A WEB BASED DOCUMENT ARCHIVING SYSTEM USING INDEXING AND MACHINE LEARNING FOR RESEARCH AND INNOVATION GRANT ALLOCATION

BY

REBECCA LUPYANI

22000262

A RESEARCH PROPOSAL SUBMITTED IN PARTIAL FULFUILMENT OF THE REQUIREMENT OF A DEGREE OF MASTERS' IN COMPUTER SCIENCE

THE UNIVERSITY OF ZAMBIA

SCHOOL OF NATURAL SCIENCES

LUSAKA

COPYRIGHT

DECLARATION

I, Rebecca Lupyani do hereby declare that this dissertation is my own original work and has not been submitted to any other college, institution or university other than the University of Zambia.

Name: REBECCA LUPYANI

Sign: ……

Date: ……

APPROVAL

This dissertation, by Rebecca Lupyani has been approved as partial fulfilment of the requirements for the award of Master of Science in Computer Science by the University of Zambia.

Examiner 1

Name: …….

Signature: ……..

Date: ……..

Examiner 2

Name: …….

Signature: ……..

Date: ……..

Examiner 3

Name: …….

Signature: ……..

Date: ……..

Chairperson (Board of examiners)

Name: …….

Signature: ……..

Date: ……..

Supervisor

Name: **Professor Jackson Phiri**

Signature: **……..**

Date: **……..**

ACKNOWLEDGEMENTS

Firstly, I wish to thank my God, Elohim without whom this project and entire journey would not have been possible. I am eternally grateful for the grace, the strength, the knowledge and all that He gave me to sustain me in my studies. I owe it all to Him.

Secondly, I would like to thank my awesome husband Emmanuel Chola Mumba for the financial, emotional and academic support rendered during my study. I am grateful for the sacrifices he made, the times he would cook for me and the kids while I was attending classes and most of all for the support rendered during the development of the system. I'm so grateful, May Elohim alone reward your efforts my love. To my children Harmony, Isubilo, Ntangi and Twange, I love you for being there and enduring the times I had to be away from you. Thankyou my girls.

My sincere appreciation goes to my supervisor, Prof Jackson Phiri for the guidance, support, the constructive criticisms and useful suggestions that pushed, motivated and inspired me to achieve academic progress that I never imagined for myself. Thank you sir for your unwavering support. This work would not have been what it is without your contribution.

I am grateful to my mum for her prayers, support, and encouragement, and for always being a pillar in my life. I thank my dear sister Christine for the encouragement, the prayers and financial support for standing with me during the hard times. I am so grateful. I thank the rest of my family members Humphrey, Nancy, Rachel and Bodson Jr, Uncle Boyd Mbasela (our prayer partner) for all their support and prayers. Indeed my heart is full for having such a family.I wish to extend my heartfelt gratitude to my beautiful friend Barbara Kunda aka my Barbie, words cannot express how grateful I am to God Abba for having given me a classmate, a study mate, a prayer partner, a friend and a sister all wrapped up in one. This journey would not have been walkable without you. Our sleepless nights, our times in the Library, our spontaneous trip to Livingstone for the ZAPUC conference. Gosh, we made memories and I will forever be grateful because our hard work and support of each other have really paid off. Thank you Barbie!

Lastly, I wish to thank the Ministry of Technology and Science for the partial scholarship rendered to me. Thank you so much for lightening my financial burden. I also thank my employers Evelyn Hone College for the two months study leave granted to me. I am so grateful.

DEDICATION

To my late dad Bodson Chailusa Lupyani and my mother Maureen Muyembe Lupyani. This is for you. Dad I know you would be so proud of me. I am becoming what you said I would be and mum is here cheering me. I am forever grateful.

ABSTRACT

In today's rapidly evolving world, research forms a cornerstone of human progress leading to new products, services, and technologies, which, in turn, can stimulate economic growth and enhance the quality of life. Research enables people to learn, innovate and address the complex challenges facing a society. It is a powerful tool for making the world a better place and as such many countries endeavor to support the research landscape by providing research grants in different sectors. The process of allocating research grants plays a pivotal role in fostering scientific progress, innovation and knowledge. The traditional manual selection of grant proposals, while well-established can be resource intensive time-consuming, subjective, and prone to bias. This paper presents an unconventional strategy that leverages machine learning algorithms to enhance the fairness, efficiency, and transparency of the grant allocation process by removing human biases and prejudices that can inadvertently influence funding decisions. The study discussed the design and implementation of a machine learning-based grant allocation system using historical grant data from a reputable funding agency and provided empirical evidence of its effectiveness by selecting the best performing text classification algorithm from a comparative analysis of three models and integrating it into a web based application. The three models compared were the K- Nearest Neighbour, the Naives Bayes and the Support Vector Machine. The Support Vector Machine exhibited the highest performance metrics with an accuracy percentage of 88%, precision of 86%, recall of 87% and F1 score of 87%, whereas the K-Nearest Neighbour portrayed the lowest performance with an accuracy of 41%, precision 52%, recall of 43% and F1 score of 37%. Thus, the Support Vector Machine was integrated into a web based application to facilitate fund allocation. The developed system will promote fair review of research and innovation proposal applications by automatically categorizing them into topic categories that facilitate funding such as engineering, science and technology. The system will also assist in tracking and monitoring the progress for the research projects for which funding institutions invest in and also in digitally archiving the documents in an effective and efficient manner

Keywords: Support Vector Machine, Research grant, Web Archive, Machine Learning

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- MOTS- Ministry of Technology and Science
- NSTC- National Science and Technology Council
- SRF: Strategic Research Fund
- SYIF: Strategic Youth Innovation Fund
- CRISP-DM: Cross Industry Standard Process for Data Mining
- SVM: Support Vector Machine
- NB: Naive Bayes
- KNN: K- Nearest Neighbour

1 INTRODUCTION AND BACKGROUND

1.1 Introduction

This chapter discusses the background to the problem which provides some historical information relating to the research. Subsequently the statement of the problem statement follows the background to put the problem into perspective. The aim of the study is highlighted to identify the main goal of this research and the objectives that were used to answer the research questions. The significance of the study is given to define the beneficiaries of the study and finally a conclusion is given to summarise all discussions relating to this chapter.

Research and Innovations have become an integral part of any country's economic development [1]. Research is defined as a systematic investigation or study that seeks to discover, interpret, and expand knowledge on a particular topic or phenomenon. The goal of research is to contribute to the existing body of knowledge by developing new theories or confirming existing ones [2]. It involves the application of structured processes to collect, analyze, and interpret data such as scientific experiments, surveys, case studies, literature reviews, and theoretical inquiries [3]. It provides the building block upon which societal growth and advancement is hinged and it is a tool for building knowledge and facilitating learning [1]. Research facilitates the economic growth of a country by reducing poverty, building stronger economies and societies, and improving the quality of life of the people [4]. When countries invest in research, it helps to generate new knowledge, ideas, and innovations that can be used to address the challenges facing the country, such as unemployment, poverty and inequality [5]. The new knowledge helps countries to formulate policies that assist in making decisions and guide the development of effective solutions to complex problems [6]. Research can also help to improve the quality of education, healthcare, and other services, and to promote social and environmental sustainability. Furthermore, research can bring world recognition to a developing country, which can be very important for its growth and development [7]. An innovation is defined as a means of creating and implementing new ideas, products, processes, or services that result in significant positive change [8]. It hinges on creativity, knowledge, and the use of resources to address challenges, meet needs, or seize opportunities. Innovation influences economic growth, societal progress, and competitive advantage [9]. Thus, Research and Innovations are of paramount importance across various fields and disciplines. Research identifies and analyses problems, seeks solutions, whereas innovations lead to new products, services, and technologies

which, in turn, can stimulate economic growth [1]. Apart from economic growth, Research and Innovation impacts the following areas:

Global Competitiveness: Cutting edge Research and Innovation enhances the global competitiveness of countries by facilitating a country's ability to stay ahead in various fields, promoting international recognition and attracting investments [10].

Policy formulation: It supports the formulation of policies that are essential for addressing challenges, solving complex problems and promoting sustainable practices and supporting the well-being of people in a country [11].

Human Capital Development: Research and innovation contributes to the development of skilled and knowledgeable human capital. A country's investment in research activities produces professionals and experts who can address complex challenges in diverse fields [11].

Advancements in Healthcare: research and Innovation plays a vital role in the health sector by facilitating the understanding of diseases, developing new treatments, and improving healthcare systems thereby leading to the discovery of vaccines, drugs, and medical technologies, contributing to the overall health and well-being of the population [12].

The importance of Research and Innovations to a country's development cannot be emphasized enough. Many countries have established institutions responsible for promoting societies where research and innovation can be created, used and shared by planning and supporting multi-disciplinary research projects, recommend new directions for research thereby resulting in a society with a better research culture and thus contribute to national development. Therefore, it is the aim of this study to develop a system, for funding institutions to use in their quest to support research and innovation that automatically defines the eligibility of research and innovation proposal applications and provides a means of archiving all the research and innovation projects.

1.2 Background

Research funding institutions are organizations whose mandate is to advance scientific knowledge, technological innovation, and academic research across various disciplines by providing financial support, grants, or funding to individuals, research teams or institutions that engage in research activities. Research funding institutions vary from country to country but they may include government agencies, private foundations, non-profit organizations, and international bodies.

The main goal of research funding institutions is to invest in research projects that align with their mission and contribute to the advancement of knowledge in specific fields. These institutions responsible for research and development are interested in monitoring and keeping track of funds invested in research and development activities in specific fields or topics [1]. Funding institutions such as the National Science and Technology Council under the Ministry of Science and Technology in Zambia focus on facilitating, handling, and funding multiple research projects in the education domain in order to offer support for technical and vocational skills development and application of science, technology and innovation through research and development [13]. The institution seeks to award research grants in Higher Education Institutions (HEI) and to award Innovation grants to eligible youths throughout the country, according to specific areas of discipline namely science, technology and communications. Additionally, the institution seeks to monitor and keep track of its investments in the specific research areas by determining how much funding has been invested in a particular field and also to decide whether or not submitted research or innovation proposals by grant applicants are eligible for funding. The institutions also endeavor to keep accurate records to recall or prove research or innovation projects that have been undertaken under its mandate.

Research funding organizations receive a large number of research proposal applications every year that call for the need to be classified and evaluated to check for compliance. In addition to research proposal applications, organisations also receive applications for Innovations which are also subjected to scrutiny in order to determine the innovation potential, the expected outcomes and whether the innovation supports the goals of national development. However, the received proposal application documents are rarely classified in ways that will inform policy and budget decisions. Hence, the research and Innovation proposal applications call for the need to be streamlined and classified into categories for the funding organisations to determine whether the proposals are viable ventures that can be supported and promoted.

1.3 Statement of the Problem

The benefits of research and innovations to a country's economic development justify the need for funding investments in the Research and Innovations landscape. Thus, many countries have established organisations that can promote the Research and Innovation agenda by providing grants. Research and Innovation grants play a fundamental role in advancing scientific endeavours, promoting discoveries, and nurturing talented researchers and innovators. In the realm of research funding, the allocation of grants stands as a critical decision point where the

future of scientific progress is concerned [2]. Traditionally, this process has been determined by human judgement thus making it susceptible to inefficiencies, inequalities, and biases.

The research and innovation proposal applications received every year by research funding organizations, call for the need to be classified and evaluated to check for compliance. Many funding organisations especially in Africa use tagging systems in which human operators manually tag and classify the applications based on the information submitted by the applicant [4]. Keyword searches are then used to search for awards in a specific research discipline and therefore use this information for their research and development portfolio analyses and the tagged applications are evaluated for compliance [1]. While such a system has been established as effective, it is not without limitations. It has the potential for subjective decisions, and the category assignments done by human operators may result in incorrect, inconsistent or incomplete tags or labels as the individual applying the tags may not be aware of all possible information in the document, or may not be aware under which discipline certain research documents belong to and therefore may not select the most salient tags [3]. These challenges have prompted a growing need for a more data-driven, objective, and equitable approach. Therefore, this paper proposes an automated categorization of the research and innovation proposal submissions through machine learning techniques to heighten the efficiency and impartiality of the grant allocation process. Harnessing the capabilities of text categorisation algorithms will assist funding organisations to streamline the evaluation process of research and innovation proposals, mitigate human bias, and make more well-informed decisions regarding which proposals merit support.

1.4 Aim of the Study

The aim of this study was to build a document archiving system that uses machine learning to automatically classify research or innovation fund application documents into respective disciplines using the title for the awarding of research and innovation grants.

1.5 Research Objectives

i To build a title-based machine learning model that automatically classifies research or innovation proposal documents according to discipline for the awarding of research and innovation grants.

ii To develop a mechanism that helps to keep track research progress.

iii To develop a web based prototype that uses machine learning to automatically classify documents for effective archiving and retrieval.

1.6 Research Questions

i. How can we develop a model and algorithm that will automatically classify documents according to discipline for the awarding of research grants?
ii. How can we develop a mechanism that helps to keep track of research progress?
iii. To what extent can we develop a web based prototype that helps to keep track of research progress and uses machine learning to automatically classify documents using the title for effective archiving and retrieval?

1.7 Significance of the Study

The findings of this study will be of great benefit to a number of funding institutions like the Ministry of Technology and Science that endeavour to grant research and innovation grants according to disciplines and also to digitally archive the documents in an effective and efficient manner. Additionally, the system developed will assist the funding institutions in tracking and monitoring the progress for the research projects for which they invest in.

1.8 Scope of the study

This study targeted funding institutions that give research and innovation grants to beneficiaries including the Ministry of Technology and Science which was used for data collection.

1.9 Organization of the Dissertation

The dissertation is divided into five chapters as follows.

Chapter one covers the introduction to the dissertation and information to the background of the study. The statement of the problem is discussed, clearly defining the gap that this study is intended to fill, followed by the aim and objectives. The research questions, scope and significance of the study are also covered in this chapter. Chapter two outlines the various literature done by different scholars on the subject matter, identifying findings and gaps. Chapter three highlights the methodology that was employed to carry out the study, discussing the research design, requirements of the system, the system design and development specifications. Chapter four presents the results from the experiments carried out. Chapter five interprets, discusses, and concludes the results. Conclusions and recommendations are given based on the findings of the study in chapter six.

1.10 Chapter Summary

This chapter has described the background to the study which highlights how funding institutions in Zambia allocate grants for research and innovation projects as well as the aim of the study. The objectives of the study were defined and used to answer the research questions that guided the research. Lastly the significance of the study was given to define the beneficiaries of the study.

2 LITERATURE REVIEW

2.1 Introduction

This chapter explores the literature that is significant in the understanding of the objectives and research questions outlined in Chapter One. It explains theories that facilitate the understanding of machine learning concepts and their relation to text classification, and finally give major findings and gaps between related works and a summary to the chapter.

2.2 Background to the Study

Research funding institutions are organisations that provide financial support, grants, or funding to individuals, research teams, or institutions that promote, coordinate and carry out research activities. Funding institutions play a critical role in the advancement of scientific knowledge, technological innovation, and academic research across various disciplines. The institutions can take various forms, including government agencies, private foundations, non-profit organizations, and international bodies [4]. The primary goal of funding institutions is to align their investments in research projects to their mission thereby contributing to the body of knowledge and to the social economic status of the country. Many countries have government agencies responsible for funding research initiatives. In Zambia one such institution is the Ministry of Technology and Science (MOTS).

2.2.1 The Funding Institution (MOTS)

The Ministry of Technology and Science is a body whose mandate is to formulate and implement policies on Technology, Science, Communications and Skills Development in order to contribute to economic growth. It collaborates with the industry and the wider private sector in developing relevant innovations; coordinates Research to promote investment in science and technology; and promotes advancement of knowledge and skills in science and technology in order to accelerate transformation into digital economy [13]. Additionally, the Ministry conducts science, technology and innovation impact assessment, monitoring and evaluation. The vision of the Ministry is to be; “A Smart and Value Centred Zambia where quality, coordinated and relevant Communications and Postal Services, TEVET and Science Technology and Innovation are the driving forces of the economy at all levels of national development by 2030”. Guided by this vision statement, the Ministry endeavours to guide, monitor, coordinate and evaluate the implementation of its programmes and projects to ensure the realization of its mission. Its mission therefore is “To facilitate, coordinate and promote relevant, cost-effective and sustainable technical and vocational skills development

and application of science, technology and innovation through research and development for improved productivity and quality of life." [13]

The Ministry derives its mandate from Government Gazette Notice 1123 of 2021 that outlines its portfolio functions as follows;

General Objective

- The strategic objective of the Ministry of Technology and Science is to increase access to efficient, equitable, quality, and relevant technical, vocational, and entrepreneurial skills; and to enhance research and development, commercialization, transfer, and diffusion of technology and innovation.

Specific Objectives

- To improve the quality and relevance of training through the provision of state-of-the-art equipment, upgrading of lecturer qualifications, and review and development of curricula;
- To establish a sustainable financing strategy for TEVET through the Training Levy as well as establish an efficient and self-sustaining Student Loans and Scholarships Fund for Higher Education;
- To increase female participation in science and technology courses and careers for national development;
- To facilitate the development and promotion of entrepreneurial skills to build lifelong skills and contribute to job creation;
- To strengthen basic and applied research in technical learning and research institutions by strengthening the enforcement of regulations, establishing standards in research and development, and developing a critical mass of scientists at Master of Science (MSc) and Doctor of Philosophy (Ph.D) levels;
- To commercialize innovations and/or R&D results to contribute to socio-economic development and wealth creation. [13]

In its quest to achieve one of its objectives of strengthening research and developing a critical mass of scientists at MSc and PhD Levels, the MOTS provides grants under the Science and Technology (S &T) Postgraduate Scholarship.

Science and Technology Postgraduate Scholarship

This fund was introduced to support MSc and PhD students who undertake research projects that are aligned to national development goals. This grant is only applicable to Zambians with who are in possession of a Green National Registration Card (NRC). The applicants are required to have a full grade 12 certificate with credits in 5 O levels and must possess letters of acceptance from a recognized accredited local Public Universities. PhD applicants must have a Master of Science Degree or its equivalent in a science related field which is an acceptable prerequisite to the field being applied for whereas MSc applicants must have a Bachelor of Science or its equivalent in a Science related field. The maximum acceptable age is 40 years for male applicants and 45 years for female applicant at MSc and a Maximum age of 45 years for male applicants and 50 years for female applicants at PhD. Anyone who meets the above requirements is eligible to apply. The fund includes tuition fees for the entire duration of the programme, stipend and research fund which are paid annually.

In order to further carry out its mandate, the MOTS has institutions that carry out its functions, among such that are aligned to supporting research and innovation initiatives are:

2.2.1.1 National Science and Technology Council (NSTC)

The National Science and Technology Council (NSTC) is a statutory body established by the Science and Technology Act No. 26 of 1997. The main function or mandate of the Council as prescribed in the Act is to promote science and technology so as to improve the quality of life in Zambia [14]. To carry out this mandate, the NSTC through the Ministry of Technology and Science, offers two types of grant funding; the Strategic Research Funds (SRF) and the Science, Technology and Innovation Youth Fund (STIYF).

- The Strategic Research Funds (SRF)

  The Government of the Republic of Zambia (GRZ) through the Ministry of Technology and Science (MoTS), administers the Strategic Research Fund (SRF) which is implemented by the National Science and Technology Council (NSTC). The SRF was established to support basic and applied scientific Research and Development (R&D) in identified strategic national priority areas. The Fund is further aimed at enhancing research capacity in Zambia. The overall indicative amount made

available under this call as at the year 2023 is Eight Hundred Thousand Kwacha only (ZMW 800,000.00) per proposal, for the duration of the project, which is up to three (3) years. This call is open to researchers in public and private R&D institutions and Institutions of Higher Learning that are registered with the NSTC in the funding cycle. Individual researchers that are not affiliated to any institution of higher learning are required to be affiliated to an NSTC-registered institution. Applications from institutions that are not registered with NSTC are not eligible for this type of funding [14].

- The Science, Technology and Innovation Youth Fund (STIYF).

The Government of the Republic of Zambia (GRZ) through the Ministry of Technology and Science (MoTS) also administers the Science and Technology Innovation Youth Fund (STIYF). This fund is implemented by the National Science and Technology Council (NSTC) and aims at assisting Zambian youths to develop scientific and/or technological innovations, with specific focus on innovations that are relevant to the creation of wealth and employment. This fund is only open to Zambians citizens who are resident in Zambia and are aged 35 years or below. Applicants can apply as individuals or as groups or organisations. Furthermore, female applicants and differently abled youths are strongly encouraged to apply. The overall indicative amount per approved project, made available under this call as at the year 2023 is Two Hundred and Fifty Thousand Kwacha (ZMW 250,000.00) project for the period of the project [14].

2.2.1.2 The National Institute for Scientific and Industrial Research (NISIR)

The National Institute for Scientific and Industrial Research (NISIR) is a statutory body established by the Science and Technology Act No. 26 of 1997 through Statutory Instrument (SI) No. 73 of 1998. The goal of the NISIR is to strengthen Research and Development (R&D) capacity in science and technology in order to meet needs of agriculture, manufacturing, health, water, ICT, energy and environment industries/sectors. In order to remain visible to its environment, NISIR endeavours to improve its Information management through effective utilization of Information Communication Technology (ICT). It also seeks to optimize available resources and take advantage of every opportunity whilst maintaining a balanced ecosystem with

stakeholders and cooperating partners. This centre is guided by the following strategic pillars:

- Client and Stakeholder Focus
- Financial Stewardship
- Internal Organizational Capacity
- Digital Transformation

2.2.1.3 National Technology Business Centre (NTBC)

This body was established in December 2009, it represents numerous opportunities for business development and improvement of various production processes. Its mandate lies in providing two functions namely commercialization of innovative products and transfer of technologies [15] through the seven support pillars outlined below:

1. Technology Acquisition and Transfer

   In facilitating the transfer of suitable and appropriate technologies to the Zambian Industry, NTBC aims at adding value to the manufacturing sector, MSMEs, Innovators and Researchers through productivity and quality improvements of products and services. NTBC realises its mandate of Technology Transfer (also called Transfer of Technology and Technology Commercialisation) through transferring skills, knowledge, technologies, methods of manufacturing, samples of manufacturing and facilities among Government, Academia and Industry [15]. The Technology Transfer service entails:

   - Facilitating acquisition, adoption, development, transfer and deployment of Technology to Entrepreneurs (from local and international R&D institutions and innovators)
   - Drafting of technology licenses and acquirement of Technology Transfer agreements
   - Developing more efficient processes for the identification and selection of Technology for adoption and utilization in Zambia
   - Conducting Technology Needs Assessments

- Technology brokering/ Technology valorisation – linking technology seekers or users (industry) to technology suppliers in order to improve production of goods and services.
- Providing technical advice to the commercialization processes of Research, Development and Innovation (RDI) products.
- Facilitating Intellectual Property Rights protection of Innovations/inventions and advice on Patent prosecution and utilisation
- Generating Technology landscape, technology forecasting/ and fore-sighting
- Providing engineering advisory service [15]

2. Provision of Technology Information

The Centre provides Technology Information through its Resource Centre. Through this medium, NTBC provides technology information to leverage formation of businesses that involve Innovative and efficient products and services. Its overall goal is to provide entrepreneurs and innovators with technological information in order to facilitate Commercialisation and Technology Transfer. This service entails providing information on:

- New technologies for improved production of goods and services
- Intellectual Property
- Investment opportunities
- Technology suppliers and seekers
- Technology landscapes

3. Technology Marketing and Innovation Promotion

This centre offers this service which aims to transform new ideas, innovations and technologies into products developed and commercialised by:

- Aiding Idea conceptualisation
- Facilitating product development and improvement
- Conducting market research
- Developing market strategies
- Launching new products and rebranding
- Developing platforms for innovation promotion [15]

4. Business Development Support Services

   NTBC undertakes the assessment of commercial value of technologies and research outputs to create sustainable and eco-inclusive businesses. This is delivered through:

   - Business plan evaluation and development
   - Financial linkages
   - Financial management skills
   - Business modelling
   - Operating space
   - Access to business networks
   - Coaching & Mentorship [15]

5. Technology Business Incubation Programme (TBIP)

The NTBC through the TBIP is aimed at providing Business Development Support to viable impact driven start-ups to increase their competitiveness in the market. The enterprises graduate from the programme within a defined period [15].

6. Technology Business Development Fund (TBDF)

The NTBC through the Technology Business Development Fund (TBDF) provides funding to technology based innovations to develop sustainable and eco-inclusive enterprises. The targeted beneciaries include; innovators, entrepreneurs, Small and Medium Enterprises (SMEs), Research Development institutions [15].

7. Technology Audit and Validation

In line with ensuring the quality production of goods and services by industry for increased competitiveness, the Centre has a responsibility to monitor the use and uptake of technology by industry. NTBC takes stock of technologies in use by industry and validates their use against existing standards drawn by various regulatory bodies to achieve full plant optimisation and cleaner production [15]. This service involves:

- Assessing production processes and systems
- Skills and knowledge exchange

- Developing standard operation and maintenance procedures
- Developing a database of technologies in use
- Providing intellectual property valuation

2.2.2 The Research Grant Application and Allocation process

The Ministry of Technology and Science offers two research grants. The Science and Technology (S & T) Postgraduate Scholarship and the Strategic Research Fund which is offered by the National Science and Technology Council. The application process for the two grants begins with a response to an advertisement by the Ministry. The advertisement is made in the public newspapers as well as their website. Applicants who meet the qualifications and requirements are asked to send their applications by email. The applicant fills in an application form which is downloaded from the Ministry Website or obtained from the Ministry of Technology and Science in Lusaka. The application form is accompanied by validated copies of Degree Certificates by Zambia Qualifications Authority (ZAQA), official transcripts of results and a photocopy of the green NRC. A photocopy of the acceptance Letter from a recognized accredited local Public University, together with letters of recommendation from employers are attached. This fund is only applicable to candidates in full time employment. However, those in informal employment are required to submit supporting documents showing their entrepreneurship in science and technology related field. Applicants are also required to submit their Curriculum Vitae with names, addresses and contact details of three (3) traceable referees and recent passport size photo of the candidate. All these documents are sent via email to the email address indicated in the advertisement.

Once the applications are received, they are downloaded, sorted, tagged and then classified according to subject class or category within which the topic of the research falls. This helps the Ministry to select those applicants who are eligible, in addition to other details such as age, nature of the educational institution (public or private) and funding history. The eligible candidates are contacted via email and invited to attend physical interviews for further scrutiny. Based on the results of the interview, human operators responsible for the processing of the applications select successful candidates and inform them of their successful selection.

2.2.3 The Innovation Grant Application Process

The process of processing Innovation Grant applications is similar to the processing of research grants. However, proposal applications are not only submitted in response to an advertisement call for innovations by the ministry but innovators who exhibit their innovations at fairs, Science Foras and expos who meet the interest and expectations of the ministry's funding polices may be granted funds. The applicants submit their proposal applications via email or physically submit them at the ministry. The proposals are then sorted, classified and evaluated by the technical and financial committee. Selected candidates are expected to pitch their innovations to the technical committee for final scrutiny. Successful candidates are informed of the results via email and granted the funding.

2.3 Machine Learning

One of the objectives of this study was to build a machine learning model that automatically classifies research and innovation proposal documents. Therefore, this section gives an overview of machine learning in order to understand its application to the study. Machine learning is a subset of Artificial Intelligence (AI), which is a general term used to describe development of computer systems capable of performing tasks that emulate human intelligence [16]. It involves a broad range of techniques, algorithms, and methodologies that simulate cognitive functions associated with the human mind, such as learning, problem-solving, reasoning, perception, natural language processing, and decision-making [17]. In simpler words Artificial intelligence describes machines that have the ability to perceive, logic and learn. Artificial intelligence is also defined as a science and technology based on disciplines such as computer science, biology, psychology, linguistics, mathematics and engineering [18] [19]. One of the subfields of artificial intelligence is machine learning. Machine Learning is a subset of artificial intelligence that allows systems to learn and progress automatically without explicit programming [20].This means that the computer intelligently extracts patterns from data, learns them and turns them into knowledge. This process does not does not require any explicit programming [21]. Thus, Machine learning can also be viewed as an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. Machine learning has been applied successfully in various fields such as pattern recognition, computer vision, spacecraft engineering, finance, entertainment, computational biology, biomedical and medical applications [22]. The learning process begins

with observations or data, such as examples, direct experiences, or instructions, to arrive at a pattern in the data and make better decisions based on the examples or samples provided. The idea is to allow the computer to learn automatically without human intervention and to be able to adjust its actions accordingly [23]. Machine learning occurs when we encounter a complex task or problem that cannot be solved by conventional methods or we are faced with a large amount of data and variables that cannot be processed and calculated by human resources using traditional methods, and we do not even have formulas or equations to help solve them [24]. Machine learning algorithms create a mathematical model with the aid of historical sample data or training data that aids in making predictions or judgements without being explicitly programmed. Computer science and statistics are used with machine learning to create prediction models. Algorithms that learn from past data are created by machine learning or used in it [25].

2.4 Supervised Machine Learning

Supervised machine learning is a type of machine learning algorithm where the model is trained on a labelled dataset. The algorithm learns from labelled training data, which means the input data is mapped with the correct output. Therefore the model has to learn the mapping or relationship between the input variables (features) and the target variable (output) based on the provided labelled data [26]. In this way, the learning algorithm predicts the output of the training data in different iterations, and then these outputs are corrected by an observer, and when the algorithm reaches an acceptable performance, the learning process stops. Supervised learning requires a number of input data in order to train the system [27]. Supervised learning is itself divided into two categories: regression and classification. Regression involves tasks or problems whose output is a continuous number or a set of continuous numbers, such as house price forecasts based on information such as number of rooms, to mention but a few. Classification algorithms are used when the target variable is categorical or belongs to a specific class or category [17]. It involves tasks whose output is part of a set, such as predicting whether an email is spam or predicting the type of illness a person has out of ten diseases. The diagram below shows an overview of how supervised learning works. Popular supervised learning algorithms include Linear Regression, Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, Naïve Bayes and K-Nearest Neighbour [22].

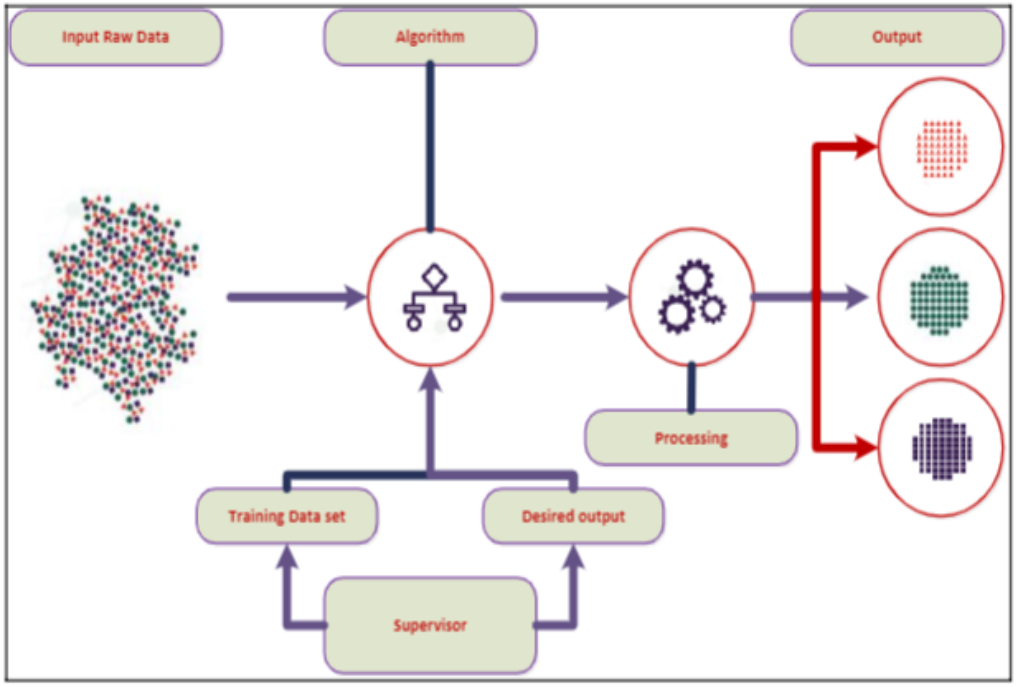

Figure 1: Supervised Learning [28]

2.4.1 Classification Algorithms

Over the recent years, a number popular classification algorithms have emerged each with its own weaknesses and strengths. The choice of which algorithm to use depends on factors such as the nature of the data, the size of the dataset, and the specific classification problem at hand. One of the classifier is the Naïve Bayes (NB) algorithm. The Naive Bayes is a simple and probabilistic model that handles its classification tasks by applying the Bayes' theorem. It makes the assumption that every pair of features that are being classified are conditionally independent of each other [29] . There are three types of NB classifiers namely the Multinomial, Bernoulli, and Gaussian Naive Bayes. Each of them can be used to handle different classification tasks but the all operate under the same principle, the Bayem theory. The Naive Bayes classifiers are known to provide insightful outcomes in handling text classification tasks including the detection of sentiments and spam in text contexts [30]. Another classifier to be reviewed is the K-Nearest Neighbour (KNN). This algorithm is a non-parametric, supervised learning classifier that is commonly used for regression and classification tasks. The performance of the KNN algorithm depends on the proximity of data points and uses the k closest training examples in a data set to classify new data points. The value of K is critical in the KNN algorithm, and it should be selected based on the input data. The algorithm works well on complex datasets as it is highly

sensitive to the structure of the data, making it suitable for datasets with complex boundaries [31]. Another classifier that is popular in handling regression and text classification tasks is the Random Forest (RF) algorithm. The Random Forest is one of the most popular and widely used machine learning algorithms as it exhibits exceptional performance when handling a wide range of text classification problems [32]. The algorithm uses the concept of building multiple decision trees, where each tree is trained on a random sample of the data and a random subset of the features. The Random Forest model considers a few key hyper-parameters which can be tuned for better performance. Some of the hyper parameters are the number of features, the number of trees, and the depth of the three [32]. Another most popular and widely used text classification algorithm is the Support Vector Machine (SVM). The SVM is a powerful machine learning algorithm that can be used for linear or nonlinear classification tasks, regression tasks, and outlier detection tasks [33]. It is therefore useful in handling tasks such as image classification, spam detection, handwriting identification, gene expression analysis, face detection, anomaly detection and text classification [33]. The Support Vector Machine (SVM) can be used for both binary and multi-classification. As a binary classifier, the SVM divides data points into two classes whereas in multi-class classification, it breaks down the multi-classification problem into multiple binary classification problems [34]. There are two main approaches for adapting the SVM for multi-class classification, namely One-vs-Rest and One-vs-One. The One-vs-Rest approach divides a multi-class classification task into one binary classification problems per class, while the One-vs-One approach divides a multi-class classification into one binary classification problem per each pair of classes. The choice of which approach to use is highly dependent on the size and nature of the dataset [34].

2.5 Unsupervised Machine Learning

Unsupervised learning is a type of learning where a computer learns patterns from data without any human intervention. The machine is trained using a set of unlabeled, unclassified, or uncategorized data, and the algorithm is required to respond independently to that data [35]. The main goal of unsupervised learning techniques is to uncover hidden patterns or structures within the data, such as similarities, clusters, or underlying distributions, without any guidance or labeled examples [22]. It involves analyzing and extracting useful information from raw, unlabeled data. Unsupervised learning uses input data to infer and

model patterns without the need for tagged results. This type of machine learning works with raw data that is not labeled and as such there is no observer to help the algorithm with learning [36]. Unsupervised learning algorithms are powerful in extracting insights from unlabeled data. They model the distribution of data so that they can learn more about the data. This means that there is no predefined outcome, input data is not mapped to the output data. Therefore, unsupervised machine learning is most suitable when training data is neither categorised nor labelled. However, assessing the performance of unsupervised learning algorithms can be more subjective compared to supervised learning due to the absence of ground truth labels [37]. Nonetheless, unsupervised learning techniques plays a critical role in data analysis, pattern discovery, and understanding complex datasets without prior knowledge of the output labels. The machines search through the vast volume of data for helpful insights [38]. The diagram below gives an overview of how unsupervised machine learning works.

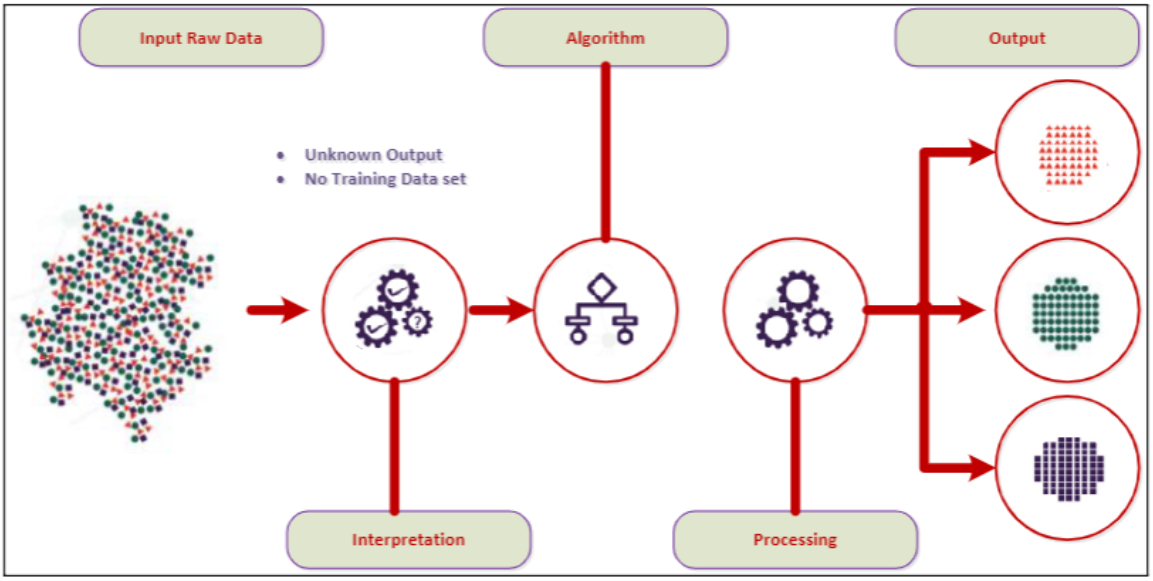

Figure 2 Overview of Unsupervised Machine Learning

Unsupervised machine learning can also be divided into two types of algorithms: clustering, association and dimension reduction. Clustering is a machine learning technique in which items that share similarities are grouped into distinct clusters. Each cluster comprising of items with attributes different from other clusters. The data items are classified based on the existence or lack of commonalities discovered by cluster analysis [38]. Association

supervised learning technique uses an association rule is to uncover the connections among the items in a sizable database. The association rule works by identifying patterns or relationships between variables in large datasets. For example, people who buy x frequently most like also frequently buy Y. This type of learning is most suitable for market basket analysis and recommendation systems [22]. Lastly, dimensionality reduction techniques are used to reduce the number of features or variables in the dataset while preserving essential information. This technique is useful in data visualisation and feature extraction applications. Applications of unsupervised learning include clustering similar documents or customer segmentation in marketing. Anomaly detection in cybersecurity or detecting irregularities in manufacturing processes, pattern recognition and exploratory data analysis, and generating synthetic data for training other models or data augmentation [39].

2.6 Reinforcement Learning

Reinforcement learning is a type of machine learning where an algorithm or software agent learns to make decisions by interacting with an environment [40]. An agent learns by receiving a reward for each correct action and receives a penalty for each incorrect activity. This feedback helps the agent to automatically learn and perform better since its objective is to accrue the most reward points [39]. The agent aims to maximize cumulative reward or performance by taking actions in the environment. It can also be said that the agent learns through trial and error, receiving feedback in the form of rewards or penalties based on its actions. Reinforcement Learning is inspired by psychology of behaviourism which focuses on the behaviours that the machine must do to maximize its reward, and by rewarding it tries to follow the best path with the goal of perfectionism [41]. Unlike supervised learning, Reinforcement learning does not explicitly define input and output. It focuses on live performance and online learning, which requires finding the right balance between exploring new things and exploiting. This property of Reinforcement Learning has made it popular in various fields such as robotics, game playing, autonomous vehicles, recommendation systems, finance, and healthcare [42].

2.7 Research Progress Tracking Mechanisms

There are various mechanisms that institutions use to track and monitor research progress thereby ensuring the effective and efficient management, evaluation, and support for ongoing research activities. Some of the most popular mechanisms are discussed below:

2.7.1 Research Information Management Systems (RIMS)

These are software applications or databases created to monitor, track and manage research-related information. Institutions also use these platforms to keep record of research documents carried out under their mandate. RIMS include platforms for tracking grants, projects, publications, collaborations, funding, and compliance [43].

2.7.2 Grant Management Systems

Institutions often utilize specialized systems to manage grant applications, funding, and the progress of awarded grants. These systems track grant proposals, budgets, deadlines, compliance requirements, and reporting obligations [44].

2.7.3 Reporting and Evaluation Mechanisms

Institutions often establish periodic reporting requirements for researchers to update on project progress, achievements, challenges, and future plans. These reports are crucial for internal assessment and external reporting purposes [45].

2.7.4 Project Management Tools

Project management tools offer tools that Institutions can use to monitor research project progress, timelines, milestones, tasks, and resource allocation. These software tools can facilitate the tracking of individual or team-based research projects [46].

2.7.5 Publication Tracking Databases

These are systems that keep track and record of all publications published by researchers. The systems facilitate the monitoring and dissemination of publications associated with researchers and institutions in terms of the number of publications and its impact [47].

2.7.6 Research Performance Metrics and Analytics

Institutions utilize metrics and analytics tools to assess research performance, including citation counts, h-index, journal impact factors, and other bibliometric indicators. These metrics help evaluate the impact and visibility of research outputs [48].

The mechanisms discusses above collectively support institutions in monitoring and managing the diverse aspects of research progress, ensuring compliance, fostering collaborations, and enhancing the impact of research outcomes. Thus, the functions of the Research Information Management Systems and the Grant Management Systems will be used as part of this study for the purpose of providing a means to monitor and keep track of research progress and grant funding.

2.8 Web-Based Archiving Systems

A web-based archiving system is a digital platform or application that enables the collection, preservation, and storage of content for archival purposes. It is accessible via a web browser and utilises web crawling technology to capture and save online materials and maintaining them in an organized manner for future retrieval and reference [49]. Web-based archiving systems come in various types, each designed to serve specific purposes and users' needs. The needs vary from general public access to specialised archiving needs. These systems are popular in preserving digital information, ensuring access to historical web content, and supporting research and cultural preservation efforts. The most common web-archiving systems are discussed below:

2.8.1 Public Web Archives

These are general public web archiving systems that are accessible to the general public and aim to capture and preserve a wide range of web content across different domains, timeframes, and geographical regions [50].

2.8.2 Institutional Web Archiving Systems

These are general public access archives that are managed by institutions such as libraries, museums, universities, and cultural institutions whose aim is to preserve web content relevant to their missions or collections [50].

2.8.3 Government Archives

These are also general public archiving systems that are managed by government bodies or national libraries to preserve web content related to legal, historical, or cultural significance of a nation or country [50].

2.8.4 Research-Oriented Archiving Platforms

These are specialised archives that manage academic or research content. These platforms are utilised by scholars and researchers, allowing them to access archived web content for academic studies, data analysis, and historical research [51].

2.8.5 Thematic Archiving Platforms

These are specialised archives that focus on capturing and preserving web content related to specific themes, topics, or subjects of interest. Some of the most popular themes archived in such platforms include to art, science and politics [50].

2.8.6 Event-Based Archives

These are archives that focus on preserving content around specific events, such as elections, natural disasters, or major cultural or historical milestones [50].

2.8.7 Business and Compliance Archives

These refer to web archiving systems that are offered by commercial entities to businesses for regulatory compliance, record-keeping, and preserving their online presence [51].

2.8.8 Personal Web Archives

There are user initiated archives that allow individual users to create and maintain their archives of web content. These tools often come as browser extensions or web applications enabling users to save specific web pages, sites of interest or personal content of interest [51].

2.8.9 Crowdsourced Archives

These are archiving platforms that allow collaborative archiving efforts where multiple users contribute to archiving web content. These initiatives often rely on volunteers or community participation to capture and preserve web materials and content [52].

From the above reviewed literature on web –based archiving systems, the system that is of interest to the study is the research oriented archiving system. The study focuses on developing a web based system to harness research and innovation documents for research or innovation proposal applications that have been given grants by funding institutions.

2.9 Related Works and Gaps in the Literature

Document archival and retrieval systems studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. According to Hatano et al [53], Digital Library Systems provide a multidisciplinary ground like digitization, information extraction, storage management and retrieval equipped with human computer interaction (HCI) and Artificial Intelligence (AI) tools. A number of authors have developed and applied a variety of techniques to digitalize, extract, store and retrieve documents applied to them.

2.9.1 Development of Electronic Document Archive Management System (EDAMS): A Case Study of a University Registrar in the Philippines

In a study carried out by Caluza, J. [54], the process of document retrieval and record keeping of the paper archives of the Leyte Normal University's Office of the Registrar was investigated and an electronic document archive and management system (EDAMS) was developed. The thought emanated from the fact that at the time, the university was experiencing challenges in the effective and efficient archiving and retrieval of documents such as Transcript of Records (TOR), Diploma, Form 137, Form 138, Honorable Dismissal, Good Moral, Birth Certificate (NSO), Marriage Certificate (NSO), Certificate of Transfer Credentials, Informative Copy of Transcript and Original Copy of TOR [54]. These challenges were as a result of using a manual based system. The study therefore, sought to investigate what issues or challenges were being faced by the office of the university registrar in dealing with documents received from the students, monitoring, and retrieval, as well as to investigate the coping strategies employed by the registrar [54]. To carry out their investigations, the researchers used an embedded single-case study using thematic analysis and utilised a qualitative design and a quantitative design using the descriptive method to determine the detailed responses of the respondents towards the processes and transactions in the University Registrar's Office. The data collection instruments used were the survey questionnaire, interviews and observation to gather more information from respondents [54]. The findings revealed that the challenges faced by the office were limited storage area, misclassification, misplacement of the document, document security, termite, and pest attacks, difficulty in monitoring, and difficulty in document retrieval [54]. In view of these findings, budgetary support, the use of a log book, periodic maintenance, and the use of maintenance technologies and infrastructure were identified as the coping strategies implemented to curb this challenges. However, despite these coping strategies, drawbacks

were still present like releasing on a budget take time, improper still happens, undermanned, and releasing of documents to the stakeholders still takes time [54]. As a result, the development of an electronic document archiving management system was made to resolve the challenges faced by the office. In the development of the system, the researcher used a fusion of the Waterfall Systems Development Life Cycle (SDLC) Model and the Sashimi model which is originated by Peter DeGrace [55]. The researcher's choice to combine two SDLC models is commended. This is because, although the of Waterfall model is one of the most widely used methodologies as it suits projects where the requirements are clear, the model is rigid and not flexible to accommodate changing requirements [56]. Thus, combining it with the sashimi model allows the changing and continual improvement of the system for further enhancement of transactions and processes into the system. The resulting system automated most of the processes and allowed faster retrieval of documents [54]. However, human intervention was still needed to enter all the details of the documents correctly to facilitate the correct classification of the documents at input level. This increases the risk of human error.

2.9.2 Web Based Document Archiving Using Time Stamp and Barcode Technologies–A Case of the University of Zambia

In another study, Mutale & Phiri [57] developed a document archiving system which was integrated with barcoding, time stamping and mobile technologies to index, archive and retrieve documents [57]. The study was an effort to establish the challenges that are faced by educational institutions in document archiving and to consequently design and develop an electronic system to curb this challenges. The study adopted the University of Zambia (UNZA), a public learning institution for its baseline study. The aim of the study was to investigate the challenges in document archiving, being faced by the institution and to consequently design and develop an electronic system to curb these challenges. Therefore, a quantitative approach using questionnaires and verbal interviews for its investigations was applied. The results indicated that 70% of the offices in the institution lacked storage space because of the high use of box files and 80% of them had difficulties in retrieving old documents. As a result of these findings, the study sought to develop a system to alleviate the archiving problems being faced the education institution; the waterfall agile model was adopted as the system development lifecycle model. Using this model, a web based prototype was developed to manage the storing and retrieving of the documents such as Memos, Contracts, Minutes and Memorandum of Understanding (MoUs) with other

Universities and organizations. Additionally, a time stamp module was integrated to the system to send reminders using short messaging service systems (SMSs), pop ups, and emails. The time stamps where incorporated to send reminders when particular documents were due for attention or submission. The documents were also stamped with a barcode before their archival to improve on document identity and quick retrieval. Whereas the system developed was used to effectively index, archive and retrieve documents as well as send reminders for critical transactions through pop-ups, emails and Short Message Service (SMS) systems, it was limited to documents such as Memos, Contracts, Minutes and Memorandum of Understanding (MoUs).

2.9.3 Design of a Digital Dissertation Information Management System

Another study carried out by Glisson and Chowdhury [58] aimed designing and archiving system that would effectively and efficiently manage dissertations produced in higher institutions of education. Dissertations contain extremely valuable information since they report the results of first-hand research information conducted by students within any specific department. The authors of this study therefore saw it imperative for a system to be developed and implemented that would allow other researchers, students and academic staff, easy and flexible access to this information. Their aim was to also develop a system that would aid students, supervisors, dissertation coordinators and support staff in electronic communication. It was hoped that the resulting system would improve correspondence between the supervisors and the students and would also be able to produce statistics and generate reports on a real-time basis [58].

Permatasari et al [59] carried out a study to address challenges that are faced in document archiving systems such as lack of confidentiality and integrity as well as poor availability of documents using block chain technology. A decentralized CEA system was developed by implementing a distributed file system to manage the archives by integrating the interplanetary File system (IPFS) and block chain technology. This system uses a smart contract to manage the functionality of the application and deploy it on an Ethereum private network. This mechanism keeps archives confidential and safe from unauthorized users [59]. Zikratov et al [60] undertook another study that employed block chain technology to address issues of low integrity, high cost, and easy tampering of documents. To address these issues, a secure data storage and recovery scheme in the blockchain-based network is developed by improving the decentralisation, tampering proof, real-time monitoring and management of storage systems, as such design supports the dynamic storage, fast repair,

and update of distributed data in the data storage system of industrial nodes. A local regenerative code technology was used to repair and store data between failed nodes while ensuring the privacy of user data.

2.9.4 OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga.

In recent times, artificial intelligence (AI) methods have been applied in document and content management to make decisions and improve the organization's functionalities. J. M. Jayoma, E. S. Moyon and E. M. O. Morales [61] developed an application that facilitates the digitization and management of paper-based records of Department of Social Worker and Development (DSWD) Caraga Field Office. At the time of the study, the DSWD, like any other organization, comprised of several departments and continuously produced a high volume of records daily. The bulk of printed documents generated were piled and archived in the record's office and concerned divisions for safekeeping and future use [61]. They used a conventional records management system that resulted in challenges in retrieving and keeping track of the record's whereabouts. Furthermore, the manual indexing of documents generated each day burdened their records officer and was susceptible to human error. To address these challenges, the study sought to achieve the following objectives; to transform the vital and permanent records on paper into a digital form, to develop a Records Management System (RMS) that can cater to digitized documents and to organise, archive and index records that no longer require large physical space and to enhance fast retrieval and tracking of records whereabouts for vital and permanent records. [61] Thus, a web application that facilitated the uploading of scanned documents, indexing and, at the same time, automated the process of classification of records into administrative, financial, legal, personnel records, and social services records was developed. This web application was developed using Django and was integrated with PyTesseract – a Python OCR Library, to manage the recognition and extraction of text from the uploaded scanned files. For text recognition and extraction, Python-tesseract, an optical character recognition (OCR) tool for python was used to recognize and read the text embedded in images. OCR involves detecting the text content on scanned materials such as images and then to translate the recognized text to encoded text that the computer can easily understand. Using OCR makes the digitization process more comfortable as the document can be scanned, processed and the text extracted stored in an editable form such as a word document [62]. However, the process may not be 100% accurate and might need human intervention to correct some

elements that were not scanned correctly. The Python tesseract is a wrapper for Google's Tesseract-OCR Engine. It is useful as a stand-alone invocation script to tesseract and it has the ability to read all image types supported by Pillow and Leptonica imaging libraries, including jpeg, png, gif, BMP, tiff, etc. Thus, using the Python-tesseract as a script in the system helped the recognized text to be printed instead of writing it to a file and the extracted text was automatically populated as the metadata form for record indexing [61]. After extracting the text from the imported file, the records need to be classified into respective categories namely, administrative, financial, legal, personnel, and social services records for easy storage and document management. The study considered two methods of classification: the first method involved using assistance from the expert for classification. This allowed the expert to identify which class the imported documents belong to as the expert is knowledgeable about all the details of the documents, including its metadata. The second method involved using the N-Grams based Linear Classifier Model. Using this model, extracted text is fed into the model and the model returns an integer value that corresponds to the classification of the records. The model predicts the class of the new record based on the classified documents. This means that the classification accuracy using the model will depend on the number of records already classified and stored in the database. The more records organized, the more accurate the classification of the model is. Therefore, even though this method may be effective for classification, it still requires expert assistance at the early stage of the record's class and the model requires frequent updating as the records increases to ensure more accuracy [61].

2.9.5 Managing and Retrieving Bilingual Documents Using Artificial Intelligence-Based Ontological Framework

In many environments, organizations store documents in Portable Document Format (PDF) form and their relevant metadata in a different storage location. This proves a challenge as organizations cannot access the document’s content without its metadata. In a quest to handle such a problem, Fahad and Sait [63] applied Artificial Intelligence technologies to document and content management systems. They developed an ontological framework for document management that employed a text extractor designed using the AI-based Tessaract module that extracts the text from the image, the Multinomial Naïve Bayes (MNB) for classifying the documents and a link-based ranking technique which is a combination of PageRank, HITS, and SALSA algorithms. The link based technique was implemented as a ranking approach to retrieve relevant documents for a

user query. The outcome revealed that the proposed ontological framework achieved adequate Precision, Recall, F1-score, and Accuracy for the bilingual documents using a user query [63].

2.9.6 Forward-looking analysis based on grants data and machine learning based research classifications as an analytical tool

In a study carried out by Christian et al [64], the application of machine-learning based models across publications, grants, and other documents was investigated in order to determine whether a consistent portfolio view across inputs and outputs could be achieved. Their study revealed that using machine-learning-based models, makes it is possible to apply the same categorization approach to different document sets, for example to grant descriptions as inputs and to publications as an output. The study concluded that this approach creates comparable data sets where natural language processing and machine learning are used to tap into the substance of the research allowing thereby allowing for immediate and deep insights [64]. A study to compare the performance of humans or machine learning (ML) classification models at classifying scientific research abstracts according to a fixed set of discipline groups was carried out by Chong et al [65]. In their study, human operators (undergraduate and postgraduate assistants) were employed for this task in separate stages, and their performance compared against the performance of support vectors machine learning algorithm at classifying the European Research Council Grant project abstracts. Their findings revealed that Machine Learning Models are more accurate than human classifiers, across a variety of training and test datasets, and across evaluation panels. They concluded that ML classifiers trained on different training sets were also more dependable than human classifiers. Additionally, they concluded that machine learning models are a cost effective and highly accurate method for addressing problems in comparative bibliometric analysis, such as harmonising the discipline classifications of research from different funding agencies or countries [65].

2.9.7 Evaluating human versus machine learning performance in classifying research abstracts

The study compared whether humans or machine learning (ML) classification models are better at classifying scientific research abstracts according to a fixed set of discipline groups. The study employed undergraduate and postgraduate assistants to classify the

abstracts and compare their performance against the performance of a machine learning algorithm called the support vectors machine [65]. The study's goals where to compare whether ML are more accurate than humans at classifying scientific abstracts, whether ML is more reliable than humans at classifying scientific abstracts and to what extent does human classification performance improve, relative to ML, through increased task training, increased prior knowledge, selection on past performance, and feedback [65]. The data that was used for this purpose were abstracts of the European Research Council (ERC) Starting Grant (StG) funded projects that were accepted between 2009 and 2016 inclusive. The ERC evaluation panel structure has been stable since 2008 [66] and thus the study chose to use them as they are considered as the ground truth. Their study design was divided in four stage, the undergraduate stage, the high performance undergraduates, high-performance undergraduates plus feedback stage and the postgraduate stage. In the Undergraduates stage, 63 undergraduate student assistants from Nanyang Technological University were recruited. They were recruited from this major research university in Singapore for a full-day task. To evaluate potential classifiers for the study, the applicants were given a short example task to classify two research abstracts.

Sixty three (63) out of a hundred (100) qualified for the task and they were given one of four training sets of abstracts in the morning [65]. Each training abstract was labelled according to the ERC evaluation panel also referred to as the 'ground truth'. This enabled the assistants to study how the abstracts ought to be classified. In the afternoon, they were given a test set of different abstracts, with the ERC evaluation panel labels removed, and were told to assign each abstract to the ERC panel so as to match the 'ground truth'. Peer discussion and internet use was prohibited to the assistants. This was to ensure that their performance was as a result of learning from their training set only. In the second stage referred to as the high-performance undergraduates stage, eight undergraduates from each training set group with the highest accuracy scores were retained [65]. The high-performance undergraduates were given another test set to classify without further training. This stage was intended to address the selection effects for human classification performance. Finally, the third stage referred to as the high-performance undergraduates plus feedback stage. Under this stage feedback on classification of the abstracts they had classified in the previous two stages was given. Then the assistants were given a third test set to classify. Finally, in the last postgraduate's stage, 26 Ph.D. students and postdoctoral researchers in STEM disciplines (postgraduates) from Nanyang Technological University

were recruited for a half-day task where they were given a test set to classify without any training or task exposure. They were also not allowed to discuss or use the internet use was not allowed [65]. This stage was intended to assess the effect of prior knowledge on human classification performance. For the Machine Learning (ML) classification, the study used the support vector machines (SVM) algorithm as it is known to exhibit the best abstract classification performance among the basic supervised classification algorithms [67]. The SVM was combined with bag-of-words pre-processing of the abstracts and the use of the text frequency-inverse document frequency (TF-IDF) for feature selection and extraction. For a fair comparison of classification performance between human and ML classifiers, the training for the ML classifiers was restricted to the same amount of training given to the human classifiers. In the undergraduate stages, the ML classifier was trained using each of the four training sets. Subsequently the Undergraduate human classifiers from each training set group were compared only to the performance of the ML classifier that had been given the same training set. In the Postgraduates stage, the classifiers had no training sets provided to them as their doctoral training was considered to have contributed a great amount of knowledge and disciplinary boundaries of Science due to the extensive period of training. To simulate this extensive background training, the ML classifiers in the Postgraduates stage for each test set were trained using all other abstracts that were left out of the test set [65]. The results revealed that ML classifiers are better at replicating the ground truth classification than human classifiers. Although some results revealed that some individual human classifiers performed as well as the Machine Learning models when trained on almost the entire corpus. The study concluded that ML classifiers, given sufficiently large training data sets, are consistently and highly accurate. Additionally, the study concluded that Machine Learning performance is robust to variations in the training and test data and that human classifiers and are less reliable than Machine Learning classifiers. In addition to having an impressive classification accuracy and reliability, Machine learning classifiers are also more efficient to train and use due to modern computing power [65].

2.9.8 Automatic text categorization and its application to text retrieval

Lam et al [21] carried out a research to determine the effectiveness of automatic text categorisation on text retrieval. They developed an approach that was derived from a combination of a learning paradigm known as instance-based learning and an advanced

document retrieval technique known as retrieval feedback. They experimented the effectiveness of this model on two real world document collections extracted from the Medline database namely the HERSH and the OHSUMED. Their findings revealed that automatic categorization of documents improves the text retrieval quality compared to using manual categorization [21]. Phiri [22] also applied the concept of automatic classification of digital objects to improve Metadata Quality of Electronic Theses and Dissertations in Institutional Repositories. The study evaluated the performance of three classification models namely, the Multinomial Naive Bayes, Random Forest (RF) and Stochastic Gradient Descent (SGD) [22]. The datasets used for conducting experiments were prepared using data harvested from the University of Zambia Institutional Repository (UNZA-IR) and from the annual MEDLINE/PubMed citations [23]. The results exhibited a highest performance accuracy from the SGD text classifier. Thus, it was concluded that automatic classification using the supervised text classifiers applied in their study have the possibility of reducing the errors that result during the preparation of metadata.

2.9.9 Applying Machine Learning to Compare Research Grant Programs

In another study, Khor, K. A., Ko, Giovanni and Theseira [67], Walter carried a study to evaluate research grant programs using machine learning. They compared the performance of three machine learning classification models, multinomial Naïve Bayes (MNB), Multinomial Logistic Regression (MLR) and Support Vector Machines (SVM) to classify the research proposals according to the research funding structure used by the European Research Council (ERC). The results revealed that the SVM model exhibited a better performance compared to the MNB and MLR in classifying the research proposals. Based on the results, the study was able to determine which funding programme was more successful than the other in a particular research discipline [67].

2.9.10 Comparison of Text Classification Algorithms

This paper presents an empirical study of three text classification algorithms using two datasets. Naïve Bayes, Support Vector Machine and C4.5 have been compared by training the dataset instances on the Weka Tool. The two datasets used were Diabetes and Calories. Diabetes dataset has a large number of training examples and attributes as compared to the Calories dataset [68]. The results were compared based on the recall and precision values that each of the algorithms were returning. Another basis of comparison was the percentage split of the dataset into training set and test set. Their results showed that out

of the three classifiers, SVM is computationally efficient. SVM has certain disadvantages which degrades its performance for small datasets. Thus, it was proposed that using Hybrid SVM may improve the existing drawbacks of SVM. Even if the approach with which SVM is applied on the dataset is changed, it can produce optimized results [68]. Other researchers have also compared text classification algorithms; Aggarwal [69] conducted a survey of text classification algorithms. Some of the classifiers surveyed by them were Probabilistic and Naive Bayes classifiers, rule - based classifiers, and Multinomial distribution. Likewise, Jimenez [70] demonstrated an example of text classification and clustering with Weka. A movie review dataset was classified into positive or negative reviews. The text of reviews was converted into vector format and Naive Bayes classifier was used for classification. Clustering was also carried out, and 65.25% instances were clustered correctly. Wilcox [71] applied classification algorithms to narrative reports. Methods such as Decision Trees, Bayesian classifiers were used to classify X-ray reports according to six attributes. They concluded that text classification algorithms were dependent on training set size. Similarly, Pandey [72] has reviewed text classification techniques for email filtering and management. Approaches used by them were Naive Bayes, Support Vector Machine (SVM), Decision Trees, Fuzzy Logic, etc. They concluded that for filtering, context based email-organization has the best potential

2.9.11 Machine-learning-based classification of research grant award records

Freyman et al [73] explored how topic co-clustering which is an approach to text analysis based on machine learning could be used to tag National Science Foundation (NSF) grant awards automatically with terms referring to scientific disciplines or to socioeconomic objectives. Their results revealed that in the case of scientific disciplines, where their language models were well-formed and had a valid comparison set for manual classification, the machine-assigned tags were a reasonable and valid means for describing the research conducted under each grant [73].

2.9.12 Blockchain implementation to verify archives integrity on cilegon E-archive

This study aimed at developing an archive management system that considered information security aspects, such as availability, confidentiality, and integrity. The system was developed to curb the challenges that the Cilegon E-Archive (CEA) system was facing at the time [59]. The Cilegon E-Archive is a centralized system running on a single cloud server and used for managing the lifecycle of archives. The centralized system

used a client–server architecture in which nodes connected to one server and were controlled by the network administrator. Thus, the challenges being faced were a single point of failure, low data availability, and difficulty in proving the originality of files [59]. A single point of failure is a situation in which the entire system stops working when the server malfunctions. This prevents clients from connecting to the server, so the system cannot process requests to provide data. To solve these problems the paper propose a new CEA system built on a decentralized architecture. The new archiving system was integrated with IPFS and blockchain technology [59]. Interplanetary File System (IPFS) is a mechanism of sharing files with other nodes in the network instead of only storing the files in one server. In addition the IPFS is incorporated with a Pinning service that ensures high availability and long-term retention of files. IPFS generates a hash to identify the content of the file. A CEA-DApp was developed as a front-end system, to make it more convenient for users to interact with the IPFS private network and CEA Network. It was built on the Truffle Suite framework, using ReactJS programming. A smart contract named CEAContract.sol was integrated on the CEA-Dapp,to store the hash of a file [59].

The CEAContract.sol comprised a state variable storedData typed string to model the hash of a file from IPFS. It also comprised set and get functions. The set function was developed to receive an input value which was then stored in storedData, while the get function displayed the storedData value. Before deploying a smart contract in CEA Network, generated accounts from CEA Network were imported into Metamask and unlocked them [59]. The results of the functionality evaluation of the system revealed that the files were well distributed among the nodes in the IPFS private network, where all nodes participated in providing data. It was therefore concluded that the CEA DApp offers a solution to the single point of failure problem and that it improves data availability. The nodes without the right swarm key cannot connect to the IPFS private network. Data confidentiality is also assured as nodes without the right swarm key cannot connect to the IPFS private network. Additionally, the hashes of files are stored as transactions on CEA Network and the transactions are unchangeable, digitally signed with Ethereum accounts, timestamped, and knowable by all participants on the CEA Network [59].

2.9.13 Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories

This study sought to demonstrate the feasibility of automatically classifying IR digital objects using the minimum possible input expected from Electronic Theses and Dissertation (ETD) manuscripts. The thought emanated from the fact that most institutional Repositories ingested digital objects either through self-archiving [52] with manuscript authors tasked with the responsibility of depositing the manuscript or by a central authority such as Library. Both instances bring about the potential to misclassify the digital objects by way of depositing them in the wrong collection or incorrectly tagging them with non-subject specific subject descriptions. The study assessed the case of the University of Zambia (UNZA), whose archiving challenges had been worsened due to the fact that the institution repository lacked self-archiving of ETDs and depended on two individuals from the Library who were responsible for the ingestion of IR objects. According to the analyses conducted in prior work [74], the findings revealed that there was a lengthy turnaround time between submission of scholarly research output and eventual ingestion into the IR. It was also revealed that the lack of use of subject-specific controlled vocabulary sets compromised the discoverability of digital objects and had an adverse effect on downstream service providers, such as the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog1 and the Open Access Theses and Dissertations portal all of which automatically harvest metadata from IRs that functioned as data providers [74]. Based on these challenges the study sought to achieve the following: devising a mechanism for reclassifying digital objects that were already ingested into the repository; Coming up with means for the implementation of tools that could potentially leverage automatic classification of digital objects in order to guarantee accurate classification of digital objects, using standard tags and also reducing the time taken to ingest digital objects into IRs [75]. Thus, the study aimed to contribute to this work by identifying a core set of features, extracted from the minimum possible input of ETD manuscripts provided by ETD authors. The study also sought to examine the performance of Classification models for automatically classifying ETD types, associated IR collections and subject headings for ETDs. The study used the Cross Standard Industry Process for Data Mining (CRISP-DM) methodology which comprises of six stages namely Business understanding, Data understanding, Data Preparation, Modelling, Evaluation and Deployment [76]. To demonstrate how standardised controlled vocabulary sets can be associated with ETDs [75]. To address the goal of implementing classification models for

automatically classifying ETDs. Three classification models were implemented in order to automatically associate structural metadata and descriptive metadata to ETDs. The models evaluated were various scikit-learn implementations of Logistic Regression, Naive Bayes (Multinomial), Random Forest and the Stochastic Gradient Descent (SGD). To train the models, ETDs from the Faculty of Medicine, that is, the annual baseline MEDLINE/PubMed citation records were used as a test case. The citation records titles, abstracts, MeSH label and combinations of the three were used to determine the most effective features [75]. When comparing the title, abstract and title+ abstract features, the best performing feature was title+ abstract, with an accuracy score of 54.6%, using the SGD Classifier. Though the accuracy score was low, the study concluded that it was comparable to results reported in similar studies involving automatic classification of MeSH subject headings. In conclusion, study proved the possibility automatically classifying ETDs in IRs, using supervised machine learning techniques, by extracting features from the minimum possible input expected from document authors: the ETD manuscript. Automatic classification of repository objects enhances the searching and browsing of content in IRs and further provides platforms for the implementation of third-party tools and extensions that could potentially result in effective self-archiving strategies [75].

2.9.14 The Elements of Statistical Learning: Data Mining, Inference, and Prediction

This paper studies and compares algorithms that are popular in data mining which is an important aspect and contribution to data archival and retrieval. The study compared the most popular text classification algorithms namely, Nearest Neighbour classifier, Bayesian Classification, Support Vector Machine (SVM), Association Based Classification, Term Graph Model, Centroid Based Classification, Decision Tree Induction and Classification model using Neural Network as some of the most popular text classification algorithms [77]. The KNN model may not be efficient for all text classification problems as it works well on smaller datasets, which need to be properly pre- processed and represented in a suitable vector space. The findings were that the KNN model was suitable for document classification problems when the dataset was small and well pre-processed. It did not perform well on large datasets. Their findings also revealed that SVM produces the highest accuracy and thus is one of the most popular models in text classification even though it has a poor recall [78].

2.9.15 Character-level convolutional networks for text classification

Text classification is a crucial step for archiving system as it facilitate the retrieval of documents. Text classification involves assigning predefined categories to free-text documents. The range of text classification research goes from designing the best features to choosing the best possible machine learning classifiers. A number of studies have revealed that convolutional networks [79] [80] are useful in extracting information from raw signals, such as computer vision applications, speech recognition and others. This study explored treating text as a kind of raw signal at character level, and applying one-dimensional Convolutional Network to it. This study incorporated a classification task as a way to exemplify Convolutional Networks' ability to understand texts. It also incorporated large data set as it historically known that Convolutional Networks usually require large-scale datasets to work. In this study, two Convolutional Networks, one large and one small, were designed. They were both designed nine layers deep with six convolutional layers and three fully-connected layers. The study then made comparisons of traditional models and deep learning models in relation to a number of datasets. The experimental comparisons of the traditional models included the Bag-of-words and its TFIDF, Bag-of-ngrams and its TFIDF and Bag-of-means on word embedding, whereas the deep learning models used included the Word-based Convolutionary Networks and Long-Short Term Memory (LSTM) recurrent neural network [81]. In order to come up with a dataset large enough to carter for all the models under experiment, the researchers built a data set from the following sources; WAG's news corpus, the AG's corpus of news articles (30,000 training and 1900 testing dataset), Sogou news corpus (90,000 training and 12,000 testing set), DBPedia ontology dataset (40,000 training and 5000 testing dataset), Yelp reviews (130,000 training and 5000 testing dataset) Yahoo! Answers dataset (600,000 training and 130,000 testing dataset).

Amazon reviews. From the experiments, the most significant conclusions drawn are that character-level Convolutional Networks could work for text classification without the need for words. This means that language is also considered as a signal no different from any other kind [81].

The other conclusion drawn was that the Dataset size forms a dichotomy between traditional and Convolutional Network models; the larger datasets exhibit better performance. Traditional methods like n-grams TFIDF remain strong candidates for

dataset of size up to several hundreds of thousands, whereas the dataset needs to have several millions of characters for the character-level Convolutional Networks to exhibit better performance. The study also concluded that Convolutional Networks may work well for user-generated data even though the performance may be affected by the degree of curation of the data. The results from the experiments show that convolutional networks performed better on less curated data [81]. According to the study, this may imply that Convolutional Networks may have better applicability to real-world scenarios. The study then suggested that further analysis is needed to assess whether Convolutional Networks are truly good at identifying exotic character combinations such as misspellings and emoticons, as the experiments carried out in this study did not show any explicit evidence. It was also revealed that that changing the alphabet in the datasets by distinguishing between uppercase and lowercase letters could effect a difference in the performance of the machine learning models. Although for million-scale datasets, omitting such distinction would work better [81]. Lastly, the study concluded that semantics of tasks in the dataset do not affect the performance of the models. The datasets used in the experiments consist of two kinds of tasks: sentiment analysis (Yelp and Amazon reviews) and topic classification (all others). This dichotomy in task semantics does not seem to play a role in deciding which method is better. From the experiments, the study also verified that there is not a single machine learning model that can work for all kinds of datasets [81]. The choice of which machine learning model would work best especially in archiving would depend on the specific problem application. This means that if convolutional networks have to be applied to text classification tasks in an archiving system, the factors drawn from the study's conclusion would need to be considered.

2.9.16 Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text of a Digital Dissertation Information Management System

Mai et al [82] proposed a document title-based semantic subject indexing approach using deep learning Multilayer Perceptron MLP, CNN and Recurrent neural networks (RNN) models. This study sought to devise better ways of deducing Semantic annotations for subject indexing in digital libraries. Semantic annotations are critical because enhance the search of scientific documents for users of digital libraries. Automatic annotation systems are also useful when classifying publications into categories, especially, when there is a

large amount of new publications [82]. However, it is challenging to generate automated annotations due poor availability of data from which recommendations may be generated. Often times it is difficult to obtain either the full-text of a publication or its abstract. In some instances where the full text is available, text mining may be prohibited due to copyright laws or regulations of the publishers. Additionally, where it is possible, collecting and processing of PDFs requires high computational requirements [82]. The study therefore sought to address these challenges by proposing an annotation method that is based on data with better availability, such as the title. To perform their experiments data was extracted from PubMed and EconBiz, where 12.83 million titles and 1.06 million titles were extracted respectively. In order to fully utilize these large amounts of data, the study developed and compared three different classifiers that have emerged from the deep learning community in recent years; convolutional neural networks (CNNs), recurrent neural networks (RNNs) and multi-layer-perceptrons (MLPs). In addition to comparing their performances, the classifiers are compared against another strong MLP baseline (Base-MLP), which has previously been known to also outperform traditional bag-of-words classifiers such as the Support Vector Machines, Naive Bayes, and k-Nearest Neighbour [83]. The findings of the study prove that title-based methods can match or even outperform the full-text performance when enough training data is available. When trained using the EconBiz dataset, the MLP title-based classifier (MLP) exhibited the same performance as the

MLP full-text based classifier (MLP) when trained with only 8× as many titles as there are full-texts available, whereas the title-based MLP outperformed the full-text based classifier when all available titles used were approximately 15× more than full-texts [82]. On the other hand the full-text based MLP performed better than the title-based RNN when trained on the PubMed dataset. The performance gap is 10.7% when using the same number of titles as full-texts available. All in all the study concluded that the MLP exhibits the best performance among the three classifiers, outperforming the RNN and CNN in three out of the four combinations of the two datasets and title/full-text. Additionally, the MLP also consistently outperforms the baseline when all titles or full-texts are used [82]. The RNN exhibited a rather poor performance on full-text, but exhibited a strong performance on titles. The CNN on the other hand, exhibited the lowest performance compared to other classifiers in all cases, even though its individual performance was generally acceptable. Therefore the study drew the following conclusions; that the title-

based methods can reach the performance of full-text-based methods by exploiting the surplus of available training data and that the title-based methods exhibit the same or outperform full-text methods when the number of training samples is sufficiently large. Based on these conclusions, the study suggests that a semantic indexing system based on the title can reach the performance of a system based on the full-text if the number of samples for training the title-based method is much larger than the number of full-text samples [82].

2.9.17 IBGA: An Incentives Based Grant Allocation Algorithm for Academic Institution

This paper proposes a scheme for allocating the grant to academic institutions. The scheme has a provision of base grant that is allocated to each faculty member/research student irrespective of his/her performance. The grant is further augmented based on the quality of research papers published and the amount of augmentation is computed based on the impact factors of journals where the research papers of the faculty member/research student are published. This scheme was devised to address the challenge that comes from the fact that the resources (in terms of grant) are limited and the funding agency or the government wishes to allocate it in the best possible manner. [84]

From the literature that has been reviewed a number of techniques to facilitate effective document management (DM) functionalities in organizations have emerged. The use of machine learning techniques in solving document archiving problems have been extensively been studied. However, such technologies have not been embraced by most developing countries including Zambia as there are challenges on how to interpret advanced technologies. This research project endeavoured to develop a document archiving system that will use machine learning to classify the research documents and to keep track of critical events in the research process. The approach used in the latter study is similar to what was used in this study. The difference is that the performance of three text classification algorithms were compared and the best performing model integrated in a web based application to facilitate the awarding of grants.

2.10 A Summary of the Related Works

The findings revealed that a number of mechanisms including the application of machine learning tools have been used in text classification problems and document archiving systems. The table below gives a summary of the literature that has been reviewed and is related to this study.

Table 1: Literature Review and Gaps

	Article	**Author (s)**	**Findings**	**Gaps**
1	Development of Electronic Document Archive Management System (EDAMS): A CaseStudy of a University Registrar in the Philippines	Caluza, J. (2017)	They developed an electronic document archiving management system that automated most of the processes and allowed faster retrieval of documents	Human intervention was still needed to enter all the details of the documents and facilitate the correct classification of the documents at input level. This increases the risk of human error
2	Web Based Document Archiving Using Time Stamp and Barcode Technologies–A Case of the University of Zambia.	Mutale, B. M., & Phiri, J (2016)	They applied Time Stamping and Barcode Technologies to index, archive and retrieve documents. Additionally they developed the system to send reminders for critical transactions through pop-ups, emails and Short Message Service (SMS) systems.	It required human intervention to manually index and classify documents according to respective category such as memo, contract or MOU.

3.	OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga.	M. E. M. O. Jayoma et al (2020)	This web application was developed using Django and was integrated with PyTesseract – a Python OCR Library, to manage the recognition and extraction of text from the uploaded scanned files.	In order to accurately classify the records, this model needs expert assistance and the model requires frequent updating as the records increases to ensure more accuracy.
4	Managing and Retrieving Bilingual Documents Using Artificial Intelligence-Based Ontological Framework	Fahad, A. A and Sait, A. R. W (2022)	They developed an ontological framework for document management that employed a text extractor designed using the AI-based Tessaract module that extracts the text from the image, the Multinomial Naïve Bayes (MNB) for classifying the documents.	The focus was on the extraction of text from images on the documents and used the information for document classification.
5	Forward-looking analysis based on grants data and machine learning based research classifications as an analytical tool	Christian et al (2020)	Machine learning models help to achieve a consistent portfolio view across inputs and outputs for different document sets.	The model was used to generate portfolio analyses

6	Evaluating human versus machine learning performance in classifying research abstracts	Yeow Chong Goh et al (2022)	Machine Learning Models are more accurate and dependable than human classifiers, across a variety of training and test datasets, and across evaluation panels.	Performance comparison of models in classifying research abstracts
7	Applying Machine Learning to Compare Research Grant Programs	Khor et al (2022)	The SVM model exhibited a better performance compared to the MNB and MLR in classifying research proposals. Hence it was used to determine which funding programme was more successful than the other in a particular research discipline	Models used to compare which funding programme is more successful than the other.
8	Machine-learning-based classification of research grant award records	Freyman, C et al (2022)	Their results revealed that in the case of scientific disciplines, where their language models were well-formed and had a valid comparison set for manual classification, the machine-assigned tags were a reasonable and valid means for describing the research conducted under each grant	Description of research carried out under each grant

9	Blockchain implementation to verify archives integrity on cilegon E-archive	Permatasari, k et al (2020)	The study used an SVM text classification Model to determine which funding programme was more successful than the other in a particular research discipline	A binary SVM classifier was used, a multi classifier would be perform better in determining the discipline a research would be more successful
10	Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories	Phiri, L. (2020)	Developed a Stochastic Gradient Descent (SGD) model to improve Metadata Quality of ETDs in Institutional Repositories. The concluded that automatic document classification using the supervised text classifiers reduce the errors that result during the preparation of metadata.	Focused on applying automatic document classification for the improvement of meta data quality
11	Automatic text categorization and its application to text retrieval	Lam et al (2023)	They developed an instance-based learning and an advanced document retrieval technique known as retrieval feedback. Their findings revealed that automatic categorization of documents improves	Instance-based learning is less effective at generalizing from the training data.

			the text retrieval quality compared to using manual categorization	
12	The Elements of Statistical Learning: Data Mining, Inference, and Prediction	Hastie, T., Tibshirani, R., Friedman, J. (2009)	The findings were that the KNN model was suitable for document classification problems when the dataset was small and well pre-processed. It did not perform well on large datasets.	The KNN model used in this study may not be efficient for all text classification problems as it works well on smaller datasets, which need to be properly pre-processed and represented in a suitable vector space
13	Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories	Nidhi and Gupta (2011)	They compared the most popular text classification algorithms namely, Nearest Neighbour classifier, Bayesian Classification, Support Vector Machine (SVM), Association Based Classification, Term Graph Model, Centroid Based Classification, Decision Tree Induction and Classification model using Neural Network as some of the most popular	Their findings revealed that SVM produces the highest accuracy and thus is one of the most popular models in text classification even though it has a poor recall.

			text classification algorithms.	
14	Character-level convolutional networks for text classification	Zhang et al (2015)	They designed a multi-class text classification using character-level Convolutional Neural Networks (CNN) on large-scale datasets. They compared the performance of this model to the performance of traditional models, such as Bag of Words (BOW), N-grams, and deep leaning such as CNN and Long Short-Term Memory (LSTM) models.	Their results concluded that the traditional models overcame the proposed deep learning models with small data sets; whereas, in the case of large-scale datasets, character level deep learning approaches were superior
15	Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text	Mai et al (2018)	They proposed a document title-based approach using deep learning Multilayer Perceptron MLP in comparison to CNN and Recurrent neural networks (RNN) models.	Their proposed deep learning model (MLP) outperformed the CNN and RNN by a large margin.

16	IBGA: An Incentives Based Grant Allocation Algorithm for Academic Institutions	Abbas, A.M. (2012)	They proposed a grant allocation scheme that considered member/research student irrespective of his/her performance and further considered the quality of research papers published.	Despite the effectiveness of the scheme, the study incorporated the use of traditional means to classify the research publications

2.11 Chapter Summary

This chapter reviewed the literature by the other scholars and researchers on the subject matter of document archiving systems. From the discussions, most studies have elaborated different approaches to the design and development of document archiving systems for different purposes. However, none of the have discussed the development of document archiving systems in relation to the awarding of research or innovation grants. This study therefore seeks to fill in this research gap by discussing, designing and developing a document archiving system for the awarding of research and innovation grants.

3 RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the research methodology that were employed in the study. The chapter includes the design methodology, requirements specifications, design specifications and system implementation. The study adopted a fusion of the Agile and the Cross Industry Standard Process for Data Mining methodologies.

3.2 Research Design Methodology

A design methodologies in information systems refers a structured approach or framework is adopted to create and develop information systems that meet specific user and business requirements. The frameworks provide a basis or set of guidelines that can be used to analyse, design, and implement information systems. The most commonly used design methodologies in information systems are the Waterfall Methodology, Agile Methodology, Rapid Application Development (RAD) Methodology and the Object Oriented Methodology.

3.2.1 The Waterfall Methodology

The Waterfall methodology is described as a traditional and linear approach to software development. It comprises of a sequence of phases, each of which must be completed before moving on to the next. The software development process follows a linear progression from beginning to end of the project meaning that, the next phase of the process cannot begin until the one that came before it has been completed [46]. The waterfall model is made up of the requirements analysis, system design, implementation, testing, deployment, and maintenance phases [85]. The waterfall methodology is popular in project management and it is suitable for projects that have clear and stable requirements, well-defined scope, and predictable outcomes. Its biggest limitation is that it is less flexible and does not accommodate changing requirements; changes that come further in the process can be time-consuming and costly to accommodate [86].

3.2.2 Agile Methodology

The Agile methodology is an iterative and incremental approach to software development that accommodates flexibility, collaboration and customer feedback. It focuses in delivering functional software in short development cycles called iterations. It is popular due to its adaptability and responsiveness to changing requirements [87].

3.2.3 Rapid Application Development (RAD)

Rapid Application Development (RAD) is also an incremental and iterative approach that provides a super flexible and adaptable process to the design and development of software. RAD consists of four phases: Plan Requirements, User Design and Input, Rapid construction, and Finalization. Unlike the agile methodology which focuses more on continuous planning and adaptation, the RAD methodology has less planning and prioritises rapid prototyping and feedback [56]. It emphasizes on delivering working software quickly and adapting to changes in requirements and technology. This makes the methodology popular as it results in delivering the software faster thereby lowering development costs. However, its lack of user involvement throughout each stage of the project is a drawback [88].

3.2.4 Object-Oriented Methodology

Object-Oriented Methodology is a software design methodology that incorporates the use of objects and classes to design and develop software systems. It is widely used for building software systems that require a modular and reusable outcome. This makes it popular in modelling complex systems. Even though this methodology promotes modularity, reusability, maintainability and scalability, it has a steep learning curve and the system developed consumes a lot of memory space. Additionally, execution efficiency may be reduced due to overheads of creating and managing objects.

3.2.5 Cross Industry Standard Process for Data Mining (CRISP-DM) methodology

CRISP-DM is a popular and comprehensive methodology used for data mining and analytics projects. It provides a structure approach that helps data professionals effectively leverage data to make data-driven decisions and solve complex business problems [76]. It comprises of six (6) stages; Business Understanding, Data Understanding, Data Preparation, Data Preparation, Data Preparation and Deployment [76].

3.2.6 Selected Methodology

The study adopted a hybrid or blended methodology approach that combined the agile and the CRISP-DM methodologies. The agile methodology was selected because of its emphasis on flexibility and adaptability to changing requirements which was the nature of

this project. The project also involved working closely with users to deliver prototypes and obtain feedback from them. Thus, the agile methodology was most suited. The system developed involved the integration of a data mining algorithm and as such the CRISP – DM methodology was also adopted. The six (6) phases of the CRISP-DM methodology were applied as follows:

3.2.6.1 Business Understanding

In this phase, the processes in the funding institution namely the Ministry of Technology and Science were understood in order to identify and define objectives. The study sought to understand the process that applicants use to apply for research and innovation grants, how funding institutions process the research and innovation grant applications and the challenges that the funding institutions face when processing applications. This step was taken to ensure that the study's objectives are aligned to the business perspective of the funding institutions.

3.2.6.2 Data Understanding

In this phase, a number of data sources were explored to understand the available data and assess its quality and relevance to the study. It involved extracting historical data from the National Science and Technology Council (NSTC) which is a funding institution under the Ministry of Technology and Science in Zambia. The extracted data was also examined to identify patterns and trends that would be relevant for the modelling phase.

3.2.6.3 Data Preparation

This stage involved pre-processing the extracted data. Pre-processing is a method used to clean the data and prepare it for modelling. Data cleaning the removal of unnecessary words known as stopwords, outliers, null and missing values, as well as transforming it into formats required for the modelling [89]. This stage involved a process called feature selection and feature extraction which involved the selection of the necessary features that will be used as inputs and selecting relevant features for modelling. The specific pre-processing functions used were TfId-Transformer, Token Count Vectorizer and NGramVectorizer. Feature selection involves filtering a set of features by selecting the relevant, informative and discriminative features while discarding the redundant or irrelevant ones [90]. This process helps to reduce noisy data and reduce the dimensionality of the dataset. Feature extraction involves summarising the original set

of features in order to create new ones. To extract the features, the Term Frequency-Inverse document frequency (TF-IDF) was used. TF counts the number of words in each document and assigns it to the feature space while the IDF assigns a higher weight to words with either high or low frequencies term in the document. Data was further transformed to bring it to a common scale or distribution. Oversampling was also applied on the dataset to deal with class imbalances in the data. Finally, the data was split into training and testing data using a ratio of 80 to 20. All these tasks were carried out to prepare the data. The preparation of data is an essential step to the performance of the model as the quality of the data directly determines the performance of the models [89].

3.2.6.4 Modelling

This phase involved training of models. Three text classification algorithms namely the K-Nearest Neighbour (KNN), the Naïve Bayes (NB) and the Support Vector Machine (SVM) were selected because of their popularity in handling text classification tasks. The three models were trained using historical data of grant applications from the National Science and Technology Council of Zambia.

3.2.6.5 Evaluation

Under this stage, the performance of the three trained were evaluated by calculating the accuracy, precision, recall and F1-score of the models. These performance metrics form the basis of evaluating the performance of any machine learning model. Out of the three models that were trained, the SVM text classifier exhibited the best performance and was therefore, selected as the text classifier to be integrated in the web based archiving system developed in this study.

3.2.6.6 Deployment

This stage involved integrating the selected SVM text classification model into the web based archival system and developing an Application Programming Interface (API) to facilitate the integrating and deployment of the system into other platforms and third party tools, To facilitate the interaction of users and the system, a Web User Interface (UI) application was also developed.

3.3 System Design and Implementation

This section describes the requirements that were defined, the development materials and tools used to develop the system, as well as the software and hardware requirements recommended for the effective and efficient use of the system.

3.3.1 Functional Requirements

These are requirements that explain what has to be done in as system by identifying the necessary task, action or activity that must be accomplished. The following are the functional requirements:

- The system is able to allow applicants to apply for research and innovation grants online
- The system is able to automatically classify the research and innovation proposal titles into respective subject classes or fields.
- The system determines the eligibility of an application in obtaining a research or innovation fund.
- The system allows an administrator to post or advertise scholarships available
- The system categorises research applications according to the educational institution affiliated to.

3.3.2 Non-functional Requirements

These are requirements that specify the qualities or characteristics that should be exhibited by the software and are used to judge the behaviour of the system or software. The qualities include performance, scalability, security, usability and reliability.

- Performance- Appropriate data structures and well optimized algorithms were implemented in the system to reduce computational overhead and memory usage.
- Usability- The system was implemented with an easy to use, intuitive and consistent User Interface (UI) to allow users to interact with the system without difficulties.
- Security- To ensure the security of the system. Multi-factor authentication and separation of user roles was implemented. Furthermore, user passwords are encrypted in the database and all database connection details are hidden. Authentication sessions are automatically closed when the user is not active after login and all the information entered through forms are sanitised before interacting with the database to prevent SQL injections.

- Reliability-To ensure reliability the system was developed using a modular approach meaning that if errors occur, they are confined to a particular module and not the entire system. Additionally, open source technologies (PHP, MySQL, and Apache) were used to develop the system. Open source technologies have adequate support from IT communities in case of any failures.
- Scalability- A modular approach was used during system development meaning that each module is responsible for a specific task. Therefore, scaling the system involves addition of modules which would not affect or disturb the performance of existing modules. Furthermore, the system can integrate with any Application Programming Interface (API) to ensure a smooth integration with other systems and platforms.

3.3.3 System Development Tools

In order to develop a Data Mining and Machine Learning Model to support decision-making for research and innovation grant allocations, a range of software and hardware tools, libraries, and frameworks are necessary. The following are the tools that were used for development using a PHP environment:

3.3.3.1 Development of the web-based system

The following are the tools and materials used to develop the web-based system.

3.3.3.1.1 PHP (Hypertext Preprocessor)

PHP is a server-side scripting language suited for dynamic and interactive web development. It was selected because it can execute on the server, generating HTML or other output that is sent to the client's web browser. This enables tasks such as processing form data, interacting with databases, file manipulation and managing user authentication and sessions, to be performed on the server. It was also selected because it has a relatively low learning curve and it is open source meaning that it has a large community of developers contributing to its improvement and providing support.

3.3.3.1.2 MySQL

It is a relational database management system that is popular for its speed and efficiency. It was selected because it is easy to use, robust and it provides optimization techniques that enhance query execution and response times, and can handle read-heavy workloads.

3.3.3.1.3 Java script

JavaScript programming language is popular for its role in web development. It is a versatile and dynamic scripting language that can be embedded within HTML to add interactivity and dynamic behaviour to web pages. For this reason and because the system developed is web based, JavaScript was selected to facilitate the creation of client-side scripts that will run in a web browser.

3.3.3.1.4 Hyper Text Markup Language (HTML)

HTML is a Markup language that was used to create the web pages. It was selected because it is the key language used in website and web application development.

3.3.3.1.5 Cascading Style Sheets (CSS)

CSS is a language is used to control and style the presentation of data on websites and web applications. It was used in order to make web pages look attractive.

3.3.3.1.6 Web hosting Server

Apache is a server software which was used for web hosting. It was selected because it is open source and it is recommended as one of the most reliable and flexible server software. Its module based structure makes it easy to configure.

3.3.3.1.7 Text Editor

Bluefish was selected and used as the text editor.

3.3.3.1.8 Hyper Text Markup Language (HTML

3.3.3.2 Building and training of the machine learning model

The tools that were used to build and train the text classification model are discussed below.

3.3.3.2.1 Libraries

Libraries are a set of pre-written, reusable code, routines, functions, classes, or any other resources that can be integrated into a software platform saving developers time to re-write the code. PHP-ML is the library that was used to handle the machine learning tasks in the project.

3.3.3.2.2 Dataset

A dataset was obtained from the historical data extracted from the funding institution under the Ministry of Technology and Science, the National Science and Technology Council (NSTC). A total number of 129987 records were extracted and they included

information such as Research or Innovation proposal details, proposed budget, previous funding history, institutional information, research or innovation area or field, collaborator information if any and project timeline. The research or innovation proposal details included the Title and abstract of the research proposal and the respective subject areas. The input features selected for the purpose of the study were the title of the research or innovation and the research or innovation subject field or class. These features were selected due to the fact that granting of research or innovation funding is primarily based on the research or innovation proposal belonging to a specific field, in this case engineering, technology and science. The field topic forms the primary determining factor of grant allocation as opposed to other factors such as previous funding history, proposed budget and project time line. After the feature selection and extraction, the dataset was split into 80% training set and 20% testing set. Training set is used to train the model whereas the testing set is used to test the model's ability to make

3.3.3.2.3 Text Classifiers

Three text classification algorithms namely the K-Nearest Neighbour (KNN), the Naïve Bayes (NB) and the Support Vector Machine (SVM) multi-classifier were employed and their performance compared to select the best performing model. The SVM Multi-classifier exhibited the best performance and thus, was selected as the model for text classification.

3.3.3.3 Hardware Requirements

The following are the recommended minimum hardware requirements for the system:

- CPU (Central Processing Unit):
  The study recommends a multi-core processor minimum of corei7 with processor speed of 1.6GHz or faster. This is recommended in order for the system to handle the computational demands of training the machine learning models efficiently.
- Random Access Memory (RAM):
  The study recommends a minimum of 16 GB RAM but recommends even a higher size in order for the system to efficiently handle large datasets, complex model training and any other extensive and computationally intensive tasks.
- Storage (Hard Disc Space)

For primary storage, the study recommends a Solid State Drive (SSD) for faster data access and model training. A minimum of 512GB SSD is recommended, but larger storage capacity might be needed in case of large dataset sizes.
For Secondary Storage: The study recommends additional HDD (Hard Disk Drive) or cloud storage for data backups and larger datasets.

The above recommended hardware requirements serve as a baseline and can be adjusted based on the specific needs and scale of the machine learning project for research and innovation grant application process and allocation decisions. Further consultation with IT experts or data scientists can be made in order to fine-tune these requirements for optimal performance and scalability.

3.3.3.4 Software Requirements

The following specifications are recommended as the minimum software requirements

- Operating System:
  The system developed is able to run on any operating system platform including Linux, MacOS and Windows 10 or better.
- Web Browser: Any web browsers such as Firefox Mozilla, Google Chrome and Safari can be used to access the Internet. A stable, consistent and high-speed internet is also recommended for effective downloading of data, model updates, and access to cloud services if required.
- Web Hosting Server:
  The study recommends Apache web hosting server due to its ease of use, flexibility and readily availability as it is an open source software.
- Database Server:
  The study recommends MySQL, PostgreSQL or any SQL Database Management System for storage, retrieval and management of structured data. Furthermore the study recommends MongoDB, Cassandra and any other NoSQL Databases for handling unstructured or semi-structured data.

The above recommended software tools and libraries provide a comprehensive environment for data processing, model development, and analysis, supporting the

creation of a machine learning model for decision-making in research and Innovation grant allocation.

3.3.4 Design Specification

This section defines the technical and design specifications of the system using varying graphical design models.

3.3.5 System High level Design

The system High level design gives a brief description of the modules of the system and their functions.

- Data Collection and Integration Module

  This module is responsible for collecting and integrating historical information about research and innovation grant allocation decisions. It accesses the data source from which the historical information is held.

- Data Preprocessing and Feature Extraction Module:

  The historical information obtained may have a lot of data inconsistencies such as stopwords, outliers, missing values and punctuations, to mention but a few. Thus, it is this module's responsibility is to prepare and clean the collected data for processing. After the data is cleaned, the module is also responsible for the selection and extraction of the features that can be used to train the model.

- Machine Learning Model Development Module:

  This module is responsible for the building and training of the machine learning model, which is a text classification algorithm. The study employed a comparison of the performance of three text classification models, the K-Nearest Neighbour, the Naives Bayes and the Support Vector Machine. The Support Vector machine emerged as the text classifier to be integrated in the system.

- Model Integration Module

  This module is responsible for the integration of the developed SVM text classifier into the grant application and allocation process. Thus, a Web-User Interface was integrated with the text classifier.

- User Interface Module

The User Interface Module provides a user-friendly interface for model interaction. It displays model predictions, statistics, and insights and allows users to use the information in the facilitation of the research and innovation grant application process.

- Testing and Quality Assurance Module

  This module ensures that the model and system built are effective, reliable and meet the specified quality standards. The module has functions that validate and verify the model's accuracy and consistency

- Security and Compliance Module

  This module ensures data and system security by using permission-based access controls and encryption of sensitive data.

The diagram below shows the overview of the system which is an integration of the text classifier and the web based system (Web Archiving System).

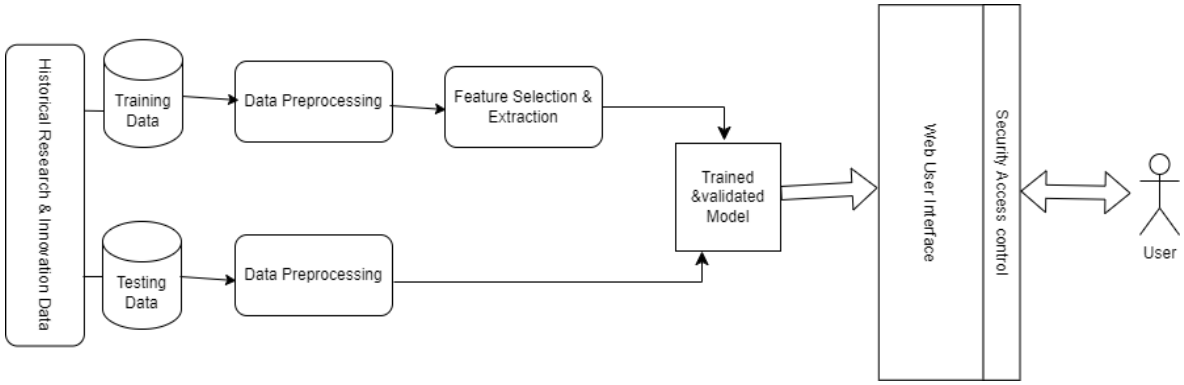

Figure 3 System High level View Design

Source: Constructed by author

3.3.6 Logical Design

A flowchart is a tool that is used to show a process, algorithm, or workflow to provide a structured way to represent the flow of information, decision points, and the order of operations within a system or process [91]. The figure below shows the logical structure if the system's research and innovation application process.

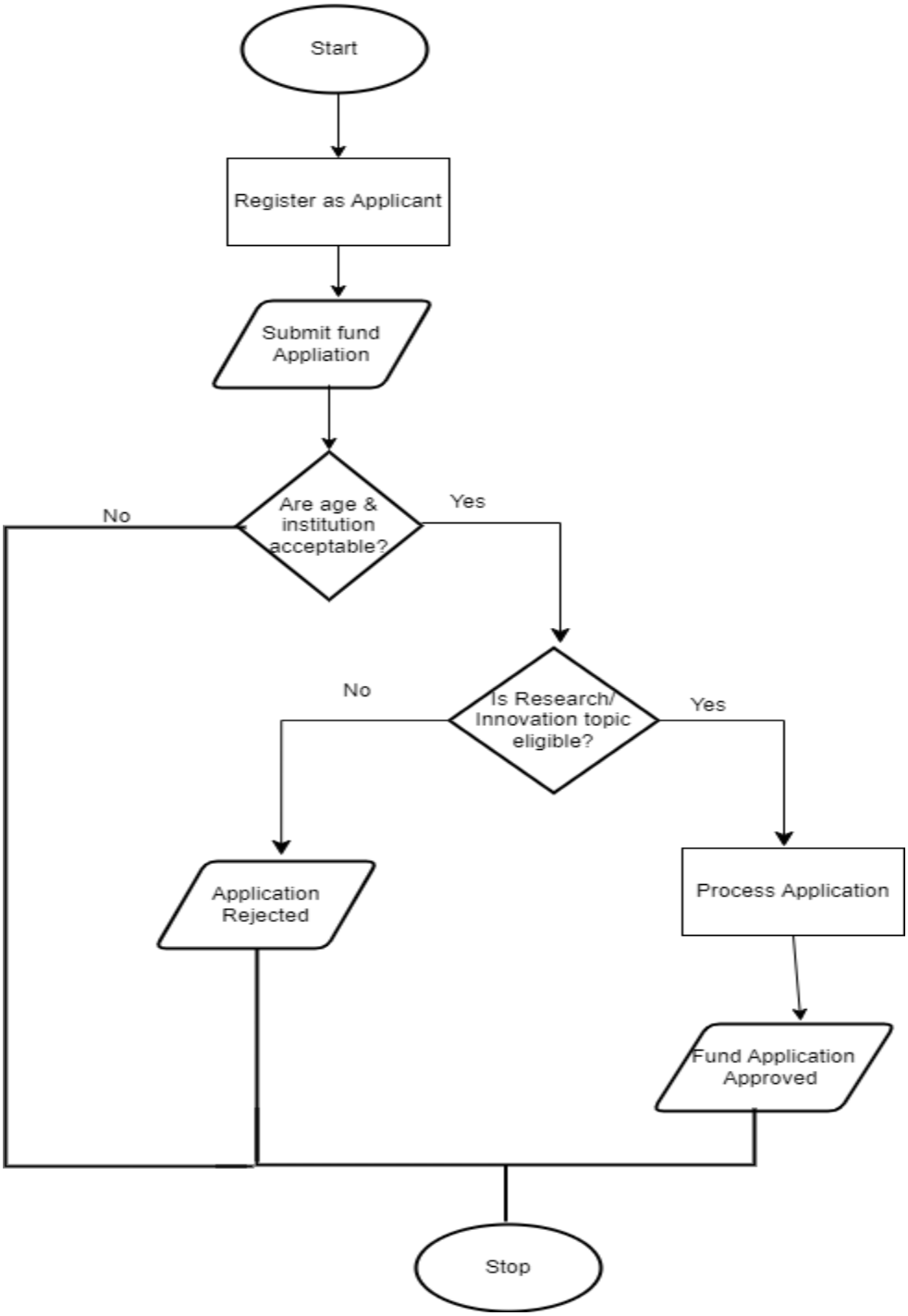

Figure 4 System Logical Design

Source: Constructed by author

3.3.7 Use Case Diagrams

A use case diagram is a Unified Modelling Language (UML) diagram that visually represents the functional requirements and interactions between users who are referred to as actors and a system. It shows the system's functionalities, user roles, and their interactions. The figure below represents the Use Case diagram for the system developed in this study [92].

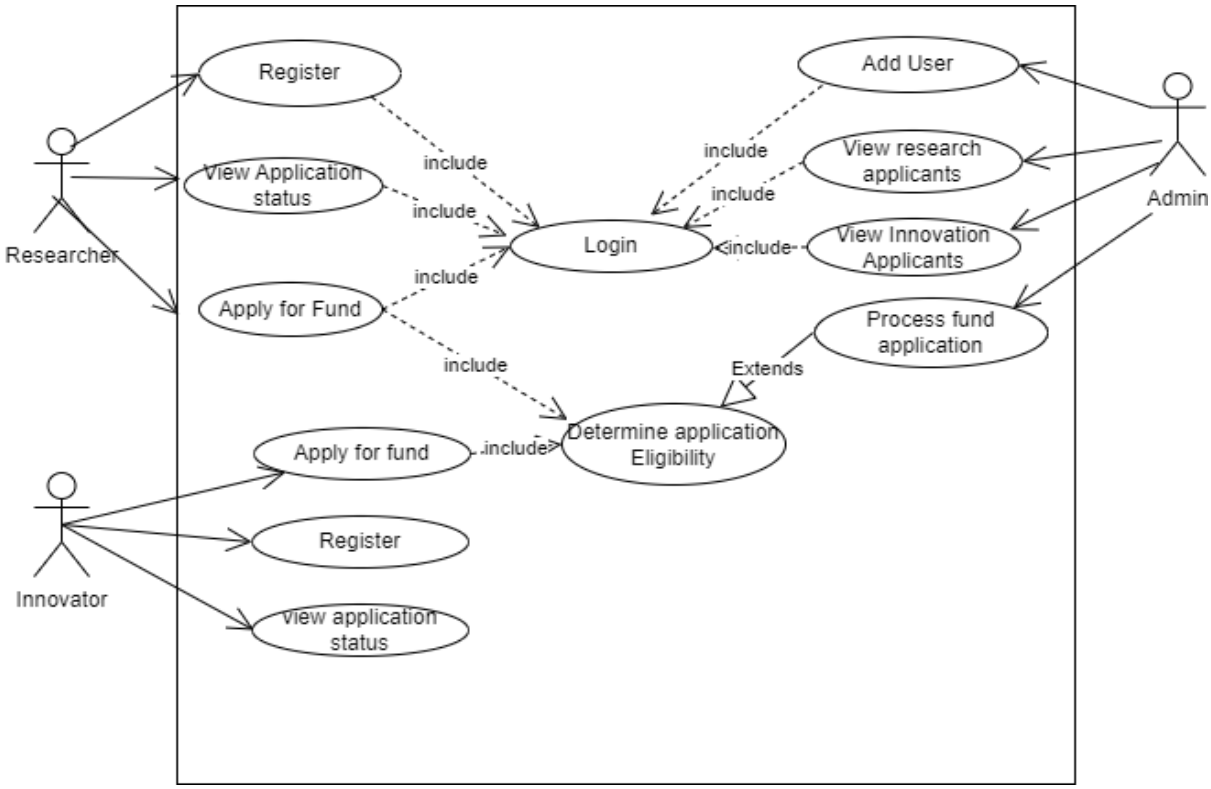

Figure 5 Use case diagram of the web Archiving system

Source: Constructed by author

The system has three (3) types of users the researcher, the innovator and the admin, even though the admin can add more users. The researcher and innovator interact with the system by registering to be an applicant, applying for any available funding and view the status of the application. The admin interacts with the system by adding more users when necessary, processing the application and viewing statistics for the application statuses.

3.3.8 Sequence Diagram

A sequence diagram is a Unified Modelling Language (UML) diagram that illustrates how operations in a system are carried out, focusing on the messages sent and the order in which they occur. It is used to show the interactions between objects in the sequential order that those interactions occur [93].

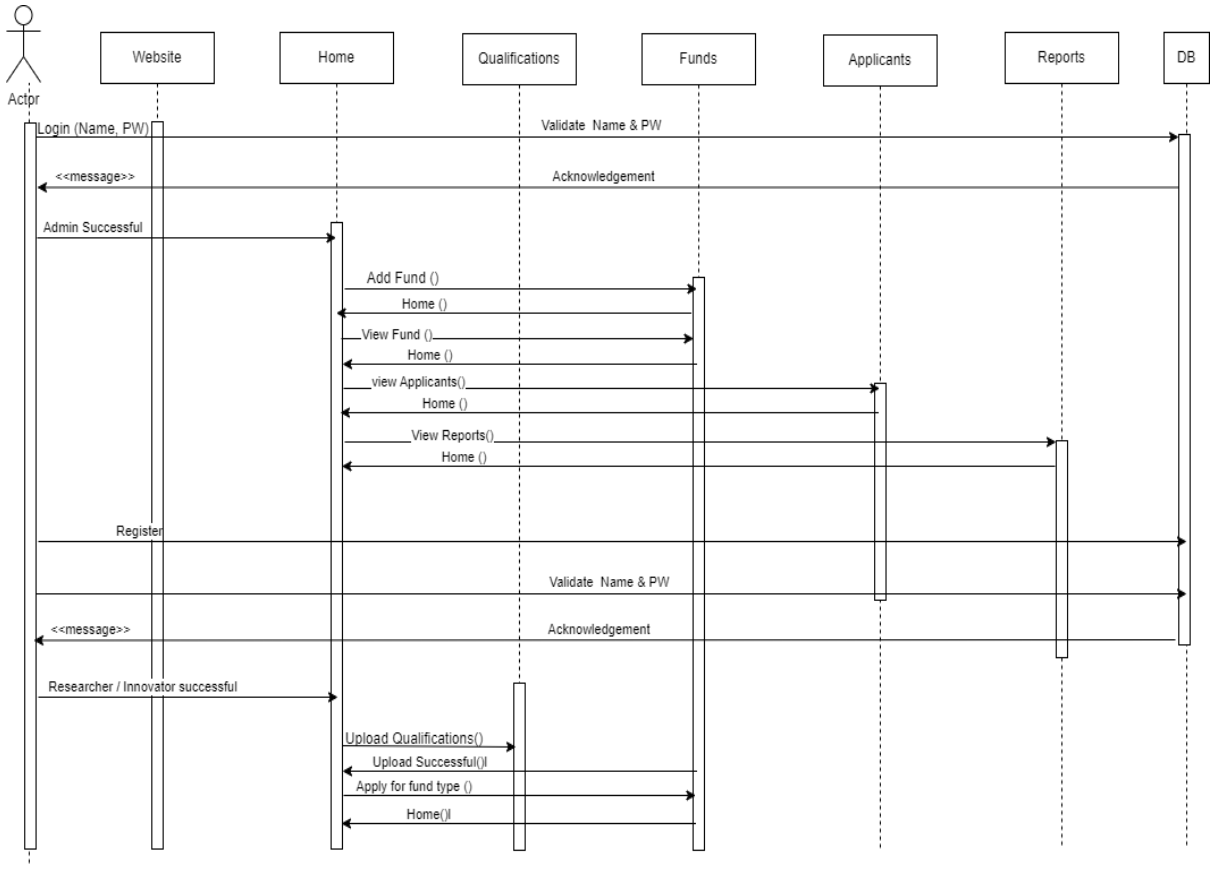

Figure 6 System Sequence Diagram

The figure shows the operations of the users of the system namely the Admin, Researcher and innovator. The Admin has access to the fund applications, applicant details and can also few various types of reports. The researcher and the innovator are able to register accounts on the system, upload qualifications and apply for the specific fund that they wanted.

3.3.9 Database Design

The database design describes the detailed data model for structure and organisation of a database. The data model used in this study is the Entity-Relationship Diagram (ERD). An ERD is a visual representation that depicts the entities (objects or concepts) within the database, their attributes, and the relationships between them. It defines the tables (entities), columns (attributes), relationships, keys, constraints, and other elements necessary to efficiently store, manage, and retrieve [94]. Below are the tables contained in the database as well as their specific functions:

- Researcher:

  The researcher table holds information about an applicant who wishes to apply for a research grant. Details such as the National Registration number (NRC), the Date of Birth (DOB), email address and gender are kept in this table.

- Innovator

  The innovator table hold information about an applicant who wishes to apply for an innovation grant. The details that are kept in this table include the InnovationID, Innovation Name, address of the applicant, cost of the innovation, the team members participating in the innovation, if any, and all the risks that are involved in the innovation.

- Research Applications

  The research applications table keeps information about the applications received for research funds. The details kept include the research topic, full name of applicant, the institution to which the applicant belongs, the fundID, the year that the fund is being applied for and the level of study.

- Fundtype

  This table keeps information about the type of fund being applied for. There are three funds that can be applied for, the Strategic Research Fund (SRF), the Strategic youth Innovation Fund (SYIF) and the Science and Technology (S &T) Postgraduate Scholarship which is a fund that comprises of tuition and research fund. The details that are kept in this table include the fundID, the status (active or inactive), the name of the fund, the year the fund is being applied for as well as the amount being applied for.

- Qualification

  The qualifications table holds details about the qualifications held by fund applicants. The details kept are the NRC number of the applicant, the full name of the applicant, the name of qualification, the name of the file, the year awarded and the field of study.

- Attachment

  Uploading of necessary qualification documents are required as part of the research fund application process. Thus, the attachment table holds meta data for the qualification documents uploaded by applicants. The details kept are the NRC number of the applicant, the name of qualification, the name of the file, the year awarded and the field of study.

- User

  The user table hold information about any member of staff that is part of handling or managing the research or innovation fund allocations. The details kept are the StaffID, the type of user (admin, Science & technology officer, etc.), the full name, the gender and email.

The tables described above are further illustrated in the Entity- Relationship figure below. The figure shows the entities, their attributes and the relationships of the entities.

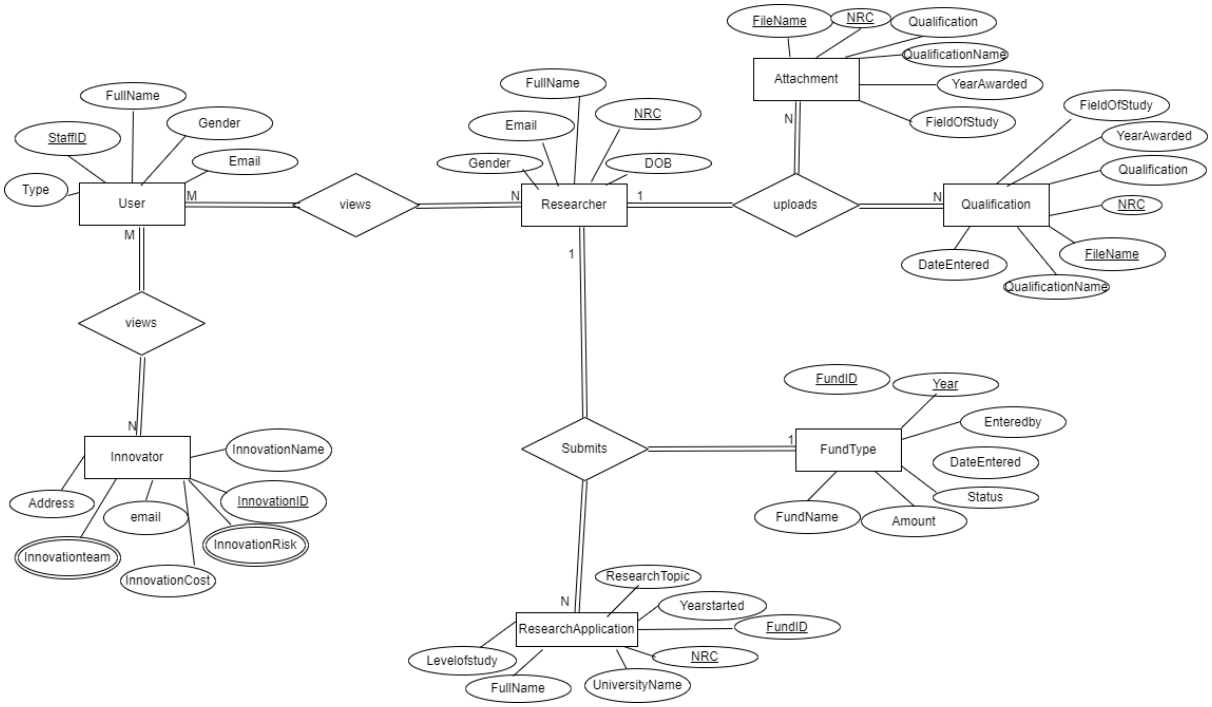

Figure 7 Entity-Relation Diagram for the research and innovation database

Source: Constructed by author

3.3.10 Security Design

This section gives a brief description of the security controls and measures implemented in order to secure the system developed and its information. A multi-layered security architecture was implemented using the AAA security model which was developed for Network and Computer Security. The AAA security model is based on three principles:

- Authentication: This is a means by which a system verifies that the user is who he or she claims to be.
- Authorisation: This is a means by which a system verifies that the user is allowed access.
- Auditing: This is means of keeping track and recording all database activity, including the username and the time in the log files.

This AAA multi-layered approach was implemented as follows:

- Server Room Access: The system will be hosted on a server and as such will need to be put in the server room. To enforce security, controls to manage the physical access of the server room is recommended. This involves installing cameras and using door access control systems such as biometric access systems or multi-factor access systems.

- Host access: access to the system was secured at the host-level by using a multi-factor authorization and authentication mechanisms including username, password and recaptcha facility to manage access to the database host.
- Database server network access: An authentication mechanism was implemented to control connections to the database server.
- Encryption in transit and host authentication (SSL): All the data including passwords and usernames — should be encrypted on the network using SSL and certificates ensure the user communicates with the intended host machine.
- SQL injection attack prevention: All the information entered through forms is sanitised in order to prevent SQL injections.
- Database authentication: A mechanism was implemented using username and passwords to authenticate users that can access the database. Furthermore, authentication has been tied with overall user management to make sure that access credentials get revoked when users move to another assignment, leave the organisation or any other reason.
- Database authorization and access control: Separation of roles was implemented and users granted permission to view and work with only data in the database that is attached to their role. Privileges are granted in order for a user to perform a job. There is no shared or group login credentials. Another mechanism was implemented in the code to ensure that all database connections are hidden.
- Data encryption: Important information in the database such as the passwords were encrypted at column level to keep the database information secure.
- Auditing: The Database Management System used (MySQL) has this feature which provides auditing capabilities that database administrators, auditors, and operators can use to track and analyse database activities, like the creation, changing, or deletion of data.
- Data redaction: The DBMS used has this feature which can be used to shield (or mask) certain data elements like NRC numbers from certain types of users.

The figure below illustrates the security measure implemented using a security architecture diagram.

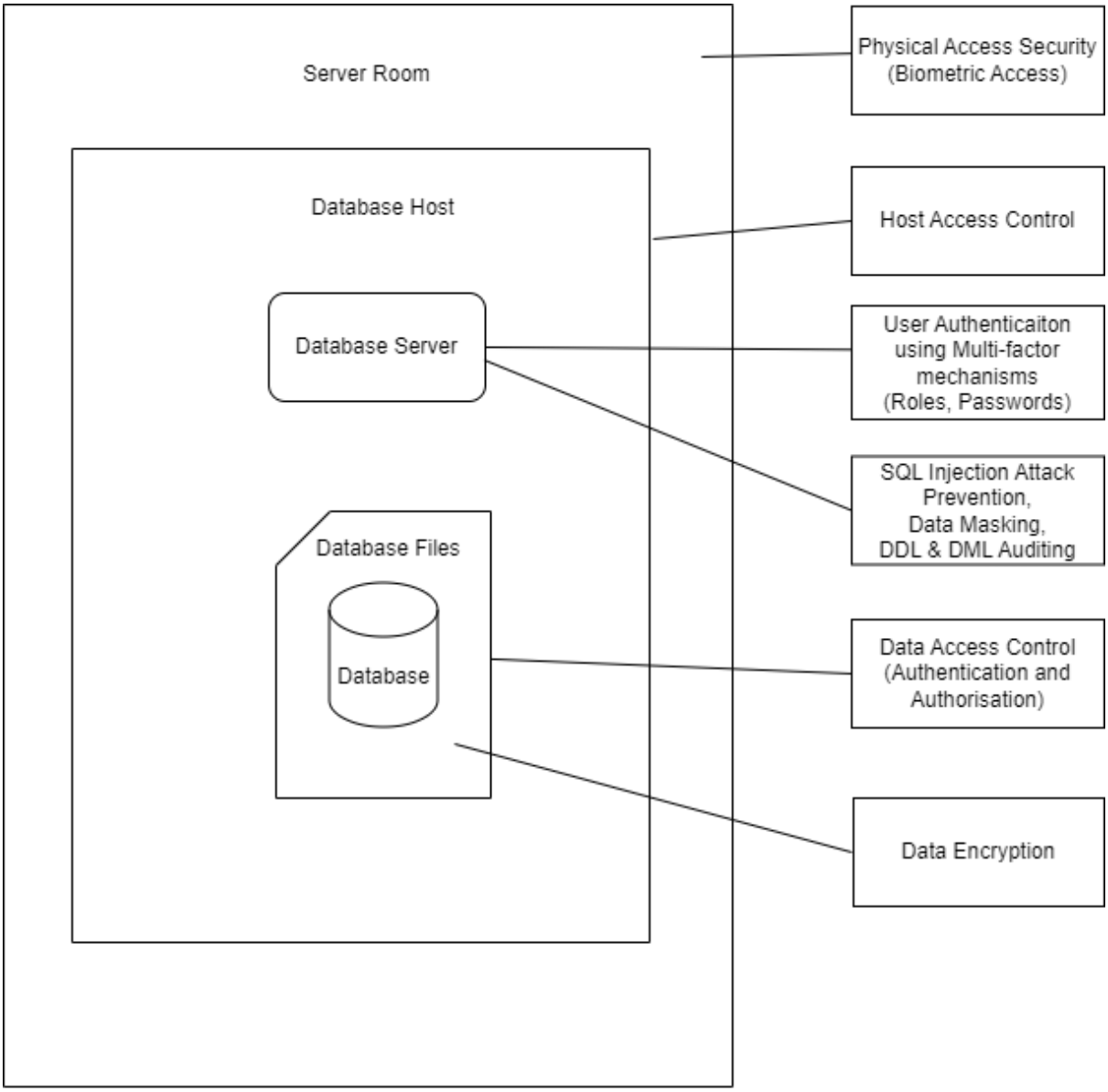

Figure 8 Security Architecture Diagram of the Web Archiving System [95]

3.4 System Implementation

System implementation is a stage at which a system is constructed, tested, and put into use within an organization. It involves deploying the system for operational use after it has been developed, tested and verified [96]. The key activities carried out under the implementation stage include:

3.5 System Development

This stage involved translating the design specifications into actual software code. The stages involved are described below:

The study employed the use of a text classifier for the classification of research and innovation application documents. Three Text Classification Algorithms namely, the Support Vector Machine (SVM) Multi-classifier, the Naïve Bayes and the K-Nearest Neighbour were selected as the models for text classification. The three models were trained using a dataset compiled from historical grant data obtained from the National Science and Technology Centre in Zambia. A total number of 129987 records were extracted and they included in-formation such as Research or Innovation proposal details, proposed budget, previous funding history, institutional information, research or innovation area or field, collaborator information if any and project timeline. The collected data was then pre-processed for the removal of unwanted and inconsistent data such as stopwords, outliers, missing values and punctuations. The research or innovation proposal details included the Title and abstract of the research proposal, keywords or subject areas related to the proposal. The input features selected for the purpose of the study were the title of the research or innovation and the research or innovation field. The dataset was split into 80% training set and 20% testing set. The figures below illustrate the code snippet used to create each of the three text classification algorithms.

```
require "vendor/autoload.php";
$dataset = new \Phpml\Dataset\CsvDataset(filepath: "./data/etd_dataset.csv", features: 1, headingRow: true);
$vectorizer = new Phpml\FeatureExtraction\TokenCountVectorizer(new Phpml\Tokenization\NGramTokenizer(1, 3), new Phpml\FeatureExtraction\StopWords\English());
$tfIdfTransformer = new Phpml\FeatureExtraction\TfIdfTransformer();
$samples = [];
foreach ($dataset->getSamples() as $sample) {
    $samples[] = $sample[0];
}
$vectorizer->fit($samples);
$vectorizer->transform($samples);
$tfIdfTransformer->fit($samples);
$tfIdfTransformer->transform($samples);
$dataset = new \Phpml\Dataset\ArrayDataset($samples, $dataset->getTargets());
$randomSplit = new \Phpml\CrossValidation\StratifiedRandomSplit($dataset, testSize: 0.2, seed:156);
$modelPath = './data/ETD_Dataset.phpml';
$modelManager = new \Phpml\ModelManager();
if( ! file_exists($modelPath) ) {
$classifier = new \Phpml\Classification\SVC(\Phpml\SupportVectorMachine\Kernel::RBF, 1000);
$classifier->train($randomSplit->getTrainSamples(), $randomSplit->getTrainLabels());
$modelManager->saveToFile($classifier, $modelPath);
}else {
$classifier = $modelManager->restoreFromFile($modelPath);
}
```

Figure 9: SVM Text Classification model code snippet

The figure below shows the code used to create the K-Nearest Neighbour text classifier

```php
require "vendor/autoload.php";
$dataset = new \Phpml\Dataset\CsvDataset(filepath: "./data/etd_dataset.csv", features: 1, headingRow: true);
$vectorizer = new Phpml\FeatureExtraction\TokenCountVectorizer(new Phpml\Tokenization\NGramTokenizer(1, 3), new Phpml\FeatureExtraction\StopWords\English());
$tfIdfTransformer = new Phpml\FeatureExtraction\TfIdfTransformer();
$samples = [];
foreach ($dataset->getSamples() as $sample) {
    $samples[] = $sample[0];
}
$vectorizer->fit($samples);
$vectorizer->transform($samples);
$tfIdfTransformer->fit($samples);
$tfIdfTransformer->transform($samples);
$dataset = new \Phpml\Dataset\ArrayDataset($samples, $dataset->getTargets());
$randomSplit = new \Phpml\CrossValidation\StratifiedRandomSplit($dataset, testSize: 0.2, seed:156);
$modelPath = './data/ETD_Dataset.phpml';
$modelManager = new \Phpml\ModelManager();
if( ! file_exists($modelPath) ) {
$classifier = new \Phpml\Classification\KNearestNeighbors(k:5);
$classifier->train($randomSplit->getTrainSamples(), $randomSplit->getTrainLabels());
$modelManager->saveToFile($classifier, $modelPath);
}else {
$classifier = $modelManager->restoreFromFile($modelPath);
}
```

Figure 10: KNN Text Classification model code snippet

The figure 11 below shows the code used to create the Naïve Bayes text classifier.

```php
require "vendor/autoload.php";
$dataset = new \Phpml\Dataset\CsvDataset(filepath: "./data/etd_dataset.csv", features: 1, headingRow: true);
$vectorizer = new Phpml\FeatureExtraction\TokenCountVectorizer(new Phpml\Tokenization\NGramTokenizer(1, 3), new Phpml\FeatureExtraction\StopWords\English());
$tfIdfTransformer = new Phpml\FeatureExtraction\TfIdfTransformer();
$samples = [];
foreach ($dataset->getSamples() as $sample) {
    $samples[] = $sample[0];
}
$vectorizer->fit($samples);
$vectorizer->transform($samples);
$tfIdfTransformer->fit($samples);
$tfIdfTransformer->transform($samples);
$dataset = new \Phpml\Dataset\ArrayDataset($samples, $dataset->getTargets());
$randomSplit = new \Phpml\CrossValidation\StratifiedRandomSplit($dataset, testSize: 0.2, seed:156);
$modelPath = './data/ETD_Dataset.phpml';
$modelManager = new \Phpml\ModelManager();
if( ! file_exists($modelPath) ) {
$classifier = new \Phpml\Classification\NaiveBayes();
$classifier->train($randomSplit->getTrainSamples(), $randomSplit->getTrainLabels());
$modelManager->saveToFile($classifier, $modelPath);
}else {
$classifier = $modelManager->restoreFromFile($modelPath);
}
```

Figure 11 Naïve Bayes Text Classification model code snippet

After training, the models were evaluated using the testing dataset to determine whether they were able to make predictions on new data. This stage is an important aspect of system

implementation as it assures quality of the model and reliability of the model. It also ensures the model's consistency and accuracy.

3.6 System Integration and Testing

In order to facilitate the interaction with users, the selected model, the SVM, was integrated in a web based archiving system. This involved developing a Web User Interface (UI) application. The Web UI provides users with a graphical user interface to facilitate their interaction with the system. The web UI application was developed using PHP language and run on apache as the web application host. PHP- ML was the library that was used to handle the machine learning tasks. After integration, the system was tested for accuracy and consistency. Testing was done at three levels unit, integration and system. Unit testing is a test that involved test the individual components of the system, that is, the model and the Web user interface application. Integration and system testing involved combining the model and the web UI application together and testing the resulting system as a whole. These tests validated the individual system components and ensured that the entire system functions effectively and efficiently in the awarding of research and innovation grants.

3.7 Deployment

Deployment involves installing the developed system in an operational environment. In this case, it involves installing the system on the premises of the funding institution(s) to facilitate the awarding of research and innovation grant allocation. The study recommends the preparation of the server environment that meets the hardware and software requirements specified. The system can then be hosted either by a web hosting provider or locally using the institution's server infrastructure. The server environment can then be set with an operating system such as Linux or Server, a web server Apache and a data base server MySQL In addition the system files can be uploaded on the server ensuring that they are placed in the correct directory. The database settings such as the username, password and hostname can be configured in the system's configuration files. Finally, other configurations involve setting up the Uniform Resources Allocator (URL), admin credentials, and other settings so that the system can be accessed using a web browser. The study recommends parallel conversion approach from the old system to the system. This means that both the old and the new system will be used until such a time as the institution completely phases out the

old system. This conversion method is recommended as it allows users to have a fall back system in case the new system malfunctions. This approach also allows the users to test the system, give feedback and adapt to using it. Any necessary data from the old system can be migrated on to the new system using the Extract, Transform Load (ETL) process.

3.8 User Training and User Acceptance Testing

User Training is another important aspect of implementation. It involves training end-users and other stakeholders on how to use the new system effectively. In addition to end users, training should also be given to users who will be in charge of operating and managing the system e.g. System Administrators. User Acceptance Testing on the other hand, is way of giving users the opportunity to use and test the system in a real-world environment to validate its functionality and usability. It helps to determine whether users have accepted the system enough to use it. From inception, the development of the system involved users from the MOTS who helped come up with requirements. This has led them exhibit willingness to use the system once it has been deployed.

3.9 Chapter Summary

This chapter discussed the research methodology applied in the study. The study adopted a hybrid methodology approach that combined the agile and the CRISP-DM methodologies. The agile methodology was selected because of its emphasis on flexibility and adaptability to changing requirements whereas the CRISP-DM was selected because the study involves the use of a machine learning model. The chapter also highlighted the functional and non-functional requirements, the design specifications using a variety of design models; flowchart, Use Case Diagram, Entity –Relationship Diagram and the security diagram.

4 RESULTS

4.1 Introduction

This chapter describes the results obtained from the experiments carried out on the system to ensure that the objectives of the study are met. The study involved the use of a supervised machine learning model to automatically categorise the proposal documents before archival. Hence, this chapter will first describe the results obtained from experiments carried out on three text classifiers and secondly the results obtained from integrating the selected text classifier into a web based application.

4.2 Text Classification Model Performance Results

The three models were evaluated by calculating the accuracy, precision and recall of each model. The accuracy measures the number of correctly classified instances out of all the instances contained in the dataset, and is expressed as a ratio between the correctly classified instances and the total number of instances in the dataset [97]. It is given by the formula:

Accuracy= (True Positives+ True Negatives)/ (True Positives +True Negatives+ False Positives +False Negatives)

The Precision measures the accuracy of the positive predictions of all instances predicted as positive and it is given by:

Precision = True Positive/ (True Positives +False Positives) [98]

The Recall measures the ability of the model to correctly identify all positive instances out of all actual positive instances and it is given by the formula:

Recall=True Positive/ (True Positives +False Negatives) [98]

Where:

True Positive = Actual class is positive and is predicted as positive

False Negative= Actual class is positive but is predicted as negative

True Negative= Actual class is negative and is predicted as negative

False Positive= Actual class is negative but is predicted as positive

The F1 score is also a vital performance metric as it provides a more realistic measure of the model's performance. It is calculated as the weighted mean of the precision and recall. [97]

The results obtained are shown in the table below:

Table 2 Text Classification Models' Performance

Text Classifier	Accuracy	Precision	Recall	F1-Score
SVM Multi-classifier	0.88	0.86	0.87	0.87
K- Nearest Neighbour	0.41	0.52	0.43	0.37
Naïve Bayes	0.86	0.86	0.81	0.84

The table above show the performance metrics of the three text classification algorithms. These are further illustrated in the figure below:

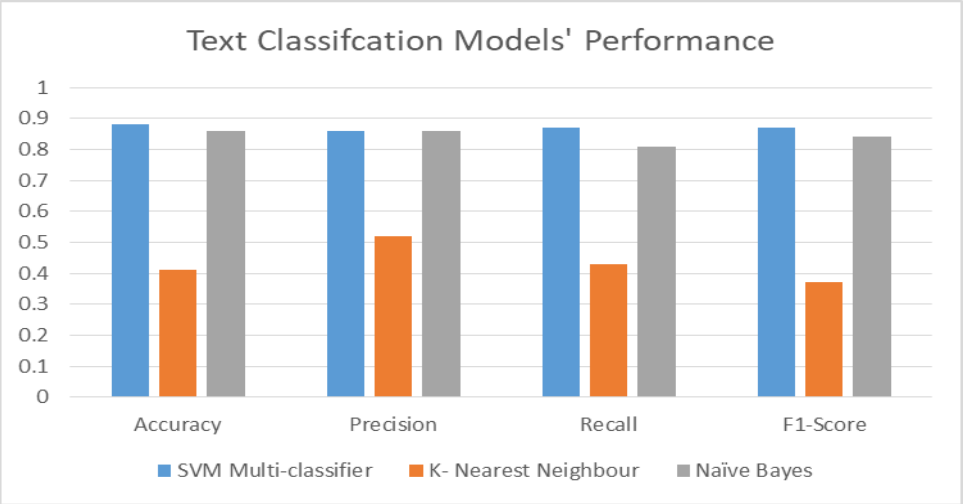

Figure 12 Text Classification Models' Performance

The results obtained reveal that the Support Vector Machine exhibited the best performance among the three models with an accuracy percentage of 88%, precision 86%, recall of 87% and F1 score of 87%. The Naives Bayes model attained an accuracy percentage of 86%, precision 86%, recall of 81% and F1 score of 84%. The K-Nearest Neighbour portrayed the lowest performance with an accuracy of 41%, precision 52%, recall of 43% and F1 score of 37%. This can be attributed to the fact that, the KNN does not perform well on large datasets [31]. Therefore, the SVM was selected as the text classification algorithm for this study.

4.3 The Support Vector Machine (SVM) Results

The SVM Multi classifier makes predictions by first calculating the probability of a research/innovation title belonging to each of the field classes. The field class that gives the highest score is selected as the predicted field class. A sample research title was tested to portray the ability of the model to make predictions. The results are shown in the table below:

Table 3 SVM prediction Results

Field Category	**Precision**	**Recall**	**F1 Score**
Business	0.91	0.91	0.91
Education	0.80	0.76	0.78
Engineering	0.80	0.86	0.88
Medicine	0.90	0.95	0.92
Sciences	0.94	0.93	0.94
Social Science	0.95	0.93	0.95
Predicted Category		*Social Science*	

The results in the table show that the social science field category exhibits the highest scores and thus is selected as the predicted category.

The experimental results revealed that the proposed SVM multi-classifier achieves superior accuracy and generalisation capabilities compared to the other models. By exhibiting a high performance, the model can be used to ensure a high level of fairness in the grant allocation process. Thus, the SVM multi-classifier was selected as the model to integrate into a web based application.

4.4 System Automation and Implementation Results

The system automation process begun with automating the application process. Even though most funding institutions have automated the process of applying for grants, the modus operandi of the application process at the Ministry of Technology and Science requires that applicants

submit their applications by electronic mail (e-mail). These applications are then sorted, classified and tagged by a human operator in order to determine which applications are eligible for funding. Once the eligible applicants are selected and granted the funding, the candidates are informed of this decision by e-mail. Thus, this study implemented a platform for online application, automatic classification of applications and granting of funds. The figures below exhibit the results of the automation process of the core functions of the application process.

4.4.1 The application Process

The application process starts by applicants creating an account on the system. Once the account has been created, the applicant can now access and apply for any active call for funding. The application includes detailing the research topic, the research abstract, institution of study as well as the nature and level of study being pursued. The figure below shows part of the research application process.

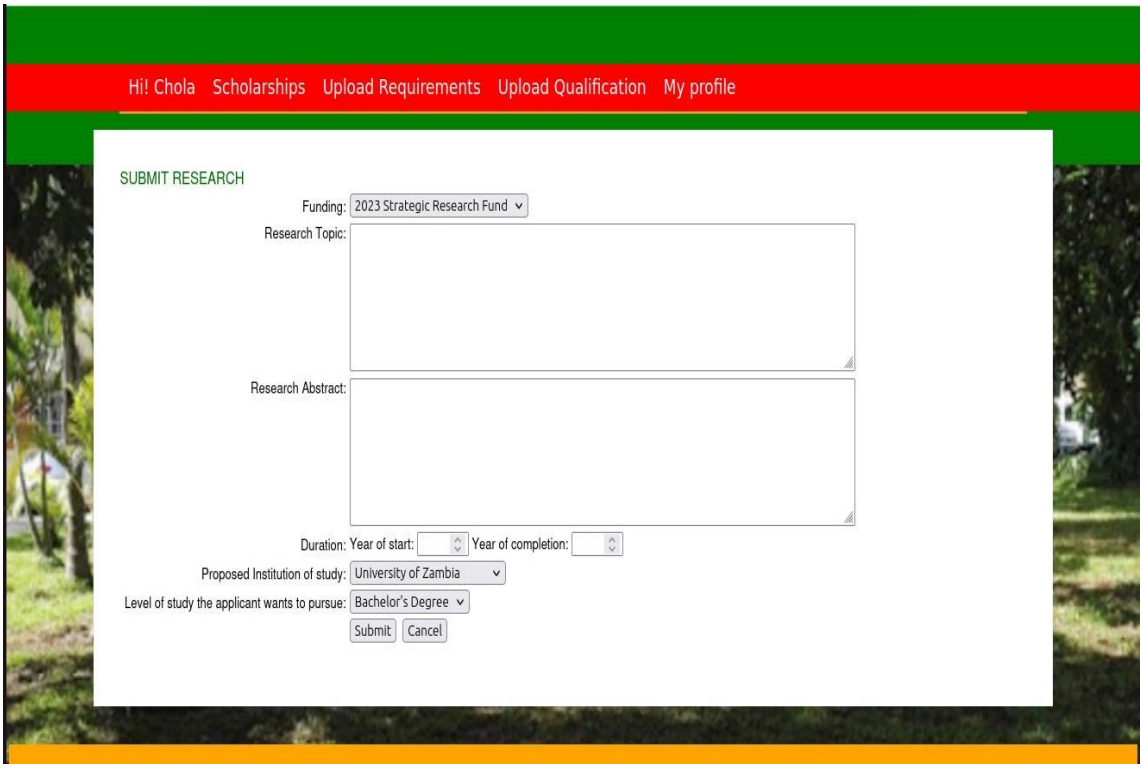

Figure 13 Research Fund Application

The diagram shows a researcher applicant selecting for type of fund to apply for, and submitting their research topic, abstract and other details needed for the application.

4.4.2 SVM Classifier Implementation Results

The aim of this study was to automatically classify research and innovation proposal applications. Therefore, the SVM multi-classifier was integrated into the web based application to help determine the eligibility of research proposal applications. Thus, after the submission of the research or innovation topic, it is automatically classified and determined as eligible or not eligible.

Figure 14 SVM Multi- classifier integrated into a web based application showing research proposal eligibility.

Figure 16 below shows the results for innovation proposal application eligibility.

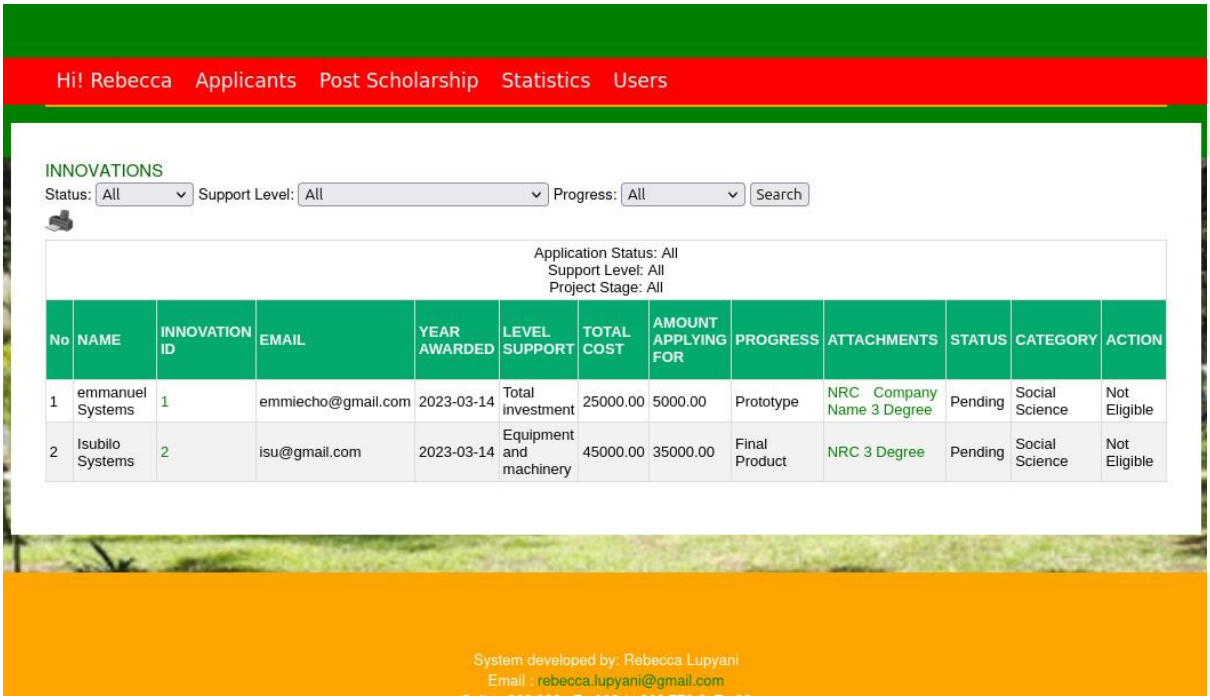

Figure 15 SVM Multi- classifier integrated into a web based application showing innovation proposal eligibility

4.4.3 The selection process

The figure below shows the process that an administrative officer can take to either reject or accept a particular research proposal application.

Figure 16 Approval/ Rejection of Research Application

Similarly the figure below shows the process that an administrative officer can take to either reject or accept a particular innovation proposal application.

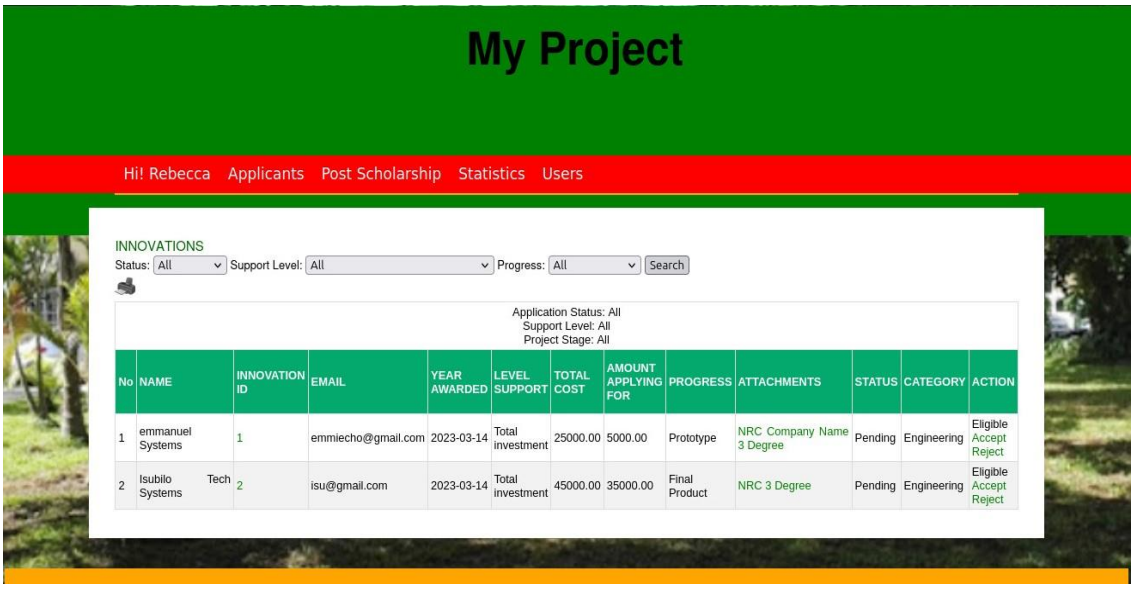

Figure 17 Approval/ Rejection of Innovation Application

The system is also able to give statistics on all the decisions that have been made on research and innovation proposal applications. It gives details on how many applications have been approved, rejected or pending a decision.

Figure 18 Statistics of the processed research and innovation proposal applications

4.4.4 System Validation

The system underwent verification and validation of the requirements as has been exhibited by the figures above. The requirements of the users, that is, to automatically classify research or innovation proposal applications, facilitate the selection process and give statistics on the status of the applications, have been met. This demonstrates that the SVM multi-classifier was successfully integrated in the web based application and exhibited exceptional results. The system also underwent unit testing, integration testing and system testing. The results of the unit testing of the text classification models are shown in tables 2 and 3 which displayed that the SVM classifier exhibited the highest performance. The results of the integration testing of the SVM text classifier into the web application and the system testing are displayed in the table below.

Table 4 Integration and System Test Case

TestID	Test Type	Test Condition	Test Input	Expected Result	Actual Result	Status
1	Usability	Check if applicants can successfully register	Users submit personal information according to rules on the form	User registration successful	User registration successful	Pass
		Check if applicants can successfully submit applications with accompanying documents	User submit application with correct format of documents (PDF)	User application successfully submitted	User application successfully submitted	Pass
2	Integration of SVM model	Check if the integrated model	An automatic research grant	Engineering-Eligible	Engineering-Eligible	Pass

	(Functionality)	will classify a topic and determine its eligibility	allocation using a multi Classifier			
3	Integration of SVM model (Functionality)	Check if the integrated model will classify a topic and determine its eligibility	Gender Based Violence against men in Zambian Compound of Choma Southern Province	Social Science- Not Eligible	Social Science- Not Eligible	Pass
4	Security	Check if Username, password and recaptcha are working	Input correct information Input wrong information	Login Successful Login unsuccessful	Login successful Login failed	Pass Pass

The results in the table show the tests for usability, functionality and security and their results which were all successful.

4.4.5 Limitation of the Prototype

The eligibility of research and innovation proposal applications for the allocation of grants is dependent on a number of factors such as the topic, the budget, funding history, age to mention but a few. However, the dataset that was used to train the SVM text classifier was limited to the topic and field category only. This means that the suggested eligibility of the applications by the system is limited to the topic category only and thus maybe biased towards that. Therefore, in order to make a comprehensive decision, other factors should still be put into consideration.

The other limitation is that the prototype lacks a mechanism for automated model re-training or updating, which is crucial for ensuring the model's accuracy over time with new data being maintained. Therefore, continuous model maintenance and improvement strategies need to be considered.

4.5 Chapter Summary

The chapter illustrates the results obtained after carrying out experiments to determine which of the three text classifiers Naïve Bayes, K Nearest Neighbour and Support Vector Machine out performs the others. The results revealed that the SVM text classifier achieved the highest performance with an accuracy percentage of 88%, precision 86%, recall of 87% and F1 score of 87%. The Naives Bayes model attained an accuracy percentage of 86%, precision 86%, recall of 81% and F1 score of 84%. The K-Nearest Neighbour portrayed the lowest performance with an accuracy of 41%, precision 52%, recall of 43% and F1 score of 37%. Thus, the SVM text classifier was selected as the machine learning model to be integrated into the web based application.

The chapter also illustrated how the SVM text classifier was successfully integrated into the application and used for decision making purposes.

5 DISCUSSION AND CONCLUSIONS

5.1 Introduction

This chapter discusses the findings of the study in reference to the research questions stated in chapter one. Furthermore, the conclusions drawn from and recommendations to the study are discussed.

5.2 Discussion

This section discusses the finding to answer the research questions developed in the first chapter.

5.3 Development of a model and algorithm to automatically classify documents

The first research objective was to develop a model and algorithm that will automatically classify documents according to discipline for the awarding of research and innovation grants. To answer this question, three text classification models namely the Naïve Bayes, the K-Nearest Neighbour and the SVM were trained using historical data from the National Science and Technology Council and compared to obtain the best performing model. The results obtained reveal that the Support Vector Machine exhibited the best performance among the three models with an accuracy percentage of 88%, precision 86%, recall of 87% and F1 score of 87%. The Naives Bayes model attained an accuracy percentage of 86%, precision 86%, recall of 81% and F1 score of 84%. The K-Nearest Neighbour portrayed the lowest performance with an accuracy of 41%, precision 52%, recall of 43% and F1 score of 37%. From the results obtained, it can be concluded that the SVM is one of the classifiers that exhibit an exceptional performance in text classification tasks. This supports the findings of Nidhi and Gupta [26] whose study also concluded that the SVM exhibits a high performance in text classification problems even though it has a poor recall. It is for these reasons that SVM classifier was selected as the model to be integrated in the web based archiving system developed in this study.

5.4 Development of a mechanism that keeps track of research or innovation progress

The second research objective was to develop a mechanism that helps to keep track of research or innovation progress. To answer this question, the web based application was developed and given a platform for researchers and innovators to submit their progress according to stipulated milestones. The milestones differ for researchers and innovators. For researcher under postgraduate, the entire project is divided into three milestones the proposal,

the midway progress and the final project submission for master's level and four milestones for PhD postgraduate researchers. For innovators, the milestones are also divided into three levels namely the proposal, the prototype stage and the final product. This information will not only help to keep track of the research or innovation progress but also help retain the actual project submissions after completion.

5.5 Development of a web based prototype that uses machine learning to classify research and innovation documents and allocate grants

The third research question was to develop a web based prototype that helps to keep track of research progress and uses machine learning to automatically classify documents using the title for effective archiving and retrieval. To address this research question, the SVM multi-classifier was integrated into a based application to help determine the eligibility of research and innovation proposal applications. The web based application was developed using PHP, the database using MySQL and deployed on an apache web hosting server. All the platforms used were selected because they are reliable and open source meaning that there is a large community of developers out there who can provide continuous support and thus provide help in cases where the system requires updates or improved functionality. The developed web based application allows researchers or innovators to first create profiles on the system, after which they have access to view active funding programs shared on the system platform. The researchers or innovators can now submit online applications which are automatically sorted for eligibility by the system. The system is embedded with the criteria for eligibility recommended by the funding institution. Once eligibility has been determined by the system, the officers in charge of processing the applications for grant approval further scrutinise the applications and give the final decision to either accept or reject the proposal. The decision made automatically appears on the applicants profiles. Furthermore the system built gives statistical reports that provide information on applicants who applied for specific funding programmes, for example, the SRF, SYIF or the S &T postgraduate scholarship, giving information as to how many were eligible, how many were granted funding and how much they were given. This makes the system to be an effective decision making tool for funding institutions.

5.6 Application of the developed model and system

The developed model has been used to improve the selection process in the allocation of funds by classifying research and innovation proposals according to the respective field and highlighting whether the proposal is eligible or not eligible for funding. Any research or innovation proposal under the fields engineering, technology and science are eligible for funding. This has eliminated the need for human operators to classify the proposal documents thereby removing human bias and human error in the classification. In other words, the model is used to shortlist eligible proposal applications as opposed to having human operators manually scrutinising and shortlisting them. The developed model and system is valuable to funding institutions that seek to allocate grants to research and innovation proposals whose eligibility is reliant on a particular field or topic. The model can help to shortlist the eligible proposals by flagging which proposal applications are eligible or not eligible for funding. Decision makers can then grant funding based on the recommendation of the system.

5.7 System Validation

The system developed was tested and validated under unit, integration and system testing levels. Different tests carried out included usability, functionality and security. To test for usability, the study checked if applicants can successfully register and submit applications with accompanying documents. The results revealed the affirmative as long as the correct details were entered and the submitted documents were in the correct format. The tests for functionality checked if the integrated SVM model would classify a topic and determine its eligibility for research/ innovation grant allocation. The results revealed that the model could automatically classify the topics and indicate whether the particular topic is eligible or not eligible for grant funding. Lastly, to test for security, access to the system through the user interface was validated by checking whether submitted credentials (username, password, recaptcha) were accepted or not accepted. The system was able to differentiate between correct and wrong log in credentials and as such passed the security test. Another test that is important is the user acceptance test which preliminary results indicate that the system will most likely be accepted as users were involved from the inception of the development of the system. These tests validated the system and deemed it ready for deployment.

5.8 Conclusions

In this study, we proved that the application of an SVM Multi-Classifier to automate the classification of research and innovation proposal applications can enhance efficiency, reduce human biases and promote fair review of the applications. This is exhibited by the performance metrics of the classifier which are 88% accuracy, 86% precision, 87% recall and 87% F1 score. The SVM model was integrated into a web based application to facilitate the grant allocation process by determining whether a submitted research or innovation topic is eligible or not eligible for a grant. This was done by assessing whether a submitted research/ innovation topic falls under a field category for which funding organisations are willing to support and fund. One the proposal is highlighted as either eligible or not eligible, the decision to accept or reject the proposal can be made. The study also illustrated the successful integration of the model into a web based application. The resulting system has enhanced efficiency and offers a promising solution to improve the fairness and transparency of the grant allocation process as human operators as eliminated from the shortlisting process.

.

5.9 Recommendations

Based on the experimental results and limitations of the model and the prototype, the study suggests the following recommendations.

5.10 Incorporating additional variables or features

The dataset used to train the text classification models used the title and field of the research or innovation as input features. This results in the model being biased towards the title only. The study therefore recommends extending the dataset to incorporate other input features such as budget, project timeline and funding history. The dataset records can also be increased so that the model can give a better accuracy result than the one exhibited in the study.

5.11 Continuous Model Refinement

The trained SVM model should undergo continuous monitoring, refinement, and updating in order to ensure the model remains adaptive and effective in managing changing data requirements and facilitate effective decision making assistance.

5.12 Future Works

Even though the SVM model exhibited a superior accuracy performance, its use in the allocation of grants maybe biased to the research or innovation topic details only. However, the awarding of grants is not only dependent on the research or innovation topics but also on other factors such as the budget, the project timeline, prior funding history to mention but a few. The dataset used in this study did not include such details and hence the possibility of bias towards the research or innovation topics only. Nevertheless, the model is still useful as the research or innovation topics are the biggest determining factors in the awarding of grants as the goal of funding institutions is to award grants to projects that are aligned to a nation's economic development goals.

5.13 Chapter Summary

This chapter discussed and concluded the results of the study. The chapter discusses how the research questions were addressed. It discusses the results from the training of three text classification algorithms; the Naives Bayes, K Nearest Neighbour and the Support Vector Machine. It discusses the successful integration of the text classifier into a web based application and its facilitation in the decision making process of the allocation of research and innovation grants.

6 REFERENCES

[1] I. R. Matthew L Wallace, “Research portfolio analysis in science policy: moving from financial returns to social benefits,” *Minerva,* vol. 53, no. 2, 2015.

[2] E. B. Alan Bryman, Business Research Methods, Oxford: Oxford Press, 2015.

[3] P. D. O. J. E. Leedy, Practical Research: Planning and Design, Essex: Pearson, 2014.

[4] OECD, “Investing in research and innovation in developing countries,” *International Development Research Centre (IDRC),* 30 June 2021.

[5] N. Sharma, “Role of Research in Nation Building,” *NOLEGEIN- Journal of Business Risk Management,* vol. 3, no. 2, pp. 9-13, 2020.

[6] T. Sibanyoni, “The Importance of Research in Economic Growth,” *Urban-Econ:NIKELA,* 2023.

[7] S. H. Q. Amita Singh, “Importance of Research in Development of Nation,” *International Journal of All Research Education and Scientific Methods (IJARESM),* vol. 9, no. 7, 2021.

[8] OECD, “Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data . .,” *OECD Publishing,* 2015.

[9] E. M. A. S. a. M. M. Q. Rogers, “Diffusion of innovations,” in *An integrated approach to communication theory and research,* Routledge, 20144.

[10] N. R. Council, “Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future.,” National Academies Press., Washington DC, 2007.

[11] D. C. R. N. .. Mowery, Paths of Innovation: Technological Change in 20th-Century America.., Cambridge: Cambridge University Press, 1998.

[12] W. H. Organization, “ Research for Universal Health Coverage: World Health Report,” 2020.

[13] MOTS, “Ministry of Technology and Science,” [Online]. Available: https://www.mots.gov.zm/?page_id=548. [Accessed 26 November 2022].

[14] N. S. a. T. Council, “National Science and Technology Council,” [Online]. Available: https://nstc.org.zm/. [Accessed November 2023].

[15] N. T. a. B. Centre, “National Technology and Business Centre,” [Online]. Available: https://ntbc.co.zm/. [Accessed November 2023].

[16] X. D. Zhang, “Machine learning,” in *A Matrix Algebra Approach to Artificial Intelligence*, Springer, 2020, pp. 223-440.

[17] J. G. N. H. S. M. E. Nahra, “Artificial intelligence and Machine Learning for Real-world problems (A survey),” *International journal of innovation in Engineering*, vol. 1, no. 3, pp. 38-47, 2021.

[18] J. Copeland, .Artificial intelligence: A philosophical introduction., John Wiley & Sons., 2015.

[19] P. Y. Z. Y. J. Y. F. Zhang, “ State-of-the-art review of machine learning applications in constitutivemodeling of soils.,” *Archives of Computational Methods in Engineering*, pp. 1-26, 2021.

[20] N. J. Nilsson, The quest for artificial intelligence, Cambridge University Press, 2009.

[21] S. N. P. Russell, Artificial intelligence: a modern approach., 2002.

[22] Shaveta, “A review on machine learning,” *International Journal of Science and Research Archive*, vol. 9, no. 1, p. 281–285, 2023.

[23] M. I. M. T. M. Jordan, “Machine learning: Trends, perspectives, and prospects,” *Science*, 2015.

[24] S. N. S. W. S. J. R Sra, Optimization for machine learning., Mit Press., 2012.

[25] F. e. a. Maleki, “ Overview of machine learning part 1: fundamentals and classic approaches.,” *Neuroimaging Clinics*, vol. 30, no. 4, 2020.

[26] V. M. V. K. A. Gupta, “An Overview of Supervised Machine Learning Algorithm,” *SMART*, December 2022.

[27] R. K. H. Choudhary, “Comprehensive Review On Supervised Machine Learning Algorithms,” in *International Conference on Machine Learning and Data Science (MLDS),* Gianey, December 2017.

[28] R. N.-M. A. aruana, “An empirical comparison of supervised learning algorithms.,” in *23rd international conference on Machine learning*, June 2006.

[29] S. Loukas, “Text Classification Using Naive Bayes: Theory & A Working Example,” *Towards Data Science,,* 2020.

[30] P. Ersoy, “Naive Bayes Classifiers for Text Classification,” *Towards Data Science,* 2021.

[31] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm,” *Towards Data Science,* 2018.

[32] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[33] S. G. Andreas C. Müller, Introduction to Machine Learning with Python, O'Reilly Media, Inc., 2016.

[34] C. M. Bishop, “Pattern Recognition and Machine Learning,” in *Information Science and Statistics* , Newyork, Springer, 2006.

[35] E. Alpaydin, Introduction to machine learning, Cambridge MA: MIT Press, 2010.

[36] F. S. S. N. J. Nouri, “Meta-heuristics algorithm for two-machine no-wait flow-shopscheduling problem with the effects of learning,” *Uncertain Supply Chain Management,,* vol. 7, pp. 599-618., 2019.

[37] P. Singh, “Unsupervised Machine Learning,” *Learn PySpark,* 2019.

[38] U. R. N. U. Hodeghatta, “ Unsupervised Machine Learning,” *Semantic Scholar,* 2017.

[39] C. M. Bishop, Pattern Recognition and Machine Learning., Springer, 2006.

[40] R. S. B. A. G. Sutton, Reinforcement Learning: An Introduction., MIT Press, 2018.

[41] N. M. F. S. N. L. K. G. A. Mehrabi, “ A survey on bias and fairness in machine learning,” *ACM Computing Surveys (CSUR),* vol. 54, no. 6, 2021.

[42] A. N. T. S. A. Singh, “A review of supervised machine learning algorithms,” in *3rd International Conference on Computing for Sustainable Global Development (INDIACom).Ieee,* 2016.

[43] M. Givens, “Keeping Up With Research Information Management Systems,” 2018. [Online]. Available: https://www.ala.org/acrl/publications/keeping_up_with/rims. [Accessed November 2023].

[44] Bloomerang, “Grant Management Software for non-profits,” [Online]. Available: https://kindful.com/nonprofit-glossary/grant-management-software-for-nonprofits/. [Accessed november 2023].

[45] T. a. Francis, “Understanding research metrics,” [Online]. Available: https://editorresources.taylorandfrancis.com/understanding-research-metrics/. [Accessed November 2023].

[46] P. M. Institute, A Guide to the Project Management Body of Knowledge (PMBOK® Guide) – Sixth Edition, Newtown Square, 2017.

[47] P. T. ,. P. D. ,. M. N. B. Thompson, “Academic Tracker: Software for tracking and reporting publications associated with authors and grants,” *National Library Of Medicine,* 2022.

[48] E. C. University, “Research Analytics and Performance,” 2016.

[49] M. L. Nelson, Web archiving., Springer International Publishing., 2016.

[50] J. F. K. M. W. M. C. Brunelle, “ Web Archiving. Synthesis Lectures on Information Concepts, Retrieval, and Services,” *Springer,* vol. 6, no. 1, p. 1–185, 2014.

[51] E. Maemura, “Towards an Infrastructural Description of Archived Web Data,” *Code4Lib Journal,* 2020.

[52] S. Harnad, “The Self-archiving Initiative,” *Nature,* vol. 410, no. 6832, p. 1024–1025, 2001.

[53] K. H. K. Y. M. &. U. S. Hatano, “Information Retrieval for XML Documents,” in *Dexa '02 Proceedings of the 13th International Conference on Database and Expert Systems Applications,* 2002.

[54] L. J. B. Caluza, “Development of Electronic Document Archive Management System (EDAMS): A CaseStudy of a University Registrar in the Philippines,” *International Journal of Digital Information and Wireless Communications (IJDIWC)* , pp. 106-117, 2017.

[55] P. T. Predrag Matkovic, “A Comparative Overview of the Evolution of Software Development Models,” *Journal of Industrial Engineering and Management* , vol. 1, no. 4, pp. 163-172, 2010.

[56] R. M. B. Pressman, Software Engineering: A Practitioner’s Approach, x., Boston: McGraw-Hill, 2014.

[57] B. M. a. J. Phiri, “Web Based Document Archiving Using Time Stamp and Barcode Technologies–A Case of the University of Zambia.,” *International Journal of Innovative Research in Science, Engineering and Technology,* 2016.

[58] W. B. ,. C. G. G. Glisson, “Design of a Digital Dissertation Information Management System,” *Emerald Sight,* vol. 36, no. 3, pp. 152 -156, 2002.

[59] I. E. M. K. H. &. J. H. Permatasari, “Blockchain implementation to verify archives integrity on cilegon E-archive.,” *Applied Sciences,,* vol. 10, no. 7, p. 2621, 2020.

[60] I. K. A. A. V. N. V. &. Y. L. Zikratov, “Ensuring data integrity using blockchain technology 20th Conference of Open Innovations Association (FRUCT),” in *20th Conference of Open Innovations Association*, Fruct, 2017.

[61] J. M. a. M. E. S. a. M. E. M. O. Jayoma, “OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga.,” in *IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology*, Phillipines, 2020.

[62] R. Smith, “An Overview of the Tesseract OCR Engine,Engine," Ninth International Conference on Document,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Parana, 2007.

[63] A. F. A. a. A. R. W. Sait, “Managing and Retrieving Bilingual Documents Using Artificial Intelligence-Based Ontological Framework,” *Computational Intelligence and Neuroscience,* vol. 2022, 2022.

[64] Christian Herzog et al, " Forward-looking analysis based on grants data and machine learning based research classifications as an analytical tool," *Digital Science,* 2000.

[65] Y. C. G. e. al, "Evaluating human versus machine learning performance in classifying research abstracts," *Scientometrics, Springer,,* 2020.

[66] E. R. Council, " ERC evaluation panels and keywords.," European Research Council., 2019. [Online]. Available: https://erc.europa.eu/content/erc-panel-structure-.

[67] Khor et al, "Applying Machine Learning to Compare Research Grant Programs," in *STI Leiden Conference on Science and Technology Indicator*, Netherlands,, 2018.

[68] M. N. S. N. S. S. N. S. Trivedi, "Comparison of Text Classification Algorithms," *International Journal of Engineering Research & Technology (IJERT),* vol. 4, no. 2, 2015.

[69] C. Z. C. Aggarwal, " A survey of text classification algorithms.," *Springer,* pp. 163-222, 2012.

[70] S. Jiménez, "Text Classification and Clustering with WEKA," *International Journal of Engineering Research & Technology (IJERT),* 2014.

[71] A. H. G. Wilcox, "Classification algorithms applied to narrative reports," *International Journal of Engineering Research & Technology (IJERT),* 1999.

[72] U. C. S. Pandey, "A Review of Text Classification Approaches for E-mail Management," *IACSIT International Journal of Engineering and Technology,* vol. 3, no. 2, 2011.

[73] C. e. a. Freyman, "Machine-learning-based classification of research grant award records," *Research Evaluation,* vol. 25, no. 4, 2016.

[74] L. Phiri, "Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia," in *2nd IEEE International Conference in Information and Communications*, Lusaka, 2018.

[75] L. Phiri, "Automatic Classification of Digital Objects for Improved Metadata Quality of Electronic Theses and Dissertations in Institutional Repositories," *International Journal of Metadata, Semantics and Ontologies,* vol. 14, no. 3, p. 234–248., 2020.

[76] R. H. J. Wirth, “CRISP-DM : Towards a Data process model for Data Mining,” *Practical Application of Knowledge Discovery and Data mining,* 1995.

[77] V. Nidhi. Gupta, “ Recent Trends in Text Classification Techniques,” *International Journal of Computer Applications,* vol. 35, no. 6, 2011.

[78] T. .. T. R. F. J. Hastie, “The Elements of Statistical Learning, Data Mining, Inference, and Prediction,” *Springer,* 2011.

[79] B. B. J. S. D. D. H. R. E. H. W. H. L. D. J. Y. LeCun, “Back propagation applied to handwritten zip code recognition,” *Neural Computation,,* vol. 1, no. 4, p. 541–551, 1989.

[80] L. B. Y. B. P. H. Y. LeCun, “ Gradient-based learning applied to document recognition.,” in *IEEE,* November 1998.

[81] A. S. H. B. L. L. Y. Conneau, “Very deep convolutional networks for text classification,” *ECACL,* 2016.

[82] L. G. A. S. Florian Mai, “Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text,” in *JCDL ’18: The 18th ACM/IEEE Joint Conference on Digital Libraries,* New York, 2018.

[83] F. M. A. S. D. B. a. A. S. Lukas Galke, “Using Titles vs. Full-text as Source for Automated Semantic Document Automation,” *K-Cap,* 2017.

[84] A. Abbas, “IBGA: An Incentives Based Grant Allocation Algorithm for Academic Institutions,” in *Wireless Networks and Computational Intelligence Communications in Computer and Information Science,,* Springer, 2012.

[85] W. W. Royce, “Managing the Development of Large Software Systems,” in *IEEE WESCON*, 1970.

[86] R. S. Pressman, Software engineering : a practitioner's approach, New York: McGraw Hill, 2010.

[87] K. e. a. Beck, “The Agile Manifesto.,” *Agile Alliance,* 2001.

[88] T. Z. Warfel, Prototyping : a Practitioner's Guide, Rosenfield Media , LLC, 2009.

[89] T. &. J. T. Dasu, “Exploratory Data Mining and Data Cleaning,” *Wiley-Interscience,* 2003.

[90] A. Kumar, “Feature Selection vs Feature Extraction: Machine Learning,” *Data Analytics,* 2023.

[91] R. S. Pressman, Software Engineering: A Practitioner's Approach., McGraw-Hill Education., 2014.

[92] J. J. I. B. G. Rumbaugh, The Unified Modeling Language Reference Manual, . Addison-Wesley., 2005.

[93] I.-Y. Song, “Developing Sequence Diagrams in UML,” in *20th International Conference on Conceptual Modeling,* 2001.

[94] T. M. &. B. C. E. Connolly, Database Systems: A Practical Approach to Design, Implementation, and Management, Pearson., 2014.

[95] M. Linster, “Creating a multi-user layered Securiy architecture for databases,” *ITOps Times,* 2019.

[96] A. W. B. H. T. D. .. Dennis, Systems Analysis and Design:An Object-Oriented Approach with UML, John Wiley & Sons, 2015.

[97] Abhigyan, “Calculating Accuracy of an MLModel,” *Abhigyan,* 2020..

[98] G. W. D. H. T. &. T. R. James, “An Introduction to Statistical Learning,” *Springer,* 2013.

[99] S. R. ,. C. J. P. P. P. Lewis David D, “Training algorithms for linear text classifiers,” in *ACM SIGIR Conference on Research and Development in Information Retrieval,* 1996.

[100] Y. Y. T. G. R. F. L. David D. Lewis, “RCV1: A New Benchmark Collection for Text Categorization Research,” *Journal of Machine Learning Research,* vol. 5, 2004.

[101] R. H. J. Wirth, “CRISP-DM : Towards,” in *Practical Application of Knowledge Discovery and Data mining,* 1995..

[102] Abhigyan, “Calculating Accuracy of an ML Model,” *Analytics Vidhya,* 2020.

[103] C. S. e. al, “A Systematic Literature Review,” in *International Conference on ENTERprise Information Systems*, 2020.

[104] G. W. D. H. T. &. T. R. (. James, “An Introduction to Statistical Learning,” in *Springer*, 2013.

[105] G. W. D. H. T. &. T. R. James, “An Introduction to Statistical Learning,” *Springer*, 2013.

APPENDICES

Appendix 1: Publications

1. Lupyani, Rebecca, and Phiri, Jackson. "Automated Document Classification for research HEI grant awards using Machine Learning." *Zambia Association of Public Universities and Colleges (ZAPUC) Conference*. Vol. 3. No. 1. 2023.
2. Lupyani, Rebecca, and Phiri, Jackson. "Automatic Classification of research grants proposals using a multi–class machine learning model." *Proceedings of International Conference for ICT (ICICT)-Zambia*. Vol. 5. No. 1. 2023.
3. Lupyani, Rebecca, and Phiri, Jackson. "From Algorithms to Grants: Leveraging Machine Learning for Research and Innovation Fund Allocation." *Proceedings of the Computational Methods in Systems and Software*. Cham: Springer Nature Switzerland, 2023. 469-480.