Session 10 April 13 - Combined Comments - 2013-04-10.xlsx	Summarised
9	Copy of SCF2 Lessons Learnt Uhde Contract - 10 Apr 2014 - Rev 0.xlsx	Summarised
10	Copy of Internal SCF2 Core Scope Lessons Learnt Consolidated - 16 May 2014 - Session 1-3 - Rev 1.xlsx	Summarised
11	Copy of SILOG Post Mortem Template Combined View 30 Nov 2010 (3).xls	Summarised
12	21 tie-in opportunity lessons learned.pdf	Summarised
13	FWSA Demolition post-morten report.pdf	Summarised
14	SasTech EPU5 Project Lessons Learned Report.pdf	Summarised

Appendix F: Turnitin Report

27724255_TII_5_Final_Submission_Thesis_

27724255:TII_5_Final_Submission_Thesis_updated__articles_-Cleopatra_Mushonga_revised_(3).docx

ORIGINALITY REPORT

7%	9%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ndl.ethernet.edu.et Internet Source	3%
2	sajbm.org Internet Source	2%
3	repository.nwu.ac.za Internet Source	1%
4	www.tandfonline.com Internet Source	1%
5	uir.unisa.ac.za Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

Appendix G: Language Editing Certificate

Mushonga Language editing confirmation.p

Maretha Botes
Independent Journalist and Language Practitioner
13 Brahms Street, Vanderbijlpark, 1911
Cell 083 401 7492
E-mail marethab@gmail.com

14-03-2022

EDITING CERTIFICATE

This certificate serves to confirm that the dissertation **The use of Text Mining to improve knowledge discovery in the project environment** by Cleopatra Tsungai Mushonga has undergone a professional language edit (including the checking of spelling, grammar, register and punctuation). The onus rests on the client to work through the proposed changes after the edit and accept or reject these changes.

Yours faithfully

Maretha Botes
Freelance journalist and language practitioner