# TEXT MINING MODEL FOR RETRIEVAL OF EXPLICIT KNOWLEDGE AT KENYA COASTAL DEVELOPMENT PROJECT, MOMBASA

**EDNAH NYAKERARIO ONKUNDI**

**ATHESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MASTERS DEGREE IN INFORMATION TECHNOLOGY OF MOUNT KENYA UNIVERSITY**



**SEPTEMBER 2022**

**DECLARATION AND APPROVAL**

**Declaration:**

This research study is my original work and has not been presented for degree in any other University or for any other award.

Ednah Nyakerario Onkundi

MIT/2014/73030

Signature _____ Date 22/08/2022

**Approval:**

We confirm that the work reported in this research study was carried out by the candidate under our supervision.

Signature _____ Date 5/09/2022

Prof. Raymond Wafula Ongus, PhD

School of Computing and Informatics

Mount Kenya University

Signature _____ Date 6-09-2022

Prof. Constantine Nyamboga, PhD

School of Computing and informatics

Mount Kenya University

## DEDICATION

To my family: I dedicate this study to my dear children Caleb, Joy, Emma, Mercy and Millie, husband Joel Onduso and parents Elijah and Dorcas, for their encouragement and support that has brought me this far.



## ACKNOWLEDGMENT

First, I thank the Almighty God for the life and good health he has bestowed upon me, through his mercies I have found encouragement to move on.

I am very grateful to my Supervisors Prof. Raymond Wafula Ongus and Prof Constantine Nyamboga who guided me through the writing of this research thesis. I am grateful to Prof. Ongus who spent his time without getting tired even when I would repeatedly make similar mistakes in the journey of research. Through his constant encouragement and guidance in my academic effort, it positively changed my way of thinking. I also thank Prof. Nyamboga for his constant encouragement and corrections in the research. I also thank Dr.John Kamau for the final corrections that enabled me to finalize the thesis. Much appreciation and special thanks goes to my family and friends for the emotional support that kept me going throughout my studies. Finally, I thank the management of KMFRI and KCDP that gave me the opportunity and financial support to pursue my Masters programme.



## ABSTRACT

The study investigated the prospects of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). The study's main objective was to establish how a Text Mining model could be used in explicit knowledge retrieval at KCDP. The study identified text-mining techniques that could be used to develop

a text-mining model, evaluate the model to be able to retrieve explicit knowledge at KCDP. The study targeted staff of the agencies that constituted the KCDP project which included, Kenya Marine and Fisheries Research Institute (KMFRI), Kenya Wildlife Service (KWS), State Department of Fisheries (SDF), Coastal Development Project (CDA), Department of Physical Planning, Kenya Forest Service (KFS) and National Environment Management Authority (NEMA). The study used the exploratory and experimental research design to be able to understand the research problem, answer the research objectives and questions. The total population of staff in the project was one hundred and fifty (150), out of which fiftytwo (52) were sampled. Purposive sampling was used to select samples from the representative groups that comprised the target population. Two methods of data collection were used namely; questionnaires and focus group discussion. The questionnaire was applied to members of staff in four major departments namely the top management, research and administration, knowledge management and finally the ICT department. The focus group discussion was applied to a special group in the knowledge management section. Content analysis was used to analyze the focus group discussions. Questionnaires were analyzed using the Statistical Package for the Social Sciences (SPSS) version 25 software. The use of questionnaires and focus groups were used to establish the current situation at the KCDP in terms of knowledge management systems in place and whether text mining could be used to retrieve explicit knowledge at KCDP. Text were collected from websites of organizations that took part in the KCDP project by using python libraries namely Python Request 2.22 and Beautiful Soup 3. The collected text was then summarized using text summarization algorithms used in the model like Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer. After summarization topic, modelling was performed on the text collected using Latent Dirichlet Allocation (LDA) topic-modelling algorithm to create topics based on patterns in text. The model was then evaluated to establish its performance by measuring the four variables identified using precision and recall to measure accuracy, topic modelling to measure rate of similarity, and perplexity to measure evaluation of the model which gave a perplexity of -6.0455 from the text analyzed and modelled. It was concluded that text analysis could be used to analyze text and create explicit knowledge from both structured and unstructured data formats using the model. Future models should incorporate artificial intelligence into machine learning, so that semantics (i.e., English grammar) are deciphered and not only syntax of the language. The system should be willing to differentiate between “willing flesh” and “good meat”. The system should detect the intrinsic difference between the phrases “weak spirit” and “bad liquor”. This will help the system to avoid getting lost in translation via the use of synonyms and will incrementally rely on semantic, as facilitated by artificial intelligence.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **BI** | Business Intelligence |
| **CBM** | Concept Based Method |
| **DM** | Data Mining |
| **ERP** | Enterprise Resource Planning |
| **HTML** | Hypertext Markup Language |
| **IBM** | International Business Machine |
| **IF** | Intermediate Form |

| | |
|---|---|
| **IR** | Information Retrieval |
| **KCDP** | Kenya Coastal Development Project |
| **KM** | Knowledge Management |
| **KMFRI** | Kenya Marine and Fisheries Research Institute |
| **NLP** | Natural Language Processing |
| **NWKNN** | Neighbor Weighted K-Nearest Neighbor |
| **PBM** | Phrase Based Method |
| **PTM** | Pattern Taxonomy Method |
| **SNA** | Social Network Analysis |
| **SVD** | Singular Value Decomposition |
| **TBM** | Term Based Method |
| **TDM** | Textual Data Mining |
| **TM** | Text Mining |

## OPERATIONAL DEFINITIONS OF KEY TERMS

**County Government:** A second level of administration as established under the current constitution that was promulgated in 2010. It is governed by the Executive head who is the Governor and County Assembly that makes laws. The study focused on the coastal counties in Kenya.

**Data Mining:** It is the study of collection, cleaning, processing and analysis of data at the KCDP to gain useful insights from the data at hand in the research study.

**Explicit Knowledge:** It is knowledge that can be accessed, verbalized, codified, and easily articulated at KCDP. It can be transmitted to others in KCDP for mentorship and succession planning.

**Indigenous Coastal Communities:** These are original settlers at the coastal Counties and they comprise of; Taita, Pokomo, Mijikenda, Holma and the Swahili. Communities. They hold 10% of the Kenyan population as per the 1999 national census.

**Kenyan Coast:** The Kenyan coast covers a 10 miles strip along the coastal region where the study was undertaken. It is the gateway to Kenya and the East African region.

**Knowledge Discovery:** Knowledge discovery involves getting information from non-retrieval or interested extraction of pattern from text, which is not structured to get information for advancement of knowledge to staff at the KCDP.

**Knowledge Retrieval:** This process consistently retrieves information in a structured way where alignment with humans is cognitive. The retrieval of Information creates knowledge and enables sharing among staff and Stakeholders at KCDP.

**Tacit Knowledge:** This is knowledge that is experience based and transferring it to another person is not easy. It can be transformed to explicit form by writing it down, verbalizing or sharing and mentorship within staff of KCDP.

**Text Mining:** Process of discovering and analyzing large amounts of unstructured text, using software and tools, which identify trends, patterns, concepts, topics, keywords among many other attributes in a database at KCDP.

# CHAPTER ONE: INTRODUCTION

## 1.0 Introduction

In this study an approach for retrieval of explicit knowledge was used to analyze text mined at the Kenya Coastal Development Project (KCDP), with a view to developing a text mining model for retrieving explicit knowledge. The areas that were covered include the following; Background of the study, Problem statement, Objectives of the study, Research Questions, and Significance of the Study.

## 1.1 Background of the study

Increase in analysis of text that is available has gone high in recent years. This is because of growth in social media, blogging and use of bulletin board systems. In addition to these, many documents, feeds from news and articles are now stored in soft copy. Classification of text documents involves an important step of classifying categories and classes of known text documents in a corpus (Shah *et al.*, 2017).

Text Mining is when knowledge is extracted from appealing and significant patterns of text documents. Text Mining is a progressive technology that promotes the understanding of clients by their enterprise. This technology assists enterprises in redefining their customer needs. Text mining techniques have grown because of the presence and demand of data in massive volumes in the web in the form of text. According to Tan (1999), text data mining has experienced challenges, which are relevant and relate to the mining of explicit knowledge. To gain great competitive advantage in today's global economy, an organization has to make effective use of its knowledge and knowledge assets. The most important resource in an organization and one that can show the strategic direction an organization takes is its knowledge base, which determines its decision making process

based on data and information available. The democratic access of knowledge by the entire organization provides for opportunities of innovation and hence giving it a competitive edge against its competitors (Jandhyala, & Phene, 2015).

An organization can easily asses the knowledge assets that can contribute to the realization of its immediate project objectives. The discovery of the knowledge assets lead to idealization of the knowledge requirements in the coastal communities of Kenya.

Text Mining is key in discovering knowledge from large volumes of data. Systems face tough competition from other systems in line with their value for financial and business growth. This is experienced in corporate sectors. Text Mining applies the principles of Data Mining (DM), which is growing and expanding at an alarming rate in relation to new technologies like big and open data.

Patterns that are difficult to discover manually can be discovered through the use of relevant trends and easily show trends with relations found within data through the application of data analysis techniques. It was noted that most of the knowledge that was at Kenya Coastal Development Project (KCDP) was not shared with staff, members and the business community at large (KMFRI, 2012).

The aim of this research study was to apply Text Mining techniques like classification, clustering and summarization to support the Kenya Coastal Development Project (KCDP) activities for explicit Knowledge Discovery, retrieval and the knowledge management process as a whole.

## 1.2 Problem Statement

This study addressed use of Text Mining for explicit Knowledge Discovery. The celebrated knowledge explosion assures increased availability of intelligence. This has been seen through the exponential expansion of scientific knowledge – producing professions and industries of research and development units (Wilensky, 2015). Advantages of automated information extraction and adoption are well understood but organizations have some significant challenges to address in the future with diverse and complex unstructured documents (Adnan & Akbar, 2019). Existing literature on big data and unstructured data does not present any model or framework to improve the information extraction (IE) procedures. There is no proper structured approach that can develop methods for specific data types, all current techniques and models are only presenting general models and techniques (Baviskar *et al.*, 2021).

Dissemination of research findings and recommendation to stakeholders and communities living at the coast of Kenya is the core mandate of KCDP. This is valued information that is held with individual scientists and not shared. If this information is centrally stored and made available, it could answer and meet some of the key objectives of KCDP that includes improvement of entrepreneurial skills and enhancement of revenue generation. The discovery and storage of large amounts of data, information and knowledge at KCDP, which are scattered and mostly kept by individual scientists and researchers, can be made useful by applying best practices in collecting, storing, archiving and sharing knowledge. These would transform tacit knowledge to explicit knowledge, which is embedded in daily activities of KCDP staff experiences, feelings and personal abilities. This knowledge is internalized in individuals through experiences and reflections; hence, an organization must invest in knowledge management systems

(Joia & Lemos, 2010). Knowledge management systems reduce the loss of intellectual capital from people leaving the institution, saves finances by not reinventing the wheel for new upcoming projects and provides faster problem solving approaches, which leads to timely decision-making processes.

The KCDP has applied the traditional methods of knowledge retrieval that only allowed retrieval of knowledge from structured databases. The current systems in KCDP have the following gaps that are hindering text retrieval and they include; inability to search and retrieve voluminous text and information, it does not have a specific document selection that is key based retrieval making definition of queries as a set of requisites a hindrance, inability for current systems in KCDP to rank documents to retrieve similarity based texts and inability to give options like Boolean retrieval or vector space retrieval. KCDP had invested in social media communication and sharing of information through the unstructured formats with a number of weaknesses, this made retrieval from this unstructured format impossible to mine though it contained the largest amount of data that touched on the objectives of the project. Data and information for the projects implemented at KCDP are both in hard and soft copies. The hard copies are retrieved manually from various files while the soft copies are in different formats that are not easily retrievable and kept in different storage sites that are accessed by authorized users (KMFRI, 2012). The KCDP was chosen because of its uniqueness involving the government of Kenya agencies and international organizations. This made the study to choose the area for its research. The dynamics of combining expertise from local and international organization to develop and improve the welfare of the local community at the coast made the area very attractive and exciting to learn and explore. Kenya Coastal

Development Project had not made use of explicit knowledge. The data storage, retrieval and sharing used legacy methods that made access to information ineffective and retrieval of knowledge at KCDP is kept by different parties.

The mining of explicit knowledge from both structured and unstructured formats is required in the project and its participating agencies for determination of its impact to the community. This study analyzed the use of Text Mining techniques for retrieval of explicit Knowledge at the Kenya Coastal Development Project in Mombasa, with a view to improve on retrieval of explicit knowledge from both structured and unstructured systems.

## 1.3 General Objective

The general objective of this study was to investigate the prospects of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP).

## 1.4 Specific Objectives

To achieve the main objective, the research study was guided by the following specific objectives;

i) To analyze how Text Mining techniques could help in retrieving explicit knowledge at KCDP.

ii) To design a Text Mining model for retrieval of explicit knowledge at KCDP.

iii) To validate the developed Text Mining model for retrieval of explicit knowledge at KCDP.

## 1.5 Research Questions

To achieve the objectives of this study, the research questions below had to be answered;

i) How could text-mining techniques help in the retrieval of explicit knowledge at KCDP?

ii) What Text Mining model could be designed for retrieval of explicit knowledge at KCDP?

iii) How would the designed Text Mining Model be validated for retrieval of explicit knowledge at KCDP?

## 1.6 Justification of the Research Study

The main purpose of the study was to investigate the prospects of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). The Information back up in most companies is unstructured. Due to this reason, it is imperative to retrieve and extract this information.

Efficiency and storage comes in handy to meet the requirements for mining this unstructured data that forms 80% of the most useful information in driving business in the world today (Gharehchopogh & Khalifelu, 2011).

The study investigated the prospectus of retrieving explicit knowledge in KCDP by mining data from all the project agencies that participated in the project. The procedure for obtaining the text-mining algorithms to use involved the use of RapidMiner Studio 9.5 tools that contained different extensions, operators and Application Programming Interfaces (API's). The mostly used RapidMiner Studio 9.5 extensions for this study were the Aylien Text Analysis extension, which provided operators like language detection, summarization, categorization and sentiment analysis.

The development of the model was dependent on the availability of data. Data was collected or mined using web scraping as a technique to get data in form of text from HTML and XML files. The procedure involved the installation of web scraping python libraries that included Python Request 2.22 and Beautiful Soup 3. Data inform of text was web scraped from websites and portals of organizations that took part in the Kenya Coastal Development Project.

Topic modelling was performed on text obtained from the kdcp.txt file. The file contained text that had been analyzed. After analysis text were grouped into words and keywords, where topic modeling was performed based on the number of times a word reappeared. Topics were modelled and visualized. The study proved that knowledge could be obtained from mining text from both structured and unstructured formats to help an organization like KCDP make better decisions that can lead to achievement of their mission and vision.

## 1.7 Significance of the study

The designed model shows the techniques of retrieving explicit knowledge at KCDP and this was expected to mostly assist managers, researchers and stakeholders in all agencies. The study could also determine the retrieval techniques that can mine text from both structured and unstructured data formats. The model that shall use text summarization could also reduce the length of a document by only selecting the most important points and assist researchers in being summarized to meaningful knowledge on the area of expertise. The model could also boost competitive intelligence in their areas of expertise and could lead to a competitive advantage. The application of the model could also identify trends and key areas in the skilled areas of Marine policy and decision-making. The managers on the other hand shall gain immensely in identifying highly productive

members of staff and enumerate them well to reduce high turnover rate of hard working members. The model could also assist managers in identifying challenges in the retrieval of explicit knowledge and employ better ways of overcoming them. The technical team would also improve their skills in designing and development of related models. The approach and use of the Text Mining techniques at KCDP would achieve high performance results and unleash hidden potential in relationships in her scattered data that would finally lead to creation and retention of knowledge. KCDP would finally have its rich asset knowledge base for future reference, where data and information would be reliable, available and provide accountability of data when retrieved.

## 1.8 Scope of the study

### 1.8.1 Geographical Scope

This study focused on the Kenya Coastal Development Project (KCDP), which is located along the Kenyan coast at 4.0552° S, 39.6821° E, according to Google Satellite Locator. The coordinates point to cement road in Mombasa city (Google satellite 2019). KCDP is a World Bank sponsored development project implemented by seven government agencies at the Coast. The location was ideal for hosting all data and knowledge documents that was required during the study. Focusing on the KCDP as an entity made the research viable when resources of time, money and availability of data were put into consideration.

### 1.8.2 Content Scope

The study focused on improving the retrieval of explicit knowledge at KCDP using text-mining techniques like information extraction, categorization, classification and text summarization.

### 1.8.3 Time Scope

This research took a duration of three months as stipulated by the School of Postgraduate studies of Mount Kenya University. The data collection covered the period between 2011 and 2018. This is because the project was initiated in the year 2011 and accumulated documents and data run up to 2018.

## 1.9 Study Limitations

This study majorly focused on KCDP project, people and documents. There was a challenge in accessing all respondents due to the natture of their roles and responsibilities in the organization. This was addresed by use of early booking and regular reminders. Due to the nature of data that was required some documents were not easy to access but confidentiality was assured in handling the documentation.

## 1.10 Summary

This chapter starts with the introduction that describes the research study and then gives background information about the study. The chapter also captures the problem statement of the study. This chapter also contains research objectives, which include the general objective and specific objectives of the study. It further outlines research questions that the research study should answer to achieve the objectives of the study and further outlines the justification of the study. The chapter is preceded with three more chapters,

chapter two that covers Literature Review, Chapter three that covers Research Methodology and Chapter four that covers research findings, analysis and presentation.

The last Chapter is on Summary, conclusions and recommendations.

# CHAPTER TWO: LITERATURE REVIEW

## 2.0 Introduction

This chapter covers the Theoretical framework of Text Mining, critical review, research gaps, the theoretical framework, Text Mining techniques, clustering techniques, classification techniques, application of Text Mining in explicit knowledge management. This research also showed the connection between the research studies compared to the existing knowledge, previous studies with relevant citations of other scholars with respect to the research study. The literature reviews also provided the study with the foundation of the research that brought out previous research on what other researchers had done on the topic, enquiries and area of interest in the area of study.

## 2.1 Kenya Coastal Development Project (KCDP)

The KCDP is a sponsored project by the World Bank and Global Environmental Facility (GEF), implemented by seven government agencies at the Coast (KCDP Project report, November 2012). It was incepted in November 11 2011. The project runs up to December 2019 with the possibility of renewal after the World Bank audit on the level of achievements for her targets.

The KMFRI hosts the project with the aim of promoting sustainable environment management of Kenya's coastal and marine resources. It focuses on strengthening government agencies, capacity enhancement of coastal communities in rural micro, small and medium-sized enterprises. The project has various types of information that comprise of fisheries information, which is collected through research on marine and aquatic life.

Secondly, the project has physical oceanographic information that comprises of Nitrate, Phosphate, Silicate and Oxygen (KCDP, 2012).

Thirdly, hydrosphere data, which comprises of all water bodies with attributes like temperature, salinity and sea surface temperature. Fourthly, atmosphere data that is a blanket of air that surrounds the earth has air temperature, sea level pressure, precipitation, cloudiness and humidity. Fifth is information on the biosphere that includes all the parts of the earth's surface, where living processes occur with the following data sets; phytoplankton, chlorophyll, zooplankton and vegetation. Staff with special skills extract, process all this information, and store it in different manual files and databases. All these information contributes to the overall project success and influence to the community, where by the information needed has to be retrieved and put in a uniform model for future use and reference. The total amount of data is estimated to be over 20 terabytes and is stored in numerous portals and databases (KCDP, 2012).

The project was designed with the aim of protecting natural resource base by developing a knowledge base. Currently the different knowledge and skills are kept in different formats that include; paper documents, soft copies individual laptops and desktops, in office report files and in various portals and websites. Program coordinators in the various programs of the project keep the various knowledge materials. The users of this knowledge are mainly government agencies, fisher communities, fish farmers and other experts in the marine fields (KCDP, 2012).

Retrieval of this knowledge in KCDP is a challenge given the fact that this knowledge is kept by different parties. The information can only be retrieved by identifying and tracing the coordinators who are the soul storage of the project information. If the coordinators

are not in a position to trace the knowledge, then it means that information is lost. If for example a laptop crashes the information is no more (KCDP, 2012).

ICT Managers working in the government agencies at KCDP need to integrate Information Systems with Business Strategies, to attain the objectives of the project. In this era of data and information, knowledge is an essential resource to an organization, to attain competitive advantage that could create knowledge management initiatives.

Many organizations have collected and stored vast amounts of data. However, they are unable to discover valuable information hidden in the data, to transform the data into valuable and useful knowledge (Mahamune & Ingle, 2014). Knowledge is an expensive commodity, which if managed properly is a major asset to the company. Knowledge is a complex and fluid concept that can be in two different forms, explicit or tacit in nature. It is for these reasons stated above that the study proposed a Text Mining approach for the retrieval of explicit knowledge.

## 2.2 Knowledge Types

Knowledge as a concept has been grouped into two, explicit knowledge and tacit knowledge. According to Nonaka (1994), explicit knowledge, is knowledge that can be easily expressed, gathered and accessed. This type of knowledge is easily transmitted. A higher percentage of the tacit knowledge can be stored in various forms of media. Examples of explicit knowledge is what is contained in encyclopedias and textbooks. It is sometimes referred to as know what and can be formalized and codified (Brown & Duguid, 1998). Explicit knowledge is easy to identify, store and retrieve because it is readily available. Knowledge Management Systems (KMS) handle explicit knowledge by facilitating the storage, retrieval and modification of documents.

Explicit knowledge possesses a great challenge to organizations in ensuring that people have access to information they need in the right format using the best medium. Organizations have to ensure that information is stored, reviewed, updated and discarded in a secure way.

The second type of knowledge is tacit Knowledge that was originally defined by Polanyi in 1966 as know-how (Brown & Duguid, 1998). Tacit knowledge is practical knowledge, obtained through experience by a person during his or her daily activities that could be based on a variety of situations faced by an individual (Fraust, 2007).

Tacit knowledge is not easy to define, since it is intuitive and is mostly experience based. Tacit knowledge is dependent on an individual's context and is personal in nature. Since it is deeply embedded in a person's dedication, actions and participation, it makes it very difficult to transfer. (Nonaka, 1994). Tacit knowledge is based on experience about certain topics and performances, like how to climb a tree, ride or talk. This makes it very difficult to explain, as it is dependent on individual's level of experience and continual practice. However, with time and more sharing it can be transformed to explicit knowledge. This knowledge is valued as the key source of knowledge that leads to breakthroughs and innovations in any growing organization (Wellman, 2009).

Every organization and institution have their own methods of demonstrating tacit knowledge with products, methods, procedures, patents, strategies and services. Tacit knowledge involves various skills, competencies, experiences and relationships. Similar

knowledge can be attached in work processes that exist in all core functions of an institution within all its systems and infrastructure.

However, explicit and tacit knowledge are not exclusive since the latter is grounded in the former through externalization (Greiner *et al.,* 2007). The most important and vital part of knowledge is the ability to understand, comprehend, use, reuse, and combine information for better results with existing knowledge (Sajjad, 2014). According to Kruger (2008), it is a reality that existence and survival of institutions is dependent on the availability of knowledge within it.

## 2.3 Knowledge Sharing

This is the process of making knowledge accessible to people in an institution (Abzari & Teimouri, 2008). According to Lichtenstein and Hunter (2006), knowledge sharing is a complex process that involves the collection, contribution incorporation and knowledge application by an organization. Knowledge sharing is the most imperative process in knowledge management (Alavi & Leidner, 2001). It increases organizational performance and effectiveness when well emphasized and articulated (Kim & Lee, 2005).

## 2.4 Knowledge Management

According to Sagsan (2006), this is a process of storing, collecting, structuring, sharing, controlling, creating, disseminating, codifying, using and exploiting knowledge in organizations. The knowledge management is an approach that is used to develop a systematic set for the creation, organization and dissemination of knowledge. This knowledge is disseminated by use of different technologies supported by the creation and sharing of knowledge (Omur, Y., Omur, N., & Engin, 2009).

### 2.4.1 Knowledge Management Process

According to (Halisah *et al*., 2021) the process of knowledge sharing in organizations leads to a culture change and improves performance and finally shapes the attitude and behavior of employees in knowledge sharing. Knowledge shared and repeated over times creates expertise in a particular field. Adoption of industry 4.0 technologies heavily rely on collaboration of different firms using different technologies to share mechanisms for adoption. This is a knowledge sharing process. (Lepore *et al*., 2021). According to (Odigwe *et al*., 2020), effectiveness in educational research is jointly contributed by the major common tasks of the knowledge management process namely, data retrieval, data storage, data security, data sharing and data re-use.

## 2.5 Evolution of Data Mining

The history and growth of data dates back to the 1960s when scholars and staticians used terms like “Data Fishing” or “Data Dredging” to refer to bad practices of analyzing data without a probability basis. The term “Data Mining” existed as from 1990 and was used by the database community. At the beginning of the 19$^{th}$ century, the expression “database mining” was trademarked by HNC, a company based in San Diego, making the researchers change the phrase to “Data Mining”. Other terms used in place of “Data mining” included Knowledge Extraction, Data Archaeology, Information Harvesting and Information Discovery. Piatetsky-Shapiro, (1990) used the term “Knowledge Discovery in Databases” on the similar topic where the term then became common Machine Learning (ML) and Artificial Intelligence (AI) Communities. Data Mining gained popularity in the press and business communities. The two terms Knowledge Discovery and Data Mining are used concurrently (Ramírez-Gallego et al., 2018).

### 2.5.1 Data Mining Background

Extraction of data has been done manually to form patterns for some centuries in the past. Other methods that were earlier used to identify trends in data are Bayes' Theorem (1700s) and Regression Analysis (1800s). With the increase of computer processing speeds, power and technology, there has been an increase in data collection, storage and manipulation. Complexity has gone to higher levels due to increase in volume sets of data. There has been improvement in direct and indirect data processing assisted by new ideas in computer science through cluster analysis, neural networks, cluster genetic algorithms (1950s), decision trees (1960s), and support vector machines (1990s). To discover the patterns that are hidden voluminous data sets, Data Mining comes in handy with the application of these methods. Through this Data Mining has bridged the gap in both artificial intelligence and applied statistics, which usually provides mathematical background to database management. Through further exploration on different ways on how data is indexed and stored in databases, the data mining process is easily achieved. This leads to execution of learning and actual discovery of more algorithms, making such approaches to be applied to larger data sets (evolution of Data Mining tutorial, 2018). This has aided in the discovery of new knowledge (Ramírez-Gallego et al., 2018).

### 2.5.2 Data Mining and Knowledge Discovery

Various studies have been conducted in the area of Data Mining and Knowledge Discovery. In this study, the researcher focused on research that is relevant, to address the objective of the study. Therefore, in this section, related studies concerning Data Mining and Knowledge Discovery are reviewed.

Text mining approaches are the real solutions that big companies and organizations use to create a great impulse for sourcing opportunities that are as a result of big data and

unstructured data to improve in achieving strategic business activities and elevate the process of decision making in the organizations (Xie et *al*., 2020).

Text Mining (TM) tasks, which were found to be useful for knowledge management in empirical research carried out, showed the importance of information retrieval in building a knowledge management system (Rumanti *et al.,* 2015). Rumanti and fellow researchers, researched on models that transfers explicit knowledge at the organizational level of small and medium-sized enterprises (SME). The model the author used comprised of six concepts, which include receiver characteristics, sources characteristics, organization context, media characteristics, explicit knowledge characteristics and effectiveness of explicit knowledge. (Sihui & Xueguo, 2016), carried a review on how data mining can be used in mining explicit knowledge by the consumers of the knowledge who are mainly readers. The researchers also developed a story-telling based knowledge management system, based on relational database management system (RDBMS) for domain experts to contribute their explicit knowledge.

The contribution of textual data from domain experts is due to the extraction of information from various collections. (Rajman, 1997) (Rajman *et al*., 1997). Extraction and collection are involved in automated synthesis of documents content. One system operates on indexed documents through an association extraction exploits.

The system of association extraction leads to a basis for extracting significant keywords associations; it is followed by an increment of algorithm that permits possible set of key words to be explored by starting with the regular singletons and iteratively adding them to produce new and frequent sets. The use of Knowledge Discovery techniques to the full textual content of documents, using a prototypical document extraction algorithm, in his

dissertation (Syed, 2019) investigated how to apply and interpret topic models to large collection of documents proved how the steps of processing the data collected can greatly affect the quality derived from latent topics. The results indicated better results when the extraction process operated on abstract concepts represented by the keywords rather than on the actual words contained in the documents (Rajman *et al*., 1997). The authors support the need of applying Natural Language techniques to identify more significant terms.

In the application of Natural Language Processing (NLP) techniques to extract explicit and implicit concepts with their semantic relations between concepts, poses a problem in Knowledge Discovery from text. The techniques of Text Mining and Data Mining are equivalent though Data Mining tools are designed to handle structured data from databases, where areas Text Mining works with unstructured or semi-structured data sets such as emails, full-text documents and HTML files (Bolasco, *et al*. 2005).

A team of researchers as per Bolasco discovered that text clustering is used for document clustering in a system which clusters a set of documents based on the user typed key term. A similar research was done by Sudhahar, *et al*. (2015) and was particular on network analysis of narrative content in large corpora. The results indicated that the automatic parsing of textual corpora had enabled the extraction of actors and their relational networks on a vast scale, turning textual data into network data.

Wyskwarski (2020) carried an analysis to determine the competencies a company required in recruiting a project manager. The author applied the use of text mining in analyzing various elements that were describing fragments for the job description that

involved the analysis of text mining which had applied text processing in retrieving the required competencies for the job.

The team also in their wisdom recommended that utilizing information retrieval and Data Mining techniques can be of great importance in facilitating knowledge management in virtual communities of practice.

A recent research carried out recently by Wang and Lo (2021) discovered the easiest method of searching, reading and summarizing on information related to COVID-19 using automated text mining techniques to inform the public on the best summarized literature that can inform them on measures to prevent the pandemic.

Tan (2005) also carried out a research and in his study, he recommended the application of the neighbor-weighted K-nearest neighbor algorithm to deal with uneven text sets, (Yang *et al*., 1998) the team of researchers studied on the scheme for improving access to such informal design information using hierarchical wordlists overlaid on generic information retrieval (IR) tools. Singular Value Decomposition (SVD) technique was applied to aid in the automated thesauri generation (Yoon *et al.,* 2008). The team also carried out a study on methodology based morphological analysis to develop a technological road map through identification of key words and their relationships for new products. Chang *et al.* (2009) did a study on developing marketing strategies through discovery of hidden knowledge within the databases of a company and he concluded that sharing and mentorship increases customer retention and improvement of new customers to the company hence increasing sales.

### 2.5.3 Data Mining and Text Mining

A powerful and current respected tool in utilizing the potential of unstructured textual data in big data analytics is by use of the text mining technique to bring out the new knowledge and unearth the important patterns and correlations that is concealed in the data (Hassani, *et al.* 2020).

Knowledge can be extracted from legal textual documents with huge amount of information to discover reality and relationships used by ordinary people and practitioners in the legal domain. Through open information extraction a relationship is built without applying a dataset that has a relation to that particular domain (Kadu, and Ashwini 2020).

Data mining as a technique can be used to analyze large data sets, obtained from big data to extract trends and statistics behind the data. Data Mining has been successful in business situations, military, social and government settings.

According to Gupta and Lehal (2009), 20 % of data in the World Wide Web is in form of numbers, and 80% is inform of text. This places Text Mining as an important concept in getting data and information from different settings where it occupies most of the data formats in both the intranet, World Wide Web and in this case in an organization's setting at KCDP.

Text Mining uses the Natural Language Processing (NLP) and Machine Learning (ML) to collect data and statistics to provide us with a valid conclusion from the unstructured text. To conclude on the two, there is a strong correlation between Data Mining and Text, where by Text Mining uses Data Mining techniques to discover connotative knowledge in text.

## 2.6 Machine Learning and Text Mining

Text Mining applies Data Mining technique of Machine Learning to extract text using computer algorithms and models used by computer systems. In Machine Learning, using algorithms and models, computer systems are able perform specific tasks without the use of explicit instructions given to computers by software or end-users. Machine learning relies on patterns and inference in data identified by machine learning algorithms like Naive Bayesian Classification, K- Nearest Neighbor, Random forest and Neural networks algorithms just to mention but a few.

These Machine Learning algorithms will enable the study to develop an optimal model for the research study at KCDP, to achieve the objectives of the study and answer the research question from a technical perspective.

### 2.6.1 Naive Bayesian Classification

The principle of Naïve Bayes applies that to determine probability of certain situations of content features appearing under controlled conditions, classification of algorithms for a given data have to be classified. The probability of relatively large word features match characteristics of corresponding labels (Yaduang *et al.*, 2015).

In discrete text classification, NB classifier tops as the most powerful technique. The technique applies probability as a principle. Due to its capability to select occurrences of text by selecting key words, text is classified discretely and produces less semantic relation through statistical classification of text. (Yadav & Parne, 2015). Naive Bayes classification has two models namely Bernoulli model and Multinomial model. The

Bernoulli model has no binary word features and word dependencies, whereas the Multinomial model uses integer word counts making it a unigram language.

The Bernoulli model produces indicators of Boolean form for every word of the vocabulary irrespective of its presence or absence. It takes into account words that are not part of the document. Multivariate Bernoulli executes better with larger vocabulary sizes giving an average of 27% reduction in error but performs best with small vocabulary sizes (Sagar *et al.*, 2014). Multinomial Naive Bayes (MNB) is a well-known classifier, a simple and easy to implement method, and still gives very good results (Baldwin & Lui, 2010, as cited in van Dam & Zaytsev, 2016).

### 2.6.2 K-Nearest Neighbor

This method is labor intensive when dealing with large training sets. The k-nearest neighbor method gained popularity in the 1960s when there was an increase in computing power. It has overseen its wide use in the areas of pattern recognition (Han & Kamber, 2006). The idea of KNN is extremely simple but effective in many applications such as text classification (Liu, 2007).

The K-Nearest-neighbor classifiers assign equal weight to attributes, where it uses distance-based comparisons. The Nearest-neighbor principle classifiers can therefore suffer challenges in accuracy, when subjected to noisy or irrelevant attributes. This has led to modification of the method to include attribute weighting and the pruning of noisy data. K Nearest neighbor method has been widely used in text classification methods. It is effective on pattern recognition based on the statistics and can achieve higher

classification accuracy for non-normal and unknown distribution. Its advantages are its robust, and its concept is very clear (Aizhang & Tao, 2015).

KNN has been popular because of its unique learning capabilities and has achieved good performance in many different applications (Suh, 2016). However, the limitation of KNN is that when it finds a category of text to be classified, it calculates the similarity of all the text samples in the training sample set. This increases training samples, therefore classification performance drops.

### 2.6.3 Neural Networks

Neural networks (NN) can be made to perform text categorization. To classify documents, feature weights are loaded into input nodes and propagated forward through the network thus producing output nodes that determine the categorization decisions (Feldman & Sanger, 2007).

Zaghoul and Al-Dhaheri (2013) experimented on an Arabic set of texts. Results demonstrated that the Artificial Neural Networks (ANN) model is effective in representing documents in Arabic.

Artificial neural networks allow researchers to run quick tests and the same time produce reliable results in complex domains. A weakness that Neural Network has is that training through it is very slow. The learned results of neural networks are highly complex for users to translate or interpret as compared to learned rules (Andreas, 2005, Korde and 94

Mahender, 2012 as cited in Rana *et al*., 2014).

The above models are of great importance in the research, where they provided the study with the building blocks for Text Mining and some of the technologies and procedures applied in Text Mining like topic tracking, information extraction, summarization, categorization, clustering, concept linkage, information visualization and question answering. The above models guided in the development of an optimal model for the research study.

## 2.7 Empirical Literature

These literatures explain how the three objectives have been used and applied in various regions of the world to retrieve explicit knowledge using various Data Mining models. Different regions and countries have applied different approaches in the retrieval of explicit knowledge as discussed in the preceding subtopics.

### 2.7.1 Text Mining

Text Mining is the process of extracting important and exciting trends from textual databases. It is sometimes referred to as Text Data Mining or Knowledge Discovery. The science of Text Mining has evolved over years in various fields. In the medical field, Text Mining has been applied in making sense of raw text. This has been applied in the biomedical area, where it was discovered that there was an increase in the large amounts of biomedical literature at a rate that was difficult to locate, retrieve and manage. The framework for semantic representation of textual information made of conceptual models came in handy in representation of text. (Spasic *et al.*, 2005).

Text Mining and qualitative research are compatible epistemologically. Qualitative research approaches are applied in grounded theory of text mining since it supports openmindedness and discourages preconceptions. It approves and enables liberty in manipulation of initial categories in an iterative fashion. Text Mining applies extraction of common themes and threads using computer algorithm. Content analysis and text mining count words by extracting common themes (Yu, Jannasch-Pennell & DiGangi, 2011).

In the legal perspective, Text Mining has been applied in the extraction of arguments and identification of complex information in case factors and participating roles. The use of automated tools provides detailed properties and relationships in cases and their identification. Text Mining has enabled legal researchers to investigate information on new cases added on the case base, resulting to identification of legal arguments (Wyner, *et al.,* 2010).

### 2.7.2 Text Mining Techniques in Explicit Knowledge Retrieval

The discovery of new, unknown information of previous research by computers and the automatic extraction from different sources is what is referred to as Text Mining. The main and key element in it is the art of linking extracted information to form new facts and hypotheses to be discovered through experiments using conventional means (Hearst, 2003).

Italian researchers did discovery and integration of explicit knowledge and learning by example in recurrent networks by use of a unified approach. The results from their study led to the evolution of intelligent systems based on connection models (Bianchini *et al.*,

*1998).*

In Tokyo University, multi-layered neural networks, which included the extraction of knowledge and its use, were applied to train explicit knowledge extraction in Information-Theoretic Supervised Multi-Layered (SOM). Connection weights were obtained at the knowledge extraction phase and were used to train. The method applied was the spam identification problem. The experiment results showed that the information theoretic Supervised Multi-Layered (SOM) improve learning and performance (Kamimura, 2014).

Researchers in India implemented a Document Management System (DMS) that captured explicit knowledge in documents of the organization. The researchers concluded that Document Management System (DMS) contained useful information that was integrated the on-line existing Human Resource Information System (HRIS) database (Khan *et al.*, 2015).

An engineering department in Indonesia assessed the impact of tacit and explicit knowledge on small and medium enterprise; the result indicated that tacit and explicit knowledge could be shared to create more knowledge (Rumanti, Samadhi, & Wiratmadja, 2016).

Researchers in South Korea carried out a study in the medical field to identify the relationship between gait-Parkinson's diseases (PD) from PD based research articles using a Text Mining approach. The study applied text pre-processing, clustering, categorization and visualization of text. It was resolved that for assessing Parkinson's disease, gait related analysis was most important (Aich *et al.*, 2017). This was

knowledge adopted and transferred to clinicians to improve on analyzing the disease preference.

In Pakistan researchers carried a study on technological verdicts using implicit and explicit knowledge mining of crowdsourced communities. The framework utilized the Text Mining techniques that supported software development teams to get a wider spectrum of opinions in form of knowledge patterns discovered by the crowsourcing knowledge mining framework (Mushtaq *et al.,* 2018).

### 2.7.3 Challenges Concerning Explicit Knowledge Retrieval

Retrieval of both tacit and explicit knowledge is not yet well researched on especially on the process of retrieval. Retrieval problems are dependent on the utility of a document based on their ranking and retrieval methods applied (Zhai *et al* 2015). In some instances, skills gained do not give any impact due to lack of knowledge sharing making projects fail, where their sustainability and continuity are hampered (Conger, 2015).

Ethical challenges during data collection have been an issue when considering issues to do with privacy, storage and location of data, data interpretation, informed consent, identification, classification and management of data. Other ways taken to understand similar approaches depend on various ideologies and assumptions (Arnold & Pistilli, 2012).

## 2.8 Text Mining Retrieval Techniques

Most techniques were developed some years back with a combination ranging from natural language processing, information retrieval, information extraction, and Data

Mining. Some of the techniques such as retrieval involve a group of algorithms that make it possible to search data as per user requirement.

Continual improvement is key for any organization to move from one level to another. Adoption of and use of Text Mining may give breakthrough in promoting effective KM and hence improve service deliveries and decision-making. In respect for continual improvement, organizations are collecting large volumes of data and storing them in various databases for future reference. Key issues identified and discussed within textual data and its classification would unearth and help the policy, knowledge workers and decision makers to better manage their activities (Ur-Rahman & Harding, 2012).

### 2.8.1 Clustering

Clustering is a tool for automatic grouping of similar objects into sets for data analysis, which solves classification problems. The major aim for clustering is to distribute cases for example people or objects into groups, so that the similarity degree can be strong between families of the same cluster and weak between groups of different clusters (Hajizadeh, *et al.* 2010). There is no pre-classified data in clustering and it does not distinct between independent and dependent variables. Groups of records similar to each other have been developed that include K-means (K-Nearest Neighbors), hierarchical, fuzzy C-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and Self-Organized Maps (SOM).

### 2.8.2 Classification

The most common learning models applied in the world today is classification and it is commonly used as a learning model in Data Mining techniques. Its main objective is to

build a model that identifies the category an object belongs to *(*Ngai *et al.,* 2009*).* According to Dunham (2000), classification is viewed as a mapping from the database to the set of classes. Classification produces predefined, none overlapping and partitioned classes in an entire database. It uses approaches for training a set of objects that are associated with known class labels. The classification algorithm learns from the training set and builds a model. There are several classification techniques including support vector machine, nearest neighbors, decision trees, and neural networks.

### 2.8.3 Summarization

Summarization or text summarization reduces the length and detail of a document and at the same avoids the distortion of data and information contained in a document. It is a challenge for computer system software to be able to summarize a document. Text summarization tools can be used for sentence extraction from a document. An example can be where summarization tools may extract sentences that follow key phrases like “in conclusion” where in most cases there is where the main knowledge is contained in form of a summary or conclusion. Summarization applies models like LDA topic models that are probabilistic for analyzing text that are meaningful and useful (Jelodar *et al.*, 2019).

The figure1 below shows an example of text summarization procedure in a computer system.

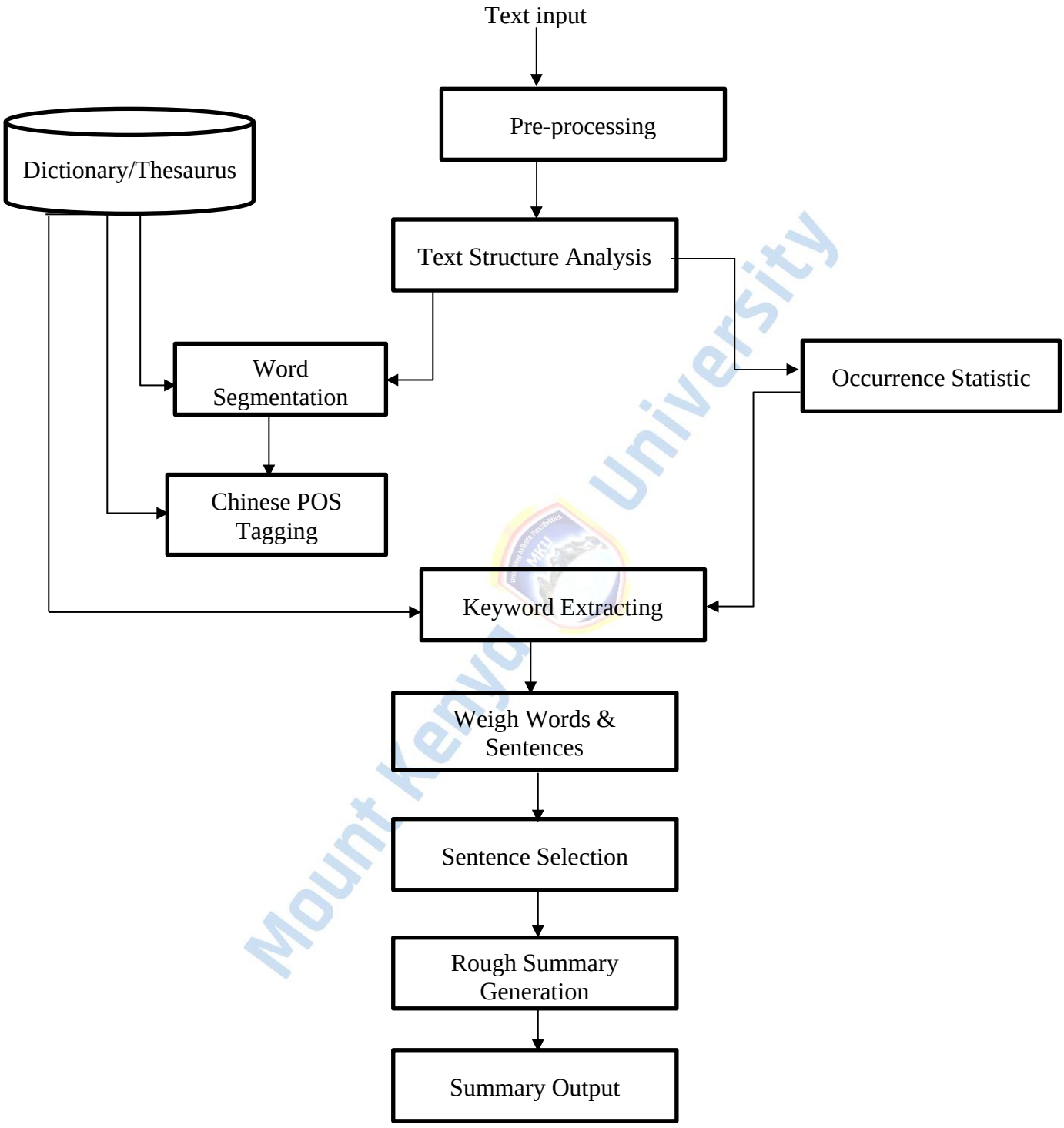



*Figure 1*: *Text Summarization Model Source: Gupta & Lehal (2009)*

## 2.9 Theoretical Framework

Interpretation of information depends on the ability of an individual, which further translates to experience and this is defined as knowledge. (Smith & Bollinger, 2001). This leads to the understanding that the power of interpretation is directly linked to the experience one has in a particular subject. Organizations can develop and handle competitive advantage if only they are able to create, share and use the knowledge gained over the years to sustain them and make better decisions. (Richard, 2004). The moment an organization realizes on the importance of its knowledge, it is easier to maintain competitive advantage. This is the reason why tacit knowledge is a kind of knowledge which is highly personal, difficult to form, communicate and share with others, and he gave full attention to explicit knowledge from the perspective of oriental cognitive science (Nonaka, 1991). The father of knowledge management believes that explicit knowledge includes mental model and skills. KCDP's explicit knowledge resources exist in the articles and all levels of the organizations, and it can be obtained through the flow and sharing from researchers in KCDP. This tacit knowledge is what is later transformed to explicit knowledge.

Where areas tacit knowledge is a process the practical knowhow, obtained by every individual in daily activity and could be based for varieties of the situation that he/she faced. (Fraust, 2007). Knowledge would be divided into three groups, namely tacit Knowledge, it is Knowledge that is the most difficult to transfer to others. Tacit knowledge is the knowledge based on experience about certain topics like how to ride a bike or how to talk. It cannot be fully explained, because it is fully realized in the individual level that is rooted in practice and experience. But with time and more sharing it can be transformed to explicit. Second is the implicit knowledge, knowledge that's relatively quite difficult to be accessed, but it is still possible to be revealed. This type of

knowledge can only be detected by observing or asking for more detail. Third is the explicit knowledge that is realized by the bearer of knowledge. Explicit knowledge is a kind of knowledge that could be written in books, journals or memos and easy to be transferred upon (Rumanti, 2011).

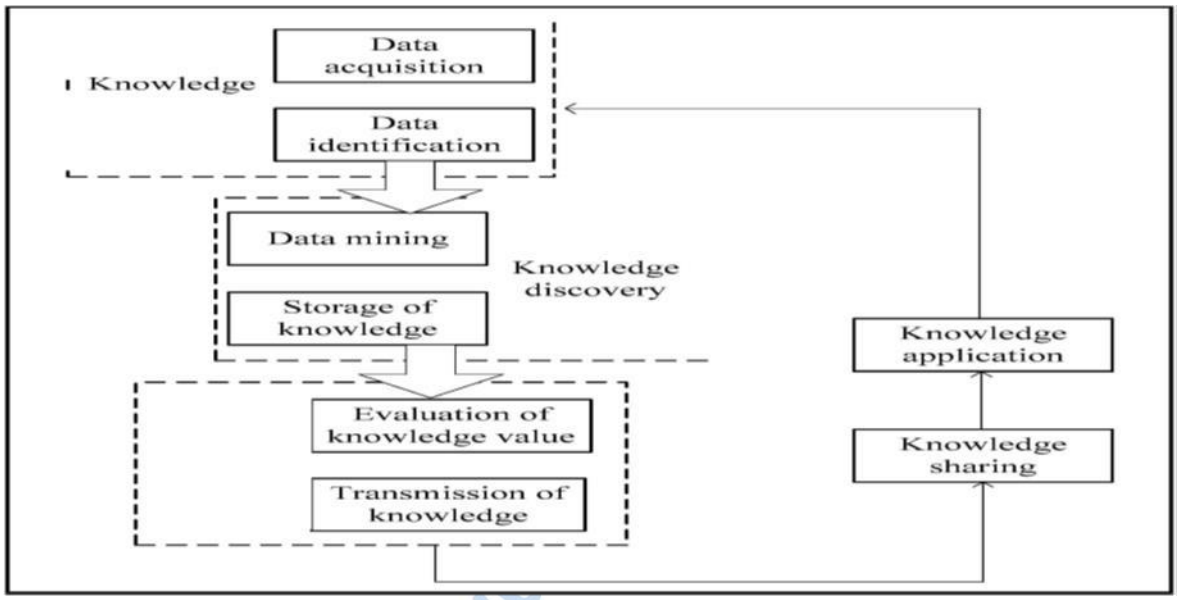



*Figure 2: Theoretical Framework of Text Data Mining in Knowledge Nonaka (1994)*



## 2.10 Conceptual Framework

INDEPENDENT INTERVENING DEPENDENT

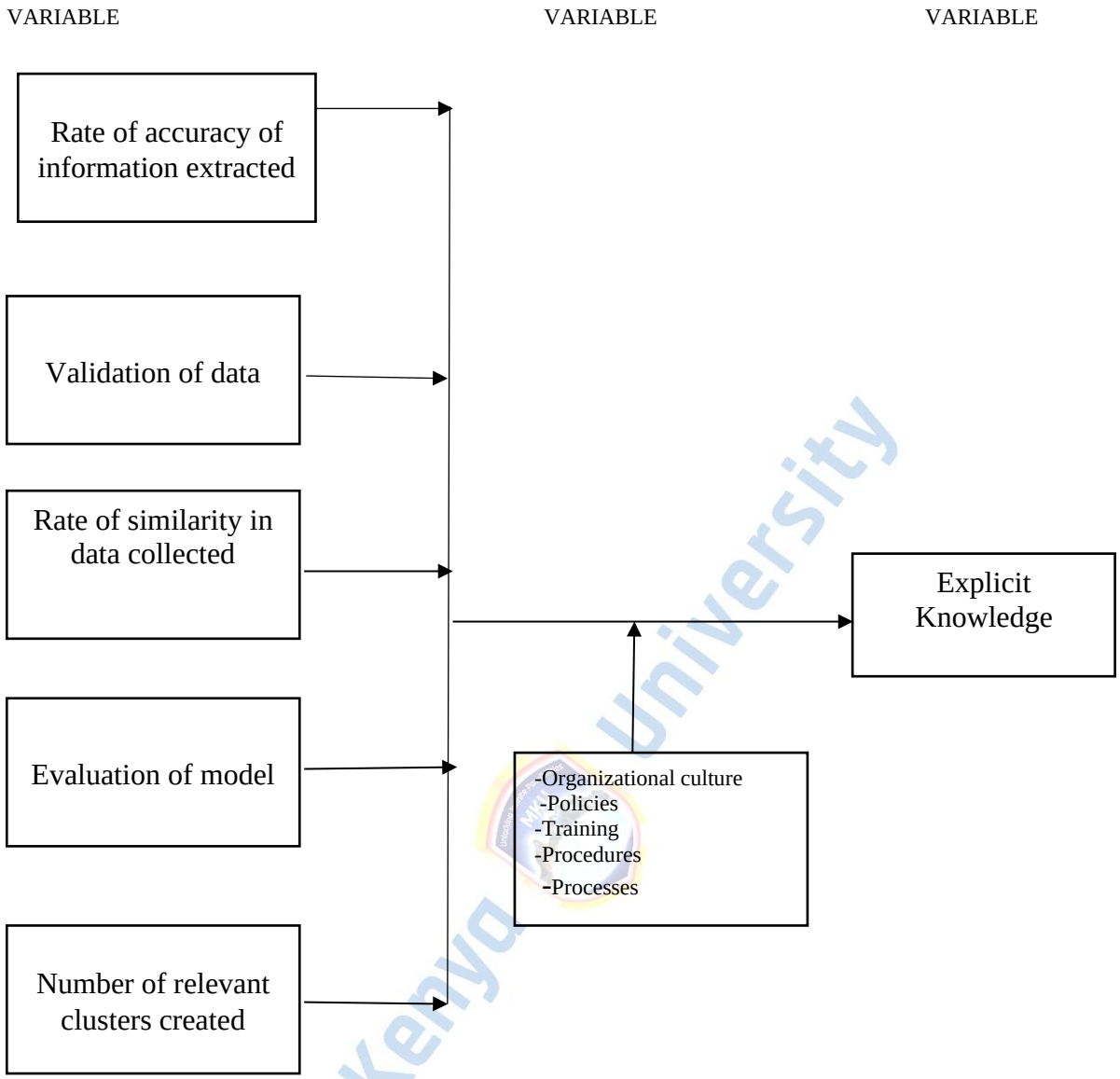



*Figure 3: Text Mining Framework Model for Retrieval of Explicit Knowledge at KCDP*

*Source: Self*

Figure 3 shows a schematic summary of the research problem and the proposed solution. The conceptual framework shows how data and information in documented format could be extracted, summarized and represented in visual formats, using text mining technologies. The text mining technologies used were information extraction, summarization and topic modelling/visualization.

These technologies represented independent variables in the study and contained different parameters and indicators that could be used to determine the accuracy for each independent variable. The conceptual framework for the text mining used in the study is diagrammed in Figure 3. The different variables and how they were measured is illustrated in the main stages for text mining implementation, which are: (i) Rate of Accuracy was the actual retrieval of information from various websites, this was done by use of information that was collected from different websites. The study investigated the prospects of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). The study's main objective was to establish how a Text Mining model could be used in explicit knowledge retrieval at KCDP. The study identified text-mining techniques that could be used to develop a textmining model, evaluate the model to be able to retrieve explicit knowledge at KCDP. The study targeted staff of the agencies that constituted the KCDP project which included, Kenya Marine and Fisheries Research Institute (KMFRI), Kenya Wildlife Service (KWS), State Department of Fisheries (SDF), Coastal Development Project (CDA), Department of Physical Planning, Kenya Forest Service (KFS) and National Environment Management Authority (NEMA). The study used the exploratory and experimental research design to be able to understand the research problem, answer the research objectives and questions. The total population of staff in the project was one hundred and fifty (150), out of which fifty-two (52) were sampled. Purposive sampling was used to select samples from the representative groups that comprised the target population. Two methods of data collection were used namely; questionnaires and focus group discussion. The questionnaire was applied to members of staff in four major departments namely the top management, research and administration, knowledge management and finally the ICT department. The focus group discussion was applied to

a special group in the knowledge management section. Content analysis was used to analyze the focus group discussions. Questionnaires were analyzed using the Statistical Package for the Social Sciences (SPSS) version 25 software. The use of questionnaires and focus groups were used to establish the current situation at the KCDP in terms of knowledge management systems in place and whether text mining could be used to retrieve explicit knowledge at KCDP. Text were collected from websites of organizations that took part in the KCDP project by using python libraries namely; Python Request 2.22 and Beautiful Soup 3. The collected text was then summarized using text summarization algorithms used in the model like Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer. After summarization topic, modelling was performed on the text collected using Latent Dirichlet Allocation (LDA) topic-modelling algorithm to create topics based on patterns in text. The model was then evaluated to establish its performance by measuring the four variables identified using precision and recall to measure accuracy, topic modelling to measure rate of similarity, and perplexity to measure evaluation of the model which gave a perplexity of -6.0455 from the text analyzed and modelled. It was concluded that text analysis could be used to analyze text and create explicit knowledge from both structured and unstructured data formats using the model. The study recommended that more research should be done in the development and evaluation of proposed models and that text analysis should support other languages other than English like Kiswahili and Arabic, which are collected by the organization. Finally, it was established that text analysis was dependent on the availability and accuracy of data, for creation of explicit knowledge and that data should be reliable and accurate to get the best results, for the best decision making process. Data was extracted via Web scrapping and Data Collection from URLs, through precision and recall (precision being the proportion of the retrieved documents that are relevant, and

recall being the proportion of the relevant documents that have been retrieved), the performance of the system was measured; (ii) Data validation was achieved through Preprocessing/Data Cleansing using Tokenization and Lemmatization, this was the actual breaking of long strings of text into smaller pieces, or tokens, this led to the achieving of key words that could describe the data in the document which finally led to improving machine learning performance and validation of data; (iii) Rate of similarity was achieved through Text Summarization Algorithms using Latent Semantic Analysis (LSA), LuhnSummarizer, EdmondSummarizer and LexRankSummarizer, this was attained by measuring the content similarity between the original document and the summary made by the model; ; (iv) Results and Evaluation of the Model was done by using Perplexity algorithms, this was achieved by quantifying how uncertain the model was in the predictions it made and through the evaluation a low perplexity was achieved which guaranteed the model confidence and performance; and (v) Number of relevant clusters created was achieved through Topic Visualization using Latent Dirichlet Allocation (LDA), this was achieved by importing documents and using LDA, sentences were tokenized, cleansed and frequency calculated. The final results was a visualization in the form of a bar chart that represented the term frequency for each of the words.

Information extraction and summarization are text-mining techniques widely used because of their flexibility and performance with different python text summarization algorithms like Latent Semantic Analysis (LSA), LuhnSummarizer, EdmondSummarizer and LexRankSummarizer. Topic modelling algorithms/libraries like Genism and Latent Dirichlet Allocation (LDA) models are then used to extract topics from the text. The text mining technologies used were information extraction, summarization and topic

modelling/visualization. To determine the accuracy of information to be extracted both the precision and recall metrics were used, as follows:

$$\text{Precision metric} = \frac{\text{(Number of relevant items retrieved)}}{\text{(Number of retrieved items)}}$$

$$\text{Recall metric} = \frac{\text{(Number of relevant items retrieved)}}{\text{(Number of relevant items)}}$$

## 2.11 Critical review of the current models and theories

According to (Sharma and Bansal, 2015) theory that states 'Different data mining tools have got their own pros and cons. The main consequence of this fact is formulated by the 'no-free lunch theorem', which states that there is no universally best data mining tool'. To overcome the no-free lunch theorem, the KCDP model developed was tested in different databases with different platforms and it worked, it can be applied in both structured and unstructured data formats, making it a universal model for text retrieval across all platforms. This was tested when testing the model on different databases and websites.

According to (Bjerva *et al.,* 2016) one neural algorithm was applied in determining byte based language identification with deep convolutional networks and the accuracy of the model from the results was 0.9069. The KCDP text retrieval model guarantees accuracy since it compiles various techniques of summarization algorithms to give the user or reader the best summary from a comparison of four algorithms, each user can pick from any of the summarized information, which creates more meaning according to their requirements. (Gamallo *et.al.* 2017) used a single classification algorithm to test the discrimination on perplexity based method and the results from the experiments gave the

best accuracy of 0.927. On the other hand, the developed and tested KCDP retrieval model uses multiple classification based on summarization to produce summaries from text documents. This was achieved by the extraction of features that are very important from text documents. The features with best attributes were captured and hence bringing out the best quality summarized sentences with specific relevant topics. The latest model by (Khachatryan & Muehlmann, 2020). The model was used to measure the draft alignment of patent documents using text mining and from the results run several times the best accuracy from the model was 0.95 and that was the best perplexity. The developed and validated text mining model used a metric that ensured independence of the size of the data set selected. This was achieved by using probability to normalize the test set by the total number of words and that gave a result which gave a per-word measure and thus a lower perplexity of -6.0455,

## 2.12 Identified Research Gaps

This study investigated intra organizational knowledge sharing and discovery using Data Mining techniques. It has both theoretical and practical contributions. The theoretical contribution is that the study explores current concepts and trends of Text Mining approaches and explicit knowledge to support organization's explicit knowledge management. The study identifies the applications of Text Mining approaches to support explicit knowledge management to promote knowledge sharing within organizations.

Different authors have made tremendous contribution to this process of knowledge sharing and management as highlighted in the literature review section above. However, there are gaps identified by various authors; More than 80 percent of today's data is composed of unstructured or semi-structured data.

The discovery of appropriate patterns and trends to analyze the text documents from massive volume of data is a big issue, Sumathy & Chidambaram, (2013). Gelman *et al., (2013)* contribution to the probabilistic approach was capable of analyzing text from massive documents but only from structured data formats. It lacked the way of taking the iterative and interactive nature of the Data Mining process. The new developed model applies the use of topic modelling that helps in discovering hidden topical patterns that are available in all collections presented in any collection, it also has a capability of annotating documents according to a specific topic and it is able search and summarize texts based on annotations. Agrawal, Imisliniki and Swani (1993) contributed in the data compression approach of compressing data and later sharing it, this technique achieved the reduction of data though it lost the quality of information transferred by losing originality, however, they stated in their research that not all data could be compressed. This left very important information on what needed to be shared. To address the gap, the KCDP model applied tokenization and lemmatization to maintain quality and integrity of information during the preprocessing stage of cleansing data collected.

This study has addressed the gaps identified by different authors by developing a retrieval model where all data will be retrieved, summarized and shared even though it will be in different formats. This study has also a practical contribution in terms of the designed model using Data Mining techniques for knowledge sharing and discovery to promote inter project knowledge sharing and discovery among KCDP members and its stakeholders.

## 2.13 Summary

The entire chapter showed that the literature reviewed supports the importance of intelligent decision making in many contexts of an institution and business in general. It also showed how various studies have been used in retrieving explicit knowledge in many parts of the world. It also showed the challenges of knowledge retrieval and finally described the proposed model for retrieving explicit knowledge in Kenya Coastal Development Project. Decision support tools researched for various for business intelligence can be interpreted as decision support tools and are of great importance to organizations for their break through within specialized markets that have a similar competitive advantage. The value of the tools can be of great importance and of value in terms of quality and relevancy that will be based on high levels of integrity. Warehouses, databases, files and various web sources are the main sources of knowledge that may contain different formats of data in structured and or unstructured forms. Therefore, the need to further exploit the use of Text Mining in discovering knowledge and retaining knowledge assets in an organization.



# CHAPTER THREE: RESEARCH METHODOLOGY

## 3.0 Introduction

Data collection and analysis to achieve the objectives of the study is presented in this chapter. This chapter identified the target population and described the methodology the study used to investigate the research problem. It contains information on the research design, target population, sample size, sampling techniques, data collection instruments, pilot study, procedures and ethical consideration. The research methodology focused on Generation of valuable knowledge using Text Mining and its techniques could be shared and improve decision-making across all cadres in an organization.

## 3.1 Research Design

The study adopted text summarization models as per Jelodar et al 2019 that employs extraction, summarization and topic modelling in retrieval of text to create explicit knowledge. The study employed an experimental and exploratory research design. The research design was flexible and aided the use of text mining tools for data collection and analysis, which guided the development of the model.

The exploratory research design was best suited for the study because of the limited research that had been done in the area of text mining for retrieval of explicit knowledge. According to Oates (2005), exploratory research design is suitable for areas of study where little or no literature has been done on an area of study. The study targeted a total number of one hundred and fifty (150) staff at KCDP. That is the total population of staff in the project as per the administrative officer of the project, (I. Githaiga, Personal Communication, and January 11, 2019).

### 3.1.1 Experimental Research Design and Plan

The data was collected using a computer as a desktop research design. The design followed a process that could support in answering the research questions as explained below:

i) **How could text-mining techniques help in the retrieval of explicit knowledge at KCDP?**

Data was collected through web scrapping from the websites of organizations that took part in the KCDP project by using python libraries namely Python Request 2.22 and Beautiful Soup 3. The collected text was then summarized to bring out meaning and content as per the objective by using text summarization algorithms like Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer. Through web scrapping text that contained various themes of the project were collected for summarization. The different algorithms summarized the text and brought out a simple summary that could be understood easily.

ii) **What Text Mining model could be designed for retrieval of explicit knowledge at KCDP?**

This involved the evaluation of different text mining techniques like summarization, classification, and categorization and information extraction. Summarization gave the best results. After summarization, the next step was to determine the best model for retrieval of explicit knowledge. In the design of the text-mining model, the study narrowed down to three text-mining techniques namely information extraction and text summarization and topic modelling. Analyzed text were represented as topics and visualized as graphs to be consumed in the simplest way possible to determine the accuracy of the text mining analysis performed.

Topic modelling was performed on the text collected using Latent Dirichlet Allocation (LDA) topic-modelling algorithm to create topics based on patterns in text. This finally retrieved the best summaries as per the topics requested.

**iii) How would the designed Text Mining Model be validated for retrieval of explicit knowledge at KCDP?**

The model was designed and evaluated on the Google Colaboratory platform. "Colab" as popularly known, a free Jupyter notebook environment. The process involved running data that had been web scrapped from the different websites through the developed model. Data used were two forms. The kcdp.txt and speech.txt. The data was run through the new designed model and met the expectations of the research as per the results presented in chapter 4. The model was then evaluated to establish its performance through testing of perplexity. The results showed low perplexity that proved that the model performance was good.

### 3.1.2 Experimental Design

The model was designed and evaluated on the Google Colaboratory platform. "Colab" as popularly known, a free Jupyter notebook environment rich in computing resources mainly RAM, GPU and storage that required no setup and run entirely in the cloud. Data was run, executed, analysed and saved as a whole in the colab notebook. The first step used in evaluting model was creating a notebook on the colab environment where code could be read, executed, saved and analyzed.

The different variables in the research were measured using different techniques;

i) **Rate of accuracy** – This was measured using the precision and recall metrics. It showed the number of relevant items retrieved to the number of items retrieved.

ii) **Rate of Similarity** - This was measured using text mining technologies like topic modelling and information extraction. The technologies had to reduce dimensionality, inflectional endings and relationship of words to give a summary of meaningful words and sentences.

iii) **Evaluation of the model** – The model perfomance was measured through determination of perplexity. Best perfoming models have the the lowest perplexity and this was confirmed as per the perplexity results in chapter 4.

iv) **Number of relevant items created** – This was measured by the number of words and characters summarised by the model. Different summaries were produces as per the results in figure 13 to figure 16 of the experimental results.

### 3.1.3 The Experimental Plan

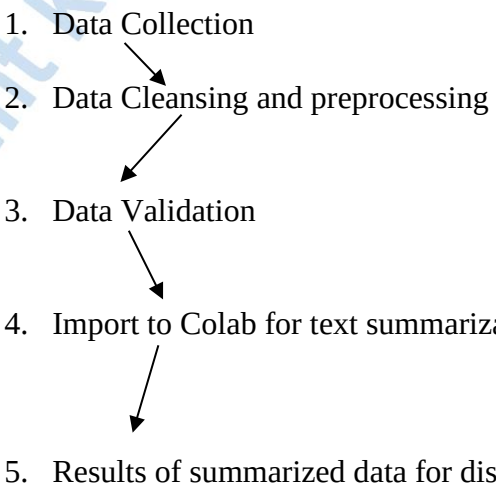



*Figure 4: Experimental Plan*

*Source: self*

## 3.2 Target Population

This research targeted a total number of one hundred and fifty (150) staff. That is the total population of staff in the project as per the administrative officer of the project, (I. Githaiga, Personal Communication, and January 11, 2019).

## 3.3 Sampling Technique

This research used purposive sampling to ensure that the data sampled, collected and analyzed could assist in achieving the objectives of the study. Purposive sampling also referred to as judgment or selective sampling, involved selecting a sample suited to the purpose of the study (Williamson & Johanson, 2013).

The criteria used in purposive sampling involved sampling of departments namely top management, research and administration, ICT and the knowledge management department. The sampled departments would provide the required data and information and would benefit mostly from the research study.

The knowledge management department provided great insights on how data and information acquired from previous and current research were managed and the challenges that needed to be addressed. The department also provided most of the information through focus group discussions and practical challenges that needed immediate solutions. This technique was chosen due to the unique subject matter in the research area to provide quick insights on explicit knowledge retrieval. The staff sampled are those that have information rich in data and knowledge management.

### 3.3.1 Sample Size

The departments sampled were those that contained much data and information about the KCDP project, which would enable the researcher meet the objectives of the project. The total population of staff in the Kenya Coastal Development Project was one hundred and fifty (150) and the sample size was fifty-two (52). Factors such as the population size and the objective of the study determined the sample size. Other criteria considered were the precision level, confidence level and the variability degree.

i. The precision level (sampling error) provided the range in percentage points where the true value of the population was estimated. The level of precision in this study was taken to be 11% due to the size of the population.

ii. The confidence level referred to the percentage of all possible samples that were distributed normally considering the true value. Since this was a normal distribution, a 95% level of confidence was selected for the research study. iii. The degree of variability referred to the manner in which attributes of the population were distributed. A small sample size implied less variability in the population. A proportion of 50% was selected as it indicated the maximum variability in the population.

The following formula was then used to calculate the sample size:

$$n_0 = \frac{z^2 pq}{e^2}$$

Where $n_0$ is the sample size, $z$ is the abscissa of the normal curve that cuts off an area $\propto$ at the tails, $e$ is the precision level, $p$ is the variability degree and $q = 1 - p$. Thus:

$$n_0 = \frac{(1.96)^2(0.5)(0.5)}{(0.11)^2} = 80$$

However, since the population is small, the sample size was slightly reduced and adjusted using:

$$n = \frac{n_0}{1 + (\frac{n_0}{N} - 1)}$$

$$n = \frac{80}{1 + \frac{(80 - 1)}{150}} = 52$$

*Table 1 Target population, sampled size and sampled technique*

| DEPARTMENT | TOTAL POPULATION | SAMPLED POPULATION |
|---|---|---|
| ICT | **40** | **10** |
| Research & Administration | **50** | **20** |
| Knowledge Management | **30** | **12** |
| Top Management | **30** | **10** |
| | **150** | **52** |

*Source: Research Data*

## 3.4 Data Collection Methods

The data collection methods used in this study involved the use of data collection tools like RapidMiner Studio 9.5 with its extensions Aylien Text Analysis, Rosette Text Analysis and web Application Programming Interfaces (API) like the Twitter Search API Open Source web scraping tools, which include Beautiful Soup 3 and Python Request 2.22, were also used. Text was collected from websites and portals through web scrapping or using web application programming interfaces (Kobayashi *et. al*, 2018).

Questionnaires were used in the study to create a foundation and a reference point for the study. This assisted the study in the design and development of the model). Questionnaires are a cheaper method of gathering data from potentially large numbers of respondents. It is a well-known technique to collect data and people's opinions (Preece *et al.,* 2002).

The Focus group discussion was applied in the collection of data from the Knowledge Management and ICT departments. The departments were of great importance to the study because they provided data sources, access, status and formats used in the project, making the study viable. Most of the information provided by these groups gave data management insights in the project that formed part of the results and discussion section of the study.

## 3.5 Pilot study

This was a small scale of the experiment carried out prior to the main study to test the feasibility, cost, time, validity and reliability of the instruments of data collection. It

helped to predict the appropriate sample size and improve the research design before the main study.

The pilot study was carried out in one of the Kenya Coastal Development Project agencies, the Kenya Marine and Fisheries Research Institute, involving 15 staff (10%) of the study's sample size. The members of staff considered for the pilot study were in the sampled population but did not participate in the data collection.

The pilot study conducted was used to test the selected data collection and analysis methods for the study. A pilot study establishes face and content validity of questionnaires alongside opinions sought from professionals and experts in the field of investigation (Mugenda & Mugenda, 2003).

## 3.6 Data Analysis

Data collected from portals and websites that took part in the KCDP project were analyzed using text summarization algorithms like Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer and visualized using a topic-modeling algorithm, the Latent Dirichlet Allocation (LDA). Questionnaires were analyzed using Social Sciences (SPSS) version 25 software. The procedure involved validation of the questionnaire to check for clarity, legibility, relevance and appropriateness of the data. The questionnaires were then edited for completeness and consistency, coded using descriptive statistic and finally analyzed in SPSS. Information collected and obtained from focus groups complimented the data and information collected and analyzed from both SPSS and the Text Mining tools.

## 3.7 Ethical Considerations

In this research study, confidentiality was of high concern, as the information relevant to the study was of strategic importance. Data sources obtained for the development of the

model did not involve illegal access of e–mails, databases, organizational confidential documents and cloud platforms, but data and information accessible to the public considered to be open data in websites and portals related to KCDP. The names of the respondents were not revealed. The responses attributed to specific individuals, institutions were kept in strict confidentiality.

# CHAPTER FOUR: RESEARCH FINDINGS/RESULTS AND DISCUSSION

## Introduction

This chapter contains the results of data and information that was collected and analysed using the data collection methods. The data was collected and analysed using statistical and analytical tools, which have been interprated for better understanding.

## 4.1 Research Presentation, Interpretation and Discussions

From the data collected and analyzed, it was established that there are several forms of databases at KCDP based on the department an employee works in. An Enterprise Resource Planning (ERP) Dynamics SL 2011 Software is used at KCDP to store data and information in form of a database for the respective departments. It was also established that some of the other forms of databases at KCDP included KOHA, an open source Integrated Library System, MYSQL, Microsoft Excel, Mat lab and the use of Geonetwork 2.0.3 software for data storage and manipulation.

For some of the departments data was stored physically in files, cabinets and in hard copy form. Based on the study it was proposed to have data and information that is in hard

copy form in its duplicate soft copy form for purposes of backup, data centralization, storage and archiving.

The tables below and graphs present the findings obtained after questionnaires and data were analyzed by SPSS.

## 4.2 Descriptive Statistics

*Table 2 Target population, sampled size and sampling technique*

| DEPARTMENT | TOTAL POPULATION | SAMPLED POPULATION |
|---|---|---|
| ICT | 40 | 10 |
| Research & Administration | 50 | 20 |
| Knowledge Management | 30 | 12 |
| Top Management | 30 | 10 |
| | 150 | 52 |

*Source: Research Data*

Table 2 represents the total target population, the sampled population and departments that were involved in the study. 52 respondents from the sampled population provided the information for the study as interpreted in the corresponding tables below.

## Database availability

As indicated in Table 3 majority of the respondents, 96.2% indicated that they had databases at their place of work while 3.8% indicated that they did not have a database.

Table 3 Availability of a database

| | Frequency | Percent |
|---|---|---|
| Yes | 50 | 96.20 |
| No | 2 | 3.80 |
| Total | 52 | 100% |

*Source: Research Data*

## Knowledge of retrieval process

The results in Table 4 indicated that 96% who were the majority of the respondents were not aware of the retrieval process while 3.8% indicated that they were aware of the knowledge retrieval process.

*Table 4 Use of knowledge retrieval*

| | Frequency | Percentage |
|---|---|---|
| Yes | 2 | 3.80 |
| No | 50 | 96.20 |
| Total | 52 | 100% |

*Source: Research Data*

## Application of text mining techniques

Table 5 results showed that 32.7% **of** the respondents indicated that they used information extraction, 15.38% used summarization and classification, 13.46% used information extraction, summarization and classification, 11.54% information classification, 11.54% did not provide their choice, 7.69% used information extraction and summarization,

5.77% used classification and the least 1.92% applied summarization as a text mining technique.

*Table 5 Application of Text Mining Techniques*

| | Frequency | Percent |
|---|---|---|
| Information Extraction | 17 | 32.69 |
| Summarization | 1 | 1.92 |
| Classification | 3 | 5.77 |
| Information Extraction Summarization | 4 | 7.69 |
| Information Extraction Classification | 6 | 11.54 |
| Summarization Classification | 8 | 15.38 |
| None Selection | 6 | 11.54 |
| Information Extraction Summarization Classification | 7 | 13.46 |
| | 52 | 100% |

*Source: Research Data*

## KCDP Encourages Knowledge Management

Table 6 results showed that 61.54% of the respondents agreed that the project encouraged knowledge management. 21.2% of the respondents understood very well the benefits of

knowledge management while 17.31% of the respondents did not have any idea of knowledge management at KCDP.

*Table 6 How KCDP encourages knowledge management*

| | Frequency | Percentage |
|---|---|---|
| Strongly Agree | 11 | 21.15 |
| Agree | 32 | 61.54 |
| Don't know | 9 | 17.31 |
| Total | 52 | 100% |

**Key**: Strongly Agree 5; Agree 4; Don't Know 3; Disagree 2; Strongly Disagree 1 *Source: Research Data*

**KCDP Needs a Model for Knowledge Retrieval**

Table 7 results showed that the majority of respondents 50% and 42.3% respectively, agreed that KCDP needed a model for retrieval of explicit knowledge. 7.7% of the respondents did not know what the model could do and how it could benefit the project.

*Table 7 Need for a Model at KCDP*

| | Frequency | Percent |
|---|---|---|
| Strongly Agree | 26 | 50.00 |
| Agree | 22 | 42.31 |

| | | |
|---|---|---|
| Don’t Know | 4 | 7.69 |
| Total | 52 | 100% |

**Key**: Strongly Agree 5; Agree 4; Don’t Know 3; Disagree 2; Strongly Disagree 1

*Source: Research Data*

## 4.3 Focus Group Discussion

The focus group discussion was done to get an in-depth view and validate the data that was collected and analyzed using questionnaires especially on the open-ended questions. It was used as a qualitative approach to gain an in-depth understanding of social issues. The focus group discussion also captured on how data and information at KCDP could be well maintained and retained through text mining on explicit knowledge retrieval to provide an end result of having a Knowledge Management System. The focus group discussion took place at the KCDP offices on 19 November 2019. The main objective of the focus group discussion was to highlight where KCDP was in terms of data and information management and where it intended to be. The discussion focused on the Knowledge Management and ICT departments in the organization.

The sampled team showed understanding of what knowledge retrieval entailed. The team gave different meaning of knowledge retrieval which included; getting information from a specific storage, recovery and restoration of data from archives, obtaining data from a management system and extraction of data and information using queries. They also gave a brief on what knowledge management means that included; sharing of information for learning, access to expert information, availability of information when needed and building of information assets within an organization. While all this information was

given some felt that more sensitization, training and subsequent audits on Knowledge Management Systems in place needed to be carried out in all departments at KCDP. The study narrowed to the two groups because they possessed an understanding of the area of study. Their understanding was corroborated through their response in the questionnaires collected and analyzed. For instance, Participant No. 25 said that knowledge retrieval is a process of obtaining information system resources that are relevant to an information need while participant No.28 was of the feeling that it is an automation process. The majority understood the whole concept of information and knowledge retrieval from the process of collection, processing, storage and sharing. Therefore, the analysis from this group confirms that a uniform model to compile this information shall create a positive impact in creating explicit knowledge in KCDP.

## 4.4 Discussion of Individual Objective Results: General Objective

The general objective of this study was to investigate the prospects of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). To achieve the general objective several specific objectives had to be met by illustrating the use of text mining tools using the different operators captured as independent variables in the conceptual framework, which included information extraction, categorization, classification and summarization.

## 4.5 Specific Objective 1

### Validation and application

The first objective was to analyze how Text Mining techniques could help in retrieving explicit knowledge at KCDP. This involved the evaluation of different text mining techniques like summarization, classification, and categorization and information extraction. These are some of the text mining techniques used in the development of text mining algorithms and models.

The procedure for obtaining the text-mining algorithms to use involved the use of RapidMiner Studio 9.5 tools that contained different extensions, operators and Application Programming Interfaces (API's). The mostly used RapidMiner Studio 9.5 extensions for this study were the Aylien Text Analysis extension, which provided operators like language detection, summarization, categorization and sentiment analysis.

The use of Rosette Text Analysis extension involved the use of operators like entity extraction, entity linking, entity sentiments, name matching, name translation and morphology just to mention but a few.

RapidMiner Studio 9.5 contains a twitter search application-programming interface. Figure 4 shows how this twitter application operator was used to mine top 100 tweets of the Kenya Marine and Fisheries Research Institute, an institute that majorly took part in the Kenya Coastal Development Project.

| Row No. | Created-At | Id | From-User | From-User-Id | To-User | To-User-Id | Language | Text |
|---|---|---|---|---|---|---|---|---|
| 33 | Jul 26, 2021 ... | 1419732351... | Big Ship CBO | 1397867851... | ? | -1 | en | RT @KmfriResearch: Earlier today, Senior Research Sci |
| 34 | Jul 26, 2021 ... | 1419702842... | Caroline Njeri | 1050373529... | ? | -1 | en | RT @Manyunyu_Corg: We celebrated the #WorldMangro |
| 35 | Jul 26, 2021 ... | 1419702822... | Caroline Njeri | 1050373529... | ? | -1 | en | RT @BigShip_CBO: PICTORIAL; |
| 36 | Jul 26, 2021 ... | 1419701996... | Caroline Njeri | 1050373529... | ? | -1 | en | Planting over 300 propagules,barefooted, sticking to the |
| 37 | Jul 26, 2021 ... | 1419694636... | Big Ship CBO | 1397867851... | ? | -1 | en | PICTORIAL; |
| 38 | Jul 26, 2021 ... | 1419693274... | Big Ship CBO | 1397867851... | ? | -1 | en | RT @Manyunyu_Corg: We celebrated the #WorldMangro |
| 39 | Jul 26, 2021 ... | 1419689609... | George Maina | 4254734482 | ? | -1 | und | #WorldMangroveDay @Nature_Africa @KmfriResearch ( |
| 40 | Jul 26, 2021 ... | 1419686053... | Manyunyu Community Base... | 9428012321... | ? | -1 | en | We celebrated the #WorldMangroveDay2021 in Mirironi J |
| 41 | Jul 26, 2021 ... | 1419661837... | Kenya Projects | 1306126684... | ? | -1 | en | RT @WeDontHaveTime: The shores of Lake Victoria are |
| 42 | Jul 26, 2021 ... | 1419661466... | KMFRI | 9021390783... | ? | -1 | en | Earlier today, Senior Research Scientist at KMFRI Dr Juc |
| 43 | Jul 26, 2021 ... | 1419660924... | KMFRI | 9021390783... | ? | -1 | en | RT @WeDontHaveTime: The shores of Lake Victoria are |
| 44 | Jul 26, 2021 ... | 1419648506... | George Maina | 4254734482 | ? | -1 | und | #WorldMangrovesDay @NRT_Kenya @CarolineLumosi |
| 45 | Jul 26, 2021 ... | 1419643931... | AgamirQueen Fashions | 1253910014... | ? | -1 | en | RT @WeDontHaveTime: The shores of Lake Victoria are |
| 46 | Jul 26, 2021 ... | 1419619109... | Joel Swagman | 2807365255 | ? | -1 | en | RT @WeDontHaveTime: The shores of Lake Victoria are |

*Figure 5: RapidMiner Studio 9.5 extraction from tweets*

*Source: Research Data*

| Text |
|---|
| RT @KmfriResearch: Earlier today, Senior Research Scientist at KMFRI Dr Judith Okello following proceedings of the international day for th... |
| RT @Manyunyu_Corg: We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund togeth... |
| RT @BigShip_CBO: PICTORIAL; |
| Planting over 300 propagules,barefooted, sticking to the mud,is fun full, though tiring,but, knowing what it's actually doing to the environment,it's totall... |
| PICTORIAL; |
| RT @Manyunyu_Corg: We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund togeth... |
| #WorldMangroveDay @Nature_Africa @KmfriResearch @EmilyCLandis https://t.co/GIK8G4Te26 |
| We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund together with the Chief Admini... |
| RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and... |
| Earlier today, Senior Research Scientist at KMFRI Dr Judith Okello following proceedings of the international day for the conservation of the mangrov... |
| RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and... |
| #WorldMangrovesDay @NRT_Kenya @CarolineLumosi @NRT_Kenya @KmfriResearch https://t.co/bRAQTy8L7W |
| RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and... |
| RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and... |

*Figure 6: Top Tweets*

*Source Data: Research Data*

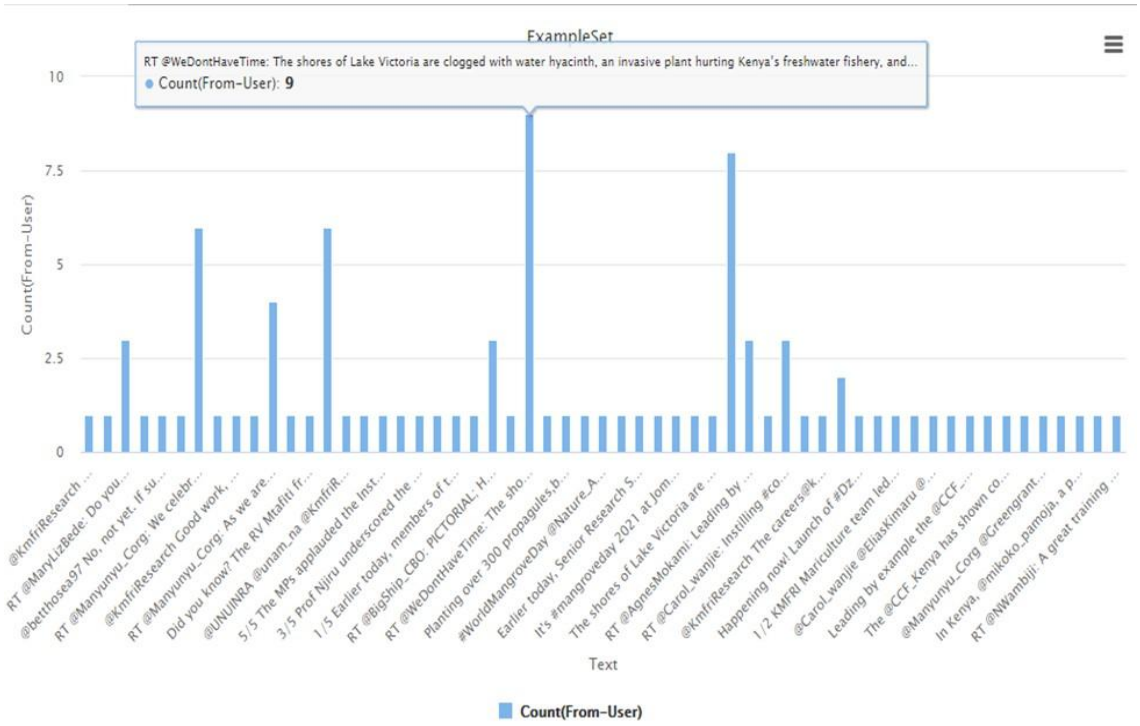



*Figure 7: Top User*

*Source: Research Data*

Figure 5 shows the user who tweeted the most on a certain topic. More information could be obtained on tweets like geo-location information, followers of a certain user and users a certain individual was following, date of joining twitter and website URL for an individual or company. The data sourced from the tweets provided great information on different topics that could be great sources of knowledge for decision-making, business intelligence, event prediction and analysis. RapidMiner Studio 9.5 as a tool was capable of reading, writing, analyzing and updating data in different formats like CSV, Excel, XML, MySQL, Access, Mails and Cloud sources like Amazon, Azure, Dropbox and Google Storage. Custom filters could be used to mine tweets examples could be retweets that are greater than a certain number, exporting tweets to a certain document like an excel document. The above twitter extension in rapid miner provided the researcher with

insights in the development of the proposed model, especially in mining unstructured data.

## 4.6 Specific Objective 2

Through the hands-on exposure from RapidMiner studio 9.5 and its tools, it was easier to design a suitable Text Mining Model, which could be used to retrieve explicit knowledge at KCDP. In the design of the text-mining model, the study narrowed down to three text-mining techniques namely information extraction and text summarization and topic modelling. Analyzed text were represented as topics and visualized as graphs to be consumed in the simplest way possible to determine the accuracy of the text mining analysis performed.

## 4.7 Text Mining Model

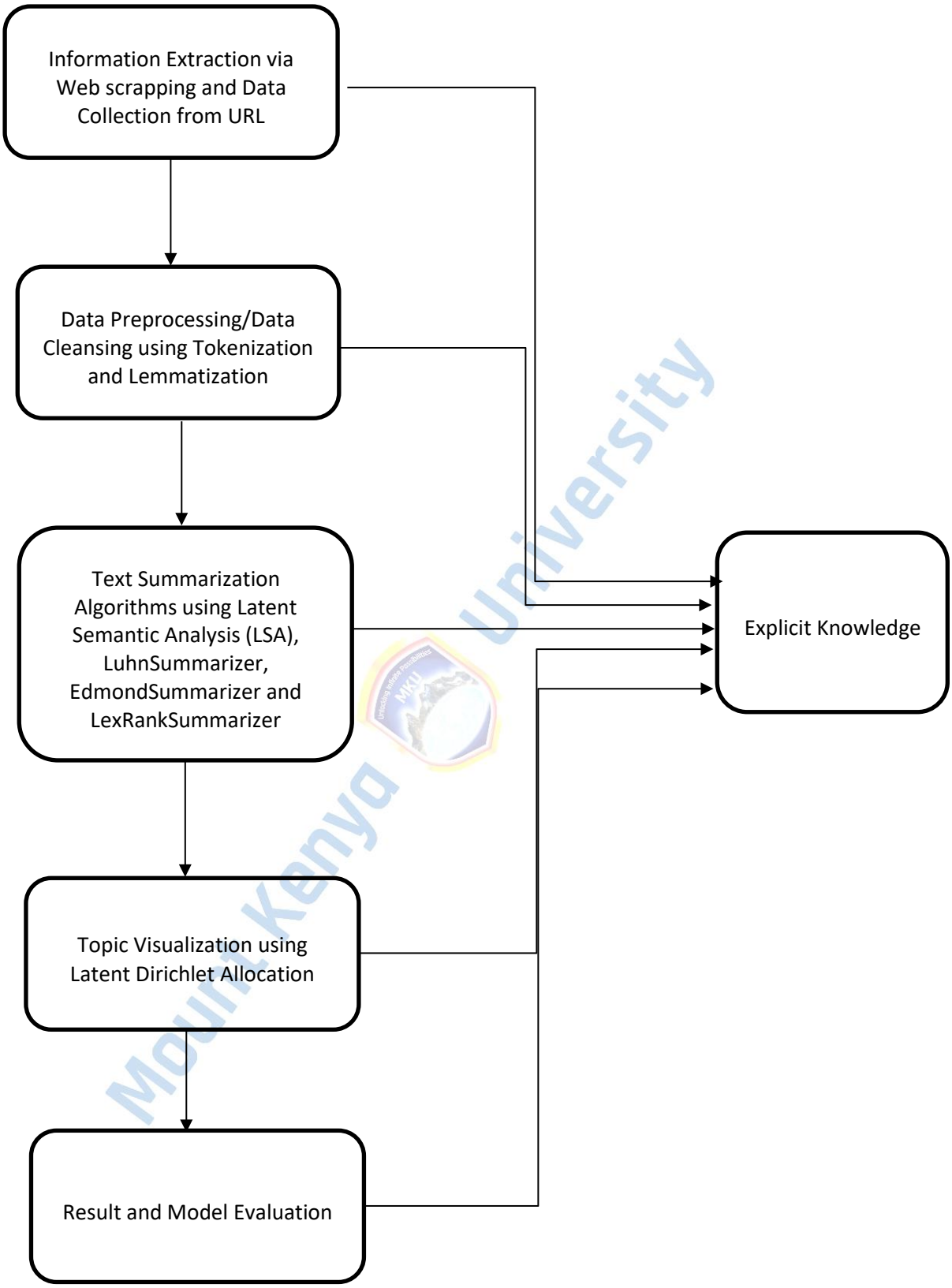



*Figure 8: Text Mining Model for Retrieval of Explicit Knowledge at KCDP*

*Source: Research Data*

## Framework Model Implementation

Information extraction and summarization are text-mining techniques widely used because of their flexibility and performance with different python text summarization algorithms like Latent Semantic Analysis (LSA), LuhnSummarizer, EdmondSummarizer and LexRankSummarizer. Topic modelling algorithms/libraries like Genism and Latent Dirichlet Allocation (LDA) models were used to extract topics from text analyzed.

The development of the model was dependent on the availability of data. Data was collected or mined using web scraping as a technique to get data in form of text from HTML and XML files. The procedure involved the installation of web scraping python libraries that included Python Request 2.22 and Beautiful Soup 3.

Data inform of text was web scraped from websites and portals of organizations that took part in the Kenya Coastal Development Project. Information extraction involved using web scrapping and URL screening tools to obtain information from different URL resources as shown below:

'https://www.kmfri.co.ke','http://kws.go.ke','http://www.kalro.org',

https://planning.go.ke' https://www.kefri.org/ http://www.nema.go.ke/ https://cda.go.ke/

http://www.kenyaforestservice.org/https://www.wiomsa.org/

Text obtained from these websites were from web pages, web links, research papers, reports, articles and policies. The text was saved in a kcdp.txt file for text analysis by the model developed.

Web scraping was the best method for information extraction since it was user friendly, faster and did not require much of human intervention only knowledge on how to install,

configure and use python web scrapping libraries and the process of information extraction. The web scrapping process is demonstrated in the code.

## Data extraction related to KCDP in the following websites:

'https://www.kmfri.co.ke',
'http://kws.go.ke',
'http://www.kalro.org',
'https://planning.go.ke'
https://www.kefri.org/
http://www.nema.go.ke/ https://cda.go.ke/
http://www.kenyaforestservice.org/ https://www.wiomsa.org/

Using Php based web crawler.

Algorithm;

For each site:
    Get site index:
        Get list of all links:
            Open link (cache when possible):
                Add link to list of scanned urls to avoid double scanning
                Check for content related to kcdp:
                    If it has such content:
                        Save content under <p> tag to file

```
<?php require
"vendor/autoload.php";

use PHPHtmlParser\Dom;

$links = ['https://www.kmfri.co.ke', 'http://kws.go.ke', 'http://www.kalro.org',
'https://planning.go.ke'];
file_put_contents(__DIR__ . "/data.txt", "");
```

```
function save($line)
{    if (trim($line) != "") {        file_put_contents(__DIR__ . "/data.txt", "------\n" .
$line . "\n", FILE_APPEND);
  }
}

if (!function_exists('dd')) {
  /**
* Dump the passed variables and end the script.
   *
* @param mixed
* @return void
   */
function dd()
  {
     array_map(function      ($value)     {                                    if
(class_exists(Symfony\Component\VarDumper\Dumper\CliDumper::class)) {
           $dumper = 'cli' === PHP_SAPI ?                new
Symfony\Component\VarDumper\Dumper\CliDumper :                new
Symfony\Component\VarDumper\Dumper\HtmlDumper;
           $dumper->dump((new Symfony\Component\VarDumper\Cloner\VarCloner)-
>cloneVar($value));
        } else {
var_dump($value);
        }
```

```
    }, func_get_args());
die(1);
  }
}

function parse(&$master, $base, $link)
{
  $dom = load($base, $link);
  $save = false;
  $keywords = ['kcdp', 'KENYA COASTAL DEVELOPMENT'];
  foreach ($dom->find('p') as $p) {       $text =
strtolower($p->text());       foreach ($keywords as
$word) {          if (strpos($text, strtolower($word))
!== false) {
            $save      =      true;
break;
        }
    }
  }

  foreach ($dom->find('p') as $p) {
    $href = $p->text();
    if         ($save)         {
save($href);
    }
  }
```

```
  foreach ($dom->find('a') as $a) {
$href = $a->getAttribute('href');
if ($href) {            if ($href[0] == '/')
{
            $href = rtrim($href, '/\\');
            $href = $base . $href;
        }           if ($href[0] != '#') {              if (!in_array($href, $master))
{                   if (parse_url($href, PHP_URL_HOST) == basename($base))
{
                    if (strpos($href, ".pdf") === false) {
echo "following $href\n";
$master[$href] = $href;                             try {
parse($master, $base, $href);
                        } catch (\Throwable $e) {
                        }
                    }
                }
            } else {
                $master[$href] = $href;
            }
        }
    }
  }
  return $master;
```

```
}

function load($base, $link)
{    echo "loading $link\n";    if
(!file_exists(__DIR__ . '/.cache')) {
mkdir(__DIR__ . '/.cache');
  }
  $dom = new Dom;
  if (file_exists(__DIR__ . '/.cache/' . md5($link))) {
     $dom->load(file_get_contents(__DIR__ . '/.cache/' . md5($link)));
  } else {
     $client = new \GuzzleHttp\Client(['base_uri' => $base]);
     $res = $client->get($link);        if ($res->getStatusCode() == 200)
{           $contents = $res->getBody()->getContents();
file_put_contents(__DIR__ . '/.cache/' . md5($link), $contents);
        $dom->load($contents);
     } else {
        $dom->load('');
     }
  }
  return $dom;
}

$refs = [];
```

## 4.7 Specific Objective 3

To evaluate the developed Text Mining model for retrieval of explicit knowledge at KCDP the process involved running data that had been web scrapped from the different websites through the developed model. Data used were two forms. The kcdp.txt or speech.txt file and data obtained from a URL i.e. https://www.capitalfm.co.ke/business/2019/04/full-speech-president-kenyatta-state-ofthe-nation-address-2019/. The kcdp.txt was text about the KCDP project, while the URL contained 14 page, 6775 text speech by President Kenyatta of Kenya on State of the Nation Address 2019. These two formed the data that was used as input to evaluate the model.

The model was designed and evaluated on the Google Colaboratory platform. “Colab” as popularly known, a free Jupyter notebook environment rich in computing resources mainly RAM, GPU and storage that required no setup and run entirely in the cloud. The researcher was able to write, execute, save, analyse the code and model as a whole from the Google Crome browser. The first step used in evaluting model was creating a notebook on the colab environment where code could be read, executed, saved and analyzed as mentioned earlier. The figures below describe the process of evaluating the model.

Data Collection

```
[ ]  from bs4 import BeautifulSoup
     from urllib.request import urlopen
     from nltk.tokenize import sent_tokenize
     import numpy as np
     import pandas as pd
     from nltk.corpus import stopwords
     from sklearn.metrics.pairwise import cosine_similarity
     import networkx as nx

     import nltk
     nltk.download('punkt')
     nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

*Figure 9: Data Collection*

*Source: Research Data*

The first step was collection of data from an uploaded file or URL. These involved the use of two python libraries Pyhton Request 2.22 and Beautiful Soup 3 to collect data from either a .txt file **or** URL, then open a file for reading.

```
# Read uploaded data
def read_text():
  with open('speech.txt', 'r') as f:
    dat = []
    txt = f.readlines()
    dat.append(txt)
    return dat

# Process the list
text = read_text()
for txt in text:
  print(len(txt))
```

```
455
```

*Figure 10: Code for uploading data*

*Source: Research Data*

Data was read from the uploaded resources using the code in Figure 6, where it showed the number of text in the speech.txt file that had 455 text, where a list containing text was then generated and processed.

```
# Loading the actual text
sentences = []
for s in text:
    sentences.append(s)

sentences = [y for x in sentences for y in x]
sentences[30:40]
```

```
['Mr. Speaker,\n',
 '\n',
 'Following the Country's first General Election under the New Constitution, I took the Oath of Office as the first P
 '\n',
 'My first term laid the foundation for a better Kenya by building on the promise and aspirations of the new Constitu
 '\n',
 'Mr. Speaker,\n',
 '\n',
 'The National Values and Principles of Governance epitomize the Vision that Kenyans have for their Nation.  The fort
 '\n']
```

*Figure 11: Loading actual text*

*Source: Research Data*

Actual text were loaded as shown in Figure 7 from the uploaded speech.txt file and a condition provided to summarize the text between 30: 40 sentences.

Data Preprocessing

```
# remove punctuations, numbers and special characters
clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ")

# change to lowercase
clean_sentences = [s.lower() for s in clean_sentences]
stop_words = stopwords.words('english')
# function to remove stopwords
def remove_stopwords(sen):
    sen_new = " ".join([i for i in sen if i not in stop_words])
    return sen_new
clean_text = [remove_stopwords(r.split()) for r in clean_sentences]
len(clean_sentences)
# clean_sentences
```

```
455
```

*Figure 12: Data preprocessing*

*Source: Research Data*

Figure 11 shows the data preprocessing process that involved cleansing of the uploaded text where tokenization and lemmatization were performed. The preprocessing process involved the structured representation of original text to reduce dimensionality, inflectional endings and relationship of words by eliminating stop words like a and the, and removal of punctuation marks.

The elimination of the stop words and punctuation marks is what is referred to as tokenization. The other preprocessing process was lemmatization. This process involved the representation of a word to its basic form. I.e. a word like Corruption, Corrupted, and Corrupting were represented to their basic dictionary form, which is Corrupt. The basic dictionary form is what is known as Lemma.

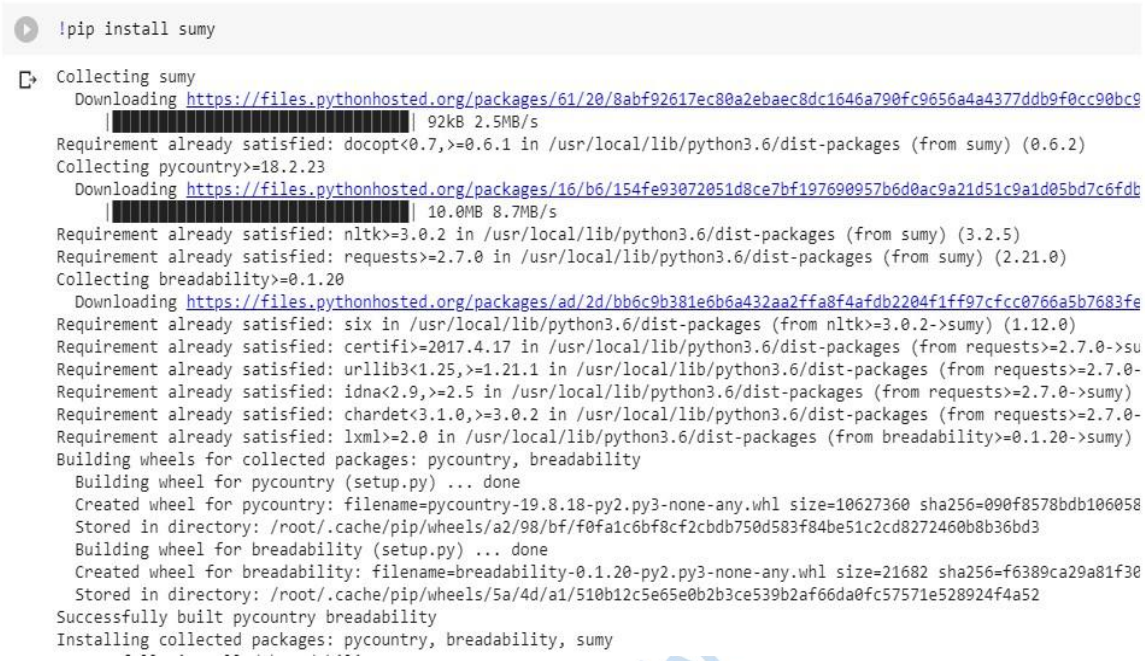



*Figure 13: Installation of Sumy library*

*Source: Research Data*

The preprocessing process was followed by the installation of sumy library, a python library used for extracting plaint text and HTML pages. In simple words, sumy was a text analysis library that provided the study with the ability to use different text summarization algorithms like Luhn, Latent Semantic Analysis, Edmondson and LexRank.

```
[ ] import nltk
    nltk.download('punkt')
    nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from __future__ import absolute_import
from __future__ import division, print_function, unicode_literals

from sumy.parsers.html import HtmlParser
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.lsa import LsaSummarizer
from sumy.nlp.stemmers import Stemmer
from sumy.utils import get_stop_words
from sumy.summarizers.luhn import LuhnSummarizer
from sumy.summarizers.edmundson import EdmundsonSummarizer
from sumy.summarizers.lex_rank import LexRankSummarizer
```

*Figure 14: Importing of text for analysis*

*Source: Research Data*

Figure 13 shows the process of importing text into the different text summarization algorithms namely LsaSummarizer, LuhnSummarizer, EdmondSummarizer and LexRankSummarizer for the text summarization processes to take place.

```
[ ]  nt ("--LsaSummarizer--")
     marizer = LsaSummarizer()
     marizer = LsaSummarizer(Stemmer(LANGUAGE))
     marizer.stop_words = get_stop_words(LANGUAGE)
      sentence in summarizer(parser.document, SENTENCES_COUNT):
      print(sentence,'\n')

     --LsaSummarizer--
     My first term laid the foundation for a better Kenya by building on the promise and aspirations of the new Constituti

     Leading the string of innovators is Roy Allela who garnered global accolades for inventing smart gloves that convert

     The consideration and approval by Parliament of various Protocols, Treaties and Sessional Papers continue to enhance

     This Exercise, together with the National Integrated Identity Management System (NIIMS), will ensure that all persons

     His Excellency Yoweri Kaguta Museveni, the President of the Republic of Uganda and a Great Statesman and Pan-Africani

     As an island of peace in a conflict-prone and fragile region, Kenya nevertheless faces challenges from transnational

     Key to this is continuing to strengthen our legal tools against these groups so that they are unable to take advantag

     Corruption and Impunity create social distortions and divisions, fuel inequity and poverty, destroy the fabric of soc

     This they did during the National Anti-Corruption Conference held in January this year, where they tasked me, the Spe

     That is why we look to the Judiciary to do their part, to apply the law firmly and fairly; and for Parliament to upho
```

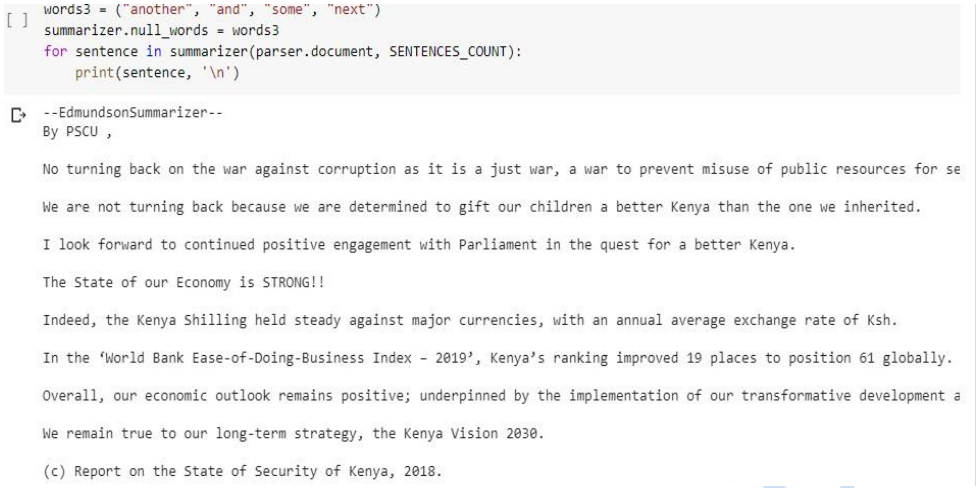
*Figure 15: Text Summarization by LsaSummarizer*

*Source: Research Data*

```
[ ]  print ("--LuhnSummarizer--")
     summarizer = LuhnSummarizer()
     summarizer = LuhnSummarizer(Stemmer(LANGUAGE))
     summarizer.stop_words = ("I", "am", "the", "you", "are", "me", "is", "than", "that", "this",)
     for sentence in summarizer(parser.document, SENTENCES_COUNT):
         print(sentence, '\n')

     --LuhnSummarizer--
     In accordance with Article 132 of the Constitution, I am honoured to report to Parliament the measures taken and prog

     My first term laid the foundation for a better Kenya by building on the promise and aspirations of the new Constituti

     On behalf of a grateful Nation, I thank all of those Men and Women who serve the Republic in whatever capacity, who u

     Devolution has received the full and firm support of my Administration, and, together with an enabling and supportive

     The consideration and approval by Parliament of various Protocols, Treaties and Sessional Papers continue to enhance

     My Administration has spearheaded the implementation of various environmental initiatives including: Interventions fo

     We do so conscious of the fact that fidelity to international law and commitment to our international obligations is

     Kenya's election to the AU Peace and Security Council in 2019 and our strategic decision to vie for a non-permanent s

     That is why we look to the Judiciary to do their part, to apply the law firmly and fairly; and for Parliament to upho

     In saying this, I do not presume to direct the Judiciary or Parliament, that is certainly not my constitutional place
```

*Figure 16: Text summarized by Luhnsummarizer, a text analysis algorithm*

*Source: Research Data*



```
[ ]  words3 = ("another", "and", "some", "next")
     summarizer.null_words = words3
     for sentence in summarizer(parser.document, SENTENCES_COUNT):
         print(sentence, '\n')

     --EdmundsonSummarizer--
     By PSCU ,

     No turning back on the war against corruption as it is a just war, a war to prevent misuse of public resources for se

     We are not turning back because we are determined to gift our children a better Kenya than the one we inherited.

     I look forward to continued positive engagement with Parliament in the quest for a better Kenya.

     The State of our Economy is STRONG!!

     Indeed, the Kenya Shilling held steady against major currencies, with an annual average exchange rate of Ksh.

     In the 'World Bank Ease-of-Doing-Business Index – 2019', Kenya's ranking improved 19 places to position 61 globally.

     Overall, our economic outlook remains positive; underpinned by the implementation of our transformative development a

     We remain true to our long-term strategy, the Kenya Vision 2030.

     (c) Report on the State of Security of Kenya, 2018.
```

*Figure 17: Text summarised by EdmondsonSummariser*

*Source: Research Data*

## 4.8 Results and Discussion

### 4.8.1 Rate of Accuracy in Summarization Using Algorithms

The uploaded speech.txt file contained 6,775 text. After preprocessing and cleansing the total text was reduced to 455. The different algorithms performed well where text were analyzed into a single page and information captured in each was meaningful. To ensure that the algorithms can be applied in different portals, text was mined from various websites and portals. All results were amazing and made a lot of meaning as per the algorithm.

Luhnsummarizer text algorithm analyzed the text to 470, Lsansummarizer to 370 and EdmondSummarizer to 149 text. Overall EdmondSummarizer algorithm performed the best by summarizing the text in the best way possible and capturing the important topics of discussion as in form of the analyzed text below. In Kenya, the issue of corruption has always been at the forefront. A war on corruption has been declared to curb the vice, which has taken the country hostage. Using the EdmondSummarizer this topic was evident and was put into consideration to address it as shown in the president's speech.

The other algorithms equally captured important information. The importance of information depends on the recipients' point of view having in mind that the society has different views, opinions and expectations.

The uploaded text from the UN environment that works with the KCDP project from its website 'https: //www.unenvironment.org/news-and-stories/press-release/improvedclimate-action-food-systems-can-deliver-20-percent-global' file contained 8,575 words. After preprocessing and cleansing the total text was reduced to 386 words by LexRankSummariser, 391 words by LuhnSummariser, 257 words by LsaSummariser, and 131 words by EdmondSummarizer.The different algorithms performed well where text were analyzed into a single page and information captured in each was meaningful.

**By EdmondSummarizer**

No turning back on the war against corruption as it is a just war, a war to prevent misuse of public resources for selfish interests by those we have entrusted to manage them. We are not turning back because we are determined to gift our children a better Kenya than the one we inherited. I look forward to continued positive engagement with Parliament in the quest for a better Kenya. The State of our Economy is STRONG!!

Indeed, the Kenya Shilling held steady against major currencies, with an annual average exchange rate of Ksh. 101 to the US dollar. In the 'World Bank Ease-of-Doing-Business Index – 2019', Kenya's ranking improved 19 places to position 61 globally.
Overall, our economic outlook remains positive; underpinned by the implementation of our transformative development agenda. We remain true to our long-term strategy, the

Kenya Vision 2030. (c) Report on the State of Security of Kenya, 2019.

**By EdmondSummarizer – from website**

Nairobi 1 September 2020 policymakers improve chance achieving climate goal limiting global warming $15^0$c making specific commitment transforming national food system enhancing nationally determined contribution nods food system new report published today wwf un environment programme unep eat climate focus find country missing significant opportunity reduce greenhouse gas emission identifies 16 way policymakers could take action farm fork currently diet food loss waste widely ignored adding national climate plan policymakers improve mitigation adaptation contribution food system much 25 percent 2015 Paris agreement country expected revise resubmit nods every five year therefore policymakers opportunity adopt food system solution set ambitious target measure reduce greenhouse gas emission turn improve biodiversity food security public health food system – gather element activity relate production processing distribution preparation consumption food – account 37 percent greenhouse'.

The results from the model determined the rate of accuracy in information extracted by validating the data and determining its content through the summarized information processed that was able to give meaning.

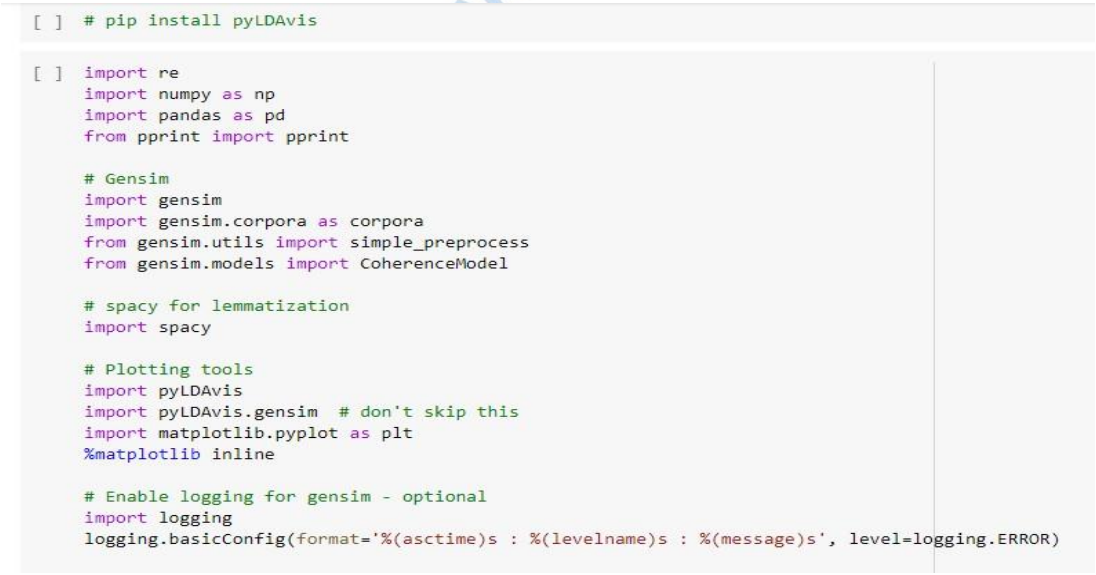
## 4.9 Validation of Data Collected and Rate of Similarity

The development of the model captured additional functionalities on the text that was analyzed. This involved capturing the key topics in text through the process of topic modelling. According to Jacobi, Attevedlt & Welbers (2016), topic modelling also

known as information visualization in Natural Language Processing, is a statistical model in machine learning that creates topics based on patterns, co-occurrence of words in text that have been analyzed.

Topic modeling is mostly used in text mining for discovery of hidden semantic structures in a text body. The algorithm used for topic modelling was Latent Dirichlet Allocation (LDA). LDA was used together with the Genism python library to perform topic modelling in the study.

Topic modelling was performed on text obtained from the kdcp.txt file. The file contained text that had been analyzed. After analysis text were grouped into words and keywords, where topic modeling was performed based on the number of times a word reappeared. Topics were modelled and visualized as shown in the figure 14 process.



```
[ ]  # pip install pyLDAvis

[ ]  import re
     import numpy as np
     import pandas as pd
     from pprint import pprint

     # Gensim
     import gensim
     import gensim.corpora as corpora
     from gensim.utils import simple_preprocess
     from gensim.models import CoherenceModel

     # spacy for lemmatization
     import spacy

     # Plotting tools
     import pyLDAvis
     import pyLDAvis.gensim  # don't skip this
     import matplotlib.pyplot as plt
     %matplotlib inline

     # Enable logging for gensim - optional
     import logging
     logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.ERROR)
```

*Figure 18: Installation of Topical modelling*

*Source: Research Data*

Python libraries pyLDAvis and gensim were used for topic modelling. First, this involved the installation of the Python library for interactive topic model visualization pyLDAvis, which was used to help users interpret topics in the topic model from text data. Gensim was used to eliminate words that occurred frequently together, two, three or four times in the corpus. These words like money_laundery, which is termed a bigram and motherin-law a trigram and words that occur more than three times frequently together as quadgrams. Similar to text analysis, lemmatization was done for data cleansing and tools for plotting the required visual graphs installed as shown in Figure 14.

```
[ ]  # Build LDA model
     lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                                 id2word=id2word,
                                                 num_topics=20,
                                                 random_state=100,
                                                 update_every=1,
                                                 chunksize=100,
                                                 passes=10,
                                                 alpha='auto',
                                                 per_word_topics=True)

[ ]  # Print the Keyword in the 10 topics
     pprint(lda_model.print_topics())
     doc_lda = lda_model[corpus]

     [(0,
       '0.009*"county" + 0.009*"kmfri" + 0.007*"fishery" + 0.004*"research" + '
       '0.003*"community" + 0.003*"project" + 0.003*"coastal" + 0.003*"management" '
       '+ 0.003*"kenya" + 0.002*"hold"'),
      (1,
       '0.004*"county" + 0.004*"kmfri" + 0.004*"fishery" + 0.002*"research" + '
       '0.002*"coastal" + 0.002*"management" + 0.002*"development" + '
       '0.002*"project" + 0.002*"community" + 0.002*"include"'),
      (2,
       '0.005*"county" + 0.005*"fishery" + 0.003*"kmfri" + 0.003*"research" + '
       '0.002*"project" + 0.002*"coastal" + 0.002*"management" + '
       '0.002*"development" + 0.002*"include" + 0.002*"committee"'),
```

*Figure 19: Model Building*

*Source: Research Data*

The topic model was then built using the python libraries and the LDA topic-modelling algorithm. This involved the printing of keywords to represent the topics from the corpus that was to be visualized.

```
[ ] # Compute Perplexity
    print('\nPerplexity: ', lda_model.log_perplexity(corpus))  # a measure of how good the model is. lower the better.

    # Compute Coherence Score
    coherence_model_lda = CoherenceModel(model=lda_model, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
    coherence_lda = coherence_model_lda.get_coherence()
    print('\nCoherence Score: ', coherence_lda)
```

```
Perplexity:  -6.045598528342255

Coherence Score:  0.32430207732963356
```

```
[ ] # Visualize the topics
    pyLDAvis.enable_notebook()
    vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
    vis
```

```
/usr/local/lib/python3.6/dist-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concatenation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.
```

*Figure 20: Computing perplexity*

*Source: Research Data*

## 4.10 Evaluating the Model Using Perplexity

Figure 20 shows how perplexity was computed. Perplexity is a statistical measure of how well a probability model would predict a topic from unforeseen data. According Jurafsky (2012), the lower the perplexity value of a model the better its performance in predicting an occurrence or a number occurrences' of words. The model applied an algorithm that used the perplexity formulae below:

$$Perplexity = N\sqrt{\frac{1}{w w(_{1\,2}...w_N)}}$$

Where N represented the number of words. The perplexity of the model was -6.0455.

```
[ ] # Visualize the topics
    pyLDAvis.enable_notebook()
    vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
    vis
```

```
/usr/local/lib/python3.6/dist-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concatenation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.
```

*Figure 21: Topical modelling statistically*

*Source: Research Data*

## 4.11 Number of Relevant Clusters Created as Topics

Figure 21. Shows the process of topic modelling using the pyLDAvis library to represent the keywords inform of topics that could be visualized in statistical formats like bar and line graphs, histograms and pie charts just to mention but a few.

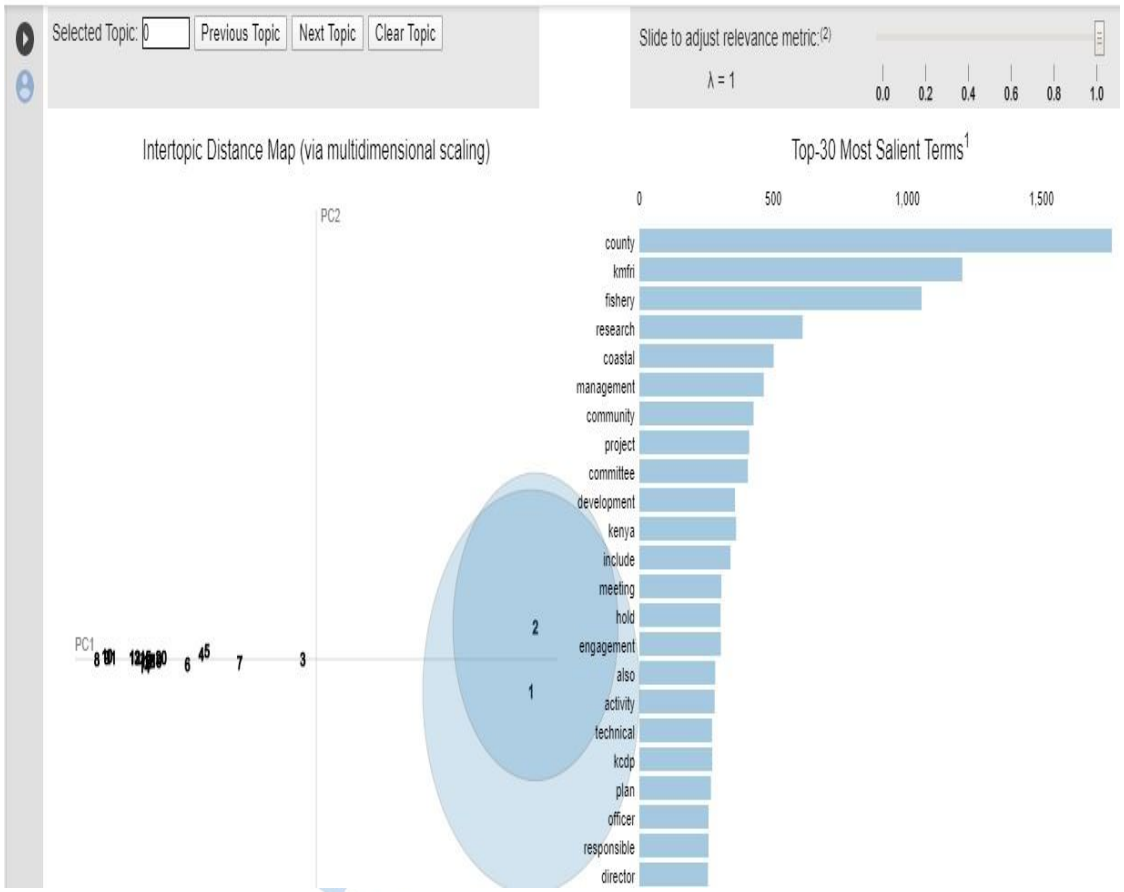



*Figure 22: Intertopic distance map*

*Source: Research Data*

Figure 22 shows a distribution of the keywords that made the topics, inform of statistics. The modelled data was web scrapped from portals and websites of organizations that took part in the KCDP project. A file that contained the text data with the name kcdp.txt was analyzed and then modelled as outlined in Figure 18. The distribution of the topics shows the most important topic to the least important topic. The visualized results show the relationship among the different topics that were key in the study. This can be used

to provide meaningful information. For instance, text analyzed from the KCDP portals showed that KMFRI and the county governments had a good working relationship in achieving the objective of the project. In addition to this, the results showed that there was more concentration on fishery research, ignoring other areas like mining and agriculture that could improve the livelihood of the coastal people.



# CHAPTER 5: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

## 5.0 Summary

The uploaded speech.txt file contained 6,775 text. After preprocessing and cleansing, the total text was summarized using the developed algorithms and the Edmondsummariser reduced the document to 455 characters. The different algorithms performed well where

text were analyzed into a single page and information captured in each was meaningful. The perplexity value achieved in accuracy of the model was

-6.0455 and this shows that a combination of algorithms improves the accuracy of a model.

The results from the tweets (unstructured format) showed additional information that included the user, user-id, text, polarity of the captured text to consider if the text analyzed was positive or negative, objective or subjective and finally geo-location information.

## 5.1. Findings of the Study

The study had three specific objectives and one general objective. The findings from the study show how each objective was achieved.

**Specific Objective 1**

To analyze how Text Mining techniques could help in retrieving explicit knowledge at KCDP. The objective was achieved through mining of tweets from websites that took part in the KCDP website. The findings in figure 5 showed the user who tweeted the most on a certain topic. More information was obtained on tweets like geo-location information, followers of a certain user and users a certain individual was following, date of joining twitter and website URL for an individual or company. The data sourced from the tweets provided great information on different topics that could be used to make decision, business intelligence, event prediction and analysis. From the findings, the study concluded that text mining from both structured and unstructured platforms can be retrieved and provide meaningful information for decision making.

**Specific Objective 2**

**To design a Text Mining model for retrieval of explicit knowledge at KCDP.** The objective was achieved by employing data collection techniques that involved involved collection of data through web scrapping. The data collected was in diffrent formats that included HTML and XML. Through pyton libraries installed, text collected was run through the model that was coded and the findings was a prove that the model could be applied in retrieving of text from the various websites and data bases of all the projects that participated in the project to give knowledge and information that could guide the managers in improving and achieving the goals of the project.

**Specific Objective 3.**

**To validate the developed Text Mining model for retrieval of explicit knowledge at KCDP**.

To validate the model, the study applied a tested platform in the google coloborator that used a free Jupyter notebook environment. Code was written, executed, saved and analysed in google chrome browser.The study was able to write, execute, save, analyze the code and model as a whole from the Google Chrome browser. The findings showed information processed and summarized to give meaning. The summarized text was run concurrently in the model and the different algorithms gave out summaries that were

## 5.2 Conclusion

The study's main objective was to establish how a Text Mining model could be used in explicit knowledge retrieval at KCDP. The study identified text-mining techniques that could be used to develop a text-mining model, evaluate the model to be able to retrieve explicit knowledge at KCDP. The main objective has been achieved fully by these findings and results from the collected text that was summarized using text summarization algorithms used in the models. After summarization topic, modelling was

performed on the text collected using Latent Dirichlet Allocation (LDA) topic-modelling algorithm to create topics based on patterns in text. The model was then evaluated to establish its performance by measuring the four variables identified using precision and recall to measure accuracy, topic modelling to measure rate of similarity, and perplexity to measure evaluation of the model which gave a perplexity of -6.0455 from the text analyzed and modelled. It was concluded that text analysis could be used to analyze text and create explicit knowledge from both structured and unstructured data formats using the mode.

Text mining is a new area of knowledge that originates from data mining where more research needs to be done. The research study established that Text Mining could be used in the retrieval of explicit knowledge. The developed model can be used in the retrieval of both structured and unstructured data, in text format by organizations that values knowledge retention and sharing. At this era of data and information, knowledge consumed depends on the availability of data. Data needs to be available to the intended users at all times. The available data should be accurate, consistent and factual to enable users acquire the best knowledge. This shall lead to knowledge transfer and better decision making across all cadres of an organization.

Organizations need to preserve data in the best way possible and have policies that protect the storage, access, use, transfer and archiving of data. Data has to be stored securely by enforcing on the pillars of security of Confidentiality, Integrity and Availability. In the study, data that was web scrapped from websites and portals from organizations that took part in the KCDP project was not of the required standard for the study to be able to perform accurate text analysis; other sources of data were used. The organizations' policies for the KCDP project were against access of data from databases, e-mails, documents and cloud platforms, which posed a challenge for the study. The study had to

resort to web scrapping as the main technique for data collection using open source python tools.

Web scrapping of websites and portals of organizations was considered illegal. These challenges were addressed by getting alternative sources of data from URL's or HTML formats. Data from the websites or portals of organizations that took part in the KCDP project was used in text modelling to evaluate the accuracy of the model.

The study demonstrated through the model that text mining could be used to retrieve explicit knowledge from both structured and unstructured text in an organization. This was dependent on the availability and accuracy of data, a clear understanding of the text mining process to develop a centralized Knowledge Management System that could store explicit knowledge.

## 5.3 Contribution of the Study

The model has combined four algorithms and text mining models to test performance in retrieval of explicit knowledge. The results from the study model proved that a combination of models gives better and improved accuracy when compared with the other models that applied one single algorithm in measuring the performance. The best model in the literature reviewed of 2020 gave a perplexity measure of 0.95 while the study model gave a perplexity measure of -6.0455 making it a better model. To determine the accuracy of information that was extracted both the precision and recall metrics were used.

The performance of topic modelling was determined by using an intrinsic evaluation metric, the perplexity metric, as covered under the results and discussion section in this study. Through the validation performed on the model, the results showed a better performance compared to earlier models that were unable to give visualization from both structured and unstructured text.

According to Khachatryan & Muehlmann (2020) using probability to determine perplexity in dominant presence of topic II in both claims and specification the results were relatively high as per the formula below:

$$1 - \frac{1}{\sqrt{2}}\sqrt{(0.78 - 0.83)^2 + (0.22 - 0.17)^2} \approx 0.95$$

The perplexity as per the results was 0.95 which is relatively high and opposed to performance of a good model that should be lower compared with the study model that gave a perplexity of – 0.60455 which is very low making it a better model using the formulae below:

$$Perplexity \square N\sqrt{\frac{1}{p\ w\ w(_{1\,2}...w_N)}}$$

Where N represented the number of words. The perplexity of the model was -6.0455.
The results showed improved accuracy of the model compared to earlier models that used single algorithms to determine the accuracy of the model, hence combined algorithms have proved to perform better and give the best accuracy.

## 5.4 Recommendations

Future models should incorporate artificial intelligence into machine learning, so that semantics (i.e., English grammar) are deciphered and not only syntax of the language.

The system should be willing to differentiate between “willing flesh” and “good meat”.

The system should detect the intrinsic difference between the phrases “weak spirit” and “bad liquor”. This will help the system to avoid getting lost in translation via the use of synonyms and will incrementally rely on semantic, as facilitated by artificial intelligence. These interventions would hopefully make text retrieval more accurate in the long run. The study also recommended further research in the area of text mining since it was still new in the area of data science and analytics. Most research done in text mining has been in explicit knowledge specifically from structured data.

From the research results and findings, the study noted that text in Arabic and Swahili could not be retrieved nor analyzed, though the text carried a large amount of unstructured data in the field covered by the study.

Analysis of such text could give meaningful results. This advocated for text mining models, techniques, algorithm that could be used in the analysis of text from different languages.

This gap therefore opens doors for scholars in the field of text mining to bridge the gap by conducting more research on techniques that can retrieve and analyze this data (multilingual).

## REFERENCES


Abzari, M., & Teimouri, H. (2008). The effective factors on knowledge sharing in organizations, *The International Journal of Knowledge, Culture and Change Management* 8(2), 105-13.

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*, 1847979019890771.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* 22(2), 207-216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, *12*(1), 307-328.

Aich, S., Sain, M., Park, J., Choi, K. W., & Kim, H. C. (2017, November). A Text Mining approach to identify the relationship between gait-Parkinson's diseases (PD) from PD based research articles. In *Inventive Computing and Informatics (ICICI), International Conference on* (pp. 481-485). New Jersey: IEEE.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics articles. In *Inventive Computing and Informatics (ICICI), International Conference* association rules. *Advances in knowledge discovery and data mining*, *12*(1), 307-328.

Baillie, L. (2019). Exchanging focus groups for individual interviews during qualitative data collection: a discussion. *Nurse researcher*, *27*(2).

Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*.

Bianchini, M., Frasconi, P., Gori, M., & Maggini, M. (1998). Optimal learning in artificial neural networks: A theoretical view. *Neural network systems techniques and applications*, 1-51.

Bjerva, J. (2016). Byte-based language identification with deep convolutional networks. *arXiv preprint arXiv:1609.09004*.

Bolasco, S. (2005). Statistical textual e-Text mining: some application paradigms Brown, J. S., &Duguid, P. (2001). The social life of information. *Harvard Educational Review*, *71*(1), 151-152.

Campos, J. R. P., Otero, P. G., & Loinaz, I. A. (2020). Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, *26*(4), 433-454.

Chandrika, G. N., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Web scraping for unstructured data over web. In *Embedded Systems and Artificial Intelligence* (pp. 853-859). Springer, Singapore.

Conger, S. (2015). Knowledge management for information and communications technologies for development programs in South Africa. *Information Technology for Development*, *21*(1), 113-134.

Desai, A. (2015). A review on knowledge discovery using text classification techniques in text mining. *International Journal of Computer Applications*, *111*(6).

Dunham, H. M. (2000). Data mining techniques and algorithms partial draft of forthcoming book from prentice hall.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82.

Faust, B. (2007, June). Implementation of tacit knowledge preservation and transfer methods. In *International Conference on Knowledge Management in Nuclear Facilities* (pp. 18-21).

Feldman, R., & Sanger, J. (2007). *The text-mining handbook: advanced approaches in analyzing unstructured data*. Newcastle: Cambridge University Press.

Fink, A. (2003). *How to sample in surveys* (Vol. 7). Sage.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text Mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Gamble, P. R., & Blackwell, J. (2001). *Knowledge management: A state of the art guide*. London Kogan Page Publishers.

Gamallo, P., Campos, J. R. P., & Alegria, I. (2017, April). A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)* (pp. 109-114).

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, Florida Chapman and Hall/CRC.

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011, October). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-4). IEEE.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1995). *Markov chain Monte Carlo in practice*. Boca Raton, Florida: CRC press.

Greiner, M. E., Böhmann, T., & Krcmar, H. (2007). A strategy for knowledge management. *Journal of Knowledge Management*, *11*(6), 3-15.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Gupta, V., & Sharma, S. K. (2014). A Survey: Issues, Challenges and Tools for Analytics of Data from Different Social Media Sources.

Gupta, V., & Lehal, G. S. (2009). A survey of Text Mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Hajizadeh, E., Ardakani, H. D., & Shahrabi, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, *2*(7), 109-118.

Halisah, A., Jayasingam, S., Ramayah, T., & Popa, S. (2021). Social dilemmas in knowledge sharing: an examination of the interplay between knowledge sharing culture and performance climate. *Journal of Knowledge Management*.

Heckerman, D. (1997). Bayesian networks for data mining. *Data mining and Knowledge discovery*, *1*(1), 79-119.

Ichise, R., Takeda, H., Koen, S., & Muraki, T. (2006, December). A mining method of communities keeping tacit knowledge. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on* (pp. 709-713). New York City IEEE.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), 89-106.

Jandhyala, S., &Phene, A. (2015). The role of intergovernmental organizations in cross-border knowledge transfer and innovation. *Administrative Science Quarterly*, *60*(4), 712-743.

Joia, L. A., & Lemos, B. (2010). Relevant factors for tacit knowledge transfer within organisations. *Journal of Knowledge Management*, *14*(3), 410-427.

Jurafsky, D., & Manning, C. (2012). Natural language processing. Instructor, 212(998), 3482.

Kadu, P., & Ashwini, V. Z. (2020). Knowledge Extraction from Text Document Using Open Information Extraction Technique. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) Volume*, *9*.

Kamimura, R. (2014, December). Explicit knowledge extraction in information-theoretic Supervised multi-layered SOM. In *Foundations of Computational Intelligence (FOCI), 2014 IEEE Symposium on* (pp. 78-83). New York City IEEE

Karanikas, H., & Theodoulidis, B. (2002). Knowledge discovery in text and Text Mining software. *Centre for Research in Information Management, Department of Computation*. 13 (4), 442-450

Kayser, V., & Shala, E. (2020). Scenario development using web mining for outlining technology futures. *Technological Forecasting and Social Change*, *156*, 120086. 30 June 2021

Khachatryan, D., & Muehlmann, B. (2020). Measuring the drafting alignment of patent documents using text mining. *Plos one*, *15*(7), e0234618.

Khan, S., Rani, U., Prasad, B. V. N., Srivastava, A. K., Selvi, S., & Gautam, D. K. (2015, March). Document management system: An explicit knowledge management system. In *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on* (pp. 402-405). New York City: IEEE.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. Organizational research methods, 21(3), 733-765.

Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85. *learning analytics and knowledge* (pp. 267-270). New York ACM.

Lepore, D., Dubbini, S., Micozzi, A., & Spigarelli, F. (2021). Knowledge sharing opportunities for industry 4.0 firms. *Journal of the Knowledge Economy*, 1-20.

Lichtenstein, S., & Hunter, A. (2006). Toward a receiver-based theory of knowledge sharing. International Journal of Knowledge Management, 2(1), 24-40.

Mahamune, M., & Ingle, S. (2014). Application of Data Mining Techniques in Knowledge Management System. ISBN 6(2)8-21.

Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., & Sundarsekar, R. (2017). Big data knowledge system in healthcare. In Internet of things and big data technologies for next generation healthcare (pp. 133-157). Springer, Cham.

Meghji, A. F., Mahoto, N. A., Unar, M. A., & Shaikh, M. A. (2020). The Role of Knowledge Management and Data Mining in Improving Educational Practices and the Learning Infrastructure. *Mehran University Research Journal of Engineering and Technology*, *39*(2), 310-323.

Morgan, D. L. (1997). *The focus group guidebook* (Vol. 1). Newburg Park: Sage Publications Inc.

Mushtaq, H., Malik, B. H., Shah, S. A., Bin Siddique, U., Shahzad, M., & Siddique, I. (2018). Implicit and Explicit Knowledge Mining of Crowdsourced Communities: Architectural and Technology Verdicts. *International Journal of Advanced Computer Science and Applications*, *9*(1), 105-111.

Mugenda, O. M., & Mugenda, A. G. (2003). *Research methods quantitative and qualitative approaches*.Nairobi: African Center for Technology Studies (ACTS) Press.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science*, *5*(1), 14-37.

Oates, B. J. (2005). Researching information systems and computing. New York: Sage.

Odigwe, F. N., Bassey, B. A., & Owan, V. J. (2020). Data management practices and educational research effectiveness of university lecturers in South-South Nigeria. *Odigwe, FN, Bassey, BA, & Owan, VJ (2020). Data management practices and educational research effectiveness of university lecturers in South-South Nigeria. Journal of Educational and Social Research (JESR)*, *10*(3), 24-34.

Piatetsky-Shapiro, G. (1990). Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. *AI magazine*, *11*(4), 68-68.

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In Multi-source, multilingual information extraction and summarization (pp. 23-49). Springer, Berlin, Heidelberg.

Polanyi, M. (1966). The logic of explicit inference. *Philosophy*, *41*(155), 1-18.

Preece, J., Rogers, Y., & Sharp, H. (2002) Interaction design: Beyond human-computer Interaction. New York: John Wiley & Sons, Inc.

Rajman, I., Desante, K., Hatcher, B., Hemingway, J., Lachno, R., Brooks, S., & Turik,

M. (1997). LY303366 single intravenous dose pharmacokinetics and safety in healthy volunteers. In *Program and abstracts of the 37th interscience conference on antimicrobial agents and chemotherapy. Abstract no. F-74.*

Ramírez-Gallego, S., Fernández, A., García, S., Chen, M., & Herrera, F. (2018). Big data: tutorial reliability. *The Qualitative Report, 16*(3), 730-744.

Richard, J. (2004). *A competitive advantage through knowledge creation*, Davenport: St. Ambrose University.

Rumanti, A. A., Samadhi, T. A., & Wiratmadja, I. I. (2016, December). Impact of tacit and explicit knowledge on knowledge sharing at Indonesian Small and Medium Enterprise. In *Industrial Engineering and Engineering Management (IEEM), 2016 IEEE International Conference on* (pp. 11-15). New York: IEEE.

Saatcioglu, O. Y., Ozmen, O. N., & Eriş, E. D. (2012). A study on knowledge management and firm performance in Turkish IT sector. *International Journal of Logistics Systems and Management*, *11*(2), 213-231.

Sağsan, M. (2006, July). A new life cycle model for processing of knowledge management. In *2nd International Congress of Business, Management and Economics* (pp. 15-18). San Francisco: Academia.edu

Shah, M., Shinde, S., Sawant, R. S., &Wagh, P. (2017). Analysis of Text Review using Hybrid Classifier. *International Journal of Engineering Science*, *10914*.

. *International Journal of Geographical Information Science*, *30*(9), 1687-1693.Kensington: Taylor and Francis

Sharma, N., & Bansal, K. L. (2015). Comparative study of data mining tools. *Journal of Advanced Database Management & Systems*, *2*(2), 35-41.

Sihui, D., & Xueguo, X. (2016, June). Research on tacit knowledge mining of university libraries based on data mining. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1-4). New York: IEEE.

Sumathy, K. L., & Chidambaram, M. (2013). Text mining: concepts, applications, tools and issues-an overview. *International Journal of Computer Applications*, *80*(4).

Smith, R.D., & Bollinger, A.S. (2001). Managing organizational knowledge as a strategic asset. *Journal of Knowledge Management*, 5(1), 8-18.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). *Text Mining and ontologies in biomedicine:* making sense of raw text. *Briefings in bioinformatics*, *6*(3), 239-251.

Syed, S. (2019). *Topic Discovery from Textual Data: Machine Learning and Natural Language Processing for Knowledge Discovery in the Fisheries*

*Domain* (Doctoral dissertation, Utrecht University).

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases* (Vol. 8, pp. 65-70).

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business-oriented approach. *Expert Systems with Applications*, *39*(5), 4729-4739.

Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artif. Intell. Res.*, *2*(1), 44-54.

Verspoor, K. M., Cohn, J. D., Ravikumar, K. E., & Wall, M. E. (2012). Text mining improves prediction of protein functional sites. *PLoS One*, *7*(2), e32171.

Wang, L. L., & Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, *22*(2), 781-799.

Williamson, K., & Johanson, G. (Eds.). (2017). *Research methods: information, systems, and contexts*. Sawston, Cambridge: Chandos Publishing.

Wilensky, H. L. (2015). *Organizational intelligence: Knowledge and policy in government and industry* (Vol. 19.: Quid Pro Books.

Wyskwarski, M. (2020). Identification of Desired Project Manager Competence Using

Text Mining Analysis. *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska*, (149), 735-749.

Xie, X., Fu, Y., Jin, H., Zhao, Y., & Cao, W. (2020). A novel text mining approach for scholar information extraction from web content in Chinese. *Future Generation Computer Systems*, *111*, 859-872.

Yang, M. C., Wood, W. H., & Cutkosky, M. R. (1998). Data mining for thesaurus generation in informal design information retrieval. *Proc. 1998 Int. Congr. Civil Engineering*, 18-21.

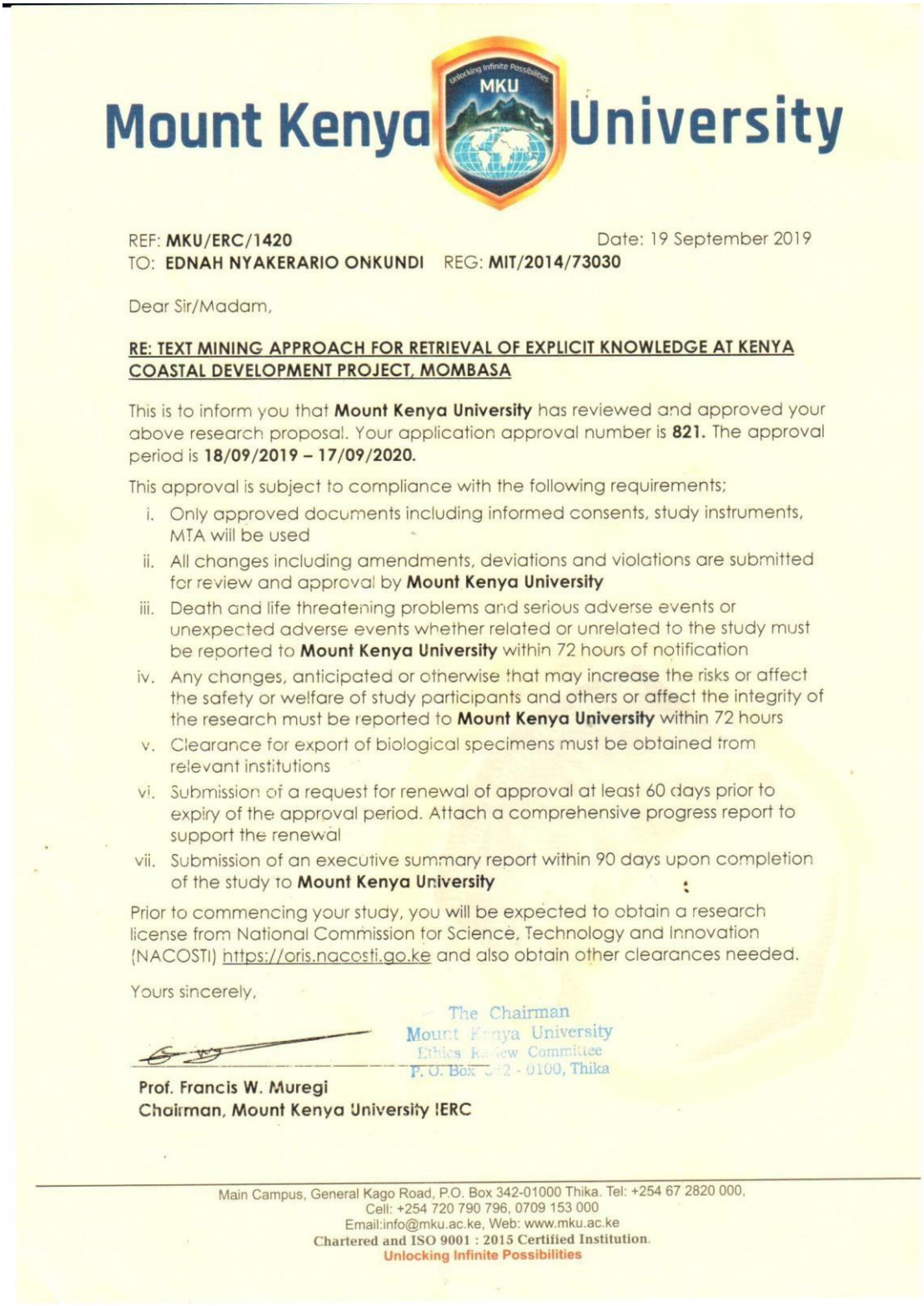
Yoon, B., Phaal, R., & Probert, D. (2008). Morphology analysis for technology roadmapping: application of text mining. *R&d Management*, *38*(1), 51-68.

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, *16*(3), 730-744.

Zhai, C., Cohen, W. W., & Lafferty, J. (2015). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR Forum* (Vol. 49, No. 1, pp. 2-9). New York ACM.

# APPENDICES

## APPENDIX 1: ETHICAL REVIEW COMMITTEE LETTER



Mount Kenya University

REF: **MKU/ERC/1420** Date: 19 September 2019
TO: **EDNAH NYAKERARIO ONKUNDI** REG: **MIT/2014/73030**

Dear Sir/Madam,

**RE: TEXT MINING APPROACH FOR RETRIEVAL OF EXPLICIT KNOWLEDGE AT KENYA COASTAL DEVELOPMENT PROJECT, MOMBASA**

This is to inform you that **Mount Kenya University** has reviewed and approved your above research proposal. Your application approval number is **821.** The approval period is **18/09/2019 – 17/09/2020.**

This approval is subject to compliance with the following requirements;

i. Only approved documents including informed consents, study instruments, MTA will be used
ii. All changes including amendments, deviations and violations are submitted for review and approval by **Mount Kenya University**
iii. Death and life threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to **Mount Kenya University** within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to **Mount Kenya University** within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal
vii. Submission of an executive summary report within 90 days upon completion of the study to **Mount Kenya University**

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://oris.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

The Chairman
Mount Kenya University
Ethics Review Committee
P. O. Box 342 - 0100, Thika

**Prof. Francis W. Muregi**
**Chairman, Mount Kenya University IERC**

Main Campus, General Kago Road, P.O. Box 342-01000 Thika. Tel: +254 67 2820 000,
Cell: +254 720 790 796, 0709 153 000
Email:info@mku.ac.ke, Web: www.mku.ac.ke
**Chartered and ISO 9001 : 2015 Certified Institution.**
**Unlocking Infinite Possibilities**

## APPENDIX 2: POST GRADUATE APPROVAL

Mount Kenya University

## SCHOOL OF POSTGRADUATE STUDIES

MIT/2014/73030

14th October, 2019

*The Director, Research Coordination Division*
*National Commission for Science, Technology & Innovation*
*Utalii House, 8th & 9th Floor*
*P.O Box 30623- 00100*
*NAIROBI*

Dear Sir/Madam,

RE: EDNAH NYAKERARIO ONKUNDI – REGISTRATION NO. MIT/2014/73030

The purpose of this letter is to introduce the above named student who is pursuing **Master in Information Technology** in the **Department of Information Technology** in the **School of Computing & Informatics.**

The title of her research is *"Text Mining Approach for Retrieval of Explicit Knowledge at Kenya Coastal Development Project, Mombasa."*

She has been cleared by the University's Ethics Review Committee (Certificate attached) and now has to proceed to the field to collect data for her research between **October and December, 2019.**

Any assistance accorded to her will be highly appreciated.

Thank you.

Mount Kenya University
Dean, School of Postgraduate Studies
P.O. Box 342 - 01000,
THIKA

**Dr. Samuel M. Karenga, Ph.D**
**Dean, School of Postgraduate Studies**
**Enc.**

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR
SCIENCE, TECHNOLOGY & INNOVATION

Ref No: **877793**

Date of Issue: **27 /October/ 2019**

## RESEARCH LICENSE

**This is to Certify that Ms.. EDNAH ONKUNDI of Mount Kenya University, has been licensed to conduct research in Mombasa**

**on the topic: TEXT MINING APPROACH FOR RETRIEVAL OF EXPLICIT KNOWLEDGE AT KENYA COASTAL**

**DEVELOPMENT PROJECT for the period ending : 27/October/2020.**

License No: **NACOSTI/P/19/2381**

**877793**
Applicant Identification Number

Director General
NATIONAL
SCIENCE,TECHNOLOGY &
INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document Scan the QR Code using QR scanner application.



**APPENDIX 3: RESEARCH PERMIT**

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 The Grant of Research Licenses is Guided by the Science, Technology and Innovation (Research Licensing) Regulations, 2014

CONDITIONS

1. The License is valid for the proposed research, location and specified period
2. The License any rights thereunder are non-transferable
3. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research
4. Excavation, filming and collection of specimens are subject to further necessary clearence from relevant Government Agencies
5. The License does not give authority to tranfer research materials
6. NACOSTI may monitor and evaluate the licensed research project
7. The Licensee shall submit one hard copy and upload a soft copy of their final report (thesis) within one of completion of the research
8. NACOSTI reserves the right to modify the conditions of the License including cancellation without prior notice

National Commission for Science, Technology and Innovation off Waiyaki Way,
Upper Kabete,
P. O. Box 30623, 00100 Nairobi, KENYA
Land line: 020 4007000, 020 2241349, 020 3310571, 020 8001077
Mobile: 0713 788 787 / 0735 404 245
E-mail: dg@nacosti.go.ke
/ registry@nacosti.go.ke
Website:
www.nacosti.go.ke

# APPENDIX 4: TURNITIN REPORT

TEXT MINING MODEL FOR RETRIEVAL OF EXPLICIT KNOWLEDGE AT KENYA COASTAL DEVELOPMENT PROJECT, MOMBASA

ORIGINALITY REPORT

| 10% | 9% | 4% | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | worldwidenewscast.com<br>Internet Source | 1% |
| 2 | dspace.lboro.ac.uk<br>Internet Source | 1% |
| 3 | en.wikipedia.org<br>Internet Source | 1% |
| 4 | mafiadoc.com<br>Internet Source | <1% |
| 5 | epdf.tips<br>Internet Source | <1% |
| 6 | pdfs.semanticscholar.org<br>Internet Source | <1% |
| 7 | lrd.yahooapis.com<br>Internet Source | <1% |
| 8 | www.codedoct.com<br>Internet Source | <1% |

# APPENDIX 5: CONSENT FORM FOR PARTICIPATION IN RESEARCH

**A Text Mining approach for retrieval of explicit knowledge at Kenya Coastal Development Project**

Dear Participant,

I invite you to participate in a research study entitled text-mining approach for retrieval of explicit knowledge at Kenya Coastal Development Project: I am currently enrolled in the *Masters in Information Technology program* at Mount Kenya University and I am in the process of writing my Master's Thesis. The purpose of the research is to determine a Text Mining approach for retrieval of explicit knowledge at Kenya Coastal Development Project (KCDP).

The enclosed questionnaire has been designed to collect information on retrieval of explicit knowledge at KCDP. Your participation in this research project is voluntary. You may decline altogether, or leave blank any questions you do not wish to answer. There are no known risks to participation beyond those encountered in everyday life. Your responses will remain confidential and anonymous. Data from this research will be kept safely and reported only as a collective combined total. No one other than the researchers will know your individual answers to this questionnaire. There are no direct benefits to you for participating in this research. However, you may find it interesting to talk about the issues addressed in the research and it may be beneficial to the field and to future clients or individuals who have experienced similar concerns.

If you agree to participate in this project, please answer the questions on the questionnaire as best as you can. It should take approximately *10 minutes* to complete. Please return the questionnaire as soon as possible to enable the researcher complete the project report. If you have any questions about this project, feel free to contact the *Investigator - Ednah Nyakerario Onkundi on 0725585956.*

## CONSENT

I have read and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost. I understand that I will be given a copy of this consent form. I voluntarily agree to take part in this study.

Participant's signature _____ Date _____

Investigator's signature _____ Date _____

# APPENDIX 6: QUESTIONNAIRE

**Questionnaire Survey for the Study on Text Mining Approach in Retrieval of Explicit Knowledge**

Dear Sir/Madam, I am humbly requesting your participation in this survey, the need for this survey is to collect data for academic purpose. This survey seeks to establish the Text Mining approach for retrieval of explicit knowledge at your organization for knowledge management. The data collected will be treated with outmost confidentiality. Kindly complete the questionnaire on or before 11 April 2019. The researcher will pick it for processing on 19 April 2019.

***Instructions: Please fill the spaces or tick as appropriate.***

**Department: _____ Date: _____**

**Signature: _____**

**PART A: GENERAL INFORMATION**

1. What is your Gender?

   i) Male ☐ ii) Female ☐

2. Highest level of Educational? (Tick only one option).

   i) Certificate ☐ ii) Diploma ☐ iii) Degree ☐ iv) Masters ☐

   (v) PhD ☐

**PART B: KNOWLEDGE ON TEXT MINING (STAFF)**

1. Does your organization have a database?

   i) Yes, ii) No

2. (a) If **yes**, kindly specify the database name and provide information on the kind of data stored in your database?

   _____

   _____

2. (b) If **no**, how is data stored in your organization? _____

3. Do you know what the knowledge retrieval process entails?

   i) Yes ☐ ii) No ☐

4. If **yes**, have you ever applied the knowledge retrieval process in your daily work activities in the office?

   i) Yes ☐ ii) No ☐

5. If **yes**, which Text Mining techniques are you familiar with and use frequently?

   i) Information Extraction ☐ ii) Summarization ☐

   iii) Categorization ☐ iv) Classification ☐

6. Do you use Text Mining when retrieving explicit information from the organizational database?

i) Yes, ii) No

7. If **yes**, how has the use of Text Mining process influenced retrieval of explicit knowledge in your organization?

   _____

8. What other ways apart from training would you recommend, to be able to transfer tacit knowledge within your organization or transform tacit knowledge to explicit knowledge in KCDP/KMFRI?

   _____

   _____

   _____

**PART C: LIKERT SCALE KNOWLEDGE ON TEXT MINING**

**INSTRUCTIONS**: ***Please tick only one option (✓) appropriately in the boxes provided***

| **2.** | Below are statements on a five point, Likert scale (1-5) related to Knowledge Retrieval? **Please tick (✓ ) appropriately in the boxes provided.** | Strongly Agree | Agree | Don't Know | Disagree | Strongly Disagree |
|---|---|---|---|---|---|---|
| | | **5** | **4** | **3** | **2** | **1** |
| a. | KCDP encourages Knowledge Management. | | | | | |
| b. | KCDP needs a suitable model that would enable it retrieve explicit knowledge. | | | | | |
| c. | Data and Information at KCDP is well maintained and retained. | | | | | |
| d. | Text Mining at KCDP will help in retrieval of explicit knowledge. | | | | | |

*Thank you for your time*

## APPENDIX 7: FOCUS GROUP

### OPINION ON USE OF TEXT MINING (FOCUS GROUP DISCUSSION GUIDE)

1. What are the advantages of using Text Mining in the retrieval of explicit knowledge?

   _____

   _____

2. What are some of the challenges you face when using Text Mining in knowledge retrieval?

   _____

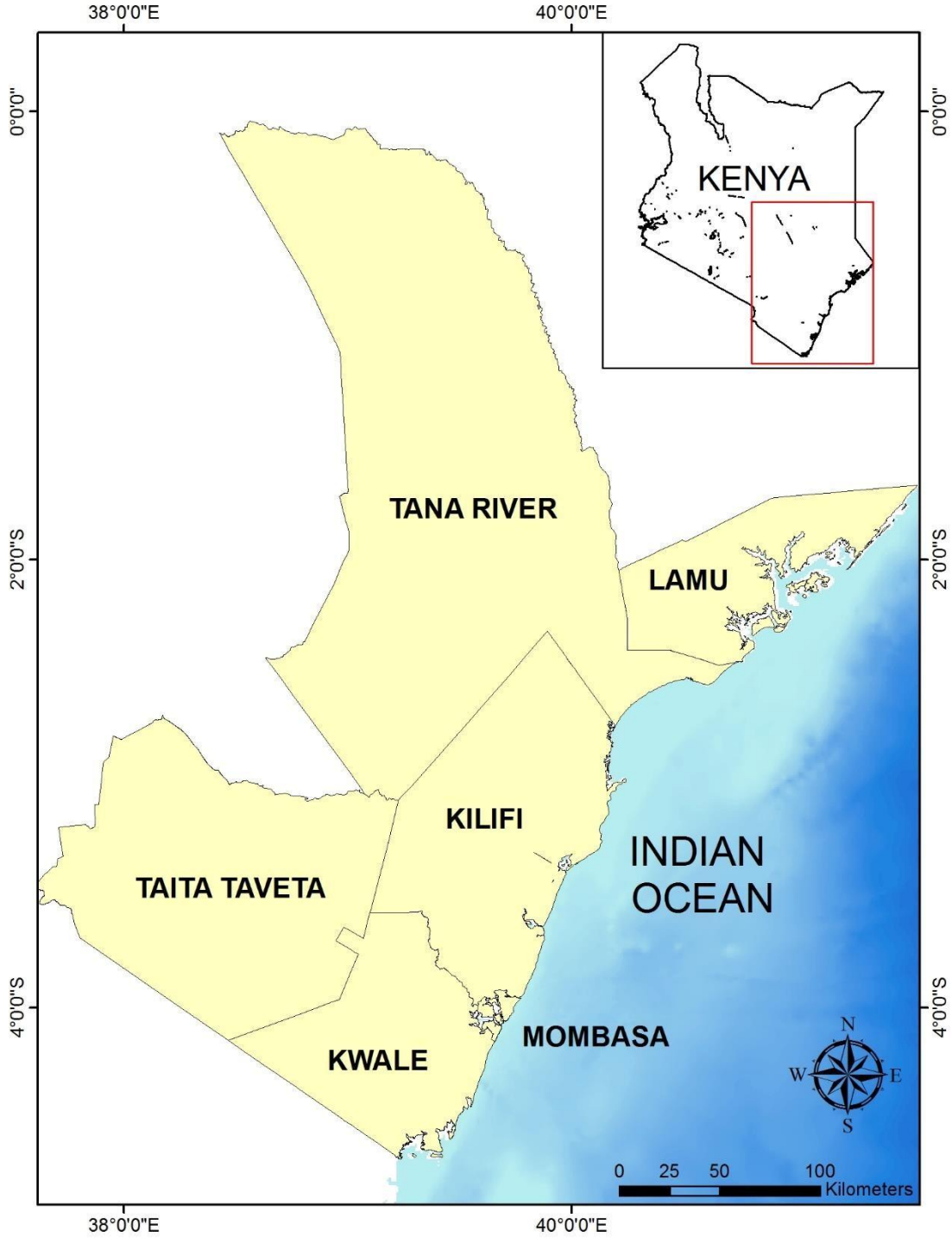
   _____

3. Suggestion for improvement for knowledge sharing, retrieval and storage in your organization.

---

*Thank you for your time.*

## APENDIX 8: MAP OF COASTAL REGION IN KENYA WHERE KCDP COVERS




*Source :( H. Ong'anda, personal communication, and November 14, 2019).*