DEVELOPMENT OF IDENTITY ATTRIBUTE METRICS MODEL BASED ON DISTANCE METRICS

By

FELIX MUSAMA LAMECK KABWE

A Dissertation submitted to the University of Zambia in partial fulfilment of the requirements for the award of the degree in Master of Science in Computer Science

THE UNIVERSITY OF ZAMBIA

LUSAKA

2020

COPYRIGHT DECLARATION

DECLARATION

I, the undersigned, declare that this has not previously been submitted in candidature for any degree. The dissertation is the result of my own work and investigations, except where otherwise stated. Other sources are acknowledged by given explicit references. A complete list of references is appended.

Name :……..…

Signature :……..……

Date :…….……

CERTIFICATE OF APPROVAL

This dissertation of **Felix Musama Lameck Kabwe** is approved as fulfilling part of the requirements for the award of the degree of Master of Science in Computer Science by the University of Zambia.

……	……	……
Examiner I	Signature	Date
……	……	……
Examiner II	Signature	Date
……	……	……
Examiner III	Signature	Date
……	……	……
Chairperson Board of Examiners	Signature	Date
……	……	……
Supervisor	Signature	Date

ACKNOWLEDGEMENT

My gratitude goes to my supervisor Dr. Jackson Phiri who provided me with guidance in my research. Whenever I needed his assistance on issues that challenged my direction on the handling of the research, he was ever present and willing to provide guidance and leadership that I needed. He was never busy for me, he was more than willing to share his thoughts when I needed that. His vast knowledge and experience in research and areas of my interest in this research were a great asset to my work; I explored various options and views at the hands of Dr. Phiri's stewardship. Dr. Phiri encouraged me to participate in the Conferences where I presented my two papers during my study and research process. I would also like to thank the Department of Computer Studies, in particular the Head of Department, Dr. Mayumbo Nyirenda, and the former Head of Department Mrs. Monde Kabemba for their unflinching support that I received during the period of my research and study. They facilitated my work during my field work and other processes that needed their offices, respectively. They were always available to attend to the needs of my research and the interventions when the Department needed to engage other Departments, like the Directorate of Research and Graduate Studies of the University of Zambia or Office of the Dean of Natural Sciences. My gratitude goes to Mr. Wakwinji Inambao whom I consulted on metric issues and in his area of research in Euclidean Geometry in information security. Mr. Inambao shares a similar interest as mine in finding solutions in problems confronting cyber security. Lastly, I say thank you to the University of Zambia for this opportunity for me to do this research and contribute to the growth of science in Zambia and world over. The environment that was provided to me helped me to successfully and professionally conduct my research to contribute to the world of knowledge in the area of Information security.

DEDICATION

My work cannot escape the attention of God almighty, my creator, who has seen me through the detail and attention that was required in this noble contribution to my community. This work would not have been a success without the unwavering support that I received from my loving wife Ina Nchimunya Mwanza Kabwe. She endured the disturbances that come with this mammoth task, the work robbed her of her time; she lost part of her valuable time where my attention was needed, the lost time was naturally hers. Much of my work was done in my son's bedroom where I reasoned from various authors' past work and explored various lines of thought. Lwiyo, I am grateful for your patience and for allowing me disturb your peace. My daughters, Mukuka, Besa and Oluse, you were not tired to support me in all ways possible for me to achieve my goals during my research. My late beloved mother Dalia Besa, I would have loved you to see my work, God knew why He took you early in 2001.

ABSTRACT

The growth in the use of online services on the World Wide Web has proliferated into cyber mischief, personality or object misrepresentation, and cybercrime. Diverse entities of different interests and intentions form a wide range of complex online identities and beneficiaries of the online activities. Fraudsters and criminals hide their online identities to steal services, assets and other valuables or harm innocent internet users. This research would help in strengthening of identity management systems as a way to arrest this growing problem and guarantee secure online services and online interactions. This research desires to identify a mathematical model that would help in improving cyber security in digital identity management. This work intends to develop metrics models based on distance metrics in order to quantify the credential identity attributes used in online services and activities. This study adds knowledge to past work on the subject matter to provide quantitative analysis to quantify the credential identity attributes in online services. The study considers major sources of identity attributes currently being used in the application and registration forms for the various services offered both in cyber and real space. The study further explores the extraction of key identity attributes that were extracted from identity tokens like identity documents, application and registration forms for the various services offered both in the cyber and real space. At the core of the research, the study seeks to establish how we would develop the identity attribute metrics model which could be used to quantify the identity attributes based on distance metrics mathematical models. The study utilized survey research with closed-ended researcher administered questionnaire. A total of 160 questionnaires were administered with a response rate of 93%. The primary data obtained from questionnaires was analysed using Statistical Package for Social Science (SPSS) and Excel. The respondents were drawn from Banks (14%), Churches (12%), Government of the Republic of Zambia (6%), Hospitals (16%), Insurance (10.7%), Mobile Phone companies (2%), and less than 1% from Pensions. Others were Schools (21%), Universities (16%), and Utility companies (1.3%). The techniques that have been used include data mining techniques and statistical analysis. The perception constructs in the research included Usefulness, Trust, Ease of use, Image, and User satisfaction. It was observed that some attributes were more important than the others in identifying entities. Statistical analysis revealed that among the constructs that were used, Usefulness, Trust and Ease of use were strongly related. Tools to text mine the identity attributes helped to generate statistical data to come up with a quantitative model metrics to assist in the identification of an online entity. Using a detailed literature review, questionnaire surveys in this area, text mining of the identity attribute from the

application forms, and the results of the study helped to develop the identity attribute metrics model. An identity attribute metrics model based on distance similarity has been proposed. The Distance similarity is based on Cosine Similarity measure. Based on this study, digital identity management in online services and activities should adopt the developed Cosine Similarity measure, the identity attribute metrics model based on distance metrics, to strengthen online identity management. This will help to curb online fraud, identity theft, and other cybercrimes. This model could be augmented to past efforts to come up with a multimodal solution and add value to the resolution of the said problem.

Keywords*: Identity Attributes, Metrics Model, Digital identity, Online services, Distance metrics, similarity measure, cosine similarity, model, normalization, term weight, data mining, entropy.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ACRONYMS

ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ATM	Automated Teller Machine
COMESA	Common Market for Eastern and Southern African countries
Covid 19	Corona Virus Disease 2019
DIMS	Digital Identity Management System
E-Government	Electronic Government
E-services	Electronic services
E-tax	Electronic Tax
FIS	Fuzzy Inference System
FISP	Farmer Input Support Programme
HTML	HyperText Markup Language
ID	Identity
IDF	Inverse Document Frequency
IDM	Identity management
ISO/IEC	International Organization for Standardization/International Electro-technical Commission
JTC	Joint Technical Committee
KDT	Knowledge-Discovery in Text
MADM	Multiple Attribute Decision Making
NT	New Technology
OS	Operating System
PAD	Personal Authentication Device
PDF	Portable Document Format
PIN	Personal Identification number
SADC	Southern African Development Community
SD	Standard Deviation
SMS	Short Messaging Service
SSN	Social Security Number
SPSS	Statistical Package for Social Science

TAM	Technology Acceptance Model
TF	Term Frequency
UNDP	United Nations Development Programme
USA	United States of America
VSM	Vector Space Model
WWW	World Wide Web
XP	eXperience
ZRA	Zambia Revenue authority
Z-scores	Standard Scores

CHAPTER ONE

INTRODUCTION TO THE RESEARCH

1.1 Introduction

This chapter introduces the purpose of this research, it will give the background of the ethos of this work, and show the significance of this study. We will outline the context and limitations of the study as well as present the actors in the area under investigation of this research. The problem statement for this research, the aim, and objectives of the study will be spelt out in this chapter. An enlistment of the research questions, that we need to address, will be done within this chapter. We will spell out the organization or roadmap of this dissertation in this part of the dissertation.

1.2 Background

The proliferation of online digital information resource utilization has presented challenges in digital identity. Examples of online activities that have presented challenges include social media, online shopping, online banking, online billing, online reservations, online medical services, online relationships, and government's electronic service provision. We will elaborate further on the examples that we have indicated above. Social media has given rise to character assassination, distortion of information, misuse of information, and breakups of relationships due to falsehoods. Electronic commerce has witnessed the rise of online theft of personal digital identity and misrepresentation of ownership. Goods are procured, online, using other people's identities; individuals and organizations get swindled by fraudsters who disguise themselves to be selling goods when in fact the real intention is usually to harm innocent people or get access to their assets. Ownership of assets and possessions of innocent people have been lost to fraudulent transactions that appear to be legitimate when in fact not. Individuals and Financial institutions have suffered the brunt of fraud as they largely rely on identification of claimants of ownership, for any transaction that takes place; hackers would harvest the credentials of the legitimate owners of accounts from either websites, applications, or electronic devices to get access to assets that are not theirs. In the case of online medical services, accurate identification of persons to have correct records is very significant for online medical services to be secure. Innocent people get hacked by fraudsters with intentions to harvest their data or information to defraud them. Such actions may also be meant to injure reputations or characters of legitimate owners of data, information, or assets. Online relationships, in terms of friendships and love affairs, sometimes end up in disappointments or tragedy. Sometimes love affairs may truly flourish

with online individuals but only to discover that the individuals presented wrong information about themselves. For instance, a person may fall in love with a wrong individual, say someone who is older than expected, or not with good looks, or has bad character; this is usually done by misrepresenting facts and misleading the other person. Such experiences lead to despondency when these online lovers meet in person. Governments across the globe are diversifying their efforts in service delivery by extending to new methods of online services. These approaches require a robust and effective management of digital identity to avoid fraud, misrepresentation, and mischief. Recent studies [1] suggest that Zambia is not an exception to this experience as it has also been rocked with identity management challenges. An example is in the area of distribution of farmers' inputs by the government to the needy is one such area. Fraudsters pose to be legitimate clients of the government so that they could receive government support even when they are not legitimate beneficiaries. Government has spent a lot of money with the understanding that the recipients were the intended beneficiaries when in fact not.

Government of Zambia, through its Ministries, experiences challenges to enlist legitimate needy for them to receive government support; this has led to unmerited government expenditure. For instance, it is reported [2] that wrong individuals had benefitted from programmes by various ministries and/or departments at the expense of more deserving poor and vulnerable women, the youth, elderly and persons with disabilities. The report further indicates in [2] that some people that deserved to be on the programme were left out.

Ministry of Community Development and Social Services provides social assistance, social security, social health insurance, livelihoods and empowerment and protection [3] contends that it is important to have accurate identification of beneficiaries in the provision of the service [3]. Furthermore, Ministry of Agriculture provides Farmer Input Support Programme (FISP) which is considered one of the largest social protection programmes in Zambia, supporting about 800,000 beneficiaries [3]. The Ministry indicates in [3] that support is given to illegitimate beneficiaries, then this huge cost is misplaced; intended beneficiaries are left out whilst illegitimate individuals get the benefits, defeating the purpose of the service. The report [3] indicates that the poor are underrepresented among its beneficiaries. It is evident that the issue of identification of legitimate beneficiaries is important, more so when service delivery is automated or is accessed online.

The advent of internet banking, electronic sending of money through Mobile phone service providers has experienced theft of digital identities in these transactions.

Developing techniques of identification of the real owners of the digital identification would help in resolving this problem [1] and enhancing security in online activities. Researchers have explored the area of digital identity management but the growth of online services shows that more efforts are required to address the security challenge in this area. This research looks at the development of a metrics model in order to quantify the credential identity attributes. This will contribute to the efforts of enhancing security in online activities by validating the true owner of a digital identity. This research will contribute to the body of knowledge on addressing challenges of online identity problems which include loss of assets and privacy. This work was carried out of interest in the trends in information security to provide solutions to problems affecting society.

1.3 Scope

This study was conducted in Lusaka, targeting organizations that generate and use identification data for the registration of individuals or customers in their course of businesses. Primary data were based on the survey conducted on the sampled organizations; these included Banks, Insurance, Government, Universities, colleges and Schools, hospitals, mobile phone service providers, utility companies (Water and Electricity), and others. The sample size for primary data collection was based on the planned financial resources to cater for the activities of the research in this geographical area. Secondary data were derived from internet pdf (Portable Document Format) documents for the same organizations, the regions for data collection in the sampled regions and organizations.

1.4 Problem statement

Activities on World Wide Web are having challenges regarding digital identity management. Fraudsters pose to own digital identities so as to have access to the owners' information or assets. These challenges are increasingly causing online users to lose trust in online services. This is despite that when online services have been appreciated to increasingly bring good solutions to the online users. If only the digital management problem is addressed, then the benefits of online services, transactions, or interactions would be appreciated. Therefore, to maximize, the benefits of online interactions and services, we need to eliminate the threat of identity abuse, identity theft, or misrepresentation. Online interaction, access to online information assets and transactions, and safety to privacy is a very topical issue regarding the World Wide Web. There are a number of incidences that have occurred online that have brought a challenge to digital identification. Research interests have tried to address the issue of digital identity management while an amount of research work has been spent on establishing tools that

would help resolve this problem. The past efforts have not done much on using some techniques of using identity metrics modeling based on distance metrics. The research focuses on getting identity attributes of an entity by using different techniques and use the

1.5 Aim

This research aims to develop an effective response to the challenge that is currently being faced by online users, regarding digital identity, digital identity theft, abuse, and misrepresentation. The research intends to contribute to the development of a framework to identify attribute metrics modeling which is based on distance metrics. The research also intends to explore how we can construct a multi-modal, user friendly authentication system that combines multiple digital identification techniques.

1.6 Objectives

To realise the aims mentioned above, the following core project objectives will need to be met during the research:

i. To identify the main services as the sources of identity attributes from both in the real space and cyber space
ii. To use data mining tools and techniques to mine the attributes in (i) above.
iii. To use mathematical models based on distance metrics to quantify the identity attributes

1.7 Research questions

This research attempts to identify techniques that would help to uniquely identify an online entity by using identity metrics models based on Distance Metrics. This area has not been explored much, and thus will be able to add value to research efforts that have been done in the past. We will attempt to answer the questions listed below in order to attend to our research question:

i. What are the sources of identity attributes that could be used in online activities and services?
ii. What are the major services that use identity attributes?
iii. How can we extract the attributes mentioned in (i) and (ii) from the sources?
iv. Based on distance matrix, how do we develop the identity attribute metrics modeling?

1.8 Significance of study

Online service delivery and online interactions of individuals have been facing challenges of identity of online users. There is need to ensure that rightful owners of identity received their required services and that interactions are secure from the growing internet challenges. It is important to improve security in online services so that owners of online assets can legitimately claim their legitimate identify and not have fraudsters gain access to their assets. The owners of the assets could be individuals or organization, and in some instances it could be machines or information systems.

This research will contribute to past efforts in research in digital identity management and will contribute to the identification of entities of online service users to help to arrest cybercrime; this will help to address the challenges that come with online identity. The study will develop a metrics model that would quantify the credentials' identity attributes in online services. This will assist in raising the probability of accurately identifying the right individuals who need to be identified in online services in the cyberspace and therefore, curtail cyber fraud, improve information integrity, and confidence in online service and interaction. This would add value to already developed solutions to come up with a multi modal solution to address the challenges of digital identity management.

1.9 Organization of dissertation

This dissertation has five (5) chapters, each of which addresses a specific area of concern.

Chapter 2 will cover the literature that will be consulted on areas that affect this research. Quite a wide range of pertinent issues will be considered and evaluated. Some areas of interest will include impact of internet, digital identity, mathematical modeling, normalization of data, data mining, Shannon information theory and entropy, distance similarity measure, and term weighting to mention but some. Related works regarding this study have been consulted so as to establish the past efforts impacting on this study.

Chapter 3 will cover the methodology of the study, present methods, tools, and techniques that will be used in attending to the needs of our research and introduce the proposed model.

In Chapter 4, we will present the findings of our research.

Chapter 5 will present closing notes of this study, this will which include discussion, conclusion, recommendations, and future study interests.

1.10 Summary

The chapter gave a snapshot of the whole picture of this research; it gives the picture of what this study is all about, the problem the research is trying to address and the key issues of the problem of the research. This chapter also indicates the organization of this dissertation.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter discusses literature that has been reviewed in relation to this research and past work that relate to this research. A number of areas have been reviewed, these include literature on internet and activities on internet. Other areas include influence of internet, benefits that have come with internet, and the challenges that have come with it. Literature on digital identity management, mathematical modeling, statistical techniques, and data mining are some other areas that have been looked at. Further areas of interest that were reviewed include Shannon's information theory - including Shannon's entropy, distance similarity measure, and term weighting techniques.

2.2 Impact of internet

2.2.1 Positive effects of internet

Jibrin *et. al*. stated that Internet has made computers around the globe to interconnect [4] and therefore, creating a World Wide Web. They further stated indicted internet is a global community, one with a very active life [4]. The interconnection of computers has brought individuals, machines and systems of different background to interact together, online. This has revolutionized the way we carry out our lives, the way we interact, the way we do business, the way we transact, and the way we secure our assets both in real space and on cyber space. The internet has brought a number of benefits; technology is seen as an important catalyst for the restructuring of commercial activities and business development strategies. Digital technologies have proven to be the drivers of economic growth and competitiveness of doing business as electronic business revolution [5]. This however, comes with challenges that may endanger the benefits that come with technology. Amongst the changes that internet has brought include social and commercial relations, change of our attitudes towards trust, creation of social networks, increased risks arising in electronic businesses, and electronic shopping (which largely is electronic commerce) [5]. Governments, Universities (and schools), hospitals, financial institutions, and other public companies offer services that the community would need to access. Other areas which are using internet immensely are research, electronic games, and remote access of services in different areas. All these areas and a lot more assume that the individuals, machines, and systems that exchange data and information are either the rightful owners, intended recipients or legitimate stakeholders in the interactions on the internet. This therefore,

implies that online identification is pivotal in cyber interaction before the online actors would benefit from the online assets.

2.2.2 Adverse effects of internet

The rapid development of information and communication networks by governments, colleges, enterprises and individuals means that they are employing more and more information systems, perhaps without clear distinctions of the persons and devices behind their use [6]. It is obvious that the need for identity that would provide complete privacy is vital [6]. A study in [7] shows that cyber-crime has become one of the fastest growing crimes in the world. It was observed in [8] that networks are subject to attacks from malicious sources; with the advent and increasing use of internet, attacks are most commonly increasing. In 2007 it was reported [9] that in Australia alone, the proceeds of identity theft was still one of the largest sources of fraud, and it was estimated to be nearly $6 billion a year. Identity theft is one of the fastest growing crimes in the world; each year, more than 10 million people fall victim to identity theft, and many do not discover the crime until it is too late [9]. When devices are lost or stolen, allowing the right user to access the devise and not losing sensitive information or data stored on mobile device are the key concerns that are necessary to be considered against threats on mobile enterprise [10].

2.2.3 A list of some benefits and challenges

A study in [11] identifies areas that are benefits that arise from internet use:

- **Social networking**

  Internet brings people from far flung areas to interact in real time. Years ago, you needed to wait for days, weeks, months or years to get a response from a letter. Today people get responses in seconds or minutes. Blogs have made people from different areas from around the globe to socialize by chatting and exchanging views online. Business activities are conducted on social platforms using different applications like Zoom, Skype, and Google meetings.

- **Electronic communication**

  Communication has been made easy through internet; there is a variety of media communication tools which include short messaging service (SMS), electronic mail, Twitter, Facebook, Skype etc. In early 2021, Zambia Election Commission communicated to the electorates to confirm their voter registration details via online access to their data base.

- **Information sharing**

Information can easily be moved from one medium to another; this could be from electronic gadget to another e.g. from iPad to mobile phone. The benefits of information sharing are experienced by Zambia Revenue authority (ZRA) and its clients. Tax returns by ZRA customers are done through online applications to reduce customer queues at their offices.

- **Electronic banking**

It is easy to have access to banking services through mobile phones and computers which access the internet. Most of the Banks have extended their business hours and interactions with their customers by providing internet banking services by giving customers access to their accounts and perform transactions in their own time.

- **Online payments, bills and shopping**

Electronic payments are done to procure various services like paying for bills for water and sewerage, electricity, data bundles, tax, etc.

- **Selling and making money**

Procuring of goods and services are being done online; gambling is done online to reach several individuals who need to gamble. This benefit of the internet is one of the reasons that prompted this research. In addition, money transfers are being done using mobile phones, this has become a very popular phenomenon as it is making life easy for customers to send and receive money to each other and to organizations.

- **Collaboration, work from home, and access to a global workforce**

The experiences around the world with the pandemic of Corona Virus (Covid 19) caused countries, schools, colleges, universities, organizations and individuals to work remotely, using internet, from homes.

- **Donations and funding**

During campaigns politicians use internet to reach their supporters through blogs (e.g. Twitter, WhatsApp, and Facebook), electronic mails. Donations and funding are made through electronic platforms.

- **Cloud computing and cloud storage**

Storage and provision of services has been extended to servers being hosted on geographically distant locations on wide area networks. This removes the burden of organizations and individuals spending time and resources on managing data as this could be outsourced through cloud computing.

2.2.4 Challenges with internet

Internet has come with its own challenges [12], [13], [14] which identify in the following thematic areas:

- **Pervasion of privacy**
  Politicians, Public individuals, popular people, celebrities, and different individuals usually find that their private space is pervaded. This brings the interest of strengthening security to accord individuals and organisations the privacy they deserve. This study takes keen interest in this area.

- **Impaired public/private boundaries**
  Poor management of security of private information and data leads to abuse of private and classified information.

- **Cyber crime**
  This has become a very big online problem, criminal activities happen on internet affecting innocent lives. This is the main reason for this research, to find a way to contribute to arrest this situation. Criminals tend to steal innocent people's lives to hide themselves as they commit criminal activities. This has caused a lot of anguish and loss of property, assets and lives by innocent lives

All these activities mentioned here touch on the key element of our research; the users of the services above would need to be uniquely identified to access these services, otherwise fraudsters would take advantage and access these services.

2.3 Identity

2.3.1 Definition of identity

Our identity is, literally, who we are: a combination of personal history, innate and learnt beliefs and behaviors, and a bundle of cultural, family, national, team, gender or other identities [15]. The report in [15] goes on to say that our identity is important because it exists in relation to others. Identity exists in relation to the economic and social structures in which we live. How we are represented in economic, political and other societal systems – and our degree of choice and control as to how we are represented in these systems – sets the parameters for the opportunities and rights available to us in our daily lives [15]. It was stated that in [15] that our identity is increasingly digital, distributed and a decider of what products, services and information we access. This identity online is not simply a matter of a website login or online avatar – it is the sum total of the growing and evolving mass of information about us, our profiles and the history of our activities online. It was reported

that in 2018, the average internet user had 92 online accounts, and is likely to have over 200 by 2020 [15].

Pfitzmann & Hansen defined identity in [16] as any subset of attributes of an individual which sufficiently identifies this individual within any set of individuals. Identification of an individual requires an entity to be unique from any other entity so that this entity is distinguished from any other entity. Ayed in [17] indicates that identity refers to a set of qualities and characteristics that make an entity definable, distinguishable, and recognizable comparing to other entities. In other words, pieces of identity include a sense of personal continuity, a sense of uniqueness from others, and a sense of affiliation [17]. Identity is defined as consisting of traits, attributes, and preferences upon which one may receive personalized services which could exist online [7].

2.3.2 Partial identity

We observe from [17] that a specific partial identity is provided for identification. The researcher in [17] further indicates that the context will determine which subset of attributes is required, or which *partial identity* will establish enough trust for a given transaction to go forward. He further indicates that partial identity refers to 'a subset of identity information. A subject can have more than one mapped identity, where each identity encompasses valued attributes within an application context [18]. We call these multiple identities partial identities of the subject and denote the completed identity as the set of combinations of these partial identities that are used for specific application contexts [18]. An entity might have various roles, depending on the situation, the context and with whom they communicate, an entity is, therefore, likely to have a number of partial identities, each containing some of the attributes of the complete identity [19].

Research in [20] submits that a person may be represented by different identities, depending on the situation and the context. A person may have different identities according to the context in which the identity is applied. For instance, a researcher may be a father, magazine columnist, human right activist, sportsman, politician, philanthropist, friend, and lecturer. He is identified differently by different people according to each respective context. Each context has respective attributes that form each respective partial identity. For each context, respective attributes make him identified accurately; these attributes may differ from one context to the other. In each context, the identity seems to be complete yet the identity is just but a subset of his universal identity. This scenario is illustrated in figure 1 below. We learn in [20] that the identity of a person comprises a huge amount of personal data with respect to individuals. All subsets of identity represent the person (or components of a person). Some of these partial identities uniquely identify the person.

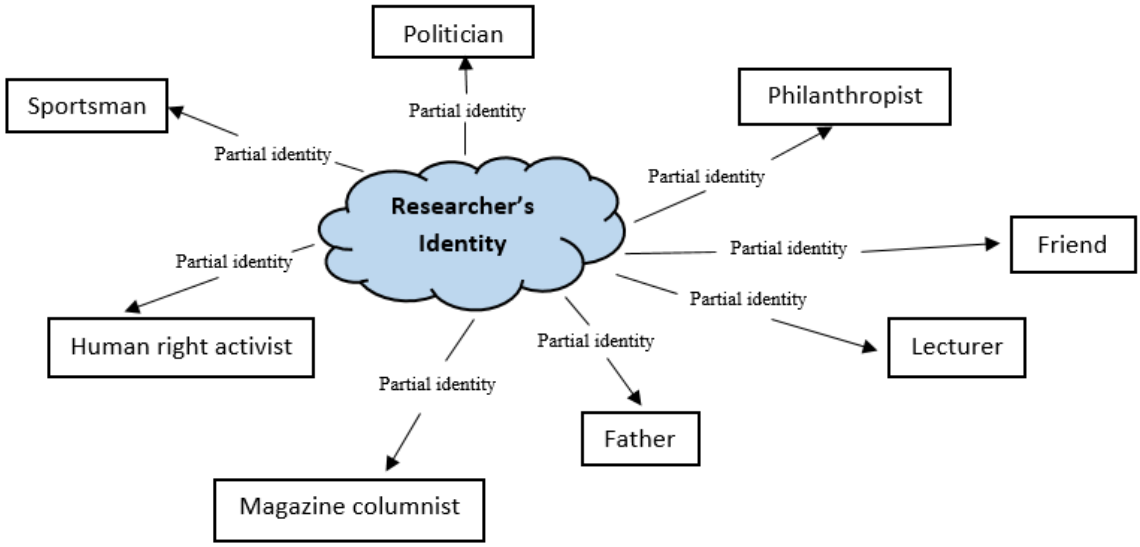

Figure 1: Illustration of Partial identity

Identity encompasses all the essential characteristics that make each human unique [7]. An identity of father of this individual may have characteristics of: father of three, kind, loving, hardworking, protective, supportive, merciful, jovial, progressive, etc. The identity of a person comprises a large number of personal properties [7], as indicated above. These properties help to uniquely identify an individual.

2.4 Identity attributes

2.4.1 Entities, identities, and identifiers

Findings in [21] indicate that the term entity refers to an active element of a system - e.g., an automated process, a subsystem, a person or group of persons that incorporates a specific set of capabilities.

An arbitrary string which is an information identifying the user in a non-ambiguous way (through identities like email address, social security number etc.), in other words, to represent an entity, we have characteristics of a set of elements that will together identify an entity [22]. It is suggested that users are generally identified by their attributes identity of any user as a set of attributes.

An entity (which could be a person or thing) can be identified using specific identification like user name, objects like electronic cards, biometrics like voice etc. The frequency of the attributes or identifiers which could be used in the identification would help in establishing metric models quantifications which could further help in uniquely identifying the entity. The figure below shows the relationships among entities, identities and attributes/identifiers contained by the identities.

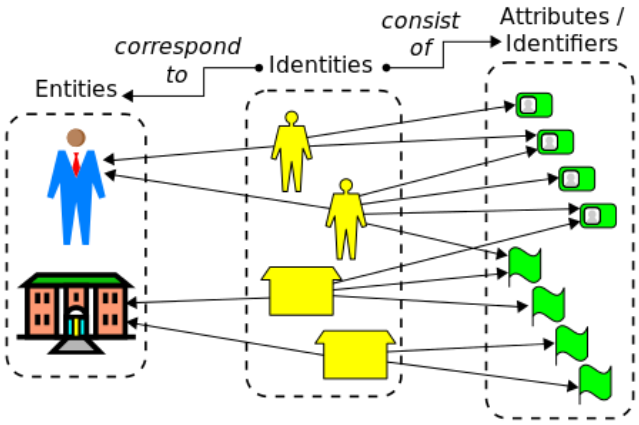

Figure 2: Entities, Identities, and Attributes/identifiers [23]

2.4.2 Relationships among entities, identities, and identifiers

An individual may have access to different services and needs that would need identification of this individual. These services could be provided by different entities. For instance, he could be an employee of a given company who had a bank account at a given bank. He may need to procure land from a government department and at the same time may be enrolled as a postgraduate student at an institution of higher learning. He may have different tokens of identities. Each institution relating with him may identify him differently.

With the growth in number of online services, a person may have access to several accounts of online services which would in turn need to confirm that they are dealing with a legitimate person. An individual may have identity tokens of various organizations like companies, banks, government Ministries or Departments, schools or universities. There could be another individual who would claim the identity that you have which would then pose a challenge in online service providers to trust online customers. The figure below illustrates what the situation would look like.

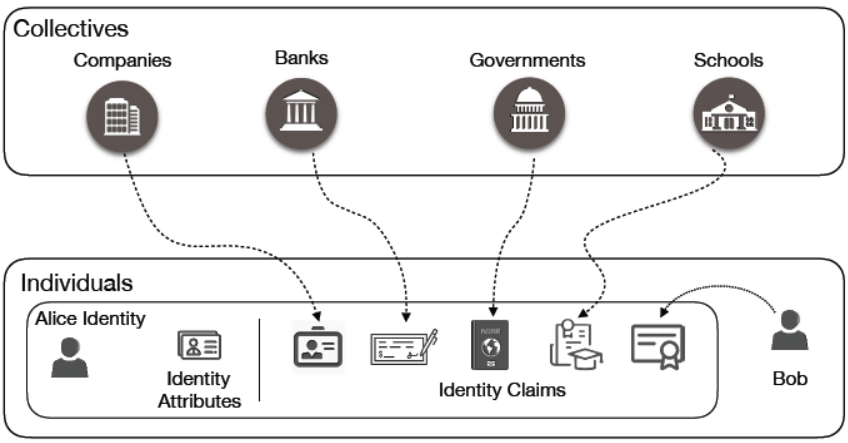

Figure 3: Example of identity claim by different individuals to same interests [24]

With the help of Partial Identities users are enabled to have a set of several identities in one community to decide for each identity which of their personal information they want to

disclose. Each Partial Identity of a user appears to other users as a unique, individual community member [25]. Credentials are often presented as evidence of identity. Credentials consist of one or more attributes. Attributes are the constituent components of credentials. During the process of establishing a person's identity, often more than one credential is required to achieve a higher level of trust [16].

It is noted in [7] that a user or object that would carry identification for access to a service would be termed as an entity. An Identity Provider is the issuer of user identity; a Service Provider is a relay party imposing an identity check. An identity will carry a set of user attributes. A Personal Authentication Device (PAD), which holds various identifiers and credentials and could be used for mobility.

2.5 Digital identity

In their study, Clauß and Köhntopp in [20] contended that Identity management systems enable the user to control the nature and amount of personal information released. In another study [26], Sittampalam stated that Identity management systems are elaborated to deal with the following core facets to reduce identity theft:

- *Reducing identity theft* - the problem of identity theft is becoming a major one, mainly in the online environment. Providers need more efficient systems to tackle this issue,
- *Management* - the number of digital identities per person will increase, so users need convenient support to manage these identities and the corresponding authentication,
- *Reachability* - the management of reachability allows a user to handle his contacts to prevent misuse of his email address (spam) or unsolicited phone calls,
- *Authenticity* - ensuring authenticity with authentication, integrity, and nonrepudiation mechanisms can prevent identity theft,
- *Anonymity and pseudonymity* - providing anonymity prevents tracking or identifying the users of a service, and
- *Organization personal data management* - A quick method to create, modify, or delete work accounts is needed, especially in big organizations.

As shown in [26], the problem of identity theft is becoming a major one, mainly in the online environment. Providers need more efficient systems to tackle this issue. Bhasker and Kapoor in [26] indicate that a digital identity is a virtual representation of a real identity that can be used in electronic interactions with other machines or people. An identity consists of traits, attributes, and preferences upon which one may receive personalized services. E-services require an effective way to manage digital identity information of the users [27]. A

digital identity is a distinguishing character or personality of an individual; it lays the groundwork necessary to guarantee that the Internet infrastructure is strong enough to meet basic expectations for security and privacy [28]. Digital identity management is a key issue that will ensure not only service and functionality expectations but also security and privacy [28].

In general digital identities can be considered and defined in terms of identity space, which can be categorised as *Real-space* and *Cyber-space* [29]. The Real-space identities are the physical identity tokens such as birth certificates, passports and driving licenses; the digital identities include the credential attributes such as usernames, passwords and Internet Protocol (IP) addresses [29]. A digital identity consists of two parts, whom the person or entity is and the attributes associated or owned by the entity (credentials and their attributes). These credentials define a digital identity and are varied, of widely differing values and have many different uses [29]. In cyber-space, when a person is interacting with another machine or person, the digital identity is usually a combination of many attributes that help to identify the other person's association [29].

Windley defines a digital identity as the data that uniquely describes a subject or an entity and the ones about the subject's relationships to other entities [30]. Further, Windley states that a digital identity is the persona that an individual presents across all the digital spaces [30]. In [31], we define digital identity as the electronic representation of personal information of an individual or organization (name, address, phone numbers, demographics etc.).

Al-Khouri indicates that digital identity is not just a number but a set of parameters that constitute a profile of the identity holder [30]. He further states that digital identities to specific persons will be authorised to perform certain actions in physical or digital forms [30]. We discover that in [31] in the digital world a person's identity is typically referred to as their digital identity [31]. It is argued in [32] that identity encompasses all the essential characteristics that make each human unique. They further argue [32] that these characteristics are the ones we call as attributes of an identity. The identity of a person comprises a large number of personal properties. It was observed that the networked environment in which we live and work requires digital identity – it is the key by which we are able to communicate, interact, transact, share reputations and create trusted relationships with people, business and devices electronically [33]. It is suggested in [33] that virtual identities, in the virtual world, can be connected to entities in the real world. Satchell *et. al*. [33] indicated that identifiers of a respective individual or entity would identify the entity online, from any context of the identity. An identifier uniquely identifies an entity (a person,

a computer, an organisation, etc.) within a specific scope. This underscores that digital identification is key in online activities of an entity on internet or computer network. It is observed that identity theft occurs when personal information is used by someone else without their knowledge; it usually supports criminal activity, including fraud, deception, or obtaining benefits and services in the person's name [33]. This is why researchers have taken a lot of interest in digital identity for the search of solution to cyber challenges that have come with online services and activities.

An individual or object, for instance a gadget, with a given identity will be represented by a token of identity to access areas where identification would be required. The token of identity could be a National identity card, passport, birth certificate, driver's license, marriage certificate, visa, or any other that fully identify an individual. Club membership cards, employee identity card, or student identity card are some other examples. ISO/IEC 29003:2013 in [9] gives a list of recognized tokens of identification. The token of identification is meant to fully distinguish the rightful owner of the token. The digital identity has attributes that help to establish an identity on online activities and services. As shown in [9], there is a list of internationally accepted attributes that have been identified that can identify an individual online. The figure below gives an imagery presentation of these details.

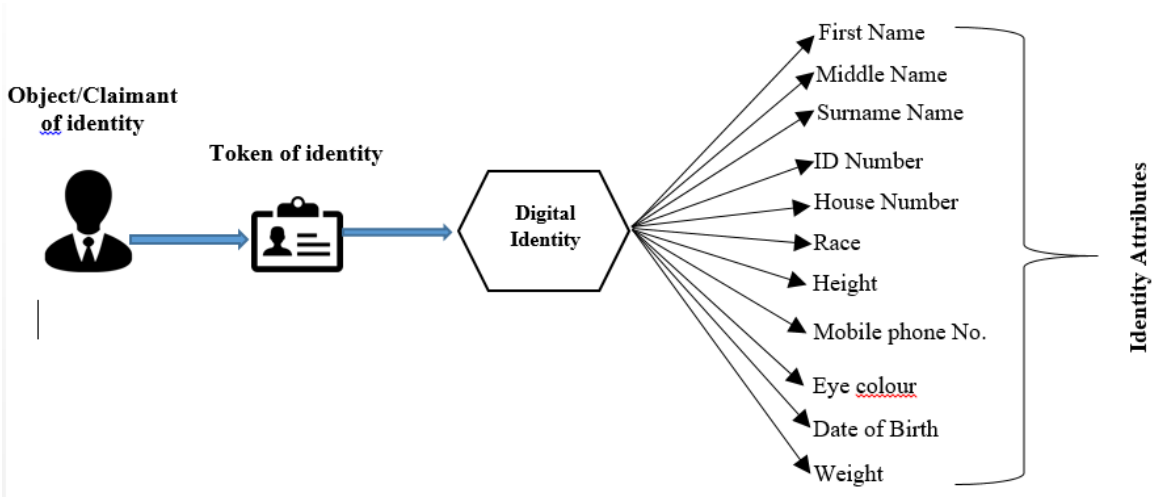

Figure 4: Representation of a digital identity by Identity Attributes

Entities in the physical world have digital identities which are represented by attributes which are the identifiers of the entity. A study in [34] shows that digital identity management presents solution to safeguarding personal information or information about an entity to protect private assets. Due to the unrestrictive nature of the Internet, without proper identification and authentication, users are becoming more vulnerable to identity fraud and theft. Online identity theft, fraud, and privacy concerns have become a huge issue now. Identity theft is big business.

2.6 Factors affecting digital identity

2.6.1 Information Security

In [35], security is defined as the quality or state of being secure - to be free from danger. The optimum state during the exchange of information or during communication would be when the actors in this environment are guaranteed of passing or receiving information freely. This is when there would be no intruder or party who would distort the intention of sending or receiving of such information. This exchange of information would imply exchange or receipt of assets. Information Security is defined in [36] as the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to ensure confidentiality, integrity, and availability. This implies that mechanisms should be put in place to deter any effort of ill intention that would harm legitimate interests in the exchange of information. This would be during the process of initiating the sending, the conveying, or receipt of information or any value attached to this information.

Smedinghoff in [37] states that the three concepts that embody the fundamental security objectives for both data and for information and computing services include Confidentiality, Integrity, and Availability. Only authorized parties should be able to access information, or assets associated with this information. To achieve this, there is need to put in place mechanisms that would limit access to confidential information and assets. For information to be deemed to have integrity, unauthorized entities should not be allowed to change any part of data or information. Upholding of integrity of data and information would help in building confidence in the data or information, the medium of exchange of these items, and the receipt of data and information. It is therefore, important to consider ways of establishing mechanisms that would help achieve this. The users or authorized individuals or entities should be able to have access at any time of their desire to access, use, or transact with these assets.

This research takes interest in finding ways of developing mechanisms to strengthen information security. This is meant to ensure that there is confidentiality, integrity, and availability of data or information or corresponding assets to rightful owners.

2.6.2 Electronic identification

Internet activities and interactions involve identification of entities. As mentioned in [38] identification process is designed to answer the question ‘who are you?’, it involves associating one or more attributes (e.g., name, height, birth date, SSN(Social Security Number), employer, home address, passport number) with a person in order to identify and

define that individual to the level sufficient for the contemplated purpose. It is proposed in [26] that entity identification would include identity proofing, identity vetting, and enrolment; this process is usually a one-time event. It typically involves the collection of personal information about a person to be identified. Gathering of identifiers occurs at enrolment; Al-Khouri in [26] defines enrollment process as registration of the entity, collecting various identifiers, and storing them for later verification and validation. The internet has come with risks on digital identity. As demonstrated in [33], the rate of identity theft and fraud has increased due to risks posed by data deluge. Data deluge poses risks to theft and fraud; for instance, disks that are full of social security data, or laptops that are loaded with tax records left in taxis, or credit-card would be stolen for fraudulent use. The owners of this data may lose properties or assets as a consequence of that loss of gadgets.

2.6.3 Technology

Soneka in [39] showed that there was a relationship among the three variables: perceived usefulness, perceived ease of use and perceived risk to E-tax adoption. In her study, these three variables were used to explain how these variables affect the adoption of E-tax system in rural Zambia. This demonstrated that there is some motivation in the use of tools by users to adopt or gain interest in the usage of such tools. This would suggest that even in the tools for digital identification would have the similar experience. Only constructs that were relevant to her research were applied. There is need to consider constructs that affect our research and observe the effect on the outcome of the relationships of such constructs.

It is stated in [40] that motivation of the use of technology, in this case identity tokens, is reflected in Technology Acceptance Model (TAM). This is tailored to information system contexts, and was designed to predict information technology acceptance and usage on the job. TAM has been widely applied to a diverse set of technologies and users. The study of Kabwe and Phiri in [41] suggests that whenever an individual perceives that using a certain technology will bring some benefits to the company, the individual will be eager to adopt it.

The research of Matikiti in [42] showed that TAM provides a basis for tracing how external variables influence belief, attitude and intention to use new technologies.

2.6.4 Identity management

Identity management elements include the user who has to sign in for issuance of a service, or for authentication for access to an electronic or non-asset device. The identification request has to be verified by a single or multi-mode verifier. An authorisation of the identification permits the requester access to the service or asset. Authentication is an

assumption of trust [43] and the enrolment of an applicant therefore, takes place. It is at this point of this process where fraud can either be prevented or allowed. We would say that the individual or person would be digitally identified at this stage.

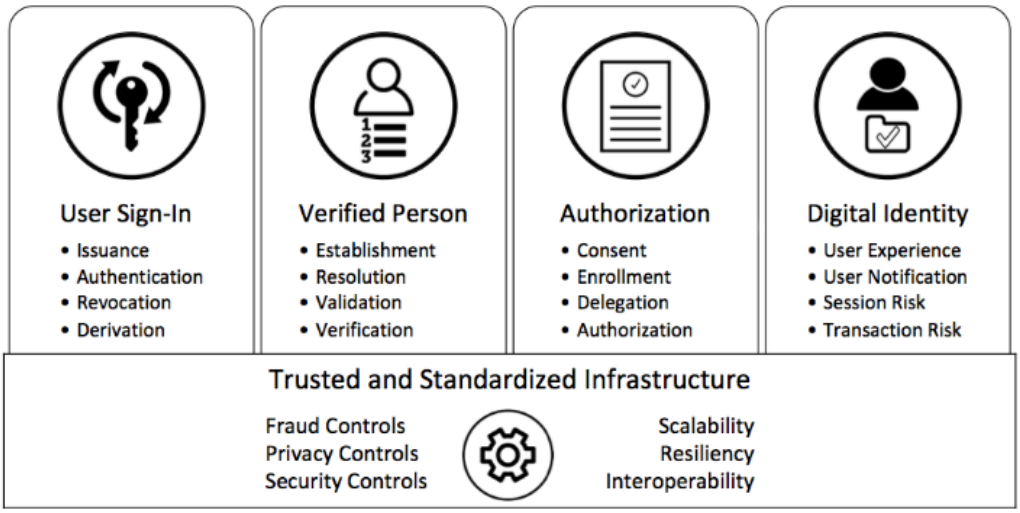

Figure 5: Identity management elements [44]

Research in [43] has showed that a Digital Identity Management System (DIMS) using multi-modal authentication would play a very big role in reducing cases of identity theft and fraud on online services. The system will incorporate the use of physical metrics (e.g. full name, passport number and race) pseudo metrics (password).

In [16], it is stated that a digital identity management system using a multimodal authenticating system was developed to address the issue of identity fraud and theft seen on most online services today.

There is a growing phenomenon of electronic service delivery by governments to their citizens. This is observed in [45] that government agencies are moving into the 'transaction stage' of Electronic Government (e-Government), it becomes clear that Identity management (IDM) more and more belongs to the core of national and international e-Government policy agendas. Miriam and Chiky observed that new questions arise as to how core IDM concepts like 'identity', 'identification', and 'identity management' can be redefined for their deployment in emerging e-Government environments [45]. They further noted that the continuing use of traditional paper-based and face-to-face public service arrangements and, with that, the use of traditional IDM means, requires not only a redefinition of these concepts for new digitised public service environments [45]. Organisation for Economic Co-operation and Development reports that e-government continues to be based on the initial principle of enabling users to access government information and services when and how they want (i.e. 24 hours a day, seven days a week) through channels including the Internet [46]. They further observed that delivery of user-

focused e-government services, on the other hand, largely involves government dealing with people in their capacity as customers or subjects, either as individuals or as part of a business [46]. The report also noted that ensuring the security and privacy of personal data that is collected and /or used in the process of electronic delivery is essential to building and maintaining users trust in online services [46].

Researchers have gained interest in this area to address the challenges that have come due to the growth in the use of internet and the problems that have evolved with it. The identity of an entity within a scope is the set of all characteristics (also called attributes) that have been attributed to this entity within that scope [47]. Identification of attributes of digital identity is of great interest in this research.

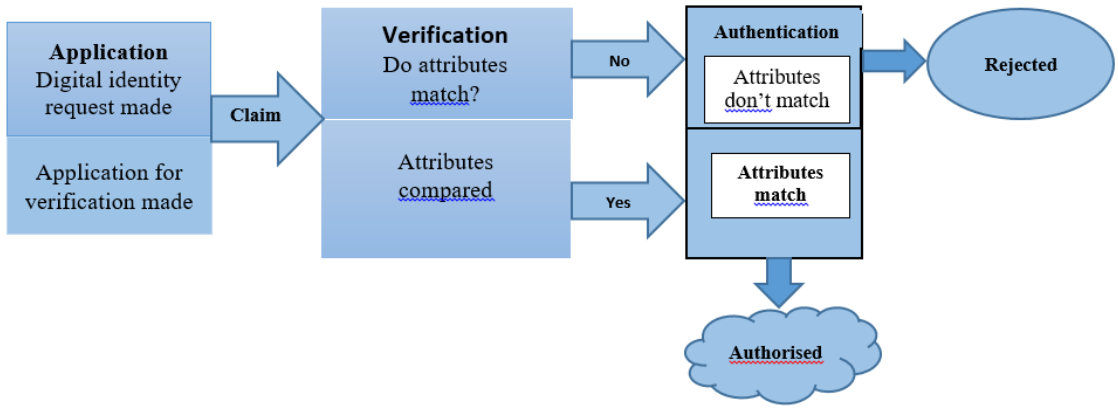

Figure 6: Digital identification process

In Digital Identity Management, enrollment of identity of an entity is by the collection of the identifiers of this entity, which we call as attributes. The attributes will be key elements in the identification proofing and verification of the applicant. At a request of an online service, by a subscriber, it is imperative that only the right subscriber should be given access to this service. Research efforts in [47] indicate that multimode authentication is likely to be the solution for most problems of fraud and identity theft seen on the cyber space today. Inambao *et. al*. in [48] showed that the use of mathematical models in the quantification of identity attributes can enhance security values to different identity attributes. They further indicated that digital identities are varied according to how sensitive or prominence the attributes are in the credential token [48].

According to [49], authentication of digital identity has three factors as the cornerstones of authentication: something you know (e.g., a password), something you have (e.g., an ID badge or a cryptographic key) and something you are (e.g., a fingerprint or other biometric data). It is stated that the strength of authentication systems is largely determined by the

number of factors incorporated by the system — the more factors employed, the more robust the authentication system [49]. This research therefore, takes interest in enhancing authentication of a digital identity; this is adding value to multifactor authentication.

2.6.5 Process of digital identification

In most cases, digital identification requires that the requester for identification must be on the database for identification. The entity must have made an application to be enrolled for recognition. The application could be done by filling in a form where details of personal identification are captured on a database. The applicant becomes subscribed and a token of identification is issued. The subscriber becomes a claimant until specific identification can be made by the applicant that is when authentication has to be made. A session of authentication takes place and verification of details that were submitted by the subscriber are thoroughly checked. The identification of the subscriber is done as verification of the personal details of the subscriber; the verifier has to be a secure mechanism. The attributes of the subscriber are the subject of authentication. It is therefore, important that a lot of attention is paid to the attributes of identification as they are the pillar of the security of identification of a subscriber. In our work, we will consider the application forms and the attributes identified which we shall apply in our work to strengthen further the process of identification. When identification is done, the subscriber can be authenticated.

A claimant of an online identity would need to apply for an online service, he would need to be authenticated through a process. Verification would need a number of techniques to be applied during this process. This includes the identification or extraction of attributes of a claimant. Other techniques that would need to be employed are data mining techniques to assess whether the attributes of a claimant match those of a real owner of the identifying attributes. Our interest in this research is on the verification of the attributes using a model we will later propose. Therefore our key area of interest is on the verification part of the process indicated in the figure below. A number of efforts have been made on improving the verification of claimants to have access to different interest on the internet. The figure below helps to present the area of interest of our research in the process of claim and grant of access to a claim of interest on the internet.

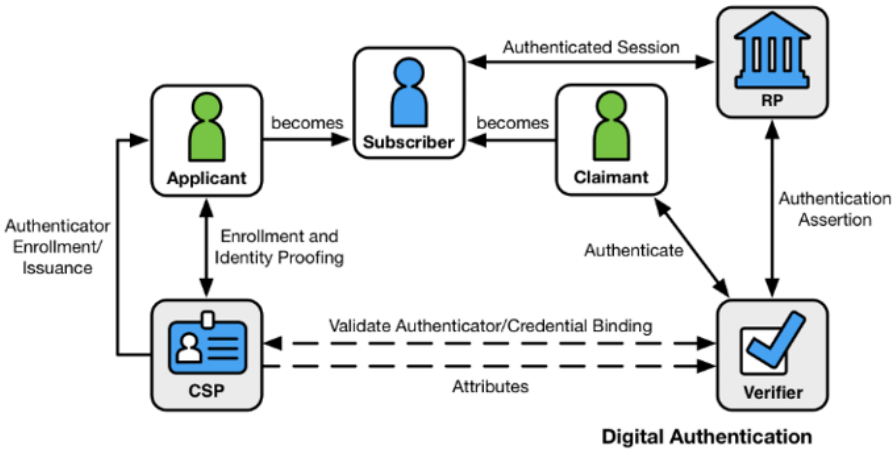

Figure 7: Online claimant verification [49]

Verification of online users requires an assessment or identification of the online users through a process as the one mentioned above. It was established in [50] that there are two ways of identifying the individual on the other side of the network session: explicit or implicit identification.

Explicit identification relates to processes in which the person is aware or even participates in the identification process by being prompted to submit a username and password; this is done so that the server could authenticate the user's identity (this is a shared secret) [51]. A person using a government service has to be fully authenticated with name, address, phone etc., whereas a person when shopping does not need to provide all these details. These changes of identity depending on the situation are represented by partial identities [51]. Implicit identification is when the user is not aware that their information is being used to authenticate them by obtaining identity information via log files, IP number of the person and visual appearance [51]. Digital Identity uses credentials of four different types and the credentials attributes mapping to segment the space of credentials into four different groups which are pseudo metrics, physical metrics, biometric and device metrics [52]. Most users are familiar with these attributes hence it is easier for the user to understand during authentication, however, this can also be a disadvantage as it is difficult to prove online whether the person is whom he or she claims to be. Hence other metrics are usually combined with physical metrics to obtain more robust user digital identity [32], [52], [53]. There are three areas that could be used to verify identification; this could be based on what a claimant is. What a claimant is would include finger print, iris, cornea, face, and voice; these are the biological phenomenon of an individual. Other form of identification would be based on what you know, this include things that could be remembered or memorized. They include names of objects, names of places, and things, these are the passwords that we

use for logging in for verification. The last kind of identification is based on what you have, this is the category that involves identification tokens to access online interests. Some examples of these tokens include smart cards, mobile phone, and security tokens. This is the category of verification of our interest in this study. The product of our study has to be developed to address the verification of tokens that fall in this category.

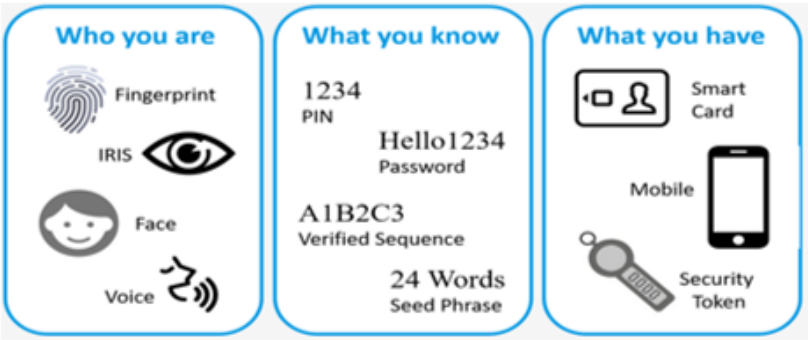

Figure 8: Types of identity verifications [54]

The figure below shows the process of verification regarding the identification token which will rely on the identity attributes of an individual or gadget.

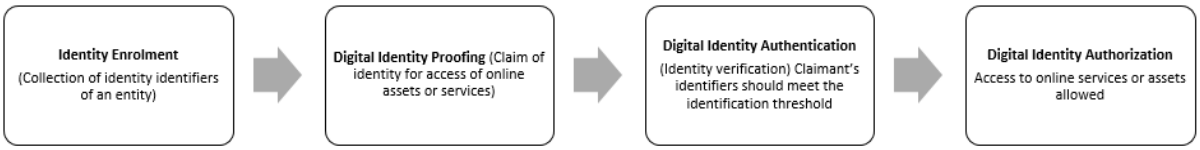

Figure 9: Digital identification verification process

Authentication of digital identity follows the identity proofing from the identity attributes (identifiers) that were collected at enrolment. Authentication can be strengthened by augmenting the authentication system with other factors that would make the identification robust. These augmented factors are the areas of interest of this research, particularly the similarity metrics.

2.6.6 Privacy requirement

Privacy is a central issue with identity and this is the reason why the official authorities of almost all countries that manage information and data of identification of entities have strict laws and policies related to identity. The reason is that privacy tends to protect the exposure of identity to theft or compromise to non-authorized access, use, or manipulation. Privacy is therefore associated with identity management of personal information and data. We observe in [25] that privacy is a complex and subjective concept that has different meanings to different people when used in different contexts. Privacy encloses four dimensions: privacy of the person, privacy of the personal behavior, privacy of personal communications, and privacy of personal data [25]. In this research, the interest we have is

the consideration of all the four dimensions. The reason being is that life on the internet or electronic interactions touch all the four dimensions mentioned above. Online interactions need to recognize the great need of privacy in all the four dimensions to enhance security of the digital identity.

Solove in [55] stated that privacy has a social value, privacy at its core relates to the integrity and autonomy of the individual, so that when privacy is compromised – no matter what type of privacy – the individual is being harmed in some way.

2.6.7 Trust framework

Trust, as defined in [56], is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another; it is a basic constituent of social life [57], [58].

This research has recognized a trust framework model which is demonstrated in [59], which is based on the communication model by Shannon and Weaver. It incorporates the sending and receiving process of information by an individual according to the three tier approach of data, information, and knowledge. The Shannon and Weaver model in [60] is one way directed and stems from the domain of information theory. In this model, a trustor places trust in the trustee [61]. It is indicated in [61] that communication consists of four major components, the sender, the receiver, the message, and the environment. The communication process can generally be distinguished in the three phases: sending, transmitting, and receiving. The phases of sending and receiving are concerned with the process of the message formation and comprehension by the sender and the receiver respectively. The communication framework guides the interaction of entities that interact or transact in the cyberspace. For two individuals to interact or transact, trust has to be involved. There has to be a trustor and a trustee, their relationship is based on the experience they have with each other in their interactions or transactions. Each actor has to verify the identity of the actor that one is relating with. The objects involved in their interactions can be accepted subject to verification of the identity of the actors, respectively. When the trustor requests for an object during communication, the trustee would need to verify the identity of the requestor. Verification would depend on the identifiers that resident on the data repository.

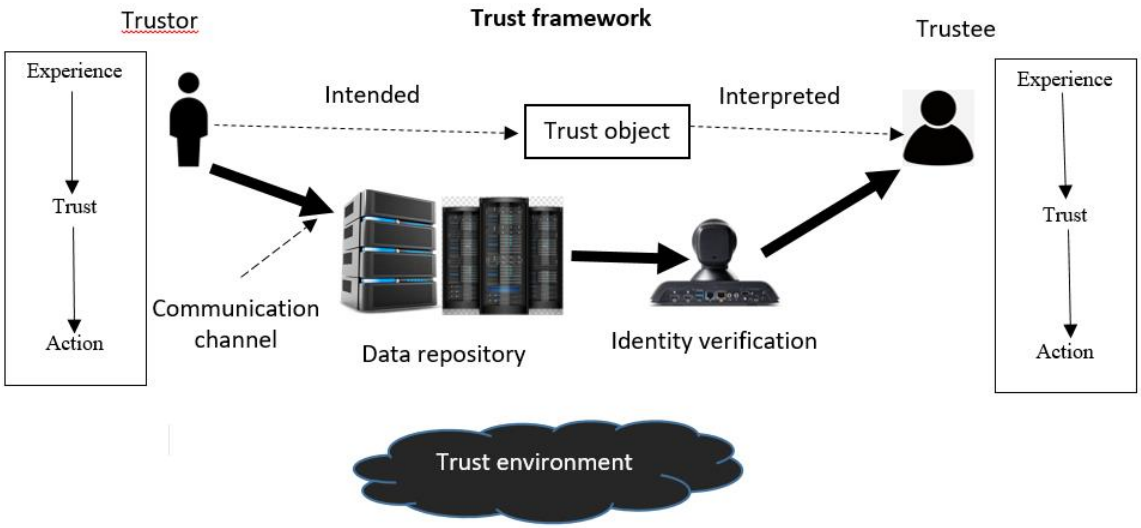

Figure 10: Communication trust framework

2.7 Mathematical modeling

Haines and Crouch (2007) characterize mathematical modeling as a cyclical process in which real-life problems are translated into mathematical language, solved within a symbolic system, and the solutions tested back within the real-life system. This demonstrates how mathematical modelling can present a mathematical model that would help in solving a real life situation using mathematics. It is the interest of this research to establish a model that would help in presenting a solution to the problem of this research. Mathematical models comprise a range of representations, operations, and relations, rather than just one, to help make sense of real-life situations (Lehrer & Schauble, 2003).

From the scientific and mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are [61].

2.8 Data standardization

A statistical standard normal distribution (X), normally termed as the ***z curve***, has a mean, i.e. $\boldsymbol{\mu} = 0$ and a standard deviation $\boldsymbol{\sigma} = 1$. This standard deviation is a measure of the extent to which it spreads about its mean [62].

As it is mentioned in [62], we often want to compare scores or sets of scores obtained on different scales. Standardizing data that comes from different sources would help us to eliminate the unit of measurement by transforming the data into new scores with a mean of 0 and a standard deviation of 1; these transformed scores are known as Z-scores [62]. Considering that this research has interest to compare with the performance of other metrics, it is prudent that we have a common ground of comparing the performance of the metrics as it is asserted by Abdi in [63] to improve our ability to discover knowledge [63]. Abdi [63] further indicated that this transformation includes normalising of data. In data

normalization, data from input sources is scaled to fall within a specified range such as 0.0 to 1.0 [64]. Olson and Delen in [65] indicate that the main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. It was noted in [66] that normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. It was further observed that incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because it was not considered important at the time of entry [66]. J. Han and M. Kamber in [66] further note that incorrect data may also result from inconsistencies in naming conventions or data codes used, or inconsistent formats for input fields. They also stated that some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies [66]. It was observed that having a large amount of redundant data may slow down or confuse the knowledge discovery process [66]. It was found that if some of the object's attributes are measured along different scales, attributes with larger scales of measurement may overwhelm attributes measured on a smaller scale. To prevent this problem, the attribute values are often normalized to lie between 0 and 1 [67]. We normalize data so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains [68]. We discover that a direct application of geometric measures (distances) to attributes with large ranges will implicitly assign bigger contributions to the metrics than the application to attributes with small ranges. The attributes should be dimensionless because the numerical values of the ranges of dimensional attributes depend on the units of measurements and, therefore, the choice of the units of measurements may greatly affect the results of clustering. One should not use distance measures without normalization of data [69].

To normalize our data, we had to find the Z-scores of the dataset of our interest; we subtract the mean of the scores from each score prior to dividing by the standard deviation [70]. A set of N scores each denoted tf_n (Term Frequency) and whose mean is equal to **M** and whose standard deviation is equal to SD is transformed in Z-scores [64].

$$Z_n = \frac{tf_n - \boldsymbol{M}}{SD} \tag{2.1}$$

Aksoy and Haralick indicated that this normalization procedure (Z_n) transforms the feature component tf_n to a random variable with zero mean and unit variance [71]. Researchers in

[72] show that the goal of all normalization procedures is the normalization of each feature component to the [0, 1] range, i.e. the min–max normalization [72].

2.9 Related Works

Campbell *et. al.* state that in the simplest case, the components of the sparse vectors are the raw frequency counts of each term in each document [72]. They also observed that search engines of the World Wide Web (WWW) are based on certain information retrieval models like Boolean model, Probabilistic model, and Vector space model [72]. Our interest is in the vector space model; Campbell *et. al.* indicate that the main purpose of information retrieval models is to retrieve relevant documents specific to a search [72]. It was observed in [72] that vector space model uses a storage matrix where columns represent the documents in a collection and whose rows represent term frequencies among the documents [72]. They also stated that for ad-hoc querying, dynamic queries are compared against a static document database in order to find documents closest to the query [72]. Simplistically speaking, a search engine has static database of documents, a query processor, to convert incoming (dynamic) queries into a format compatible with the representation model, and a relevant measure to compare converted queries against documents [72]. The researchers indicate that when conducting a query, one method is to search through the storage matrix and match the query terms with row terms producing the document closest to the query [72].

Researchers in [73] have established that Shannon's entropy method is one of the various methods for finding weights. In this study [73] it was observed that multiple attribute decision making (MADM) would require an evaluation, prioritization, and selection over the available alternatives that are characterized by multiple, usually conflicting, attributes. Lotfi and Fallahnejad also observed that since each criterion has a different meaning, it cannot be assumed that all the attributes have equal weights, and as a result, finding the appropriate weight for each criterion was necessary [73]. From this study [73], they discovered that in MADM, the greater the value of the entropy that corresponds to a special attribute (implying that the smaller attribute's weight), the less the discriminating power of that attribute in decision making process. It was established in [72] that the original procedure of Shannon's entropy can be expressed in a series of steps as follows:

i. Normalize the raw data,
ii. Compute the entropy of this data,
iii. Check the degree of diversification of data, and
iv. Set the degree of importance of the attributes.

In these steps, the raw data are normalized to eliminate anomalies with different measurement units and scales. This process transforms different scales and units among various criteria into common measurable units to allow comparisons of different criteria [73]. A study in [74] indicates that the entropic-weight method, from Shannon's entropy theory, was applied for the purpose of quantification of information in the respective variables under consideration. Vajapeyam in [75], summarizes Shannon's entropy as a direct measure of the number of bits needed to store the information in a variable, as opposed to its raw data. He adds that entropy is a direct measure of the 'amount of information' in a variable [75].

Inambao *et.al.* in [48] came up with a digital identity model that would supply trusted digital identities; this model would identify and extract various forms of identity attributes from various forms (identity tokens) [48]. The model was established on Euclidean Distance metric based on Euclidean geometry. This model identified attributes that were very key as identifiers of an entity, in other words, these are attributes that can closely identify an entity. This model helps in quantifying, implementing, and validating of the attributes from application forms (or identity tokens) [48].

We observe in [76] how to secure biometric data whilst at rest and or in motion so as to deter attackers in public organizations. Biometric identification contributes immensely to a person's identification and can therefore, contribute to the collection of digital identity attributes for individual identification.

Ibou *et.al.* in [77] indicated that attribute-based digital identity modelling needed to take into account privacy issues; A model was proposed in [78] that takes into consideration three fundamental aspects, namely security, privacy and identity theft.

The work of Phiri *et. al.* in [79] introduced a multifactor authentication system based on two identity attributes metrics models. This was in quest to demonstrate that an authentication system which has more than one factor would be improve in improving methods of identity management. Multifactor authentication broads the scope of digital identification in an Identity Management system; we could employ different modes of identification to make the digital identification robust and effective. Strengthening of the security of digital identity would include the developing of multi-modal authentication. This would include a combination of different authentication methods. For instance, [79] indicates that in the case of using an ATM bank card, in addition to the PIN number the user may be requested to submit a biometric feature such as a fingerprint in order to withdraw a certain amount of money above a given limit. Furthermore, a combination of biometrics, token based

credentials and pseudo metrics will most likely form a very effective defense against imposters [79]. The researchers where hoping that an additional fourth category of inputs would take into account identity attributes such as the name, date of birth, address and other acquired identity attributes for consideration [79]. Our research efforts are building on these past research efforts.

Phiri *et. al.* in [77] argued that a multifactor authentication system (in this case four factors of authentication) reduces the cases of cybercrime since it becomes difficult to forge all the authentication factors. In [77], one of the factors of authentication included biometrics. They went on to demonstrate the performance of the three fuser block technologies namely Artificial Neural Networks (ANN), Fuzzy Inference System (FIS) and Adaptive Neuro-Fuzzy Inference System (ANFIS) using the term weight and entropy identity attributes metrics [77].

The interest of this current research was an effect of the work of Phiri *et. al.* The aspect of expanding factors of online identification to metrics modelling became an area of interest. Phiri *et*. *al*. proposed in the close of their paper a consideration of authentication factors that would include metrics modelling methodologies [77].

2.9.1 Important data considerations

2.9.1.1 Term frequencies

Attributes are words (terms); word frequency (term frequency) is the number of times terms occur in a document itself and in a collection. Term frequency is the frequency with which a term occurs in each document [80]. In other words, Term frequency is the frequency with which a term appears in the whole document collection [81].

For us to determine which entity document contained the identity texts for an entity, we needed to ensure that we recognized text that was important for the required representation of a particular identity.

2.9.1.2 Term weights

Terms that appear in many documents in the collection should be discounted compared with the terms that appear in only a few documents. Thus, it is usual to take into account a *term weight* (also known as an *inverse document frequency* or *IDF*), denoted here as *idf* [82].

Jones in [83] stated that frequently-occurring terms in a corpus are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific terms are of greater value than matches on frequent terms.

2.9.1.3 Shannon Information Entropy

In information theory, as espoused by [84], the entropy of a random variable is the average level of *information*, *surprise*, or *uncertainty* inherent in the variable's possible outcomes.

Shannon Information Entropy represents the information content in a given variable in a data set; in this case the variables are the identity attributes. The Shannon information of a random variable *X*, with possible outcomes $x_1,\ldots,x_n$, which occur with probabilities $p(x_1),\ldots,p(x_n)$, the entropy of *X* is formally defined as [85]:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \quad (2.2)$$

where

$\sum$ denotes the sum over the variable values and *log* is the logarithm, the choice of base varying between different applications.

Base 2 gives the unit of bits (or Shannons); a bit (a binary digit) is either 0 or 1, the entropy H(X), expressed in bits, can take any positive value [86]. The base of the logarithm in (2.2) above can be chosen freely [86]. Since a change of base amounts to a multiplication by a constant, it specifies a certain unit of information. Then Shannon's entropy is simply the logarithm of the number of possible values [86].

We need to recognize that in logarithmic functions, it is a principle that for a variable *x*

$$-\log_a x = \log_a \frac{1}{x} \quad (2.3)$$

where *a* is a base

It is also worth noting that for a given function,

$$\sum_{i=1}^{n} k a_n = k \sum_{i=1}^{n} a_n \quad (2.4)$$

It then follows that Shannon Entropy function can be written as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) = \sum_{i=1}^{n} p(x_i) \log_2 \left[\frac{1}{p(x_i)}\right] \quad (2.5)$$

The text-mined documents in our research had statistics of text frequencies (represented by *tf*, for all *x* in *X* - a corpus of n documents) noted and applied in our computations. We apply Shannon Entropy function (2.5) above,

where

$$x = tf_i \quad (2.6)$$

and $tf_i > 0$

for $i = 1, 2, 3, \ldots, n$ for our corpus of *n* documents.

tf_i is an event, term frequency, of an attribute; it is determined from the Shannon entropy function as follows:

$$H(x) = \sum_{i=1}^{n} p(x_i) log_2 \left[\frac{1}{p(x_i)}\right] \quad (2.7)$$

where H is the entropy with the information x, while x_i is the i^{th} digital identity attribute of a given vector of a data set; $p(x_i)$ is the probability of occurrence of the i^{th} symbol.

2.9.1.4 Data Normalization

Term frequencies of the attributes of the subject dataset are collected according to organizations and countries, respectively. To standardize the data, we follow the Shannon entropy series.

Studies in [87] indicates that data integration needs to be transformed into forms suitable for mining. Data transformation involves smoothing, generalization of the data, attribute construction and normalization. Data transformation such as normalization may improve the accuracy and efficiency of mining algorithms involving clustering classifiers. Normalization is particularly useful for classification algorithms involving distance measurements such as lustering.

From the research in [88], it was stated that for distanced-based methods, normalization helps prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges. Such methods provide better results if the data to be analyzed have been normalized, that is, scaled to specific ranges such as [0.0, 1.0]

We have opted for Z-Score standardization since our data are not structured and we can identify the minimum value and maximum value from any given vector of our dataset. We would also be able to compute the mean value of each vector of a data set. We can also determine the standard deviation for the data that we have. Normalization by Z-score standardization goes with the function:

$$Z_n = \frac{tf_n - \mathbf{M}}{SD} \quad (2.8)$$

where Z_n is the normalized data, tf_n is the text frequency of an identity attribute, within the data set on *n* elements. **M** is the mean of all the attributes in a data set of a given organization, and SD is the standard deviation of the data set (the set of scores).

The mean of a given vector of attributes is obtained by the following function:

$$\mathbf{M} = \frac{\sum_{i=1}^{n} tf_i}{n}. \quad (2.9)$$

The Standard Deviation (SD) of a given data set, of *n* elements, is computed by the function:

$$SD = \sqrt{\frac{\sum(x-m)^2}{n}} \quad (2.10)$$

where *x* is an observation per data, *n* is the sample size, and *m* is population mean.

Text frequencies of the identity attributes from the four organizations in our research formed the vectors of our consideration. There were 19 attributes as our sample size, we had four organizations, namely: Banks, Government, Insurance Companies, and Universities and Schools.

2.9.1.5 Term weights using Shannon Entropy

Term weight of the identity attributes can also be determined using Shannon's Entropy. This would help to classify which attributes have more frequency in the documents than the others; the entropy would equally help to disregard the differences that would be in measurements of different attributes. The noise in the information of the identity attributes would be eliminated. The computations of the entropy would lead to classification of the different identity attributes according to importance in identifying an entity.

The weight for each attribute could be obtained by computing the entropy of each desirable property of the identity attribute by following the steps that are given in section 2.9.1.3.

From the Entropy function, we could deduce that:

$$\textit{Weight of an attribute} = W_i = p_i \log_2 \left(\frac{1}{p_i}\right) \tag{2.11}$$

Simplistically, we would represent the Shannon's Information Entropy by

$$H = \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right) \tag{2.12}$$

2.10 Distance metrics

A cluster is a collection of data objects that are similar to objects within the same cluster and dissimilar to those in other clusters. Similarity between two objects is calculated using a distance measure [89]. Charulatha *et.al* point out that clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar [59]. As pointed out by Backer and Jain, in cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create 'interesting' clusters) [49]. Researchers note that it is natural to ask what kind of standards we should use to determine the closeness, or how to measure the distance (dissimilarity) or similarity between a pair of objects, an object and a cluster, or a pair of clusters [49].

In order for the distance metrics to make sense, good data transformation or normalization is required. In data normalization methods, the objective is usually to ensure that the computed distance metric or similarity measure will reflect the inherent distance or similarity of the data [50].

Documents and clusters are represented as points in space, we can compare them using vector cosine. Clusters include a "center" or centroid vector that is the weighted average of the documents or clusters they contain. To prevent longer documents from dominating centroid calculation, we normalize all document vectors to unit length. To compare a document to a cluster, we simply calculate the cosine between the document vector and the cluster's centroid vector [90]. *Similarity function or measure* is a real-valued function that quantifies the similarity between two objects [91].

Cosine similarity measure is a similarity that measures the cosine of the angle between two vectors projected in a multi-dimensional plane [91].

A study by Weller-Fahy *et. al*. in [92] shows that a distance measure is a function *dist*() which takes as input two distinct variables A and B, and returns the value as distance.

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors as the cosine similarity. Cosine similarity is one of the most popular similarity measures applied to text documents, such as in numerous information retrieval applications and clustering [93].

It is mentioned in [94] that the cosine of 0° is 1 and it is less than or equal to 1 for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0 and two vectors that are diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is restricted to the first quadrant of the Cartesian coordinate system. Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter [94]. This distance metric will give us a number from the closed interval [0, 1], where the two overlapping vectors (i.e. 0^0) would denote 0 and vectors having an angle of 90 ° (which is the highest difference between the vectors in the first quadrant) would denote 1 [95]. Cosine Similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of an angle between them [95]. Cosine similarity is an efficient distance metric that compares the difference between vectors [96].

2.11 Term weighting scheme

The frequency term means the raw frequency of a term in a document. Moreover, the term regarding inverse document frequency is a measure of whether the term is common or rare across all documents in which can be obtained by dividing the total number of documents by the number of documents containing the term [97].

The term frequency factor (TF) has a substantial importance in a term-weighting system [98].

Term frequency alone is not enough to achieve plausible performances in retrieval systems. There are situations where the query terms are spread in the entire document collection, making the system retrieve all these documents and consequently affecting the precision of the results. This means that in order to fill the precision gap, a new factor must be introduced. That factor is the inverse document frequency (IDF) [99].

Words that appear often in a collection of documents do not provide much information as words which occur occasionally. IDF is given by the equation 3.19 and is given by the logarithm of the inverse proportion of a word over the entire document corpus. In equation 3.19, *N* is the total number of documents in the collection and n_i is the number of documents which contain the query term *i* [99].

The Inverse Document Frequency component (*idf*) of the function is expressed when we multiply original *tf* factor by an inverse collection frequency factor [52]. The study in [100] shows that the inverse document frequency, can be calculated by

$$idf = \frac{The\ number\ of\ total\ docments\ in\ the\ corpus}{The\ number\ of\ documents\ that\ include\ the\ term} \tag{2.13}$$

Each word in a document has weights, these weights can be of two types i.e. local and global weights. If local weights are used, then term weights are normally expressed as Term Frequencies (TF). If global weights are used, Inverse Document Frequency (IDF) gives the weight of a term. It is possible to do better term weighting by multiplying "TF" values with "IDF" values, by considering local and global information. Therefore, total weight of represented by "Total weight of a term = TF-DF". This is commonly referred to as "TF-IDF" weighting [101].

The Term Frequency-Inverse Document Frequency (=TF-IDF) is a numerical statistic which reflects how important a word is to a document in the collection or corpus [101]. This method is often used as a weighting factor in information retrieval and text mining. In [102], Buckley stated that one class of term weights has proven itself to be useful over a wide variety of collections; this is the class of *tf*idf* (term frequency times inverse document

frequency) weights. The function "TF-IDF" is also one of the most popular term-weighting schemes for user modeling and recommender systems [56].

Researchers indicate that term frequency (*tf*) factor is represented by the logarithm of the term frequency to scale the effect of unfavourably high term frequency, where *N* is the total number of documents in collection, and n_i is the number of documents to which a term is assigned [52], for a term *i* in a document.

Chen and Chang indicate that the Term Frequency (TF) and inverse document frequency (IDF) form an *algorithm TF-IDF* (Term Frequency-Inverse Document Frequency) *weight* which is widely applied to count the weight of a term [100]. TF represents the number of times a term occurs in a document in a given corpus, while IDF is the inverse document frequency, IDF, indicates the general importance of a term in overall documents [100].

Umadevi's work in [103] shows that the cosine similarity between two vectors (or two documents in a Vector Space) is a metric which measures orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we are not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents.

We will represent the term frequency by *tf*, where $tf > 0$. A lecture in [104] indicates that a log frequency weight of term *i* in document *d* in a corpus is defined as follows

$$w_{i,d} = \begin{cases} 1 + tf_{i,d}, & if \;\; tf_{i,d} > 0 \\ 0 & otherwise \end{cases} \tag{2.14}$$

It follows that the term frequency of a document *d* would be expressed as

$$TF = 1 + \log tf \tag{2.15}$$

In the weight scheme *tf*idf*, the *idf* factor is the inverse document frequency, which estimates the document frequency of the term *I (*the numbers of documents that contain *i*).
df_i is an inverse measure of the informativeness of *i*,
while
$df_i \leq N$, where *N* is the total number of documents in the corpus.
We define inverse document frequency *(idf*) by

$$idf_i = \log \frac{N}{df_i}, \text{ where } df_i > 0. \tag{2.16}$$

The combination of the term frequency measure and the inverse document frequency forms the well- known weighted scheme, TF-IDF, which is given by (2.17). In this equation, *tf* (*i*) is the number of times that the term *i* occurs in the document collection, *N* the total number of documents and n_i the number of documents that contain the terms q in their contents.

$$TF * IDF(i) = tf(i)\ X\ \log\frac{N}{n_i} \quad (2.17)$$

Researchers in [105] state that TF-IDF can be interpreted as the total quantity of information needed in order to compute the mutual information between documents and query topics. This means that TF.IDF can be thought as the reduction of the uncertainty about a random variable.

Table 6 represents the term frequencies (TF) of the corpus. Considering the TD-IDF term weight scheme, from our findings above, we would have the weighting computational outcomes to be as indicated in the tables 18 and 19 below. Since W_i = *TF-IDF*, then from the function below (which has already been stated above),

$$w_{i,d} = \begin{cases} 1 + tf_{i,d}, & if\ \ tf_{i,d} > 0 \\ 0 & otherwise \end{cases} \quad (2.18)$$

it follows that when $tf_{i,d} > 0$, the metric would be represented by:

$$W_i = tf_i * idf_i = (1+\log tf_{i,d}) * \log\frac{N}{df_i} \quad (2.19)$$

Otherwise we would have:

$$W_i = tf_i * idf_i = 0 * \log\frac{N}{df_i} = 0 \quad (2.20)$$

We obtain the weighting of the attributes (terms) by considering the function above of which the outcomes are indicated in the tables 19 and 20 below.

Cosine similarity using tf-idf vectors showed accuracy in the research conducted in [91].

Summary

Literature related to areas of interest for this study was reviewed so as to establish what has been covered by other researchers. Pertinent issues that touch on the issues of this study were considered and looked at in the literature involving these areas. Some of the areas include the impact of internet on the lives of the users, benefits and challenges associated with the use of internet. We looked at the concept of identity and attributes that are identifiers of an object or individual. Online identification was considered, reflecting on contributing elements of entity identification. We consulted literature to consider the effect of trust and privacy on online users as they interact with each other. Literature covering mathematical modelling and the standardization of data was reviewed. Literature on distance metrics and how it could be of use to identification of online users was reviewed. In general, the literature covering past work that was related to this study was looked at. The gaps between past efforts related to this study and the interests of this research were identified.

The table below is a summary of some literature that has been consulted which is relevant to the study.

Table 1: Summary of literature review

No.	Title	Year	Author	Findings	Gap
1	Design and Implementation of Multimodal Digital Identity Management System Using Fingerprint Matching and Face Recognition	2006	J. Agbinya, N. Mastali, R. Islam, and J. Phiri	A digital identity that distinguishes character or personality of an individual consists of traits; digital identity management is a key issue in online service, security and privacy	Identity Attribute Metrics Model based on Distance Metrics were neither considered nor developed
2	Identity Attributes Quantitative Analysis and the Development of a Metrics Model Using Text Mining Techniques and Information Theory	2012	J. Phiri and T Zhao	Identity attributes can be quantified to develop a Metric Model using Text Mining techniques and Information Theory	
3	Identity Management and its support of multilateral security, Computer Networks	2001	S. Clauß and M. Köhntopp	The identity of a person comprises a large number of personal properties. All subsets of the properties represent partial identities of the person and may relate to roles the person plays	Identity Attribute Metrics Model based on Distance Metrics were neither considered nor developed
4	Credit Card Frauds and Measures to Detect and Prevent Them	2010	K. Christian, B. Katja, T. Markus, H. Stephan and R. Kai	The process on identity management includes authentication. This is a process of verifying claims about holding specific identities	
5	A model for improving E-Tax Systems Adoption in Rural Zambia based on the TAM Model	2019	P. N. Soneka and J. Phiri	There is a model which reflects that factors that influence the level of e-Tax systems adoption in Zambia using Technology Adoption Model (TAM) are useful, easy to use, and secure	i. Only the importance of some constructs were considered ii. Identity Attribute Metrics Model based on Distance Metrics were not even looked at
6	User Acceptance of Information Technology:	2003	V. Venkatesh et. al	TAM has been widely applied to a diverse set	

	Toward a Unified View			of technologies and users	
7	Application of the Technology Acceptance Model and the Technology-Organisation-Environment Model to examine social media marketing in the South African tourism organization	2018	R. Matikiti, M. Mpinganjira, and M. Roberts-Lombard	TAM provides a basis for tracing how external varriables influence belief, attitude and intention to use new technologies	
8	Digital Identity Modelling for Digital Financial Services in Zambia	2018	W. Inambao, J. Phiri and D. Kunda	The study was able to extract and identify the identity attributes that were closely related in identifying an entity. This would largely help in ranking key attributes that would be required in identifying an entity.	The focus was on identifying attributes that ranked high in identifying an entity. Identifying a legitimate claimant from multiple claimants was not emphasized.
9	A Comparative Analysis of Eucledian Distance and Cosine Similarity Measure for Automated Essay-Type Grading	2018	O. E. Oduntan *et. al.*	Cosine similarity gives a useful measure of how similar two documents are like to be in terms of their subject matter	This area has not been focused on as a means to consider the identity attributes that could be close to identify an entity
10	Choosing a Distance metric for automatic word categorization	1998	E. Korkmaz and G. Üçoluk	Distance metric will give us a number from the closed interval [0, 1], 0 denoting that the two vectors are overlapping and 1 denoting that there the highest difference between the vectors	Previous studies on digital identity did not look at using distance metric to identify an authentic claimant. We could employ distance metric to test if a claimant could be used in digital identity

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter covers the setup of the research and the approach that was taken to conduct the research to respond to the research questions; this includes the strategies, methods, and the design for the collection of data. During the research, target population had been identified and sample size of the research was considered. Sampling procedure and research instruments for this research were considered. Data gathering methods and techniques for extraction of data that had been established. The procedures and techniques that were used to gather, analyse, and implement the model were considered, and the the model development was proposed.

In our methodology we addressed our research objectives which included the finding of major sources of identity attributes. These are used in the application and registration forms for various services offered both in the cyber space and real space. We showed ways of extracting the key identity attributes from the application and registration forms. These were for various services that were offered both in the cyber space and real space using data mining techniques. The research established a mathematical model which was based on distance metrics to quantify the identity attributes.

3.2 Research setup

A study in [106] states that research design is intended to provide an appropriate framework for a study. In research design process, the choice is made regarding research approach since it determines how relevant information for a study will be obtained.

Research design is the conceptual presentation within which research is conducted. It includes the blueprint for the collection, measurement, and analysis of data [107].

Our research design was intended to help in addressing the research problem and guide us to attend to the research objectives. The design included the research methods and techniques that helped us collect data, extract identity attributes from identity entities, analyse the attributes, and come up with a model that helped us to quantify the attributes for further analysis.

Figure 11, below, shows our general approach that was employed in this research. We chose our research approach, identified our research population and sample, identity documents for the entities were represented in real space or in cyber space. We then extracted data from

the entity documents. We had two sets of data that we considered in our research, primary data and secondary data; Key identity attributes from these entity documents were then identified for analysis. We established a mathematical model that helped us to quantify the identity attributes which helped us to identify the claimant of a digital identity in a given space, either real or cyber. The figure below shows the phases that were used in this research.

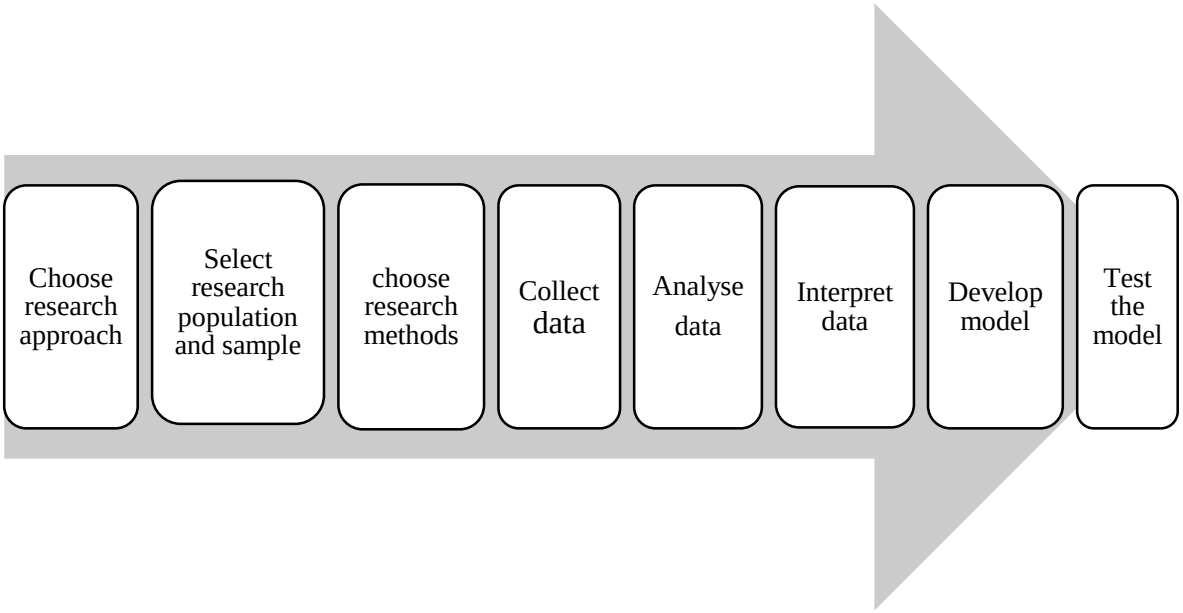

Figure 11: Research phases

3.3 Research Approach

In our research, we took a quantitative approach; Leedy in [108] explains that this is a research method that deals with numbers and anything that is measurable in a systematic way of investigation of phenomena and their relationships. It is used to answer questions on relationships within measurable variables with an intention to explain, predict and control a phenomena.

The research was quantitative since there was need to quantify the identity attributes using a mathematical model on distance metrics. A study in [1] showed that we could use the Distance metrics to quantify the identity attributes. The quantification would identify attributes that are very key as identifiers of an entity, in other words, these are attributes that can closely identify an entity in online activities.

The research considered different variables of different identity attributes and observed from the mathematical model the relationships amongst attributes.

The reasoning of our adoption of the quantitative approach was based on the following:

- The research quantified the identity attributes using distance metrics mathematical models,
- The identity attribute model that was established was used to quantify the identity attributes is quantitative in nature,
- The results of quantification and computations assisted us to establish relationships between phenomena of different variables,
- The study was based on representative sample, results were generalizable to the population and this helped us to make inferences on the population.
- We had to compare the relationships between and among different variables of identity attributes.
- The research had used numerical data and therefore, needed to employ quantitative analysis.

3.4 Research methods, techniques, and tools

A number of methods and techniques appropriate for this research were employed. Some of the techniques included data collection techniques mentioned below, text mining techniques and statistical analysis. Key identity attributes from identity tokens like identity documents, application and registration forms for the various services offered both in the cyber and real space were extracted using data mining techniques and use of data mining tools. More details have been supplied at respective appropriate portions of this work below.

3.5 Research Population

The entire set of cases from which research sample is drawn is called a population. Since, researchers neither have time nor the resources to analyze the entire population, they apply sampling technique to reduce the number of cases [109].

The research population was conveniently selected as Lusaka, as it is the National Capital of Zambia, which is a hub of commerce and administration of most institutions in Zambia. This is supported by position of a study in [110] which indicated that Lusaka provides services including administrative functions to Zambia as a whole. Another study by the United Nations Development Programme indicates that Lusaka province is relatively more industrialized and economically diversified and has a larger share of productive, and predominantly nonagricultural and employment [111].

Most of the sampled institutions deal with entity documents of identification in their day-to-day businesses for registration of business accounts and electronic commerce. Most of these organizations also have online information systems that run on internet and electronic

information systems that can be accessed by end-users on computer networks. These institutions include Banks, Government Ministries, Universities, colleges, schools, Insurance companies, Mobile phone companies, Utility companies, and Churches. Some information systems in Lusaka that would be found in some of these institutions include farmers input system, mobile money system, Zambia Revenue Authority Tax payer system

3.6 Sampling

Sampling is the process of selecting a statistically representative sample of individuals from the population of interest [112]. We would therefore, say that a sample is a representative unit (or section) from a population for study. Sampling is an important tool for research studies because the population of interest usually consists of too many individuals for any research project to include as participants. A good sample is a statistical representation of the population of interest and is large enough to answer the research question [113].

Choosing a study sample is an important step in any research project since it is rarely practical, efficient or ethical to study whole populations. The aim of all quantitative sampling approaches is to draw a representative sample from the population, so that the results of studying the sample can then be generalized back to the population [114].

Our research sample for primary data was picked from organizations in Lusaka using probability; this included clustering and systematic sampling techniques. Organizations were identified according to their groupings in their business areas, lists from online information were made per each business area. Systematic sampling was then applied on the list of each grouping. Secondary data sample was equally picked using the same sampling techniques to identify the regions and countries whilst maintaining same clusters of institutions.

Clustering involves the task of dividing data points into homogenous classes or clusters so that items in the same class are as similar as possible [60]. We observe in [60] that clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters.

160 closed-ended self-administered questionnaire were distributed to the identified organizations.

3.7 Eligibility criteria

For primary data, organizations that were considered needed to be Lusaka based; respondents needed to be those who had first-hand experience and were involved in the registration of accounts for the services that their organizations offered. Research was being

conducted in English, therefore, respondents needed to be able to respond in English since the research was being done in English. Secondary data were gathered from countries and regions that were English speaking with organizations that were in the similar clusters of the primary data organizations.

3.8 Delimitation of the study

The field work for this research was conducted in Zambia and was restricted to Lusaka, according to the resources that were available. Economic activities in Zambia are largely concentrated in urban areas like Lusaka. The research's focused on tokens that involve registration for identification, it would therefore be suitable to be done in Lusaka as it is the capital of Zambia where most registration of identity in many organizations would be found.

Secondary data were collected from internet and are discussed further in the succeeding sections of this dissertation.

3.9 Ethical Attention

There were no ethical issues in the research since there was no sensitive information or issues that were encountered. Permission was sought from the University to conduct a research in Lusaka and an introductory letter of the researcher to the targeted organizations was issued.

3.10 Research Limitations

The researcher was self-sponsored, therefore, only a sample that was within the budget of the researcher, time, and available logistics was considered. However, the sample had to be sufficient to address the requirements of the research to deliver representative results. Respondents to the questionnaires answered at their own time due to their busy schedules, only few of them did not return questionnaires that were issued to them.

3.11 Conceptual framework

We adopted Technology Acceptance Model (TAM) as our conceptual framework for analysing data involving different variables (constructs) of our data. This helped us to identify the constructs that influenced research respondents in the use of identity tokens. These identity tokens contained the identity attributes of service applicants and service users in their respective organizations.

The TAM is an information technology framework for understanding users' adoption and use of emerging technologies particularly in the workplace environment and has been tested in older populations [115], [116].

The theory posits that a person's *intent to use* (acceptance of technology) and *usage behavior* (actual use) of a technology is predicated by the person's perceptions of the specific technology's *usefulness* (benefit from using the technology) and *ease of use*. Simply, users are more likely to adopt a new technology with high-quality user experience design (i.e., usable, useful, desirable, and credible) [117].

In the context of the TAM, Davis et al [118] argued that people form attitudes and intentions toward learning to use a novel technology, which are associated with uncertainties, before starting efforts aimed at performing. As an early form of acceptance, intention to use represents a well-established predictor of behavior

This framework was applied on primary data that were collected since the questionnaire was designed to collect data that would be useful for our analysis. Secondary data had no such provision for collection of such data, therefore, primary data were adequate to attend to this part of analysis that we needed.

In our survey for this research, we asked respondents to indicate their perception of the identity tokens that were used in their day-to-day operations in delivering services. There were five constructs as shown in figure 2 above.

Respondents had to choose from a rating scale from the following options: Strongly disagree (score was 1), Disagree (score was 2), Neutral (score was 3), Agree (score was 4), Strongly Agree (score was 5), and Not Applicable (score was 0).

The variables in this analysis were the constructs that included Usefulness, Ease of use, Image (status), Trust, and User satisfaction.

The sources of identification documents (identity tokens) that were considered could be noted from the questionnaire which is on Appendix 3 of this dissertation, these were Banks, Insurance, Government, Universities and Schools, Hospitals, Mobile phones, and Utility Bills.

The data pertaining to how these constructs affected the acceptance of the identical tokens (documents of identity and service application forms) was analysed and the outcomes are reflected in figures 18 and 19 of Chapter 4 of this dissertation. The use, acceptance of these documents, and effect of the constructs are also discussed in Chapter 4. The outcome of statistical analysis are indicated on figures 18 and 19 and the corresponding discussion thereof. We looked at all the variables (constructs) that are involved in this part of the research and considered what the motivations were there on these variables. We subjected the responses from the questionnaire to statistical analysis. Using SPSS, we analysed the

data using statistical methods to analyse and determine which variables were in key in the use of the identity tokens.

3.12 Data collection

According to Faryadi [119], data collection in research is a long process of gathering, measuring and establishing meaning so that you have answers to your questions. It is useful to have a systematic road map for gathering relevant and current data to answer your hypotheses and research questions.

The method of collecting data, analyzing data, and evaluation of data were tailored to respond to the needs of this research. Data collection methods, data collection techniques, data analysis methods and techniques of this research were chosen with a hindsight of the type of this research, research question, and objectives.

3.12.1 Data collection methods

The researcher had to decide which sort of data would be used (thus collecting) for the study and accordingly select the method of data collection. The methods of collecting primary and secondary data differ [120], we therefore, chose the methods according to the purposes of the data and the objectives of this research.

This research involved primary data and secondary data, therefore, the methods of collecting data were those for primary and secondary data. Our research questions and research objectives required primary data collection due to the following reasons:

- We needed to sample from research population and involve respondents who had first-hand experience with identity tokens that had identity attributes.
- We needed data that were original, as originality increased authenticity.
- There was need to formulate specific questions which needed specific responses from questionnaire respondents. These questions had to attend to our research questions and objectives

Our research also involved secondary data because of the following reasons:

- There was data which were already available from documents that had identity attributes which were subject of our research.
- This method of data collection did not need much resources for collection. Costs for collecting and analyzing this data were expected to be lower than what would be spent on primary data.

- We would gather a lot of data in a short space of time to meet the required data of our research.
- Data were already cleaned and stored in electronic form, therefore, much of the time was spent on analysis than collection of data.
- It was easy to gather data to represent the chosen sample.

The primary data and secondary we collected was meant to help us attend to our first objective of our research. We gathered the data which helped us to identify the key identity attributes from the application forms to identify applicants. These application forms were used as the source of identity attributes for various services offered by both in the real space and cyber space. We used secondary data to text mine data from pdf documents that had identity attributes from application forms for required services. This helped us attend to the second objective of our research. We used the results of text mining to attend to the third objective of our research.

3.12.1.1 Primary data

Kathari in [120] defines *primary data* as the data which are collected afresh and for the first time, and thus happen to be original in character.

To collect primary data, a closed and structured questionnaire was developed and distributed to various organizations of our selected research sample. The study considered the major sources of identity attributes currently being used in the application and registration forms for the various services offered both in the cyber and real spaces. The organizations that deal with registration of identity tokens include Banks, Driver license registration organization, Insurance companies, Mobile phone networks, Government Departments, Pensions organizations, Power utility companies, National Tax Registration organization, Universities, Schools, Hospitals, Clinics, Water utility companies, Churches, and Clubs. Selection of organizations for consideration was cast to cover a wide range of organizations that are very active in identity registration.

3.12.1.1.1 Primary data collection approach

To collect primary data, the following steps were followed and factors:

i. A closed questionnaire was prepared which was based on the research question and research objectives,
ii. Each questionnaire was answered by one person who was familiar with opening of an account for service delivery in an organization,
iii. Questionnaires were self-administered,
iv. Respondents answered the questionnaire at their own time,

v. The respondents answered all sections and all questions,
vi. Confidentiality of responses on the questionnaires was upheld,
vii. Organizations to participate in the survey were sampled for the distribution of questionnaires, and
viii. Questionnaires were delivered to organizations together with an introductory letter from the University of Zambia.

3.12.1.1.2 Primary data collection and instrument (questionnaire)

Kathari elucidated in [120] that a questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms. The questionnaire is sent to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own [120].

The questionnaire was developed and a survey was conducted within the sample that was adopted. This questionnaire had four parts, of which the first part was focusing on demographic information; the second part was on perceived importance of identity tokens. The third part of the questionnaire was on the level of importance of identity attributes, and lastly, the fourth part was assessing the frequency of use of identity documents.

3.12.1.1.3 Demographic information

Part one included questions on gender, marital status, age, highest level of education, type of employment of the respondent, and occupation.

We need to see how different groups within the population, would relate with each other with respect to identity attributes of a given entity. Different demographic groups in a population could help us get insights of particular ways of understanding or perceptions on different topics.

3.12.1.1.4 Perception of importance of Identity document

The second part of the questionnaire was considered the perceived importance of the identity tokens that are used in organizations. Respondents from various organizations gave their responses according to their perception. Five constructs were assessed regarding how respondents perceived these identity tokens against these five constructs, respectively. The constructs that were assessed were Usefulness, Ease of use, Image (status), Trust, and User satisfaction. These data were helpful to identify which documents were more important in identification of entities than the others according to the perception of users. The data were also important to identify the perception of different organizations in the level of importance of different constructs, respectively. It was also important to observe how these constructs

relate with each other for us to develop a framework of in determining constructs that would be of great importance in choosing constructs of a digital identity document. The definitions of these constructs were as follows:

Table 2: Description of constructs on perception of identity tokens

Construct	**Meaning of construct**
Usefulness	The degree to which a person believes that using the particular document would help his or her job in identifying an individual/thing.
Ease of use	"The degree to which the document is perceived as being difficult to use"
Image (social status)	"The degree to which use of the document is perceived to enhance one's image or status in one's social system"
Trust (secure)	How would the attributes on the document of the individual/thing being identified enhance trust
User satisfaction	How satisfied with the use of the document

The constructs were being assessed within the following options:

Table 3: Key used in rating the technology constructs

OPTION	MARK
SD = strongly disagree	1
D = Disagree	2
N = Neutral	3
A = Agree	4
SA = Strongly Agree	5
NA= Not Applicable	6

3.12.1.1.5 Identity attributes' level of importance

We found it important and necessary to consider what International Standard Organization (ISO) already has on recognized important identity attributes. It was important that we apply the already established standards, on identity attributes that are known to be important, in our investigations for certain parts of our research. In this case, our interest at this stage was to observe the performance of the standard attributes in our research regarding their importance. Attention was paid to which sources were major for the identity attributes from our chosen sample from secondary data.

Part three had a list of identity attributes that are based on the International Standard Organization (ISO/IEC JTC 1/SC 27) [121]. Users had to consider these attributes based on their identity tokens how they would rank them with respect to their perception of importance. The ranking was from the range of 1 to 5, of which a choice of 1 would be that of a least identity attribute important whilst 5 would be very important.

3.12.1.1.6 Identity Document's frequency of use

Part four considered the use of identity tokens in the respective organizations. This part assessed how long a respondent had been using the document. The range of choices was from the following list:

- Under 1 year
- 1-2 years
- 3-4 years and
- More than 4 years

The other consideration was to assess how often in a week a respondent used the document; the options were as follows:

- Not at all,
- Once a week,
- 2-3 times,
- More than 3 times.

3.12.1.2 Secondary data

According to Green [122] secondary data were data which the researcher did not collect for him/herself directly from respondents or subjects. This means that secondary data were not collected with the researcher's purpose and objectives in mind. It may have been collected by institutions, whose job is to collect data (e.g. government or regional offices of statistics and information, international bodies whose purpose is information collection).

Secondary data were gathered from the pdf application forms that had been placed online by organizations for customers to print and fill in their identity particulars for registration of accounts for application of services. The pdf documents had text for the identification of respective applicants for the requested services or accounts for sampled institutions, and regions.

This was done by searching for these enrolment application forms on internet from the areas of interest that were sampled.

3.12.1.2.1 Secondary data collection framework

We followed the following steps to collect secondary data:

i. Countries and regions were sampled using clustering and systematic sampling techniques,
ii. Sampled countries and regions had to be English speaking,
iii. The pdf documents needed to be able to be text mined. Documents that were not able to be text mined would not have the text for identity attributes separated for statistical analysis,
iv. If document was in MS Word it needed to be converted to pdf for readiness for text mining as the required by the text mining tool,
v. The pdf document needed to be converted to text file as the text mining tool could only deal with pdf documents for text mining,
vi. The names of the electronic files needed not to be long for ease of processing on the computer,
vii. Files needed to be saved in the folders on the computer according to their clusters for ease of processing according to respective clusters.
viii. Use text mining tool to mine the text from the text files

3.12.1.2.2 Secondary data collection, techniques, and tools

A research population was established by considering only the regions and countries that use English in their official functioning of their governments, since the researcher was English speaking. Lists of countries from these regions were listed, respectively, random sampling and clustering techniques were employed to establish samples from internet lists. To avoid any bias, countries that occupied odd number positions on the lists were selected. From the regions of the world that were picked, countries were selected following the same method as indicated above. The resultant regions and countries included the following: African Countries, Common Market for Eastern and Southern African countries (COMESA) countries, Southern African Development Community (SADC) countries, European countries, Asian countries, United States of America, Australia, Canada, and New Zealand. The sampling was done at random from listed countries of respective given regions. Table below gives the list of countries and regions that had their data extracted.

Table 4: Sampled Regions and countries for secondary data

AFRICA (general)	COMESA	SADC	ASIA	EUROPE	OTHER REGIONS
Botswana	Ethiopia	Botswana	Bahrain	Cyprus	Australia
Ethiopia	Kenya	Lesotho	Brunei	Gibraltar	Canada
Ghana	Malawi	Malawi	Burma	Iceland	New Zealand
Lesotho	**Rwanda**	**Mauritius**	**India**	**Ireland**	**United States of America**
Libya	Swaziland	Namibia	Israel	Malta	
Mauritius	Uganda	Seychelles	Malaysia	Monaco	

Namibia	Zambia	Tanzania	Palestine	United Kingdoms	
Rwanda	Zimbabwe	Zambia	Philippines		
Seychelles		Zimbabwe	Singapore		
Uganda			Thailand		
Zambia					

There was need to consider the identity attributes from different regions of the world, therefore, literature for International Standard Organization was consulted to identify attributes that are recognized as standard in the enrolment of diverse online services. Therefore the already identified attributes by International world standards, ISO/IEC JTC 1/SC 27, were considered and used in this research. These standards have identified a list of attributes that could be collected from individuals during the time of enrolment of individuals; we learn from [121] that entities from identity tokens can be validated during Identity Proofing, Identity Information Verification and Verification regarding. A list of attributes that ISO/IEC JTC 1/SC 27 indicates as elements that would form identifiers to identify an individual include:

Table 5: A list of standard attributes based on ISO

ATTRIBUTES (ISO/IEC JTC 1/SC 27)	
First name	ID Number
Middle name	Issuing authority
Last name	Expiry date
Date of Birth	Home email address
Place of Birth	Work address
Race	Work telephone number
Gender	Work email address
Home address	Bank account details
Home Unique Property Reference Number (House Number)	Height
Home telephone number	

Documents in pdf format in the four organizations namely Banks, National Governments, Insurance, and Universities and Schools were collected from internet from the institutions of these respective organizations in the sampled countries.

The identity documents and applications documents for consideration are documents that fell in the category of the international standards of ISO/IEC 29003:2013 [121]. A list of such documents are listed as standards.

Secondary data for this research had to be gathered from PDFs of identity tokens, which are used to collect personal data of individuals during identity enrolments of respective services.

Areas of interest for the research were areas which are active in enrolment of individuals for different services in different organizations. Collection of data, as indicated above was done from the internet by searching using the search engines from internet web browser. The browsers that were used in searching for the pdf documents of the account or customer service application forms were Google Chrome and Mozilla Firefox; Google search engine was the search engine that was used in searching for the documents. The data collection was done from computers which had pdf writers.

Documents were searched from the internet satisfying different queries at different times. A search query would include a given country and given sampled organization at a given time. The organizations included Banks, Governments, Insurance, and Universities and Schools. The folders that were containing these organizations that were sampled were placed in a folder of a country that was under consideration, respectively.

There was need to separate pdf files for this research so that there was good management of data for this research. Folders were created on the computer to represent different sampled regions, respectively. These regions included Africa, COMESA, SADC, Asia, Europe, Australia, Canada, New Zealand, and United States of America. The folders for these regions contained folders for countries that were randomly sampled from the regions mentioned above, respectively. These folders for the countries, in turn, contained folders containing pdf files for sampled organizations for this research. These pdf files were harvested from internet after searching for the areas of the organizations that were of interest in this research.

The collected data were to be subjected to online identification for online services; these are the attributes that are used as identifiers of individuals who wish to enroll for some specific online services. Attention was to be paid in the ranking of the commonly used attributes in the assessment of the performance of the attributes during the research. The hierarchy of importance as perceived from different parts of the world was closely observed. The same organizations were maintained for consideration in the nominated countries from different geographical areas, which were sampled. Collection of pdf documents and conversion of the same to text files was then followed by text mining. This was for the extraction of the key attributes, based on the established International world standards, from the online documents.

3.13 Data analysis

Data analysis refers to the computation of certain measures along with searching for patterns of relationship that exist among data-groups. In the process of analysis, relationships or

differences supporting or conflicting with original or new hypotheses should be subjected to statistical tests of significance. This is to determine with what validity data can be said to indicate any conclusions [123].

Analysis of data in a general way involves a number of closely related operations which are performed with the purpose of summarizing the collected data. The analysis should organize these in such a manner that they answer the research question(s) [120].

Data analysis for primary data and secondary data focused on addressing the research objectives and the research questions. Attending to these issues was the crux of our research. To keep an eye on these elements, it was worth restating them at this point once more. The objectives of our research were as follows:

1. To identify the key identity attributes from the application forms used as the source of identity attributes for the various services offered both in the real space and cyber space
2. To mine the key identity attributes from the sources in (1) using data mining tools and techniques
3. To use mathematical models based on distance metrics to quantify the identity attributes.

The research questions that we were addressing were as follows:

1. What are the major sources of identity attributes currently being used in the application and registration forms for the various services offered both in the cyber and real space?
2. How can we extract the key identity attributes from the application and registration forms for the various services offered both in the cyber and real space using data mining techniques?
3. Based on distance metrics mathematical models, how do we develop the identity attribute metrics model used to quantify the identity attributes?

3.13.1 Data analysis methods

Since we had primary data and secondary data, we analyzed primary data and secondary data separately.

In general terms, to analyze data we used statistical analysis (including descriptive statistics and inferential statistics in our primary data), Shannon Information Entropy, and text analysis at appropriate stages of the data analysis, respectively. We used statistical analysis to summarize and compare the data variables and observe visual presentation; we used Shannon Information Entropy to evaluate the amount of information and level of importance

of the text in the identity attributes, that was contained in the data variables that were under consideration. We used text analysis to analyze the text that was mined from pdf documents for our secondary data; text frequencies were observed and coupled with some text-weight schemes, we were able to rate the level of importance of identity attributes by using Shannon's Information Entropy. Our proposed model was able to determine the uniqueness, quantify the identity attributes, and determine the level of importance of identity attributes that were involved in the data that were collected.

3.13.1.1 Statistical analysis

According to [124], statistical Analysis shows "What happened?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data; it analyses a set of data or a sample of data. There are two categories of this type of analysis - Descriptive analysis and inferential analysis.

Under descriptive statistics, we used graphical methods and numerical methods. Numerical methods that were applied were measures of central tendency like mean, frequency and, percentages; measures of variability that we applied were Standard deviation whilst Normal distribution method that we used in this research was standardized scores (also known as z-scores).

It is stated in [125] that inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. The goal of inferential statistics is to draw conclusions from a sample and generalize them to a population.

Regardless of the specific test, an identical process is used for all inferential statistical tests:

- Select a sample representative of the population.
- Calculate the appropriate test statistic.
- Test the Null hypothesis.

The statistics that we used in our research that are inferential in nature include ANOVA, Chi-square, and correlations. According to [126], *ANOVA (analysis of variance)* is an inferential statistic used to determine whether the value of a single variable differs significantly among three or more levels of a factor. *Chi-square* is an inferential statistic used with qualitative data to determine if differences between frequency distributions are statistically significant. *Correlation* is an inferential statistical test that is used to determine whether there is a statistically significant connection, or a relationship, between two variables.

The study in [127] indicates that Chi-square test is a nonparametric test used for two specific purpose: (a) to test the hypothesis of no association between two or more groups, population

or criteria (i.e. to check independence between two variables) and (b) to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). Chi-square is used to analyze categorical data.

3.13.1.1.1 Primary data analysis

Statistical analysis was used as a technique for data analysis using descriptive statistics and inferential statistics. Descriptive statistics tries to describe the relationship between variables in a sample or population. The study in [128] mentions that descriptive statistics provide a summary of data in the form of mean, median and mode. Inferential statistics use a random sample of data taken from a population to describe and make inferences about the whole population. It is valuable when it is not possible to examine each member of an entire population.

3.13.1.1.2 Analysis tools

In our approach, we used computer software called Statistical Package for Social Sciences (SPSS) and Microsoft Excel as data analyse tools.

Shannon Information Entropy was used to identify the identity attributes that were key in identifying entities.

We used the Shannon Entropy method to analyse both the primary data to determine the level of importance of identity attributes as perceived by respondents in the primary data.

3.13.1.2 Data analysis using Shannon's Entropy

We used Shannon Information Entropy to determine the level of importance of identity attributes of service applicants from text mined service application forms for the secondary data. The computations that were undertaken for this method of analysis are covered under section 4.4 of this research dissertation.

An assertion was made in [75] that the entropy of a variable is the "amount of information" contained in the variable. Shannon's entropy quantifies the amount of information in a variable, thus providing the foundation for a theory around the notion of information.

A study in [129] shows that the entropy weight method, grounded on Shannon's entropy theory, was applied for the calculation of the objective weights or to apply a data analysis. The particular purpose of this approach is to obtain a ranking by means of a weighted sum

3.13.1.3 Text analysis

Text Analytics is the discovery of new, previously unknown information, by automatically extracting information from different written resources [130]. The studies suggest in [130] suggest that Text Analytics is an extension of data mining, that tries to find textual patterns

from large non-structured sources, as opposed to data stored in relational databases. Text Analytics, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting non-trivial information and knowledge from unstructured text. It indicated in [130] that Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc.

We utilized Text analytics to analyse data that was text mined from the pdf documents that were extracted and collected, in our research sample, from the World Wide Web as secondary data. The text mining tool had inherent algorithm for clustering of data from the text mined data. Documents were clustered according to the respective queries on specific texts that were made, this was the text of identity attributes from identity tokens.

Research in [130] shows that clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given document to a given category, clustering is unsupervised learning: there are no previously defined topics or categories. We further observe in [130] that a basic clustering algorithm creates a vector of topics for each document and assigns the document to a given topic cluster.

Studies show that clustering technique can be used for corpus summarization by providing coherent summary of the collection in the form of word cluster [131], [132].

3.14 Identity Attribute text mining

This part of the research helped us to address the need of the second objective of our research. In our second objective of this research, we mined the key identity attributes from the application forms as a source of identity attributes for the various services offered in the real space and cyber space.

The term text mining is a special form of data mining and is referred to as "*text data mining*" or "*text analytics*". The text data mining refers to computer-based methods for the semantic analysis of texts, which support the automatic or semi-automatic structuring of texts, especially very large amounts of texts [133].

Text data mining aims to identify and extract knowledge that is implicit in the text that the user of the information system does not know [134].

Text mining process is the same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-

structured data sets such as emails, HTML files, and full text documents etc. Structured data is data that reside in a fixed field within a record or file; these data are contained in relational databases and spreadsheets. The unstructured data usually refers to data that do not reside in a traditional row-column database and it is the opposite of structured data. Semi-Structured data is the data that are neither raw nor typed in a conventional database system [135].

3.15 Text mining tools and text mining process

3.15.1 Text mining tools

The Pdf documents were converted into text files respectively using a pdf converter tool, TalkHelper PDF Converter version 2.2.9.0. This was followed by text mining of these text files using an open source AntConc 3.5.8, a Natural Language Processing text mining and analysis toolkit. Since AntConc 3.5.8 can only read text files, we had to convert the pdf files to text files which can be read in the Notepad software. The text files were then loaded into AntConc 3.5.8 text mining tool when there were required for text mining process and text analytics. It is stated in [136] that AntConc 3.5.8 is a freeware corpus analysis toolkit for creating concordance and text analysis. It runs on any computer running Microsoft Windows (tested on Win 98/Me/2000/NT, XP, Vista, Win 7), Macintosh OS X (tested on the versions of 10.4.x, 10.5.x, 10.6.x), and Linux (tested on Ubuntu 10, Linux Mint). It is developed in Perl using various compilers to generate executable software for the different operating systems.

AntConc 3.5.8, the text mining tool was used to get the text frequency of the corpus files that were gathered from different organizations and regions. The gathering of data and text mining of this data was done as discussed above. Data were imported into the tool for processing from respective folders which were holding these files.

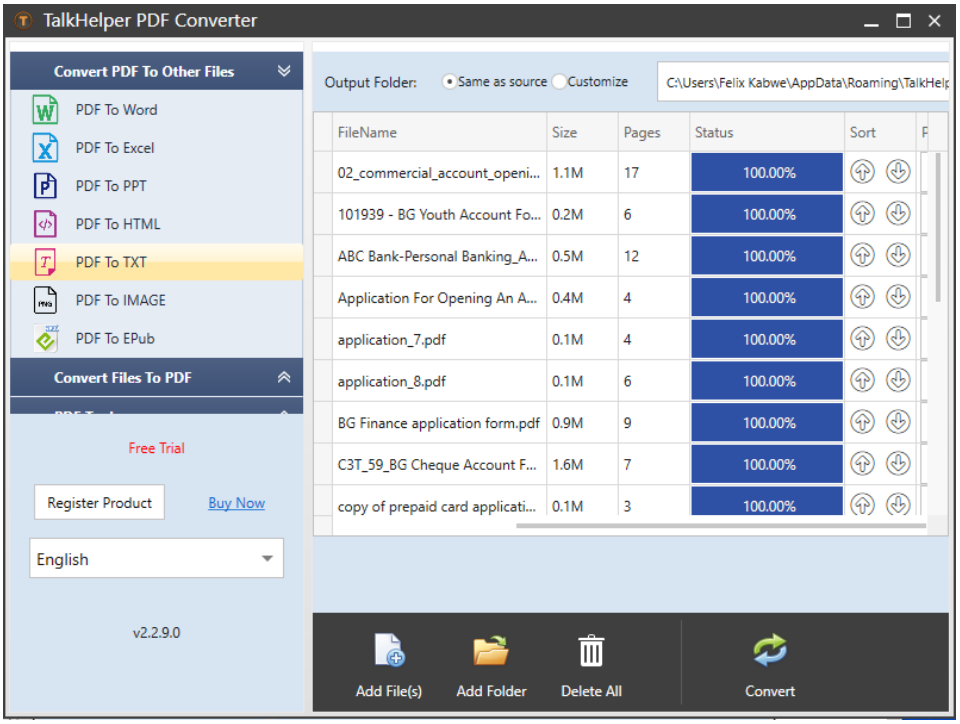

Figure 123: TalkHelper PDF Converter version 2.2.9.0

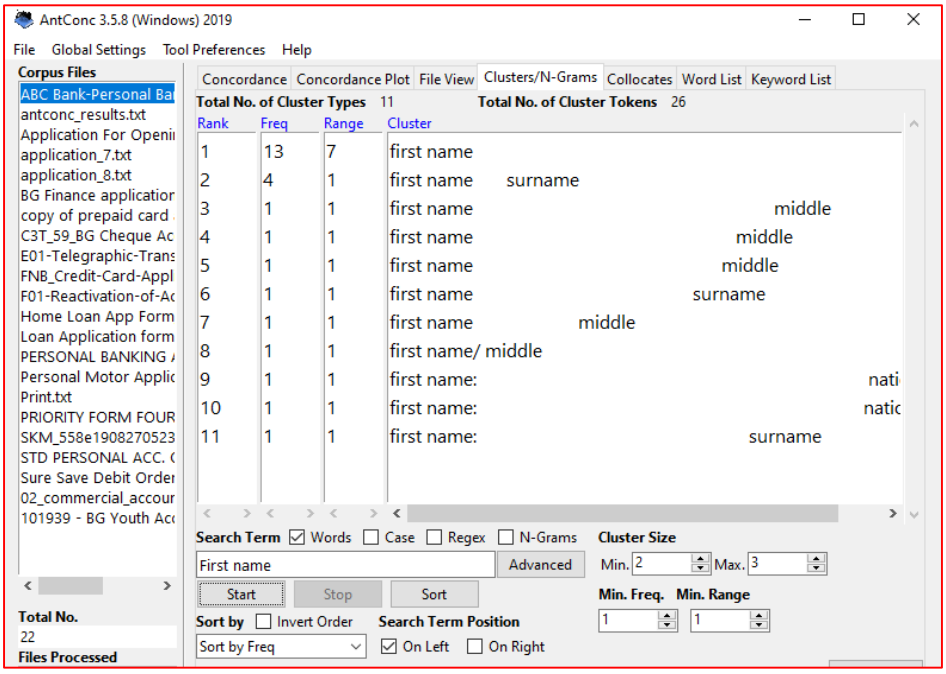

Figure 134: AntConc 3.5.8, a corpus analysis toolkit for data mining

3.15.2 Text mining process

There are five steps under text mining process: Data collection, Data pre-processing, and Data transformation, Data analysis and Result evaluation [137].

3.15.3 Document gathering

In the first step, the text documents are collected which are present in different formats. The document might be in form of pdf, word, html doc, css etc. [137].

As indicated above, text documents from different sources, containing data, were collected; the documents were from different unstructured data of different formats including pdf, MS Word, HTML documents, etc. Data collection was followed by Data pre-processing; during this process documents that were not in pdf were converted to pdf before a text mining tool was used. Documents that were in image form were eliminated as these could not have text on the document separated or extracted for text mining. Documents that could not be converted to text files were equally eliminated from the corpus. Data preprocessing was meant to prepare documents for text mining before analysis of text was done. The AntConc 3.5.8 text mining tool is designed to carry out the stages of Data pre-processing (partially), Data transformation, Data analysis and Result evaluation of the text mining process. AntConc 3.5.8 [138] is widely used tool in linguistics and corpus linguistics. AntConc 3.5.8 has a dedicated keyword-extraction module, that extracts keywords (consisting of one word). AntConc 3.5.8 can be used to extract single-word and multi-word terms by using the "Word List" and "N-Grams" modules respectively, which list the words and multi-word expressions sorted by the frequency of occurrence in the corpus. AntConc 3.5.8 evaluates and reflects the role of pure frequency for term extraction.

3.1.5.4 Document pre-processing

In this step the collected data were preprocessed for removing redundancies, inconsistencies, separate words and stemming. In the tokenization, the data were divided into single words i.e. tokens [137]. In this process, the given input document was processed for removing redundancies, inconsistencies, separate words, stemming and documents were prepared for next step, the stages performed included tokenization, removal of stop word, and stemming [137]. With tokenization, the given document was considered as a string and identified single word, character, or phrase, in a document i.e. the given document string was divided into one unit or token. Removal of Stop word removed usual words like a, an, but, and, of, the etc. Stemming described the base of particular word. Inflectional and derivational stemming were two types of methods. One of the popular algorithms for stemming was porter's algorithm, e.g. if a document pertained word like resignation, resigned, resigns then it would be considered as resigned after applying stemming method [137].

AntConc 3.5.8 [138] used the Concordance tool to allow the researcher to see how words and phrases were commonly used in a corpus of texts. It removed redundancies, inconsistencies, separated words, stemming and documents were prepared for next step. The Key word List tool of the AntConc 3.5.8 allowed us to identify characteristic words in the corpus, for example, as part of a genre.

Data transformation was the next phase in the text mining process; data transformation was meant to convert text document into the bag of words or vector space document model notation, which was used for further effective analysis. In feature extraction, the useful meaning words were extracted from the document whilst in feature selection, relevant words were selected. There were two methods in feature selection i.e. filtering and wrapping methods [137].

AntConc 3.5.8 converted text documents into a bag of words which were easily identified, ordered and listed. Useful meaning words which were relevant were extracted from the documents.

3.16 Text mining techniques

3.16.1 Vector Space Model

We would need a standard to determine the closeness, or to measure the distance or similarity between a pair of objects, an object and a cluster, or a pair of clusters. A data object is described by a set of features, usually represented as a multidimensional vector [139].

Vector space model (VSM) [140], [141] is a model based on a vector space, which represents information objects (e.g., terms, images, documents, queries, etc.) by vectors in a vector space. We can think of a vector space in general, as a collection of objects that behave as vectors do in multi dimension real space ($\boldsymbol{R}^n$) [142], where n is the dimension of the space (n = 1, 2, 3, …,∞). The objects of such a set are called vectors.

Each dimension of a vector space represents a feature of an information object, corresponding to a basis element of a vector space of the VSM [143].

For the weighted information-object vectors, distance functions are often used to determine how to measure the similarity between information-object vectors [144].

One common similarity measure between two information-object vectors is the cosine similarity, measuring the cosine of the angle between two information-object vectors in a vector space of the VSM [145].

3.16.2 Statistical Methods

In the work of [146] it was mentioned that most of text mining tools use statistical methods in conjunction with other methods.

We observe from [138] that the Keyword List tool in AntConc 3.5.8 used statistical methods like Chi-Squared and Log-Likelihood. The Collocate tool in said text mining tool ordered words by the value of a statistical measure. The collocates were ordered either by total frequency, frequency on the left, frequency on right of the search term, or the start or end of the word. The Clusters tool in AntConc 3.5.8 ordered the word clusters by frequency, the start or end of the word, the range of the cluster (number of files in which the cluster appears), or the probability of the first word in the cluster preceding the remaining words

3.16.3 Text Clustering

Research in [147] indicates that text clustering involves document representation, document similarity measure and clustering techniques. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity)

As indicated in [138], the Cluster tool in AntConc 3.5.8 allows to search for a word or pattern and group (cluster) the results together with the words. The clusters can be ordered by frequency, the start or end of the word, the range of the cluster (number of files in which the cluster appears), or the probability of the first word in the cluster preceding the remaining words. The Cluster tool in AntConc 3.5.8 produces a set of clusters from a corpus.

3.17 Data analysis and interpretation

Secondary data from a given corpus were standardized by normalization, anlysed using statistical methods, and interpreted. A weighting scheme, of inverse document frequency (IDF) was used on the term frequencies (TF) from identity tokens to remove noise (anomalies) in the data to improve accuracy of data. Data were further analysed to determine the weighting of the text that was mined from a corpus of identity tokens, and then interpreted. The purpose of analyzing the text weighting of identity attributes was to identify the identity attributes were key from the identity tokens. Shannon information entropy was also used in the analysis of text of the identity attributes. This was used to evaluate the level of importance of identity attributes and the quantification of information in the text corpus of the identity attributes.

3.18 Identity attributes Quantification

3.18.1 Proposed model

The third objective of our research required us to *develop* the identity attribute metrics model which could be used to quantify the identity attributes of an entity. This model is based on the distance metrics mathematical model.

We considered the word "*develop***"** in the English dictionaries, the following were the meanings we adopted in our research:

1. To make available or usable of something [148].
2. Change the use of and make available or usable of something [149].
3. To bring out the capabilities or possibilities of; to evolve [150].
4. Convert to a new purpose by making other use of the object's resources [151].

Our application of the word "develop" was based on the definitions above; we considered the word *development* as the making available or usable of an existing mathematical model which had capabilities of quantification of identity attributes and would be applied in options that were available. In other words, we utilized an existing mathematical model in purposes it would not have been much applied or used; this model was to meet our objective.

The studies in [152], [153] indicated that vector space model (VSM) is a mathematical model which represents information objects (e.g., terms, images, documents, queries, etc.) by vectors in a vector space. Each dimension of a vector space represents a feature of an information object, corresponding to a basis element of a vector space of the VSM [154]. For the weighted information-object vectors, distance functions are often used to determine how to measure the similarity between information-object vectors [155]. One common similarity measure between two information-object vectors is the cosine similarity, measuring the cosine of an angle between two information-object vectors in a vector space of the VSM [156]. In the vector space model after representing the documents as a vector, we can find out the similarity of documents with each other by measuring the angle between two vectors [157].

Clustered documents form subgroups on the basis of an often subjectively chosen measure of similarity [158]. It is mentioned in [159] that clustering algorithms require a metric to quantify how different two given documents are. This difference is often measured by some distance measure such as Euclidean distance and Cosine similarity.

Documents that cluster together (are very similar to each other) will have a similar relevance profile for a given query. Much research work has been focusing on browsing of the World Wide Web and search of documents [160], [161].

In the context of document classification and clustering, there has been numerous researches on the effectiveness of different similarity measures [162]. Subhashini *et. al*. [163] evaluated the clustering performance of different similarity measures on three web document collections. Cosine similarity performed better than Euclidean similarity distance and Jaccard similarity measure in this evaluation.

Singhal in [161] indicated that Cosine similarity is most commonly used in high-dimensional positive spaces. In information retrieval, a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter. We saw in [164] that Cosine similarity compares two documents with respect to the angle between their vectors. Document vectors for similar documents generally point in the same direction [165].

Several recent studies [166], [167], [168], [169] have showed how Cosine similarity has extensively been used in matching and retrieving text documents and web pages. This showed how effective and undoubtedly had the focus of Cosine similarity measure been in *document* retrieval from online information internet.

The study in [170] shows that the Cosine Similarity Measure could be used to compare documents or give a ranking of documents, with respect to a given term of query words. The cosine similarity measure was superior to the other measures such as Jaccard measure and Euclidean measure for text documents.

A study in [171] showed that Cosine Similarity was a popular and commonly used similarity measure. This could be derived directly from Euclidean distance, however, Euclidean distance was generally not a desirable metric for high-dimensional data mining applications. Cosine similarity is one of the most popular text similarity measures. Cosine similarity is most commonly used in high-dimensional positive spaces.

In research from [172], we learnt that Clustering algorithms required a metric to quantify how different two given documents were. This difference was often measured by some distance measure such as Euclidean distance and Cosine similarity

Cosine similarity measure is a mathematical model which is based on a distance metric which can be used to quantify the identity attributes. Cosine similarity measure is not new

but has largely been applied in data mining for clustering of text documents to retrieve similar documents.

Our interest is to avail and use this similarity measure of retrieval capabilities of text from documents. We would want to utilize these capabilities and redirect this distance metrics resource into text retrieval capabilities for the quantification of the identity attributes. Our focus in our research will be on text clustering than document clustering using Cosine similarity measure.

In text mining, a document is represented as a vector in which each component indicates the value of its corresponding feature in the document [173]. We learn in [174] that text similarity aims to find the commonality existing among text documents. It was argued in [175] that Cosine similarity was a more favourable distance measure in text mining than Euclidean, Manhattan, and Jaccard Index. It was found that Cosine similarity handles both continuous and categorical variables. Cosine similarity measure is useful when finding the similarity between two text documents whose attributes are word frequencies [176]. In the research [177], it was found that Cosine similarity measure outperformed Euclidean distance in high dimensional space, while [176] suggests that Euclidean distance is only appropriate for data measured on the same scale.

It was stated in [178] that Vector space model (VSM), helps us to convert the original string text within a document into a vector of numbers. In VSM, each document is considered as a vector in a vector space. Assume $D = \{d_1, d_2, ..., d_n\}$ is a data set that has n number of documents and that $T = \{t_1, t_2, ..., t_n\}$ is a set of distinct terms, which occurs in D. Then the It was indicated in [178] that vector representation of document D is defined as:

$$v_d = \{tf(t_1, d), tf(t_2, d), (t_3, d), \dots tf(t_n, d)\} \quad (3.1)$$

where $tf(t,d)$ denote the frequency of term $t \in T$ in document $d \in D$.

From [178], we find that the vector representation of two documents d_1 and d_2 is as follow:

$$v_{d_1} = \{tf(t_1, d_1), tf(t_2, d_1), tf(t_3, d_1), \dots tf(t_n, d_1)\} \quad (3.2)$$

$$v_{d_2} = \{tf(t_1, d_2), tf(t_2, d_2), tf(t_3, d_2), \dots tf(t_n, d_2)\} \quad (3.3)$$

where $tf(t_n, d1)$ denote the frequency of the term $t_n \in T$ in document d_1 and $tf(t_n, d_2)$ denote the frequency of the term $t_n \in T$ in document, d_2 [178].

In the vector space model after representing the documents as a vector, we can find out the similarity of documents with each other by measuring the angle between two vectors [179].

According to [178], Cosine similarity is an angle based measurement. It calculates the cosine of the angle between two vectors and helps us to find out how related two documents are. The cosine similarity of A and B is defined as:

$$\cos\theta = \frac{A.B}{\|A\|\|B\|} \quad (3.4)$$

or

$$Cos_{sim}(d_1, d_2) = \frac{d_1.d_2}{\|d_1\|\|d_2\|} \quad (3.5)$$

$$d_1.d_2 = [tf(t_1, d_1)*tf(t_1, d_2)] + [tf(t_2, d_1)*tf(t_2, d_2)] + \ldots + tf(t_n, d_1) * tf(t_n, d_2)] \quad (3.6)$$

$$\|d_1\| = \sqrt{tf(t_1, d_1)^2 + tf(t_2, d_1)^2 + \cdots + tf(t_n, d_1)^2} \quad (3.7)$$

$$\|d_2\| = \sqrt{tf(t_1, d_2)^2 + tf(t_2, d_2)^2 + \cdots + tf(t_n, d_2)^2} \quad (3.8)$$

The cosine value varies between [-1, 1]. If documents are similar, their vectors will be in the same direction from origin, thus, they form a relatively small angle, which its cosine value will be near 1. On the other hand, when two vectors are different direction from origin, they form a wide angle and the value of the cosine is near to -1, consequently, the documents are dissimilar, and they map no similarity [178].

Given two documents $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ with the angle between them in vector space beingθ, the Cosine similarity measure between $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ would be represented by:

Similarity $= S(\boldsymbol{d_1}, \boldsymbol{d_2}) = Cos\,\theta = Cos\,(\boldsymbol{d_1}, \boldsymbol{d_2}) = \frac{\boldsymbol{d_1}.\boldsymbol{d_2}}{\|\boldsymbol{d_1}\|\|\boldsymbol{d_2}\|}$ [180]

For n-dimensional vectors, this could be generalized as

Similarity $= S(\boldsymbol{d_1}, \boldsymbol{d_2}) = Cos\,\theta = \frac{\sum_{i=1}^{n} \boldsymbol{d}_{1,i}\boldsymbol{d}_{2,i}}{\sqrt{\sum_{i=1}^{n} \boldsymbol{d}_{1,i}^2}\sqrt{\sum_{i=1}^{n} \boldsymbol{d}_{2,i}^2}}$ [180]

The proposed model Identity Attribute Metric Model based on the Distance Metrics in this research is the Cosine Similarity measure presented above. This mathematical model would be used as an identity attribute model to quantify the identity attributes. It will be used for the identifying of a unique service applicant or unique identity ownership.

In summary, we could cultivate our model based on outcomes of research work from diverse researchers alluded to in this section of our research. The following were the points that made us identify Cosine Similarity measure as the best candidate in coming up with a model which we could apply in this research:

- Cosine similarity measure is a mathematical model that is one of the most popular similarity measures applied to text documents in information retrieval applications and clustering.

- This similarity measure can be used in text similarity with the aim of finding the commonality existing among text documents and the text within the documents.
- Cosine similarity measure is a real-valued function that quantifies the similarity between two objects
- It is suitable to be used for quantification of identity attributes utilizing its capabilities as the cosine of the angle between vectors as the cosine similarity.
- It is useful when finding the similarity between two text documents whose attributes are word frequencies
- Cosine similarity has extensively been used in matching and retrieving text documents and web pages. This showed how effective and undoubtedly had the focus of Cosine similarity measure been in *document* retrieval from online information internet.
- Cosine Similarity Measure could be used to compare documents or give a ranking of documents, with respect to a given term of query words. The cosine similarity measure was superior, in text mining than and in high dimensional space, to the other measures such as Jaccard measure, Manhattan measure, and Euclidean measure for text documents.
- It is a more favourable distance measure in text mining than similarity measures like Euclidean, Manhattan, and Jaccard Index.
- Cosine similarity measure outperformed Euclidean distance in high dimensional space. Euclidean distance is only appropriate for data measured on the same scale.
- Cosine similarity measure is useful when finding the similarity between two text documents whose attributes are word frequencies
- Cosine similarity handles both continuous and categorical variables.
- The measure can compare a document to a cluster, by calculating the cosine between the document vector and the cluster's centroid vector
- This similarity measure measures the cosine of the angle between two vectors projected in a multi-dimensional plane
- Cosine similarity is accurate using tf-idf vector scheme.

We will take advantage of the widely application in document retrieval efficiency of Cosine similarity measure to direct our attention on how it could be used in information security, particularly in Digital Identity Management. In this case, our consideration is how we could use it in the quantification of identity attributes of a digital identity of an entity.

3.18.2 Model quantification

The computations from Cosine similarity distance metric gave us outputs, as numbers, from the closed interval [0, 1] [95]. We were interested to observe outcomes of our computations when two documents (entities) were similar and when they were dissimilar, based on the outputs of the quantifications. Two documents would be similar when the angle of orientation between the two vectors of the two documents was 0° (zero degrees), this was when the similarity measure of the two documents was 1 i.e. *Cos* $0^{\circ} = 1$. On the other hand, when documents were not similar, the extreme measure would be when the two documents were diametrically opposed. In a logical sense, the extreme circumstance would be when one document had text and the other one being compared to, was blank. This would be the extreme case of comparison of two documents hence being diametrically opposed. An angle of diametrically opposed to 0° would be 180°, whose cosine measure would be *Cos* $180^{\circ} = -1$. Therefore, similarity of two documents could be fully represented between 0° and 180°, i.e. in the first and second quadrants of the Cartesian plane. However, term weights of a document could never be negative, therefore, cosine similarity measure which was negative would not be of use in this instance. This ruled out the consideration of the second quadrant, and therefore, left us only with the first quadrant of the Cartesian plane, i.e. between 0° and 90°, in our application (this gives Cosine similarity range of zero to one, [0, 1]).

The study in [59] suggests that cosine similarity measure is a measure between two vectors by measuring the cosine of the angle between them. As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer, meaning that the similarity of whatever is represented by the vectors increases.

In addition to the points that played a role in identifying Cosine Similarity measure as the best candidate in coming up with a model, the following factors were pinnacle in the consideration and quantification of the identity attributes:

1. For verification of ownership
    a. When we are specifically interested in attending to one applicant for verification of ownership claim of a particular digital identity with known identity attributes.
    b. When multiple applicants make claims of ownership claims of a particular digital identity with known identity attributes and we need to verify.
2. The principle of orientation of two similar vectors in a metric space that is inherent with the cosine Similarity distance.

Cosine Similarity measure is used in data mining as a technique to compare two documents that are similar based on the text that these documents contained. This metric is used in considering those who share same tags on a blog, persons who viewed same documents, customers who bought similar items online.

We used the data that we collected to compute, evaluate, and quantify the uniqueness of any given two points in a vector space and ascertain their uniqueness.

3.18.3 Identity verification

Verifying online identity for claimants could either be that of an individual applicant or for multiple applicants. The metrics and mathematical computations would still be the same. For the sake of checking the metrics, we considered applications of multiple applicants in different organizations. We shall therefore, use sampled data from Zambian organizations in this research to test our model, i.e. the Cosine Similarity measure that has been elaborated above. We checked if this model could be used in the verification process of an applicant or applicants in the Digital Identity Management System.

We were keen to observe which identity attributes were key in identifying an entity that was under consideration. We were interested to consider this phenomenon from different regions that were sampled. A comparison of the important identity attributes which were observed from primary data and secondary data was of particular interest. Our research placed great importance in verifying if our model could show the uniqueness of the identify attributes that could identify an entity online. In other words we were interested to establish the identity ownership of an entity.

3.18.4 Testing the model

A record of the term frequencies of respective attributes from the tokens was made per subject organization, country and region. Each attribute had its term frequency recorded as indicated, from corpus analysis toolkit. Frequencies for each respective identity attribute from respective organizations in this research were recorded for further analysis. Examples of the tables that have these records include Table 6, Table 8, Table 14, Table 15, Table 18, Table 21, and Table 24. Since the data collected were quite enormous from the countries and regions that were mentioned, for the sake of computations, we used a representative country to show calculations and quantify the identity attributes as a sample for the other countries and regions. In this case, we used Zambia, the residence of the researcher in our computations.

To test the proposed model, we got a set of documents at random from our selected area, Zambia. We picked ten (10) documents from one of the organizations in our research, from

departments' category, of the Government of the Republic of Zambia. Our model had focused on identifying the set of attributes that would identify a claimant of a digital identity that sufficiently matched the entity for identification. Matching of a claimant could be done on one claimant or multiple claimants. In simple terms from our documents, if one document was an entity that owned the digital identity which was being claimed by the claimant, we could compare the attributes of digital identity of this entity to those of the claimant. For the purposes of this research, the ten documents would suffice, of which one was the object and the nine others were claimants of the digital identity. We subjected all the ten documents to be claimants of the object.

As indicated, ten (10) documents were picked from a set of documents from the Government of Zambia. These included:

1. Airspace application form
2. Residence Permit application form
3. Visiting Visa application form
4. Consent Form application form
5. Farm small holding application form
6. Residential Land acquisition application form
7. Aquaculture Fund application form
8. Borehole Form application form
9. Health Professional Council membership application form
10. Immovable Property application form

Documents in PDF format from the corpus of the Government of the Republic of Zambia documents were searched and harvested from the internet. Text mining was done on these documents using the same techniques as discussed above, based on the nineteen (19) existing attributes that we were used above. The following table shows the term frequencies (*tf*) of each of the respective attributes after the text mining.

*Table 6: Term frequencies of ten documents for computing TF*IDF Weighting*

ATTRIBUTE	Term (tf_i)									
	Airspace (d_1): tf_1	Residence Permit (d_2): tf_2	Visiting Visa (d_3): tf_3	Consent form (d_4): tf_4	FarmSmallholding (d_5): tf_5	Residential land (d_6): tf_6	Aquaculture Fund (d_7): tf_7	Borehole Form (d_8): tf_8	Health Prof Council (d_9): tf_9	Immovable Property (d_{10}): tf_{10}
First name	5	5	4	2	4	4	2	3	1	2

Middle name	5	5	4	2	4	4	2	3	1	2
Last name	5	5	4	2	4	4	2	3	1	2
Date of Birth	0	5	2	0	0	0	0	0	1	0
Place of Birth	0	3	3	0	0	0	0	0	0	0
Race	0	0	0	0	0	0	0	0	0	0
Gender	0	3	3	0	2	2	0	0	1	0
Home address	1	1	2	3	4	1	3	0	1	3
House Number	1	0	0	0	0	0	0	0	0	1
Home telephone number	1	0	0	1	1	1	1	0	0	0
ID Number	1	1	1	0	1	1	1	0	1	0
issuing authority	1	0	0	1	1	1	0	1	1	1
Expiry date	0	1	1	0	0	0	0	0	0	0
Home email address	0	0	0	1	1	1	1	0	1	0
Work address	1	1	0	2	4	1	0	0	1	3
Work telephone number	1	0	0	1	1	1	1	0	1	0
Work email address	0	0	0	1	1	1	1	0	1	0
Bank account details	0	0	0	0	0	0	0	0	1	0
Height	0	0	0	0	0	0	0	0	0	0
Sum	**22**	**30**	**24**	**16**	**28**	**22**	**14**	**10**	**13**	**14**

For us to identify the hierarchy of importance of attributes in the corpus, we needed to consider the term weight of each attribute within the corpus of the ten (10) documents. We had represented the ten (10) documents in our functions as d_1, d_2, d_3,…, d_{10}. The general expression of d_i, represents the same ten documents ranging from d_1 to d_{10}.

In text mining each document is represented as a vector. The elements in the vector reflect the frequency of terms in documents, and each word is a dimension and documents are vectors [103].

3.18.5 Term importance

Jiao *et. al.* established that a classic way to assess the importance of a term is the so-called *tf-idf* (term frequency - inverse document frequency) term weighting scheme [57]. They further indicated that the term importance is based on two assumptions:

a. *idf* assumption: rare terms are more informative than frequent terms,
b. *tf* assumption: multiple occurrences of a term in a query document are more relevant than single occurrence [57].

After sorting the outcomes of the computations of the weighting in the product *tf*idf* we were able to arrange in order of which attribute was more important than the other.

3.18.6 Euclidean distance based similarity

Past efforts in [16] have showed that Euclidean Distance Geometry could improve the authentication in digital identity management system and particularly improve the security in digital financial services. Therefore we would need to test Euclidean Distance measure on the same dataset.

The Euclidean distance between two points or terms (a and b), from a corpus, in an n-dimensional space is represented by the function

$$d_{a_i,b_i} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \quad (3.9)$$

where $i = 1, 2, 3, \ldots, n$

3.19 Summary

This chapter looked at the research process and methods that guided our study; other areas that were looked at included process of data collection, research methods, techniques, and data analysis. Data were analysed using different tools and techniques; two types of data were collected, these were primary data and secondary data. Primary data were collected through a survey, in a sample of a population, using methods and tools appropriate for the quantitative research. Secondary data were extracted and text mined using adopted methods and tools. We discussed limitations that were in this study and also the ethical considerations regarding this research. Data were standardized and data that had more weight than the other were analysed using different methods and techniques. Major sources of identity attributes that were used in the application and registration forms of various services offered both in the cyber and real space were identified. Identity attributes were extracted from documents that were gathered from the internet using text mining techniques. We established a mathematical model based on distance metric, to develop an identity attribute metrics model to quantify the identity attributes.

The model for this research was proposed, tested, and was used to quantify identity attributes for the verification of ownership identification.

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter gives the outcome of the research; this will be presented in parts by considering the assessment of the primary data on the research, this will be followed by the assessment of the secondary data. A framework influencing research based on primary data will be considered. Statistical analysis of different techniques will be considered. Results on standardized data will be reflected on. The assessment of the proposed model will follow thereafter by verifying the outcome of the metrics on the standardized data.

4.2 Statistical analysis results

4.2.1 Results on primary data

Data were subjected to the analysis of using Statistical Packages for the Social Sciences (SPSS) software and Microsoft Excel spreadsheets for graphical presentation. Results of the analysis have been given below.

The outcomes of the analysis has been presented in section 4.2.2 of this report.

4.2.2 Mean score

The following mean score for each organization across the 5 dimensions were computed and used for ranking. These scores were mean scores from the frequencies of part two of the questionnaire; this part of the questionnaire was rating each identity attribute how it was perceived by the respondents from the five constructs: usefulness, trust, ease of use, image (status), or user satisfaction. The results below are the means of the outcomes of the survey for that section.

Table 7: Mean score of the five constructs on perceived importance of identity tokens

Institution and tokens	Constructs				
	Usefulness	Trust	Ease of use	Image	User satisfaction
Banks	4.59	4.41	4.18	3.80	4.09
Insurance	4.06	3.88	3.88	3.49	4.11
Churches	4.91	4.97	4.97	4.69	4.69
Government	4.94	4.82	4.80	4.50	4.71
Hospital	4.64	4.40	4.53	4.25	4.59
Mobile Phone Companies	4.83	3.58	3.50	3.50	4.83
Schools	4.80	4.66	4.80	4.56	4.49
Universities	4.60	4.42	4.32	4.16	4.49
Utility Bills	5.00	4.00	3.50	4.00	4.50

The chart below gives the average scores for the five constructs, namely Usefulness, Trust, Ease of use, Image (status), and User satisfaction.

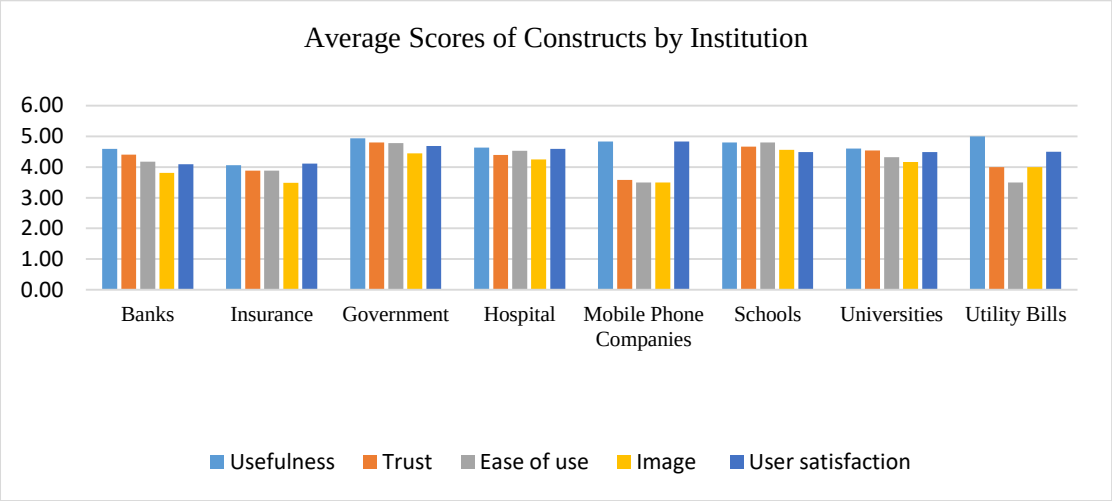

Figure 14: means scores for the eight organizations across the 5 constructs

From figure 15 above, we observe that data from all institutions including government, hospitals, schools, Banks, insurance, and Mobile phone companies considered a large amount of identity attributes from our list to be very important and scored an average of more than 3 out of five score level. Where 1 was for least important and 5 was for very important attributes.

4.2.3 Key identity attributes

4.2.3.1 Primary data

Mean score of identity attributes were as follows:

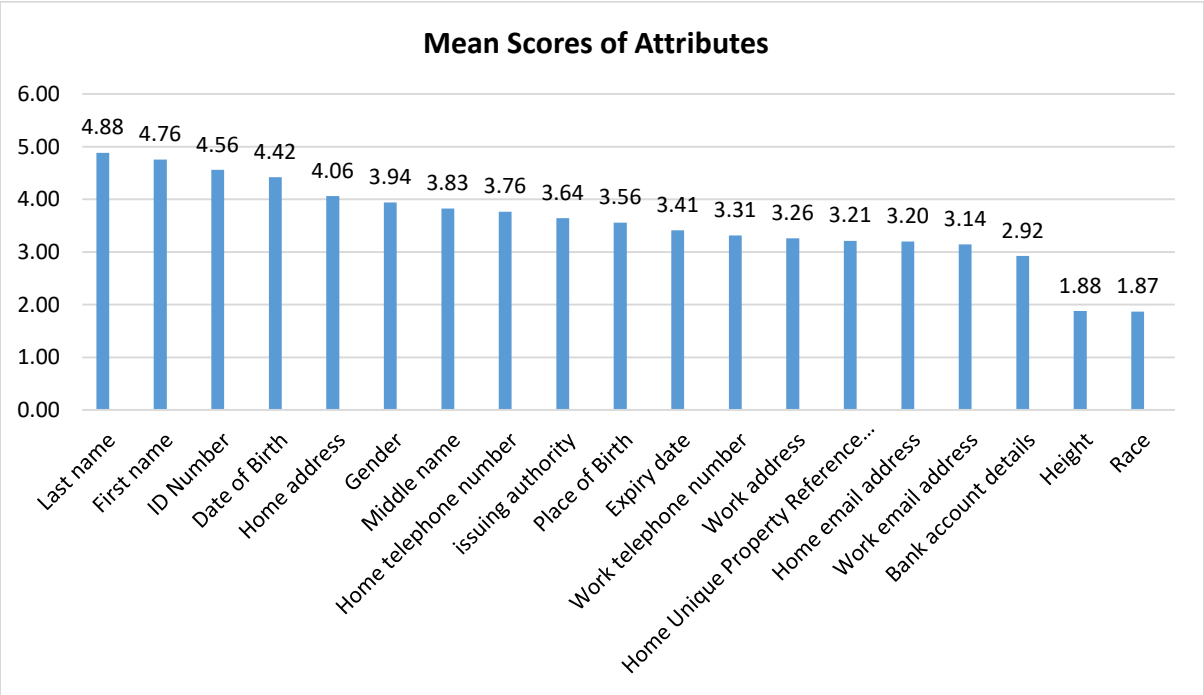

Figure 15: Mean scores for the identity attributes

The analysis of the primary data (using statistical methods, techniques, and MS Excel) revealed that 89.47% (rounding the score to one significant figure) of the attributes that were obtained from ISO/IEC JTC 1/SC 27 for use in this research were important for identification of an entity. From the survey, each identity attributes scored at an average of either 3 or above, out of 5, respectively. It was observed that key identity attributes that were on application forms included the following identity attributes with their respective outcome performances in the analysis:

Table 8: Level of importance of key attributes

Item	Attributes	Mean Score (in one significant number) on the importance of Attributes
1	Last name	5
2	First name	5
3	ID Number	5
4	Date of Birth	4
5	Home address	4
6	Gender	4
7	Middle name	4
8	Home telephone number	4
9	issuing authority	4
10	Place of Birth	4
11	Expiry date	3
12	Work telephone number	3
13	Work address	3
14	Home Unique Property Reference Number (House Number)	3
15	Home email address	3
16	Work email address	3
17	Bank account details	3

Only “height” and “race” attributes from the ISO identity list were below average of importance of identity attributes, as per responses from our survey. The survey results from our sample confirmed that these 17 identity attributes were very key in identifying an entity in real space.

4.2.3.2 Secondary data

Table 9: Key identity attributes from secondary data

Zambian Institutions

ATTRIBUTE	Organisations				Shannon Information Entropy
	Average term weights of attributes				
	Banks W_1	Gov W_2	Insurance W_3	Univ & Sch W_4	
Place of Birth	0.5193	0.5279	0.3824	0.1643	1.5939
Bank account details	0.4814	0.5307	0.4721	0.0820	1.5662
ID Number	0.4788	0.5205	0.4890	0.0256	1.5139
Date of Birth	0.4690	0.4193	0.4619	0.0403	1.3905
Gender	0.4253	0.4347	0.4980	0.0324	1.3904
Last name	0.4692	0.3425	0.4885	0.0323	1.3326
Home telephone number	0.5302	0.0000	0.5305	0.1739	1.2346
First name	0.5263	0.1373	0.5075	0.0591	1.2302
Work telephone number	0.4422	0.0000	0.5300	0.2107	1.1829
Middle name	0.5079	0.4581	0.0000	0.0363	1.0023
Work email address	0.5000	0.0000	0.0000	0.0000	0.5000
Expiry date	0.1993	0.0000	0.0000	0.2804	0.4797
Home address	0.2755	0.1228	0.0000	0.0000	0.3983
Work address	0.0000	0.0533	0.0000	0.0000	0.0533
Race	0.0000	0.0000	0.0000	0.0000	0.0000
Home Unique Property Reference Number (House Number)	0.0000	0.0000	0.0000	0.0000	0.0000
issuing authority	0.0000	0.0000	0.0000	0.0000	0.0000
Home email address	0.0000	0.0000	0.0000	0.0000	0.0000
Height	0.0000	0.0000	0.0000	0.0000	0.0000

The selected sample from our secondary data from text mined pdf documents was subjected to weighted schemes of computations to consider the level of importance. It was observed that the “Race” and “Height” identity attribute were found to be insignificant in identifying an entity. Two other identity attributes were found among the least important identity attributes to identify an entity. The identity attributes that were found to be key or highly important were about 77% of the entire list as highlighted and shown on table 9 above. We had to normalize the data to remove errors from data, we used the Shannon Information Entropy to establishing the weighting of the identity attributes.

4.2.4 Demographic analysis

Table 10: Results on demographic analysis

		Frequency	Percent	Valid Percent	Cumulative Percent
Gender	Male	76	51.0	52.8	52.8
	Female	68	45.6	47.2	100.0
	Total	144	96.6	100.0	
Missing	System	5	3.4		
Total		149	100.0		
Marital status	Single	58	38.9	40.3	40.3
	Married	80	53.7	55.6	95.8
	Divorced	1	.7	.7	96.5
	Other	5	3.4	3.5	100.0
	Total	144	96.6	100.0	
Missing	System	5	3.4		
Total		149	100.0		
Age	20 years & below	4	2.7	2.7	2.7
	21-30 years	71	47.7	48.0	50.7
	31-40 years	48	32.2	32.4	83.1
	41-50 years	19	12.8	12.8	95.9
	61+ years	6	4.0	4.0	98.0
	Total	148	99.3	100.0	
Highest education level	Grade 12 & below	15	10.1	10.1	10.1
	Diploma	54	36.2	36.5	46.6
	First degree	52	34.9	35.1	81.8
	Master's degree	25	16.8	16.9	98.6
	PhD	2	1.3	1.4	100.0
	Total	148	99.3	100.0	

The respondents from the sample had 53% male and around 40% were single. Close to 50% were aged below 30 years. In terms of education level, only 10% had either grade 12 or less in qualifications – with 37% having a college diploma, another 35% with university degree with the rest having a masters or PhD. Almost 88% were salaried employees.

For the banking organization usefulness was ranked highest (means score 4.2), followed by Trust & Ease of use (3.8), user satisfaction (3.7) and the least was image or status (3.6).

4.2.5 Correlation Analysis

The table below highlights the results for the correlation analysis. Testing at 0.05 and 0.01 levels of two tailed-test, it was found that there was a positive significant relationship between how often one uses the document and perceived ease of use, perceived usefulness, and security (trust). Other studies have shown similar trends around such variables.

The results show that relationships that exist between the following variables in the TAM model are positive significant ones. The relationship of perceived usefulness with trust gave us the following outcomes: r (133) = 0.759, p < .001. Similarly, there was a relationship between usefulness and ease of use, r (138) = 0.759 and so was the case of between trust and ease of use, r (136) =0.751, p=0.00. How often the use of the documents was also related to perceived usefulness, r (136) =0.201, p<0.05), trust - r (135) = 0.203, p <0.05) and ease of use - r (133), p<0.01.

Table 11: Correlation on perceived use of the research constructs

			Often	**Usefulness**	**Trust**	**Ease of use**
Spearman's rho	**Often**	Correlation Coefficient	1.000	-.201*	-.203*	-.251**
		Sig. (2-tailed)	.	.019	.018	.004
		N	142	136	135	133
	Usefulness	Correlation Coefficient	-.201*	1.000	.759**	.759**
		Sig. (2-tailed)	.019	.	.000	.000
		N	136	141	139	138
	Trust	Correlation Coefficient	-.203*	.759**	1.000	.751**
		Sig. (2-tailed)	.018	.000	.	.000
		N	135	139	139	136
	Ease of use	Correlation Coefficient	-.251**	.759**	.751**	1.000
		Sig. (2-tailed)	.004	.000	.000	.
		N	133	138	136	138

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

4.2.6 Regression Analysis

Table 12: Correlation of how documents are used

		Often	**Usefulness**	**Trust**	**Ease of use**
Pearson Correlation	**Often**	1.000	-.230	-.226	-.270
	Usefulness	-.230	1.000	.771	.738
	Trust	-.226	.771	1.000	.760
	Ease of use	-.270	.738	.760	1.000
Sig. (1-tailed)	**Often1**	.	.004	.005	.001
	Usefulness	.004	.	.000	.000
	Trust	.005	.000	.	.000

	Ease of use	.001	.000	.000	.
N	**Often1**	132	132	132	132
	Usefulness	132	132	132	132
	Trust	132	132	132	132
	Ease of use	132	132	132	132

From the table above, it can be seen that ease of use (p=0.001) is the largest predictor of how often documents are used, followed by usefulness (p =0.004) and lastly trust (p = 0.004).

The table below shows a relation between the dependent variable – How often documents are used and the predictor variables ease of use, usefulness and trust. Both the model summary and the analysis of variance shows a good relation between how often documents are used with the predictor variables, with p =0.018.

4.2.7 Constructs relationships

Table 13: Framework affecting this research

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.274[a]	.075	.054	.39392	.075	3.469	3	128	.018
a. Predictors: (Constant), Ease of use , Usefulness , Trust									
b. Dependent Variable: Often1									

In the model summary (above) of the regression, although test R Square = 0.075 which is 7.5% variance in how often this accounts for use, there is a relationship which can be accounted for by perceived ease of use, trust and usefulness. Using the Anova table test and taking alpha as 0.05 refer to Table 5, the regression model for the variance of the predicators was significant, F (3.4) = 7.5, p < 0.005. The variance of the independent variables as a whole was therefore significant as it was less than the alpha.

4.2.8 ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.615	3	.538	3.469	.018[b]
	Residual	19.862	128	.155		
	Total	21.477	131			
a. Dependent Variable: Often1						
b. Predictors: (Constant), Ease of use , Usefulness , Trust						

4.2.9 Chi-Square Test

When Chi square test is conducted for each construct in relation to how often they use the identity documents, there was strong statistical association across the number of times identity documents are used and their perceived usefulness, perceived trust and perceived ease of use. Those who perceived the identity documents more useful (50.5%) were more likely to use them two or more times in a week compared to those whose perceptions were less (37%). This was similar for trust perceptions (41.7%) and ease of use (44.3%) against 33.3% and 29.6% respectively.

Table 14: Relationship between use and constructs

	Usefulness	Trust	Ease of Use
At most once a week	37.0%	33.3%	29.6%
2 or more times a week	50.5%	41.7%	44.3%
Chi-Square	X^2=0.001	X^2=0.0003	X^2=0.00014

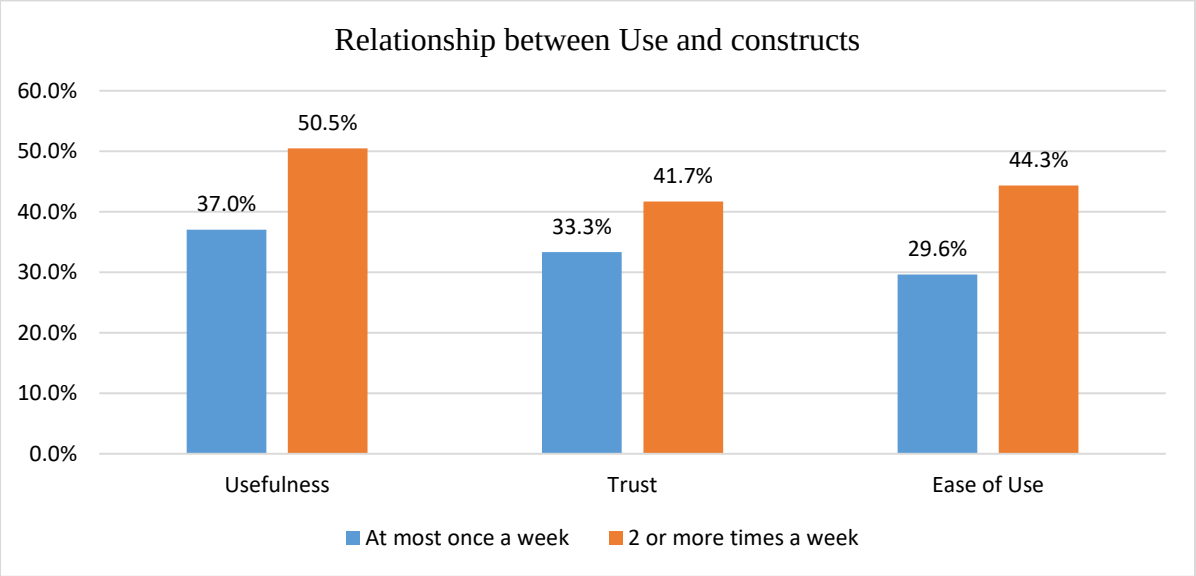

Figure 16: Use of identity tokens against the constructs for this study

Methodology: Mean scores for each organization across the 5 dimensions were computed and used for ranking.

Results: For the banking organization, usefulness was ranked highest (means score 4.2), followed by Trust & Ease of use (3.8), user satisfaction (3.7) and the least was image or status (3.6).

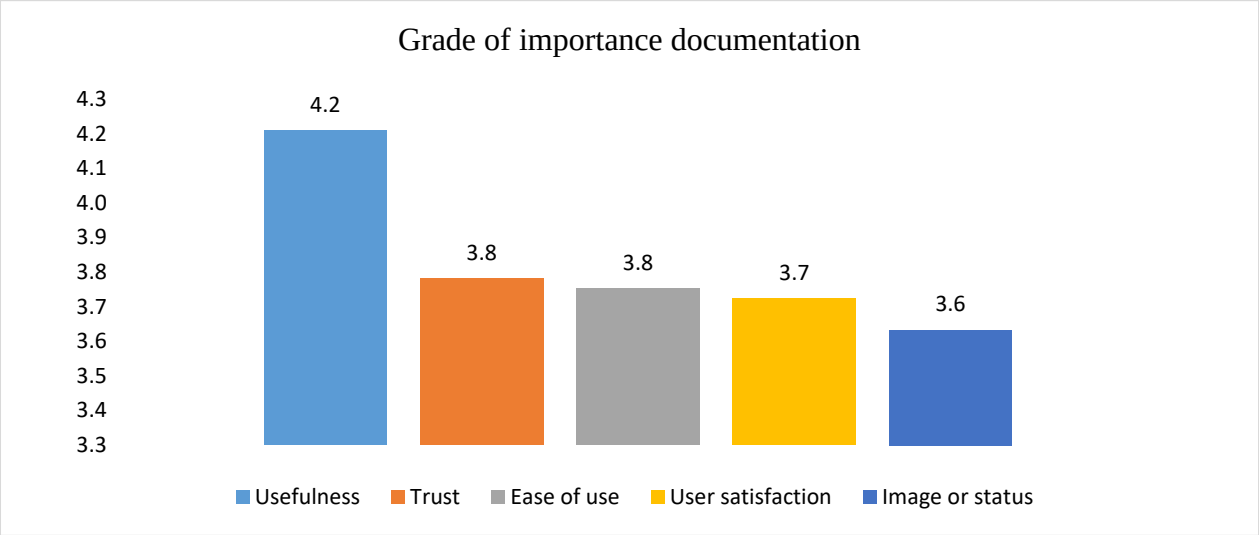

Figure 17: Rating of identity documents against importance

4.2.10 Conceptual framework

The model that was influencing our research was adopted as the Technology Acceptance Model (TAM), as indicated below.

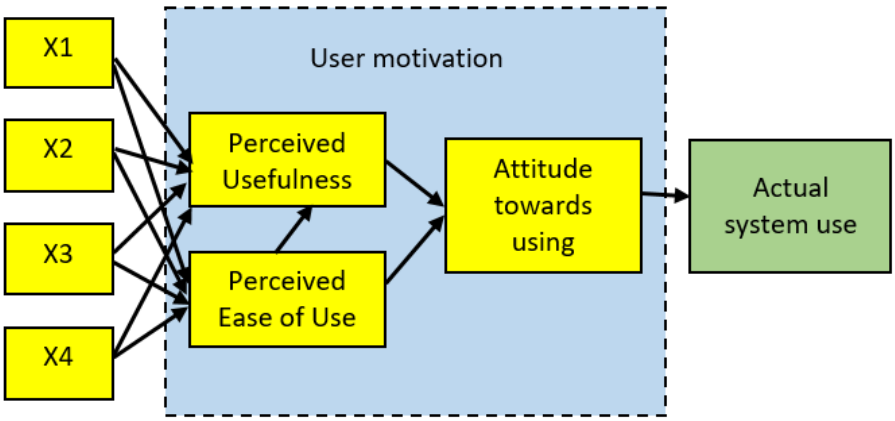

Figure 18: Model influencing this research

The five variables that were key in the use of the identity tokens, included Use of use, Trust, Ease of use, User satisfaction, and Image (status). We applied our adopted framework, TAM model, to the analysis of these variables to find out how these variables related to each other. These variables were considered as considerations involved in accepting to use these documents that contained identity attributes of a user or applicant of a required service. Subjecting these variables to statistical analysis, we were able to find the relationships among these variables. The relationship amongst these variables was the motivation of the use of the identity tokens. Figure 19 shows the interaction of the five variables and the resultant variables that motivated the actual use of the identity tokens.

The following was the resultant framework from analysis and was proposed as our research framework for the study:

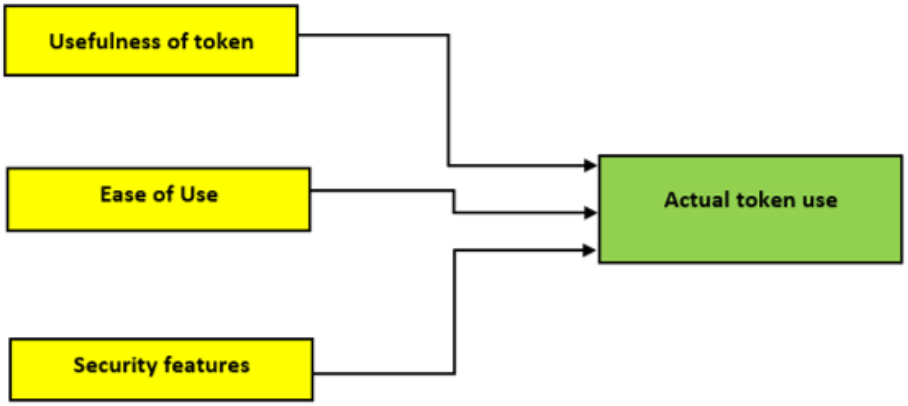

Figure 19: Proposed research framework

Statistical analysis in correlation analysis, regression analysis, and chi-square test revealed that the variables of usefulness, Ease of use, and Trust (security features) were key motivations in the use of identity tokens. They were influential in motivating a user of the identity tokens that bear the identity attributes of an entity.

4.3 Results on Secondary data

The term frequencies of the selected attributes and organizations for Zambia were as given in table 14.

4.3.1 Text analysis

4.3.1.1 Identity's Term Frequencies

The table shows a term frequencies of identity attributes from four different institutions in Zambia from a corpus of secondary data. This was the basis of our analysis of secondary data using different statistical methods and other methods of finding the amount of information from the extracted text for the identity attributes.

Table 15: A sample of Term frequencies

ZAMBIA

ATTRIBUTE	ORGANIZATION							
	Banks		Government		Insurance		Universities & Schools	
	No. of Tokens	Freq.	No. of Tokens	Freq.	No. of Tokens	Freq.	No. of Tokens	Freq.
First name	22	18	23	1	16	18	36	19
Middle name	22	13	23	7	16	5	36	24
Last name	22	22	23	9	16	40	36	36
Date of Birth	22	23	23	14	16	43	36	32
Place of Birth	22	5	23	5	16	1	36	6
Race	22	0	23	0	16	0	36	0
Gender	22	17	23	15	16	37	36	36
Home address	22	29	23	8	16		36	

Home Unique Property Reference Number (House Number)	22	0	23	2	16	0	36	1
Home telephone number	22	5	23	0	16	3	36	5
ID Number	22	32	23	51	16	15	36	49
issuing authority	22	0	23	1	16	0	36	0
Expiry date	22	17	23	0	16	0	36	3
Home email address	22	0	23	0	16	1	36	0
Work address	22	17	23	14	16	0	36	0
Work telephone number	22	12	23	0	16	3	36	5
Work email address	22	3	23	0	16	3	36	0
Bank account details	22	43	23	14	16	5	36	18
Height	22	0	23	0	16	0	36	0
	Total	256		141		174		234

The term frequencies needed to be standardized through normalization so as to remove errors or noise which could affect the accuracy of the data. The data in Table 14 were normalized in preparation for further for further analyses of data. Table 15 below shows that outcome of standardization of data in Table 14. Data were normalized using the z-score Normalization (zero-mean Normalization) technique.

Table 16: Normalised data in the four organizations in Zambia

ZAMBIA

ATTRIBUTE	Organizations				Total	Desirable properties of the attribute (Normalised data)			
	Banks (1)	Government (2)	Insurance (3)	Universities & Schools (4)	$\sum tf_i$	$P_i = tf_i/\sum tf_i$			
	Term freq.	Term freq.	Term freq.	Term freq.	Sum of tf_i	P_1	P_2	P_3	P_4
First name	18	1	18	19	56	0.3214	0.0261	0.4820	0.9582
Middle name	13	7	5	24	49	0.2653	0.1930	0.1697	0.9745
Last name	22	9	40	36	107	0.2056	0.1056	0.5242	0.9773
Date of Birth	23	14	43	32	112	0.2054	0.1569	0.5706	0.9717
Place of Birth	5	5	1	6	17	0.2941	0.4067	0.1299	0.8784
Race	0	0	0	0	0	0.0000	0.0000	0.0000	0.0000
Gender	17	15	37	36	105	0.1619	0.1701	0.5046	0.9773
Home address	29	8	0	0	37	0.7838	0.9108	0.0000	0.0000
House Number	0	2	0	1	3	0.0000	0.6667	0.0000	0.6000
Home telephone number	5	0	3	5	13	0.3846	0.0000	0.3578	0.8707
ID Number	32	51	15	49	147	0.2177	0.4426	0.2320	0.9821
issuing authority	0	1	0	0	1	0.0000	1.0000	0.0000	0.0000
Expiry date	17	0	0	3	20	0.8500	0.0000	0.0000	0.7792
Home email address	0	0	1	0	1	0.0000	0.0000	1.0000	0.0000

Work address	17	14	0	0	31	0.5484	0.9623	0.0000	0.0000
Work telephone number	12	0	3	5	20	0.6000	0.0000	0.3488	0.8405
Work email address	3	0	3	0	6	0.5000	0.0000	0.8571	0.0000
Bank account details	43	14	5	18	80	0.5375	0.3730	0.2091	0.9414
Height	0	0	0	0	0	0.0000	0.0000	0.0000	0.0000
Sum	256	141	174	234					

4.4 Shannon Information entropy

4.4.1 Primary data analysis

From the Table 16 above, tf_i are the term frequencies for each identity attribute where *tf* is a term frequency for a respective attribute and i = 1, 2, 3,…, n in a document vector.

$\sum tf_i$ (4.1)

This is the sum of term frequencies for each identity attribute in a corpus.

$$P_i = \frac{tf_i}{\sum_{i=1}^{n} tf_i} \quad (4.2)$$

P_i is the mean frequency on normalized data of each identity attribute in the corpus.
is a desirable property of the attribute and tf_i is the Term Frequency (*tf*) of a given i^{th} attribute in a given organization.

$P(tf_i)$ is the probability of the occurrence for each term from the set of identity attributes.

An example of computations from the formula above

tf_1= Banks = 0.3214

tf_2= Governments = 0.0261

tf_3= Insurance = 0.4820

tf_4= Universities and schools = 0.9582

It follows that for $P_i = \frac{tf_i}{\sum_{i=1}^{n} tf_i}$, when i = 1, then

$P_1 = \frac{18}{56}$ = 0.32143

Computing for each term frequency of the different organizations in a similar way for the same attribute and calculating the entropy function on the attribute would yield the following:

$$H(x = First\ Name) = \sum_{i=1}^{4} P_i \log_2 \left(\frac{1}{P_i}\right) = 1.2302 \quad (4.3)$$

This is therefore, the entropy of the attribute, First Name, which is the calculated weight of the attribute in the dataset that was obtained. Repeating the calculations for each attribute would yield the figures indicated in the corresponding table below under 4.4.2. From the

table, W_1, W_2, W_3, and W_4 are the contributing entropies of the required weight of an identity attribute. The last column h_i represents the computed information of an attribute in a given dataset. This entropy is the weight of an attribute in this dataset.

4.4.2 Identity Attribute Entropy

Table 17: Weighted data using Shannon entropy

ZAMBIA

ATTRIBUTE	Weighting				Entropy	Degree of diversification
	$P_i\log_2(1/P_i)$				$h_i = \sum P_i\log_2(1/P_i)$	$Div = 1 - h_i$
	W_1	**W_2**	**W_3**	**W_4**		
First name	0.5263	0.1373	0.5075	0.0591	1.2302	-0.2302
Middle name	0.5079	0.4581	0.0000	0.0363	1.0023	-0.0023
Last name	0.4692	0.3425	0.4885	0.0323	1.3326	-0.3326
Date of Birth	0.4690	0.4193	0.4619	0.0403	1.3905	-0.3905
Place of Birth	0.5193	0.5279	0.3824	0.1643	1.5939	-0.5939
Race	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Gender	0.4253	0.4347	0.4980	0.0324	1.3904	-0.3904
Home address	0.2755	0.1228	0.0000	0.0000	0.3983	0.6017
House Number	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Home telephone number	0.5302	0.0000	0.5305	0.1739	1.2346	-0.2346
ID Number	0.4788	0.5205	0.4890	0.0256	1.5139	-0.5139
issuing authority	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Expiry date	0.1993	0.0000	0.0000	0.2804	0.4797	0.5203
Home email address	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Work address	0.0000	0.0533	0.0000	0.0000	0.0533	0.9467
Work telephone number	0.4422	0.0000	0.5300	0.2107	1.1829	-0.1829
Work email address	0.5000	0.0000	0.0000	0.0000	0.5000	0.5000
Bank account details	0.4814	0.5307	0.4721	0.0820	1.5662	-0.5662
Height	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

The entropy of data from secondary data from different regions is compared and the top ten of the attributes are compared as follows:

a. Randomly selected countries for the sake of comparison with that of the model that has been developed. The countries include Zambia, Botswana, Namibia, Canada, Australia, New Zealand, and USA.
b. Weighted data from ten documents whose data were used to test the developed model

4.4.3 Top ten identity attributes on sampled countries, respectively

Table 18: Comparison of top ten identity attributes from different countries

ZAMBIA		NAMIBIA		BOTSWANA		CANADA	
ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$	ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$	ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$	ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$
Place of Birth	1.5939	First name	0.6980	Work telephone number	1.8446	Bank account details	1.8734
Bank account details	1.5662	Middle name	1.2829	Home address	1.7646	ID Number	1.6134
ID Number	1.5139	Last name	0.7807	Expiry date	1.7302	Expiry date	1.5662
Date of Birth	1.3905	Date of Birth	0.9268	Date of Birth	1.7060	Home address	1.5647
Gender	1.3904	Place of Birth	0.9480	ID Number	1.6739	Gender	1.5530
Last name	1.3326	Race	0.0000	Gender	1.6532	Last name	1.5294
Home telephone number	1.2346	Gender	1.1667	Last name	1.5053	Date of Birth	1.4653
First name	1.2302	Home address	1.1599	Bank account details	1.3760	First name	1.3952
Work telephone number	1.1829	Home Unique Property Reference Number (House Number)	0.0000	Place of Birth	1.3691	Work telephone number	1.2952
Middle name	1.0023	Home telephone number	0.7428	First name	1.3266	Place of Birth	1.2561

AUSTRALIA		NEWZEALAND		USA	
ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$	ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$	ATTRIBUTE	Entropy $\sum P_i \log_2(1/P_i)$
Bank account details	1.8017	Bank account details	1.5904	Work telephone number	1.6212
Home address	1.5838	Last name	1.5855	Home address	1.5913
Gender	1.4741	Date of Birth	1.4848	Bank account details	1.5820
Last name	1.4715	Expiry date	1.4499	Date of Birth	1.4773
Date of Birth	1.3971	Home address	1.4411	Gender	1.4395
ID Number	1.3827	First name	1.4271	Place of Birth	1.4062
Middle name	1.3419	Place of Birth	1.3611	ID Number	1.3328
Work telephone number	1.2705	Gender	1.2313	Last name	1.3054
First name	1.2634	ID Number	1.2029	First name	1.2679
Home telephone number	1.1512	Home telephone number	1.0386	Home telephone number	1.0528

4.4.4 Quantification of identity attributes

Weighting of the text mined identity attributes improves the accuracy of data since errors and noise get eliminated during this process. We had adopted the TF-IDF weighting scheme to weight the identity attributes in our corpus. Table 18 computes the first part of the scheme i.e. the term frequencies (TF) of the attributes.

Table 19: Term frequencies on ten documents for the metrics

ZAMBIA

ATTRIBUTE	$tf_i = 1 + \log tf_i$									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Middle name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Last name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Date of Birth	0.0000	1.6990	1.3010	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
Place of Birth	0.0000	1.4771	1.4771	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Race	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gender	0.0000	1.4771	1.4771	0.0000	1.3010	1.3010	0.0000	0.0000	1.0000	0.0000
Home address	1.0000	1.0000	1.3010	1.4771	1.6021	1.0000	1.4771	0.0000	1.0000	1.4771
House Number	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Home telephone number	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
ID Number	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
issuing authority	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
Expiry date	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Home email address	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Work address	1.0000	1.0000	0.0000	1.3010	1.6021	1.0000	0.0000	0.0000	1.0000	1.4771
Work telephone number	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Work email address	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Bank account details	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
Height	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The table below quantifies the index document frequencies (IDF) in our term weighting function (TF-IDF). Once we know the TF and IDF, then we can obtain TF*IDF so as to be able to rank the most weightier identity attributes

*Table 20: Inverse function (IDF) for the TF*IDF weighting*

ATTRIBUTE	Total No. of docs	$idf_i = \log \frac{N}{df_i} = \log N - \log df_i$									
	N	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Middle name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Last name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Date of Birth	10	0.00000	0.30103	0.69897	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
Place of Birth	10	0.00000	0.52288	0.52288	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Race	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Gender	10	0.00000	0.52288	0.52288	0.00000	0.69897	0.00000	0.00000	0.00000	1.00000	0.00000
Home address	10	1.00000	1.00000	0.69897	0.52288	0.39794	1.00000	0.52288	0.00000	1.00000	0.52288
House Number	10	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000
Home telephone number	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000
ID Number	10	1.00000	1.00000	1.00000	0.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
issuing authority	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	0.00000	1.00000	1.00000	1.00000
Expiry date	10	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Home email address	10	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Work address	10	1.00000	1.00000	0.00000	0.00000	0.39794	1.00000	0.00000	0.00000	1.00000	0.52288
Work telephone number	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Work email address	10	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Bank account details	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
Height	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

4.4.5 Weighted identity attributes

Table 20 gives the results of the computation of the weighting of the identity attributes in the function W=TF*IDF, where "W" represents the weights of respective identity attributes from the corpus of the ten documents from a Zambian government Department that we had sampled. Using these figures, we would be able to rank the identity attributes in their level of importance.

*Table 21: TF*IDF weighting of the identity attributes on ten documents*

ZAMBIA

ATTRIBUTE	$W_{i,d}=TF_i * IDFi = tf_i * \log \frac{N}{idf_i}$									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381
Middle name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381
Last name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381

Date of Birth	0.000000	0.511441	0.909381	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
Place of Birth	0.000000	0.772355	0.772355	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Race	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Gender	0.000000	0.772355	0.772355	0.000000	0.909381	0.000000	0.000000	0.000000	1.000000	0.000000
Home address	1.000000	1.000000	0.909381	0.772355	0.637524	1.000000	0.772355	0.000000	1.000000	0.772355
House Number	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
Home telephone number	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
ID Number	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
issuing authority	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000
Expiry date	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Home email address	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Work address	1.000000	1.000000	0.000000	0.000000	0.637524	1.000000	0.000000	0.000000	1.000000	0.772355
Work telephone number	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Work email address	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Bank account details	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
Height	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

From table 20, we were able to identify and rank from the ten documents which identity attributes were most important in identifying an entity from an identity token. The table below gives a sequence of the most important attribute in the corpus amongst our attributes of interest that we had listed from the corpus.

Table 22: Listing of importance of the identity attributes

ZAMBIA

ATTRIBUTE	Term (tf_i)										Total	T F_i*IDF_i $\sum_{i=1}^{n} tf_i X \log \frac{N}{idf_i}$
	AIRSPACE (d_1): tf_1	RESIDENCE PERMIT (d_2): tf_2	VISITING VISA (d_3): tf_3	Consent form (d_4): tf_4	FarmSmallholding (d_5): tf_5	Residential land (d_6): tf_6	AQUACULTURE FUND (d_7): tf_7	BOREHOLE FORM (d_8): tf_8	HEALTH PROF COUNCIL (d_9): tf_9	IMMOVABLE PROPERTY (d_{10}): tf_{10}		
Home address	1	1	2	3	4	1	3	0	1	3	19	7.86397063
First name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
Middle name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
Last name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
ID Number	1	1	1	0	1	1	1	0	1	0	7	7.00000000
issuing authority	1	0	0	1	1	1	0	1	1	1	7	7.00000000
Work telephone number	1	0	0	1	1	1	1	0	1	0	6	6.00000000
Work address	1	1	0	2	4	1	0	0	1	3	13	5.40987908

Home telephone number	1	0	0	1	1	1	1	0	0	0	5	5.00000000
Home email address	0	0	0	1	1	1	1	0	1	0	5	5.00000000
Work email address	0	0	0	1	1	1	1	0	1	0	5	5.00000000
Gender	0	3	3	0	2	2	0	0	1	0	11	3.45409156
Date of Birth	0	5	2	0	0	0	0	0	1	0	8	2.42082187
House Number	1	0	0	0	0	0	0	0	0	1	2	2.00000000
Expiry date	0	1	1	0	0	0	0	0	0	0	2	2.00000000
Place of Birth	0	3	3	0	0	0	0	0	0	0	6	1.54471062
Bank account details	0	0	0	0	0	0	0	0	1	0	1	1.00000000
Race	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Height	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Sum	**22**	**30**	**24**	**16**	**28**	**22**	**14**	**10**	**13**	**14**	**193**	

4.5 Verification of ownership

For the purposes of verification of ownership of the attributes by an online user, we will assume that the object of ownership is the user of document 2 from our corpus of ten documents. Document 2 was capturing attributes of a people applying for residence permit. It is only an individual who has entered responses that match the attributes of the specific individual. For the sake of assessment of key attributes, we will consider the attributes involved in identifying the digital identity of our object and compare with the other attributes from the other nine (9) documents. We are going to look at the attributes of the second document and compare them to each of the documents of the nine other documents, respectively. Using our proposed model of the Cosine Similarity measure we would then observe the performance on similarity of the attributes of the second document to those of the other nine.

4.5.1 Verification based on Term Frequencies

We have the following vectors from the Term Frequencies of the attributes of the ten documents of the corpus:

i. Airspace (d_1): $tf_1 = \boldsymbol{d}_1$ = (5, 5, 5, 5, 3, 0, 3, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0)
ii. Residence Permit (d_2): $tf_2 = \boldsymbol{d}_2$ = (5, 5, 5, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0)
iii. Visiting Visa (d_3): $tf_3 = \boldsymbol{d}_3$ = (4, 4, 4, 2, 3, 0, 3, 2, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0)
iv. Consent form (d_4): $tf_4 = \boldsymbol{d}_4$ = (2, 2, 2, 0, 0, 0, 0, 3, 0, 1, 0, 1, 0, 1, 2, 1, 1, 0, 0)
v. FarmSmallholding (d_5): $tf_5 = \boldsymbol{d}_5$ = (4, 4, 4, 0, 0, 0, 2, 4, 0, 1, 1, 1, 0, 1, 4, 1, 1, 0, 0)
vi. Residential land (d_6): $tf_6 = \boldsymbol{d}_6$ =(4, 4, 4, 0, 0, 0, 2, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0)
vii. Aquaculture Fund (d_7): $tf_7 = \boldsymbol{d}_7$ = (2, 2, 2, 0, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0)
viii. Borehole Form (d_8): $tf_8 = \boldsymbol{d}_8$ = (3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0)
ix. Health Prof Council (d_9): $tf_9 = \boldsymbol{d}_9$ = (1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0)

x. Immovable Property (D_{10}): $tf_{10} = \boldsymbol{d}_{10}$ = (2, 2, 2, 0, 0, 0, 0, 3, 1, 0, 0, 1, 0, 0, 3, 0, 0, 0, 0)

Replacing the variables of the documents in our model, we let the documents to be identified by $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{10}$. We then apply these vectors in the model given here below.

Similarity = $S(\boldsymbol{d}_2, \boldsymbol{d}_i) = Cos(\boldsymbol{d}_2, \boldsymbol{d}_i) = \frac{\boldsymbol{d}_2 * \boldsymbol{d}_i}{\|d_2\| \|\boldsymbol{d}_i\|} = \frac{\sum_{i=1}^{n} \boldsymbol{d}_2 \boldsymbol{d}_i}{\sqrt{\sum_{i=1}^{n} \boldsymbol{d}_i^2} \sqrt{\sum_{i=1}^{n} \boldsymbol{d}_i^2}}$ (4.4)

1. For $S(\boldsymbol{d}_2, \boldsymbol{d}_1)$:

$\boldsymbol{d}_2.\boldsymbol{d}_1$ = (5, 5, 5, 5, 3, 0, 3, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0). (5, 5, 5, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0)

=

((5x5)+(5x5)+(5x5)+(5x0)+(3x0)+(0x0)+(3x0)+(1x1)+(0x1)+(0x1)+(1x1)+(0x1)+(1x0)+(0x0)+(1x1)+(0x1)+(0x0)+(0x0)+(0x0)) = 78

This follows that

$\boldsymbol{d}_2 * \boldsymbol{d}_i$	$\boldsymbol{d}_2.\boldsymbol{d}_1$	$\boldsymbol{d}_2.\boldsymbol{d}_2$	$\boldsymbol{d}_2.\boldsymbol{d}_3$	$\boldsymbol{d}_2.\boldsymbol{d}_4$	$\boldsymbol{d}_2.\boldsymbol{d}_5$	$\boldsymbol{d}_2.\boldsymbol{d}_6$	$\boldsymbol{d}_2.\boldsymbol{d}_7$	$\boldsymbol{d}_2.\boldsymbol{d}_8$	$\boldsymbol{d}_2.\boldsymbol{d}_9$	$\boldsymbol{d}_2.\boldsymbol{d}_{10}$
Outcome	78	122	92	35	75	69	34	45	26	36

$\|d_2\|$ = $\sqrt{5^2 + 5^2 + 5^2 + 5^2 + 3^2 + 0^2 + 3^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2}$

= 122

$\|d_1\|$ = $\sqrt{5^2 + 5^2 + 5^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2}$

= 82

Therefore, Similarity = $S(d_2, d_1) = Cos(d_2, d_1) = \frac{d_2 * d_1}{\|d_2\| \|d_1\|} = \frac{78}{122\ x\ 82} = \underline{0.007797}$

It follows that for the rest of the computations we have

$\|d_i\|$	$\|d_1\|$	$\|d_2\|$	$\|d_3\|$	$\|d_4\|$	$\|d_5\|$	$\|d_6\|$	$\|d_7\|$	$\|d_8\|$	$\|d_9\|$	$\|d_{10}\|$
Outcome	82	122	76	30	90	60	26	28	13	32

The rest of the computations are as follows:

2. For $S(d_2, d_2)$:

$S(d_2, d_2) = Cos(d_2, d_2) = \frac{d_2 * d_2}{\|d_2\| \|d_2\|} = \frac{122}{122\ x\ 122} = 0.008197$

3. For $S(d_2, d_3)$:

$$S(d_2, d_3) = Cos(d_2, d_3) = \frac{d_2 * d_3}{\|d_2\| \|d_3\|} = \frac{92}{122\ x\ 76} = 0.009922$$

4. For $S(d_2, d_4)$:

$$S(d_2, d_4) = Cos(d_2, d_4) = \frac{d_2 * d_4}{\|d_2\| \|d_4\|} = \frac{92}{122\ x\ 76} = 0.009563$$

5. For $S(d_2, d_5)$:

$$S(d_2, d_5) = Cos(d_2, d_5) = \frac{d_2 * d_5}{\|d_2\| \|d_5\|} = \frac{75}{122\ x\ 90} = 0.006831$$

6. For $S(d_2, d_6)$:

$$S(d_2, d_6) = Cos(d_2, d_6) = \frac{d_2 * d_6}{\|d_2\| \|d_6\|} = \frac{69}{122\ x\ 60} = 0.009426$$

7. For $S(d_2, d_7)$:

$$S(d_2, d_7) = Cos(d_2, d_7) = \frac{d_2 * d_7}{\|d_2\| \|d_7\|} = \frac{34}{122\ x\ 26} = 0.010719$$

8. For $S(d_2, d_8)$:

$$S(d_2, d_8) = Cos(d_2, d_8) = \frac{d_2 * d_8}{\|d_2\| \|d_8\|} = \frac{45}{122\ x\ 28} = 0.013173$$

9. For $S(d_2, d_9)$:

$$S(d_2, d_9) = Cos(d_2, d_9) = \frac{d_2 * d_9}{\|d_2\| \|d_9\|} = \frac{26}{122\ x\ 13} = 0.016393$$

10. For $S(d_2, d_{10})$:

$$S(d_2, d_{10}) = Cos(d_2, d_{10}) = \frac{d_2 * d_{10}}{\|d_2\| \|d_{10}\|} = \frac{36}{122\ x\ 32} = 0.009221$$

Sorting the Cosine measure of the outcome that was calculated based on the Term frequencies of the documents d_1, d_2, d_3,…,d_{10} will give us the following:

Table 23: Results on un-weighted data on the Cosine measure

Rating	**Function**	**Item**	**How close is the document to the object (d_2)?**
1	$S(d_2,d_5)$	Document 2 compared to Document 5	0.006830601
2	$S(d_2,d_1)$	Document 2 compared to Document 1	0.007796881
3	$S(d_2,d_2)$	Document 2 ***compared to itself***	0.008196721
4	$S(d_2,d_{10})$	Document 2 compared to Document 10	0.009221311
5	$S(d_2,d_6)$	Document 2 compared to Document 6	0.00942623
6	$S(d_2,d_4)$	Document 2 compared to Document 4	0.009562842
7	$S(d_2,d_3)$	Document 2 compared to Document 3	0.009922347

8	$S(d_2,d_7)$	Document 2 compared to Document 7	0.010718789
9	$S(d_2,d_8)$	Document 2 compared to Document 8	0.013173302
10	$S(d_2,d_9)$	Document 2 compared to Document 9	0.016393443

It was observed that using term frequencies in our computations yields a result where the metrics using Cosine similarity measure gives an interesting result. Comparing a document to itself without standardizing data in the computations yields a result which ranks third on the table. This is clear indication that using term frequencies includes errors from the documents, which would include noise and other errors. Using standardized data helps in improving accuracy of results.

4.5.2 Verification based on Term Weights

We have the following weights of the ten documents:

d_2*d_i	$d_2.d_1$	$d_2.d_2$	$d_2.d_3$	$d_2.d_4$	$d_2.d_5$	$d_2.d_6$	$d_2.d_7$	$d_2.d_8$	$d_2.d_9$	$d_2.d_{10}$
Outcome	3.784715	6.239353	5.545708	2.167639	3.955580	3.978167	3.167639	1.185042	5.818119	2.939995

We also have

$\|\|d_i\|\|$	$\|\|d_1\|\|$	$\|\|d_2\|\|$	$\|\|d_3\|\|$	$\|\|d_4\|\|$	$\|\|d_5\|\|$	$\|\|d_6\|\|$	$\|\|d_7\|\|$	$\|\|d_8\|\|$	$\|\|d_9\|\|$	$\|\|d_{10}\|\|$
Outcome	7.784715	6.239353	6.066322	8.077454	8.859156	9.219310	8.077454	2.789598	13.000000	5.673987

We therefore, have the following Cosine similarity measures from the data we have above:

1. For $S(d_2, d_1)$:

$$S(d_2, d_1) = Cos(d_2, d_1) = \frac{d_2 * d_1}{\|d_2\| \|d_1\|} = \frac{3.784715}{6.239353\ x\ 7.784715} = 0.077920$$

2. For $S(d_2, d_2)$:

$$S(d_2, d_2) = Cos(d_2, d_2) = \frac{d_2 * d_2}{\|d_2\| \|d_2\|} = \frac{6.239353}{6.239353 x\ 6.239353} = 0.160273$$

3. For $S(d_2, d_3)$:

$$S(d_2, d_3) = Cos(d_2, d_3) = \frac{d_2 * d_3}{\|d_2\| \|d_3\|} = \frac{5.545708}{6.239353\ x\ 6.066322} = 0.146518$$

4. For $S(d_2, d_4)$:

$$S(d_2, d_4) = Cos(d_2, d_4) = \frac{d_2 * d_4}{\|d_2\| \|d_4\|} = \frac{2.167639}{6.239353\ x\ 8.077454} = 0.146518$$

5. For $S(d_2, d_5)$:

$$S(d_2, d_5) = Cos(d_2, d_5) = \frac{d_2 * d_5}{\|d_2\| \|d_5\|} = \frac{3.955580}{6.239353\ x\ 8.859156} = 0.071561$$

6. For $S(d_2, d_6)$:

$$S(d_2, d_6) = Cos(d_2, d_6) = \frac{d_2 * d_6}{\|d_2\| \|d_6\|} = \frac{3.978167}{6.239353 \; x \; 9.219310} = 0.069158$$

7. For $S(d_2, d_7)$:

$$S(d_2, d_7) = Cos(d_2, d_7) = \frac{d_2 * d_7}{\|d_2\| \|d_7\|} = \frac{3.167639}{6.239353 \; x \; 8.077454} = 0.062852$$

8. For $S(d_2, d_8)$:

$$S(d_2, d_8) = Cos(d_2, d_8) = \frac{d_2 * d_8}{\|d_2\| \|d_8\|} = \frac{1.185042}{6.239353 \; x \; 2.789598} = 0.068085$$

9. For $S(d_2, d_9)$:

$$S(d_2, d_9) = Cos(d_2, d_9) = \frac{d_2 * d_9}{\|d_2\| \|d_9\|} = \frac{5.818119}{6.239353 \; x \; 13.000000} = 0.071730$$

10. For $S(d_2, d_{10})$:

$$S(d_2, d_{10}) = Cos(d_2, d_{10}) = \frac{d_2 * d_{10}}{\|d_2\| \|d_{10}\|} = \frac{2.939995}{6.239353 \; x \; 5.673987} = 0.083046$$

Our main interest is to identify the text from the documents that would be the best identifier of the online user. The details of the digital object of an applicant of identity and verification, which in our case is represented by the identifying attributes, would need to accurately match attributes of verification. We therefore, consider the importance of attributes that is in the corpus of ten documents. The table below shows the documents that are sorted in the order of importance; in this case, the documents would represent the applicants that are being subjected for verification by the process of authentication.

Table 24: Results on using weighted data on the proposed model

Rating	**Function**	**Documents compared**	**How close is the document to the object (d_2)?**
1	**$d_2 * d_2$**	**Document 2 *compared to itself***	**0.160273**
2	$d_2 * d_3$	Document 2 compared to Document 3	0.146518
3	$d_2 * d_{10}$	Document 2 compared to Document 10	0.083046
4	$d_2 * d_1$	Document 2 compared to Document 1	0.077920
5	$d_2 * d_9$	Document 2 compared to Document 9	0.071730
6	$d_2 * d_5$	Document 2 compared to Document 5	0.071561
7	$d_2 * d_6$	Document 2 compared to Document 6	0.069158
8	$d_2 * d_8$	Document 2 compared to Document 8	0.068085
9	$d_2 * d_7$	Document 2 compared to Document 7	0.062852
10	$d_2 * d_4$	Document 2 compared to Document 4	0.043010

From this table we see that it was important to normalize the Term frequencies from the documents so as to remove the errors from data. Without normalizing the data, we have the rating of the document affected to a point that the document compared to itself shows deficit in the content of terms. Removing the errors through normalization done by term weighting of the data from the corpus of the ten documents gives the rating where document 2 is compared to itself becomes first in rating. This is the natural expectation of the outcome of this process.

From the computations, we have been able to show that when we apply our metrics model on ten different documents, we have different results. Applying the model on two documents that have same identity attributes, the model is able to identify the closeness of the identity attributes to each other to be 100%. The model is able to identify that the two documents had the two sets of identity attributes match by 100%. This notion is very important for matching an online applicant for identification to oneself. This implies that if an applicant has the identity attributes that match the identification by 100%, then this would be the owner of the identification that is being sought; this would be the legitimate owner of the identity. We also noticed that when we applied the metric model on two different documents that contained two different sets of identity attributes, the result was very clear that the two documents did not match. The percentage of matching was less that 100%, depending on the identity attributes contained therein.

We have just established that when an online application or applications from multiple users for authentication, Cosine Similarity measure could help us to accurately identify who the true owner of the digital identity would be. This indicates that Cosine Similarity measure could be a very strong tool in information security to add another level in authentication. Coupled with other techniques, we could build a robust system in information security for Digital Identity management.

4.5.3 Results on Metrics Model

The table below shows the top ten identity attributes from the ten documents where *TF*IDF* term weighting was applied. Picking identity attributes that have been found to be higher in terms of weighting would help us identify the owner of the identity attributes for online identity claimant. Applying developed Identity Attribute Metrics, which was developed using the Cosine Similarity measure we obtain the following results:

Table 25: List of top ten identity attribute from the proposed model

ZAMBIA

ATTRIBUTE	Term (tf_i)										Total	TF_r*IDF_r $\sum_{i=1}^{n} tf_i * log\frac{N}{idf_i}$
	Airspace (d_1): tf_1	Residence Permit (d_2): tf_2	Visiting Visa (d_3): tf_3	Consent form (d_4): tf_4	FarmSmallholding (d_5): tf_5	Residential land (d_6): tf_6	Aquaculture Fund (d_7): tf_7	Borehole Form (d_8): tf_8	Health Prof Council (d_9): tf_9	Immovable Property (d_{10}): tf_{10}		
Home address	**1**	**1**	**2**	**3**	**4**	**1**	**3**	**0**	**1**	**3**	**19**	**7.86397063**
First name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
Middle name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
Last name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
ID Number	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**7**	**7.00000000**
issuing authority	**1**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**7**	**7.00000000**
Work telephone number	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**6**	**6.00000000**
Work address	**1**	**1**	**0**	**2**	**4**	**1**	**0**	**0**	**1**	**3**	**13**	**5.40987908**
Home telephone number	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**5**	**5.00000000**
Home email address	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**5**	**5.00000000**

Table 24 gives a list of the top ten identity attributes from the quantification of the identity attributes after applying our model, Cosine similarity function.

4.6 Summary

This chapter has presented results from field work of the research where primary data were collected and analyzed. Data extracted from secondary data using text mining tools and analyzed using statistical methods, tools and techniques has been presented in this chapter. Analysis of data has extensively employed various statistical techniques. A framework influencing this research has been identified. It has been established that data would need to be standardized before we could use it to enhance accuracy on the metrics. The chapter has also presented the model that has been proposed and the output of the test of the model on the data that was text mined and had its data weighted to standardize the data. The chapter has a presentation of the outcome of the comparison of data from primary data of various organizations and countries to secondary data. The model has been tested on secondary data and compared to that of primary data. From section 4.6.2, the model has been verified to be able to identify the owner of the identity attributes that would identify an online user. The chapter has indicated that there are identity attributes that would be key to identify an online service user or digital identity claimant.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1 Introduction

This section discusses the results of this study that is in chapter four of the dissertation, it presents the answers to our research problem and objectives. The chapter concludes the outcome of this research.

5.2 Sources of key identity attributes

We observed that some identity attributes were found to be more important than the others, in a given collection, as they ranked high using statistical mean. The mean scores of the identity attributes indicated that some attributes were more commonly used than the others. The attributes that were uniquely able to identify an entity were found to be more important than the others. The major sources of identity attributes were documents that were used for collecting identity attributes or identity tokens of entities.

Identity attributes differ from organization to organization in the level of importance to identify an entity. However, some attributes rank high across different organizations. It is also observed that identity attributes of high importance differ from country to country or region to region. The identity attributes affecting perceived importance of identity tokens differ from one organization to the other.

Organizations gather identity attributes from their customers through service application forms. Once these identity attributes are collected, they are used on the tokens of identity for identification of entities or individuals. Therefore, to extract identity attributes, we have to go to documents that are used for gathering identifiers of entities. Identity attributes have different levels of importance in identifying an entity.

Analysing primary data and secondary data indicated consistency on some identtity attributes over the others. After obtaining the term weight of data on primary data and on secondary data, we find the consistency of terms that are perceived to be more imporetant by the others.

It was observed that it was important to standardise data before use in the metrics as this removes errors and at the same time improves accuracy in the computations in the metrics.

5.3 Constructs influencing token usage

Identity attributes from identity tokens could be classified into categories, this was observed in all the organizations that were surveyed. We had five variables to classify these identity attributes, we called these variables as constructs. The constructs varied in importance from one organization to the other. Some constructs had more influence in the adaptation of use of identity tokens. Ease of use, Usefulness, and Trust were key predictors of how the use of identity tokens would be influenced.

The five constructs as demonstrated by their strong mean scores, showed that the perception of the identity tokens affected their use. All the organizations tested in this research showed that the constructs were very useful. However, some organizations had stronger perception of importance of the identity tokens. For instance, for banking organization, usefulness was ranked highest (means score 4.2 out of 5), followed by Trust & Ease of use (3.8 out of 5), user satisfaction (3.7 out of 5) and the least was image or status (3.6 out of 5).

Demographics played a major role in influencing perception of importance of use of the identity tokens. Gender analysis showed that more men responded more on the perception of importance of identity tokens than women. It was also found that marital status influenced the perception of importance of identity tokens. The married ones perceived had highest in the affirmative of importance of identity documents followed by singles, then divorced, with other groups being last. The age group between 21 years and 30 years had the highest score, this implied that this age group trusted technology more than other age groups. It was observed that the older the respondents, the least they perceived importance of identity tokens. Education was found to impact perception on the importance of identity documents. Frequency of use of the identity token was affected by perception of the level of importance of a given construct.

5.4 Research data

There was a relationship between primary data and secondary data. Identity attributes that were found to be very key in primary data were also very key in secondary data. The ones that were not key in primary data were also not very key in secondary data.

Standardizing data through normalization improved in helping to uniquely identify an entity. When term weighting schemes were applied on data, we were able to identify very key identity attributes in a given corpus. The attributes that were frequent in the corpus were not necessarily the most important ones. The weighting schemes helped us to identify the

key identity attributes in identifying an entity and we were able to rank the level of importance of these identity attributes.

Comparing the term weights from primary data (using Shannon's entropy) and the TF*IDF term weights on secondary data, it was observed that 50% to 70% of the terms of the top ten of the nineteen terms from respective organizations and countries, were popular. It was also found that using the TF*IDF weighting, 80% of the top ten of the terms were more popular.

5.5 Extracting key identity attributes

Text mining of identity attributes was found to be very useful in extracting of identity attributes from identity tokens. Documents have to be prepared in the format that is suitable for extracting of text before it is analyzed.

Our developed mathematical model which is based on distance metrics was able to quantify the identity attributes and rate them in their level of importance. The model was found to be very useful in the unique identification of a claimant of an identity. The identity attributes of the legitimate identity were found to match the object identity attributes (which were a match of itself) by 100% using our model. Other sets of identity attributes could not match the object identity attributes by 100% using the developed model.

5.6 Conceptual framework

The research framework that we adopted was the Technology Acceptance Model (TAM) it was found to have a very strong influence on this research. Relationships on the constructs were tested in the study and found that out of the five constructs that we were testing, three of them had more influence in the acceptance of the use of the entity documents; an identity token would often be used once it was trusted. An identity token which is easy for use becomes popular to users. It was found that a document would become popular in use once users found it to be important as the importance of the document would draw them to the document. TAM model has shown its influence on this research by demonstrating that the actual use of a given identity token is strongly affected by the usefulness of the token, the ease of use and the security features the identity token has.

5.7 Proposed model

Testing the proposed Cosine Similarity measure as an Identity Attribute Metric Model verifies that this model can identify the document that uniquely has its identity attributes similar to itself as the the highest and hence identify a claimant of the digital identity as the

legitimate owner. The matching of the identity attribute give a 100% match, the model would be able to identify the claimant from one to multiple claimants. This would help in improving security on identifying the legitimate digital identity owner of a specific identity. Only such an owner should be given access to online assets, services, or attention.

5.2 Conclusion

The study was able to identify identity attributes that were important to identify an online identity claimant. The key identity attributes were extracted from identity tokens; for primary data, we were able to extract from the entity documents like service application forms. For secondary data, identity attributes were extracted using text mining techniques using text mining tools, which are data mining tools. Data were gathered from internet as pdf documents which were then subjected to text mining and text analysis to be able to use text analysis methods to understand the closeness of identity attributes. The study was able to develop an identity attribute metrics model, a mathematical model, using the Cosine Similarity distance measure and show that Cosine similarity measure could be used to quantify the identity attributes. The model was tested on data that were text mined and standardized; the outcome showed that the Cosine Similarity model could identify the unique owner of the digital identity attributes. The model also showed that it could identify a legitimate identity claimant from multiple claims as the results of applying the model were able to show a set of identity attributes that could uniquely match the object identity attributes. This model could add value to enhancing security in online activities by validating the true owner of a digital identity. This model could also be used in multi modal tools for a robust online digital solution to arrest the challenges of online information security.

5.3 Future research interest

There is need to develop and implement the outcome of this research and build a multimodal solution which will consolidate previous works in this area and come up with a single robust solution. Such a solution should recognize how much threat would be rid of in the online services and activities.

REFERENCES

[1] F. Kabwe and J. Phiri, "A review of Identity Attribute Metrics Modeling based on Distance Metrics," pp. 127-132, 2018.

[2] C. K. Kalila, "Committee on Health, Community Development and Social Services," National Assembly of Zambia, Lusaka, 19 June 2020.

[3] Ministry Editors, "Ministry of Community Development, Mother and Child Health, National Social Protection Policy: Reducing poverty, inequality and vulnerability," Ministry of Community Development, Mother and Child Health, Lusaka, 6 June, 2014.

[4] M. A. Jibrin, M. N. Musa and T. Shittu, "Effects of Internet on the Academic Performance of Tertiary Institutions' Students in Niger State, Nigeria," *International Journal of Education, Learning and Training,* vol. 2, no. 2, pp. 57-69, November, 2017.

[5] E.-I. Apăvăloaie, "The impact of the Internet on the business environment: Emerging Markets Queries in Finance and Business," *Elsevier,* vol. 15, no. 1, p. 951 – 958, 2014.

[6] J. I. Agbinya, N. Mastali, R. Islam and J. Phiri, "Design and implementation of multimodal digital identity management system using fingerprint matching and face recognition," *7th International Conference on Broadband Communications and Biomedical Applications,* vol. 1, no. 1, pp. 272-278, November, 2011.

[7] J. Phiri, T.-J. Zhao, C. H. Zhu and J. Mbale, "Using Artificial Intelligence Techniques to Implement a Multifactor Authentication System," *International Journal of Computational Intelligence Systems,* vol. 4, no. 4, pp. 420-430, June, 2011.

[8] M. S. Gaigole and M. A. Kalyankar, "The Study of Network Security with Its Penetrating Attacks and possible Security Mechanisms," *International Journal of Computer Science and Mobile Computing,* vol. 4, no. 5, pp. 728-735, May, 2015.

[9] J. I. Agbinya, R. Islam and C. Kwok, "Development of Digital Environment IdentiTY (DEITY) System for Online Access," *2008 Third International Conference on Broadband Communications, Information Technology & Biomedical Applications,* vol. 1, no. 1, pp. 1-12, December, 2018.

[10] S. M. Mahmood, B. M. Amen and R. M. Nabi, "Mobile Application Security Platforms Survey," *International Journal of Computer Applications,* vol. 133, no. 2, pp. 40-46, January, 2016.

[11] G. Quaglio, "European Parliamentary Research Service Blog: Empowering through knowledge," European Eunion, 18 February 2019. [Online]. Available: https://epthinktank.eu/2019/02/18/how-the-internet-can-harm-us-and-what-can-we-do-about-it/. [Accessed 29 May 2020].

[12] RANZCP, "The impact of media and digital technology on children and adolescents," The Royal Australian and New Zealand College of Psychiatrists, 29 May 2018. [Online]. Available: https://www.ranzcp.org/news-policy/policy-and-advocacy/position-statements/the-impact-of-media-and-digital-technology-on-chil. [Accessed 29 May 2020].

[13] Hope Computer, "What are the advantages of the Internet?," Computer Hope, 12 01 2019. [Online]. Available: https://www.computerhope.com/issues/ch001808.htm. [Accessed 29 May 2020].

[14] Computer Hope, "What are the disadvantages of the Internet?," Computer Hope, 31 12 2020. [Online]. Available: https://www.computerhope.com/issues/ch001810.htm. [Accessed 29 May 2020].

[15] Executive Committe, World Economic Forum, "Identity in a Digital World-A new chapter in the social," 2018 World Economic Forum, Cologny/Geneva, September, 2018.

[16] A. Pfitzmann and M. Hansen, "Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management –," 15 February 2008. [Online]. [Accessed 29 May 2020].

[17] G. B. Ayed, "Architecting User-Centric Privacy-as-a-Set-of-Digital Identity-Related Privacy Framework," *Springer Theses - Recognizing Outstanding Ph.D. Research,* vol. 1, no. 1, pp. 11-55, March, 2014.

[18] S. Pal, M. Hitchens and V. Varadharajan, "Modeling identity for the internet of things: Survey, classiffication and trends," *2018 12th International Conference on Sensing Technology (ICST),* vol. 1, no. 1, pp. 45-51, 2018.

[19] A. Hovav and R. Berger, "Tutorial: Identity Management Systems and Secured Access Control," *Communications of the Association for Information Systems,* vol. 25, no. 1, pp. 531-570, December, 2009.

[20] S. Clauß and M. Köhntopp, "Identity management and its support of multilateral security," *Computer Networks*, vol. 37, no. 2, pp. 205-219, 2001.

[21] I. Agudo-Ruiz, "Digital Identity and Identity Management Technologies," *UPGRADE - the European Journal for the Informatics Professional*, vol. 11, no. 1, pp. 6-12, February, 2010.

[22] M. Jain and M. Singh, "Identity Based and Attribute Based Cryptography: A Survey," *International Journal of Engineering, Management & Sciences (IJEMS)*, vol. 2, no. 5, pp. 88-92, May, 2015.

[23] A. Jøsang, "Relationship between entities, identities and attributes / identifiers," Wikipedia, The free Encyclopedia, 5 February 2009. [Online]. Available: https://commons.wikimedia.org/wiki/File:Identity-concept.svg. [Accessed 25 May 2020].

[24] X. Zhu and Y. Badr, "Identity Management Systems for the Internet of Things: A Survey Towards Blockchain Solutions," *sensors*, vol. xx, no. 5, pp. 1-18, 1 December, 2018.

[25] J. Phiri and J. I. Agbinya, "Using Artificial Neural Networks to Implement Information Fusion in Digital Identity Management Systems," *International Journal of Computational Intelligence Systems*, vol. 4, no. 4, pp. 420-430, June, 2011.

[26] S. Sittampalam, *Digital Identity Modelling and Management*, Sydney: University of Technology Sydney, Australia, 2005.

[27] R. Bhaskar and B. Kapoor, Information Technology Security Management, Computer and Information Security Handbook, Sydney: Morgan Kaufmann Publishers, an imprint of Elsevier, 2009.

[28] J. Vacca, Computer and Information Security Handbook, Burlington, USA: Morgan Kaufmann Publishers, an imprint of Elsevier, 2009.

[29] A. M. Al-Khouri, "eGovernment Strategies The Case of the United Arab," *European Journal of ePractice*, vol. 1, no. 17, pp. 126-150, September 2012.

[30] P. J. Windley, Digital Identity: Unmasking Identity Management Architecture (IMA), New York: O'Reilly Media, 2005.

[31] G. Roussos, D. Peterson and U. Patel, "Mobile Identity Management: An Enacted View," *International Journal of Electronic Commerce*, vol. 8, no. 1, pp. 81-100, 2003.

[32] C. Satchell, G. Shanks, S. Howard and J. Murphy, "Identity crisis: User perspectives on multiplicity and control in federated identity management," *Behaviour & Information Technology*, vol. 30, no. 1, pp. 51-62., 2011.

[33] International Standard ISO/IEC, "Text for ISO/IEC 2nd WD 29003 - Information technology – Security techniques – Identity proofing," International Standard ISO/IEC, Berlin, Germany, 2013.

[34] M. Nieles, K. Dempsey and V. Y. Pillitteri, "An Introduction to Information Security: NIST Special Publication," *NIST Special Publication 800-12 Revision 1*, pp. 1-91, 21 June 2017.

[35] Merriam-Webster SINCE 1828, "security," 25 May 2020. [Online]. Available: www.m-w.com/dictionary/security. [Accessed 2 June 2020].

[36] W. Stallings, Cryptography and Network Security Principles and Practices, Fourth Edition ed., London: Pearson Education, Inc., 2005.

[37] T. J. Smedinghoff, "Introduction to Online Identity Management," Smedinghoff2008IntroductionTO, Chicago, USA, 2008.

[38] A. M. Al-Khouri, *Federated E-Identity Management Across The Gulf Cooperation Council*, vol. 1, Abu Dhabi, United Arab Emirates: International Journal of Public Information Systems, 2013.

[39] P. Soneka and J. Phiri, "A Model for Improving E-Tax Systems Adoption in Rural Zambia Based on the TAM Model," *Open journal of business and management*, vol. 7, no. 1, pp. 908-919, January, 2019.

[40] V. Venkatesh, "User Acceptance of Information Technology, Toward a unified view," vol. 27, no. 3, pp. 425-478, September 2003.

[41] F. Kabwe and J. Phiri, "A Framework For Digital Identity Management," *Proceedings of the International Conference in ICT (ICICT2019)*, vol. 1, no. 1, pp. 127-132, 2019.

[42] R. Matikiti, M. Mpinganjira and M. Roberts-Lombard, "Application of the Technology Acceptance Model and the Technology– Organisation–Environment Model to examine social media marketing use in the South African tourism industry," *South African Journal of Information Management*, vol. 20, no. 1, pp. 1-12, 2018.

[43] L. Miriam and P. Chiky, *Identity Management in Information Age Government Exploring Concepts, Definitions, Approaches and Solutions*, Wellington, New Zealand: Victoria University of Wellington, 2008.

[44] Department of Homeland Security - USA;, "Cyber security - Identity Management," Department of Homeland Security of the United States of America, 21 7 2017. [Online]. Available: https://www.dhs.gov/science-and-technology/idm. [Accessed 8 June 2020].

[45] Organisation for Economic Co-operation and Development, "OECD E-Government Project, E-Government for better Government," OECD, Paris, March, 2005.

[46] R. Joosten, D. Whitehouse and P. Duquenoy, "Towards a meta model for identity terminology," *In Pre-proceedings of the IFIP/FIDIS Internet Security & Privacy Summer School*, vol. 1, no. 1, p. 141–146, 2008.

[47] J. Phiri and J. I. Agbinya, "Modelling and Information Fusion in Digital Identity Management Systems," *International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06)*, vol. 1, no. 1, pp. 1-6, 2006.

[48] W. Inambao, J. Phiri and D. Kunda, "Digital Identity Modelling for Digital Financial Services in Zambia," *Journal on Communication Technology (ICTACT)*, vol. 9, no. 3, pp. 1829-1837, September, 2018.

[49] Grassi, P., A.; Garcia, M., E.; Fenton, J., L., "Digital Identity Guidelines," NIST Special Publication 800-63-3, Los Altos, California, June, 2017.

[50] Lohokare, Archit, *NIST 800-63-3 Digital Identity Guidelines – A Primer*, Los Altos, California: National Institute of Standards and Technology, June, 2017.

[51] Sven Wohlgemuth;, D3.1: Structured Overview on Prototypes and Concepts, Freiburg, Germany: Future of Identity in the Information Society, September, 2005.

[52] J. Phiri, *Digital Identity Management System*, Cape Town: University of Western Cape, South Africa, 2007.

[53] J. Jackson and J. I. Agbinya, "Modelling and Information fussion in digital identity management systems," *Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies*, vol. 1, no. 1, pp. 181-186, April, 2006.

[54] SignKeys, "Digital ID is not Good Digital Identity," SignKeys, 25 January 2018. [Online]. Available: https://www.signkeys.com/products/good-digital-identity. [Accessed 15 April 2020].

[55] D. J. Solove, *Understanding Privacy*, Cambridge, Massachusetts: Harvard University Press, 2008.

[56] D. M. Rousseau, S. B. Sitkin, R. Burt and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 393-404, 1998.

[57] N. Luhmann, Trust and Power, Chichester, England: Wiley, 1979.

[58] R. C. Solomon and F. Flores, Building Trust in Business, Politics, Relationships, and Life, Oxford, Enland: Oxford University Press, 2001.

[59] B. Charulatha, P. Rodrigues, T. Chitralekha and A. Rajaraman, "A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. Special Issue, no. Special Issue, pp. 1-5, 2013.

[60] S. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematics and methods in Applied Sciences*, vol. 1, no. 4, pp. 300-307, 2007.

[61] C. D. Schultz, *A Trust Framework Model for Situational Contexts*, Ontario, Canada: International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, October 30-November 1, 2006.

[62] R. Peck, C. Olsen and J. L. Devore, Introduction to Statistics and Data Analysis, Boston, USA: Cengage Learning, 2016.

[63] H. Abdi, Encyclopedia of Research Design, N. J. Salkind, Ed., Thousand Oaks, California: SAGE Publications, Inc, 2010.

[64] G. Williams, Data Mining with Rattle and R - The Art of Extracting Data for Knowledge Recovery, R. Gentleman, K. Hornik and G. G. Parmigiani, Eds., New York, USA: Springer, 2011.

[65] A. Singhal, Data Warehousing and Data Mining Techniques for Cyber Security - Advances in Information Security, S. Jajodia, Ed., New York, USA: Springer, 2007.

[66] D. L. Olson and D. Delen, Advanced Data Mining Techniques, New York, USA: Springer-Verlag Berlin Heidelberg, 2008.

[67] J. Han and K. M., Data Mining: Concepts and Techniques, 2nd ed., Oxford, UK: Elsevier Inc., 2006.

[68] S. Prabhu and N. Venkatesan, Data Mining and Warehousing, New Delhi, India: New Age International (Pvt) Limited, January, 2001.

[69] J. Han, M. Kamber and J. Pei, Data Mining": Concepts and Techniques, 3rd ed., Massachusetts, USA: Elsevier Inc., 2012.

[70] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters,* vol. 22, no. 5, p. 563–582, 2001.

[71] M. M. Suarez-Alvarez, D. Pham, M. Y. Prostov and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," *Journal of the Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences,* vol. 468, no. 1, pp. 2631-2632, September, 2012.

[72] *Problem 4 : Term Weighting Schemes in Information Retrieval,* 1997.

[73] F. H. Lotfi and R. Fallahnejad, "Imprecise Shannon's Entropy and Multi Attribute Decision Making," *Entropy,* vol. 12, no. 1, pp. 53-62, 2010.

[74] A. Delgado and A. B., "A Computational Model Based on Shannon Entropy to Analyze Social Development in South America," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 8, no. 11, pp. 2515-2522, September, 2019.

[75] *Understanding Shannon's Entropy metric for Information,* March 24, 2014.

[76] M. K. Chinyemba and J. Phiri, "Gaps in the Management and Use of Biometric Data: A Case of Zambian Public and Private Institutions," *Zambia Information Communication Technology (ICT) Journal Journal,* vol. 2, no. 1, pp. 35-43, 2018.

[77] S. Ibou, C. A. Aziz and N. Oumar, "Toward an Attribute-Based Digital Identity Modeling for Privacy Preservation," *ArXiv,* vol. 3, no. 1, pp. 1-5, Sep 13, 2019.

[78] J. Phiri, T. J. Zhao and J. Mbale, "Identity Attributes Mining, Metrics Composition and Information Fusion Implementation Using Fuzzy Inference System," *Journal of Software,* vol. 6, no. 6, pp. 1025-1033, June 2011.

[79] J. Phiri, D. M. Zulu, J. I. Agbinya and T. Zhao, "Fuser Block Technologies Performance Based on Identity Attributes Metrics Models," *Pan African International Conference on Information Science, Computing and Telecommunications,* vol. 1, no. 1, pp. 188-193, 2013.

[80] E. Oliveira and D. B. Filho, "Automatic classification of journalistic documents on the Internet," *TransInformação,* vol. 29, no. 3, pp. 245-255, 2017.

[81] S. Büttcher, C. L. A. Clarke and G. V. Cormack, Information retrieval: Implementing and evaluating search engines, Cambridge, Massachusetts, USA: Mit Press, 2010.

[82] J. Zobel and A. Moffat, "Exploring the similarity space," *SIGIR Forum,* vol. 32, no. 1, p. 18–34, 1998.

[83] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation,* vol. 28, no. 1, p. 11–21, 1972.

[84] Wikipedia, "Information theory," May 2020. [Online].

[85] *Information Theory Primer - with an Appendix on Logarithms,* Apr, 2018.

[86] O. Rioul, "This is IT: A Primer on Shannon's Entropy and Information," *L'Information, Séminairé Poincaré,* vol. 23, no. 1, pp. 43-77, 2018.

[87] L. A. Shalabi, Z. Shaaban and B. Kasasbeh, "Data Mining: A Preprocessing Engine," *Journal of Computer Science,* vol. 2, no. 9, pp. 735-739, 2006.

[88] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, Burnaby, British Columbia, Canada: Morgan Kaufmann, 2001.

[89] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Trans. Pattern Anal. Mach. Intell,* vol. 3, no. 1, p. 66–75, 1981.

[90] B. Larsen and C. Aone, *Fast and effective text mining using linear-time document clustering,* New York, USA, 1999, pp. 16-22.

[91] P. Sitikhu, K. Pahi, P. Thapa and S. Shakya, *A comparison of semantic similarity methods for maximum human interpretability,* Kathmandu, Nepal, November, 2019.

[92] D. J. Weller-Fahy, B. J. Borghetti and A. A. Sodemann, "Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection," *IEEE Communication Surveys & Tutorials,* vol. 17, no. 1, pp. 70-91, 2015.

[93] K. Maher and M. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 7, no. 4, pp. 1715-1720, 2016.

[94] O. E. Oduntan, I. A. Adeyanju, A. S. Falohun and O. O. Obe, "A Comparative Analysis of Eucledian Distance and Cosine Similarity Measure for Automated Essay-Type Grading," *Journal of Engineering and Applied Sciences,* vol. 13, no. 11, pp. 4198-4204, 2018.

[95] E. Korkmaz and G. Üçoluk, "Choosing A Distance Metric for Automatic Word Categorization," *New Methods in Language Processing and Computational Natural Language Learning, ACL,* vol. 1, no. 1, pp. 111-120, 1998.

[96] X. Wu, Z. G. Shi and L. Liu, "Quasi Cosine Similarity Metric Learning," *Asian Conference on Computer Vision,* vol. 9010, no. 1, pp. 194-205, 2015.

[97] N. Munot and S. S. Govilkar, "Comparative study of text summarization methods," *International Journal of Computer Applications,* vol. 102, no. 12, pp. 33-37, 2014.

[98] S. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science,* vol. 27, no. 1, pp. 129–146, , 1976.

[99] *Why Language Models and Inverse Document Frequency for Information Retrieval,* March, 2014.

[100] L. Chen and C. Chang, "A New Term Weighting Method by Introducing Class Information for Sentiment Classification of Textual Data," *Proceeding of the International MultiConference of Engineers and Computer Scientists,* vol. 1, no. 1, pp. 16-18, 2011.

[101] G. Salton and C. Buckley, "Term-Weighting approaches in Automatic Text Retrieval," *Information Processing and Management,* vol. 24, no. 5, p. 513–523, 1988.

[102] C. Buckley, "The Importance of Proper Weighting Methods," *Proceedings of the workshop on Human Language Technology—HLT'93,* vol. 1, no. 1, pp. 349-352, August, 1993.

[103] M. Umadevi, "Document Comparison based on Tf-Idf Metric," *International Research Journal of Engineering and Technology (IRJET),* vol. 7, no. 2, pp. 1546-1550, February, 2020.

[104] S. Teufel, *Term Weighting and the Vector Space Model Information Retrieval, Computer Science Tripos Part II,* London, UK: Natural Language and Information Processing (NLIP) Group, 2014.

[105] *Why Language Models and Inverse Document Frequency for Information Retrieval?,* March, 2014.

[106] *Research Design and Methodology,* August, 2019.

[107] I. Akhtar, "Research Design," in *Research in Social Science: Interdisciplinary Perspectives,* 1st ed., New Delhi, New Age International , September, 2016, p. 68.

[108] P. D. Leedy, Practical research: planning and design, New Jersey, USA: Prentice-Hall, 1993.

[109] H. Taherdoost, "Sampling Methods in Research Methodology: How to Choose a Sampling Tech- nique for Research," *International Journal of Academic Research in Management (IJARM),* vol. 5, no. 2, pp. 18-27, 2016.

[110] *Urban Slums Reports: The case of Lusaka, Zambia,* 2003.

[111] Nagel, Urs, "Zambia Human Development Report Industrialisation and Human Development - Poverty reduction through wealth and employment creation," United Nations Development Programme, Lusaka, Zambia, February, 2010.

[112] F. Kamangar and F. Islami, "Sample size calculation for epidemiologic studies: principles and methods," *Archives of Iranian Medicine (AIM),* vol. 16, no. 5, p. May, 2013.

[113] W. S. Browner, T. B. Newman, S. R. Cummings and S. R. Hully, "Getting ready to estimate sample size: hypotheses and underlying principles," *Designing clinical research,* vol. 2, no. 1, pp. 51-63, , 1988.

[114] M. N. Marshall, "Sampling for qualitative research," *Family Practice - an international journal,* vol. 13, no. 6, pp. 522-525,, 1996.

[115] F. D. Davis, R. P. Bagozzi and P. R. Warshaw, "User acceptance of computer technology: a comparison of two theoretical models," *Management Science,* vol. 35, no. 8, pp. 982-1003,, August, 1989.

[116] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies.," *Management Science,* vol. 46, no. 2, pp. 186-204, February, 2000.

[117] J. D. Portz, E. A. Bayliss, S. Bull, R. S. Boxer, B. B. David, K. Gleason and S. Czaja, "Using the Technology Acceptance Model to explore user experience, intent to Use, and Use behavior of a patient portal among older adults with multiple chronic conditions: Descriptive Qualitative Study," *Journal of Medical Internet Research,* vol. 21, no. 4, pp. x-x, 2019.

[118] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly,* vol. 13, no. 3, pp. 319-340, , September, 1989.

[119] Q. Faryadi, "PhD Thesis Writing Process: A Systematic Approach-How to Write Your Methodology, Results and Conclusion," *Creative Education,* vol. 10, no. 1, pp. 766-783, 2019.

[120] C. R. Kathari, Research Methodology: Methods and Techniques, New Delhi, India: New Age International Publishers, 2004.

[121] Curry, P.; Nadalin, A., International Standard ISO/IEC WD2 29003: Information technology - Security techniques - Identity Proofing, S. -. ISO, Ed., Berlin, Germany:

International Organization for Standardization, International Electrotechnical Commission, July, 2013.

[122] S. Greener, Business Research Methods, London, London, United Kigndoms: Ventus Publishing, 2008.

[123] G. B. Giles, Marketing, London, United Kingdom: Macdonald & Evans Ltd., 1974.

[124] M. Islam, "Data Analysis: Types, Process, Methods, Techniques and Tools," *International Journal on Data Science and Technology,* vol. 6, no. 1, pp. 10-15, 2020.

[125] J. Frost, "Difference between Descriptive and Inferential Statistics: Making statistics intuitive," 11 December 2019. [Online]. Available: https://statisticsbyjim.com/basics/descriptive-inferential-statistics. [Accessed 18 March 2020].

[126] J. H. McDonald, "Inferential Statistics and Data Interpretation - in Handbook of Biological Statistics," vol. 1, no. 1, pp. 131-141, 2014.

[127] R. Rana and R. Singhal, "Chi-square Test and its Application in Hypothesis Testing," *Journal of the Practice of Cardiovascular Sciences,* vol. 1, no. 1, pp. 69-71, January-April, 2015.

[128] E. B. Satake, "Statistical Methods and Reasoning for the Clinical Sciences Evidence-Based Practice," *Plural Publishing Inc.,* vol. 1, no. 1, p. 1-19, 2015.

[129] A. Delgado and B. Ayala, "A Computational Model Based on Shannon Entropy to Analyze Social Development in South America," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 8, no. 11, pp. 2515-2522, September, 2019.

[130] A. Moreno and T. Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 3, no. 6, pp. 57-64, 2016.

[131] H. Schutze and C. Silverstein, "Projections for Efficient Document Clustering," *ACM SIGIR,* vol. 14, no. 1, pp. 74-81, 1997.

[132] R. Bekkerman, R. El-Yaniv, Y. Winter and N. Tishby, "DistributionalWord Clusters vs. Words for Text Categorization," *Journal of Machine Learning Research,* vol. 3, no. 1, pp. 1183-1208, 2003.

[133] G. Heyer, U. Quasthoff and T. Wittig, "Text Mining: Wissensrohstoff Text," *17th international Conference on Computational Linguistics,* vol. 1, no. 4, pp. 131-134, 2008.

[134] O. Azeroual, G. Saake, M. Abuosba and J. Schöpfel, *Text data mining and data quality management for research information systems in the context of open data and open science,* Rabat, Morocco: Third International Colloquium on Open Access, 2018.

[135] S. Vijayarani, J. Ilamathi and S. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *International Journal of Computer Science & Communication Networks,* vol. 5, no. 1, pp. 7-16, , 2015.

[136] L. Anthony, "AntConc Manual," Laurence Anthony, Birmingham, UK, June, 2018.

[137] M. U. Maheswari and J. G. R. Sathiaseelan, "Text Mining: Survey on Techniques and Applications," *International Journal of Science and Research (IJSR),* vol. 6, no. 6, pp. 1660-1664, 2017.

[138] D. Inkpen, T. S. Paribakht, F. Faez and E. Amjadian, "Term Evaluator: A Tool for Terminology Annotation and Evaluation," *International Journal of Computational Linguistics and Applications*, vol. 7, no. 2, pp. 145-165, 2016.

[139] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks,* vol. 16, no. 3, pp. 645-678, May, 2005.

[140] G. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. 18, no. 1, pp. 613-620, 1975.

[141] I. Silva, J. Souza and K. Santos, "Dependence among terms in vector space model," *8th International Database Engineering and Applications Symposium, IDEAS,* vol. 1, no. 1, pp. 97-102, July, 2004.

[142] J. H, *Lecture Notes, Math 2331 { Linear Algebra - 4.1 Vector Spaces & Subspaces*, Texas, USA: University of Houston, June, 2020.

[143] W. Lowe, "Towards a theory of semantic space," *in Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* vol. 1, no. 1, pp. 576-581, 2001.

[144] S. K. M. Wong and V. V. Raghavan, "Vector space model of information retrieval: a reevaluation," *in Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval,* vol. 1, no. 1, pp. 167-185, July, 1984.

[145] M. W. Berry, Z. Drmac and E. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM review,* vol. 41, no. 1, pp. 335-362, 1999.

[146] A. Kaur and D. Chopra, "Comparison of Text Mining Tools," *5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions,* vol. 1, no. 1, pp. 108-114, September, 2019.

[147] N. S. J. Math, M. Radovanovic and M. Ivanovic, "Text Mining: Approaches and Applications," *NOVI SAD J. MATH,* vol. 38, no. 3, pp. 227-234, 2008.

[148] Dictionary, Merriam Webster, "Definition of develop," Merriam Webster Dictionary, 10 January 1828. [Online]. Available: https://www.merriam-webster.com/dictionary/develop. [Accessed 12 February 2021].

[149] Vocabulary.com, "Definitions of develop," Vocabulary.com, 15 January 2008. [Online]. Available: https://www.vocabulary.com/dictionary/develop.. [Accessed 24 February 2021].

[150] DICTIONARY.COM, "Develop," DICTIONARY.COM, 14 May 1995. [Online]. Available: https://www.dictionary.com/browse/develop.. [Accessed 24 February 2021].

[151] Lexico, Oxford, "Develop," Oxford Lexico, 12 June 2019. [Online]. Available: https://www.lexico.com/definition/develop. [Accessed 24 February 2021].

[152] G. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. 18, no. 1, pp. 613-620, 1975.

[153] I. Silva, J. N. Souza and K. S. Santos, "Dependence among terms in vector space model," *8th International Database Engineering and Applications Symposium,* vol. 1, no. 1, pp. 97-102, 2004.

[154] W. Lowe, "Towards a theory of semantic space," *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* vol. 1, no. 1, pp. 576-581, 2001..

[155] S. K. M. Wong and V. V. Raghavan, "Vector space model of information retrieval: a reevaluation," *Proceedings of the 7th annual international conference on Research and development in information retrieval,* vol. 1, no. 1, pp. 167-185, 1984.

[156] M. Berry, Z. Drmac and E. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM review,* vol. 41, no. 1, pp. 335-362, 1999.

[157] A. Huang, "Similarity Measures for TextDocument Clustering," *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008),* vol. 1, no. 1, pp. 49-56, 2008.

[158] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Trans. Pattern Anal. Mach. Intell,* vol. 3, no. 1, pp. 66-75, January, 1981.

[159] K. Maher and M. S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 7, no. 4, pp. 1715-1720,, 2016.

[160] A. Griffiths, H. C. Luckhurst and P. Willett, "Using interdocument similarity in document retrieval systems," *Journal of the American Society for Information Science,* vol. 37, no. 1, p. 3–11, 1986.

[161] A. Singhal, "Modern Information Retrieval: A Brief Overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,* vol. 24, no. 4, pp. 35-43, 2001.

[162] M. Oghbaie and M. M. Zanjireh, "Pairwise document similarity measure based on present term set," *Journal of Big Data,* vol. 5, no. 1, p. 52, 2018.

[163] R. Subhashini and V. J. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," *1st international conference integrated intelligent computing,* vol. 1, no. 1, p. 27–31, 2010.

[164] J. D'hondt, J. Vertommen, P. A. Verhaegen, D. Cattrysse and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering," *Information Science,* vol. 180, no. 12, pp. 2341-2358, 2010.

[165] C. L. Tan, W. Huang, S. Y. Sung, Z. Yu and Y. Xu, "Text Retrieval from Document Images Based on Word Shape Analysis," *Applied Intelligence,* vol. 18, no. 1, p. 257–270, 2003.

[166] G. U. and B. Goradiya, "Similarity Measures of web pages using Cosine Similarity," *International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017),* vol. 5, no. 3, pp. 348-351, , 2017.

[167] J. Usharani and K. Iyakutti, "A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval," *International Journal of Engineering Research & Technology (IJERT),* vol. 2, no. 2, pp. 1-5, February, 2013..

[168] M. A. H. Al-Hagery, "Google Search Filter Using Cosine Similarity Measure to Find All Relevant Documents of a Specific Research Topic," *International Journal of Education and Information Technologies,* vol. 10 , no. 1, pp. 229-242, 2016.

[169] M. Patel, "TinySearch- Semantics based Search Engine using Bert Embeddings," *ArXiv Journal,* vol. abs/1908.02451, no. 1, pp. 1-6, August, 2019.

[170] R. i. Subhashin and V. J. S. Kumar, "Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval," *First International Conference on Integrated Intelligent Computing,* vol. 1, no. 1, pp. 27-31, August, 2010.

[171] S. Sohangir and D. Wang, "Improved sqrt-cosine similarity measurement," *Journal of Big Data,* vol. 4, no. 25, p. 25, 2017.

[172] K. Maher and M. S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 7, no. 4, pp. 1715-1720, 2016.

[173] K. Maher and M. S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *International Journal of Computer Science and Information Technologies,* vol. 7, no. 4, pp. 1715-1720, 2016.

[174] S. Sohangir and D. Wang, "Improved sqrt-cosine similarity measurement," *Springer: Journal of big data,* vol. 4, no. 24, pp. 1-13, 2017.

[175] S. Pandit and S. Gupta, "A comparative study on distance measuring approaches for clustering," *International Journal of R in Computer Science,* vol. 2, no. 1, pp. 29-31, 2011.

[176] O. E. Oduntan, I. A. Adeyanju, A. S. Falohun and O. O. Obe, "A comparative analysis of Euclidean distance and Cosine Similarity measure for automated essay-type drading," *Journal of Engineering and Applied Sciences,* vol. 13, no. 11, pp. 4198-4204 , 2018.

[177] M. Mihajlovic and N. Xiong, "Finding the most similar documents using case-based reasoning," *ArXiv,* vol. abs/1911.00262, no. 1, pp. 1-9, 2019.

[178] W. H. Gomaa and A. A. Fahmy, "A survey of Text Similarity Approaches," *International Journal of Computer Applications,* vol. 68, no. 13, pp. 13-18, 2013.

[179] A. Huang, "Similarity Measures for TextDocument Clustering," *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008),* vol. 1, no. 1, pp. 49-56, 2008.

[180] L. Zahrotun, "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Computer Engineering and Applications,* vol. 5, no. 1, pp. 11-18, February, 2016.

APPENDIX

Appendix A: Mean scores of the constructs per organization

Institution and tokens	**Constructs**				
Banks	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Mortgage application form	4.67	4.56	3.88	3.87	4.18
Bank loan application form	4.88	4.47	4.12	3.81	4.12
Credit card/Debit card application form	4.75	4.56	4.33	3.60	4.13
Student loan application form	4.23	3.86	3.93	3.54	3.31
Bank account application form and card	4.84	4.53	4.05	3.80	4.44
Employee ID	4.67	4.63	4.63	4.13	4.47
Club membership cards	4.11	4.25	4.29	3.89	4.00
Average	**4.59**	**4.41**	**4.18**	**3.80**	**4.09**
Insurance	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Life Insurance application form and card	3.15	3.08	2.92	2.50	3.46
Car Insurance application form and disc	4.71	4.31	4.29	4.07	4.38
Medical insurance application form and card	4.33	4.25	4.44	3.89	4.50
Average	**4.06**	**3.88**	**3.88**	**3.49**	**4.11**
Churches	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Baptism certificate	4.81	4.94	4.94	4.38	4.88
Club membership card	5.00	5.00	5.00	5.00	4.50
Average	4.91	4.97	4.97	4.69	4.69
Government	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
National Visa application form and card	5.00	5.00	5.00	5.00	5.00
Title Deeds	5.00	4.50	4.50	3.50	4.50
Marriage certificate	5.00	5.00	4.33	4.67	4.33
Divorce certificate	4.33	4.00	4.00	3.33	3.00
Driving license	5.00	5.00	5.00	4.50	5.00
National tax application form	5.00	5.00	5.00	5.00	5.00
National vehicle registration certificate	5.00	5.00	5.00	4.00	5.00
National birth certificate	5.00	5.00	5.00	5.00	5.00
National Passport	5.00	5.00	5.00	5.00	5.00
National ID card	5.00	4.50	5.00	4.50	5.00
Average	**4.93**	**4.80**	**4.78**	**4.45**	**4.68**
Hospital	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**

Club membership card	3.93	4.29	4.36	3.92	4.36
Employee ID card	4.90	4.45	4.58	4.76	4.74
Medical insurance application form and card	4.84	4.16	4.72	4.12	4.61
Patient Registration Card	4.88	4.70	4.48	4.19	4.67
Average	**4.64**	**4.40**	**4.53**	**4.25**	**4.59**
Mobile Phone Companies	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Sim registration application form and sim card	5.00	3.67	3.67	3.67	5.00
Internet services registration application form and service	4.67	3.50	3.33	3.33	4.67
Average	**4.83**	**3.58**	**3.50**	**3.50**	**4.83**
Schools	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Student ID card	4.92	4.72	4.96	4.58	4.81
University application form	4.75	4.71	4.71	4.57	4.57
School certificate (or Diploma/Degree/Masters/PhD)	4.94	4.76	4.89	4.72	4.89
Club membership card	4.57	4.33	4.80	4.33	3.57
Employee ID card	4.83	4.77	4.62	4.62	4.62
Average	**4.80**	**4.66**	**4.80**	**4.56**	**4.49**
Universities	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Student ID card	4.75	4.61	4.67	4.29	4.52
University application form	4.74	4.42	4.19	4.21	4.56
School certificate (or Diploma/Degree/Masters/PhD)	4.83	4.61	4.50	4.27	4.62
Club membership card	4.00	Percent	3.88	3.80	4.21
Employee ID card	4.68	4.53	4.35	4.22	4.53
Average	**4.60**	**4.54**	**4.32**	**4.16**	**4.49**
Utility Bills	**Usefulness**	**Trust**	**Ease of use**	**Image**	**User satisfaction**
Water bill	5.00	4.00	4.00	4.00	5.00
Electricity	5.00	4.00	3.00	4.00	4.00
Average	**5.00**	**4.00**	**3.50**	**4.00**	**4.50**

Appendix B: Top ten identity attributes for respective organizations and sampled countries

	Banks		**Government**		**Insurance**		**Universities & Schools**		**Overall**	
Botswana	Bank account details	3.7045	Date of Birth	2.6059	Last name	3.1464	Last name	3.3436	Last name	8.9541
	Date of Birth	0.8269	Last name	1.7523	ID Number	1.6763	First name	1.8662	Date of Birth	4.5180
	Last name	0.7118	Place of Birth	1.7523	Date of Birth	0.6963	Gender	0.4943	ID Number	3.1480
	Home address	0.3090	ID Number	0.8986	Expiry date	0.6963	ID Number	0.4943	Bank account details	2.9656
	Middle name	0.1939	Gender	0.4718	First name	0.2063	Date of Birth	0.3888	First name	2.1962
	First name	0.0788	First name	0.0449	Gender	0.2063	Middle name	-0.0333	Gender	1.1361
	ID Number	0.0788	Work address	-0.1685	Home address	0.2063	Place of Birth	-0.0333	Place of Birth	0.6359
	Expiry date	0.0788	Middle name	-0.3819	Bank account details	0.2063	Work telephone number	-0.3499	Expiry date	-0.1678
	Gender	-0.0363	Home address	-0.3819	Work telephone number	-0.0387	Bank account details	-0.3499	Home address	-0.4276
	Home telephone number	-0.2090	Expiry date	-0.3819	Home telephone number	-0.2837	Home telephone number	-0.4554	Work telephone number	-0.9795
Zambia	Bank account details	2.4147	ID Number	3.7306	Date of Birth	2.3048	ID Number	2.3470	ID Number	7.9907
	ID Number	1.5151	Gender	0.6488	Last name	2.1005	Last name	1.5153	Date of Birth	4.9065
	Home address	1.2698	Date of Birth	0.5632	Gender	1.8962	Gender	1.5153	Last name	4.4483
	Date of Birth	0.7791	Work address	0.5632	First name	0.6022	Date of Birth	1.2594	Gender	4.3487
	Last name	0.6973	Bank account details	0.5632	ID Number	0.3979	Middle name	0.7475	Bank account details	3.0584
	First name	0.3702	Last name	0.1352	Middle name	-0.2832	First name	0.4277	First name	0.8503
	Gender	0.2884	Home address	0.0496	Bank account details	-0.2832	Bank account details	0.3637	Middle name	0.3896
	Expiry date	0.2884	Middle name	-0.0360	Home telephone number	-0.4194	Place of Birth	-0.4041	Home address	-0.0923
	Work address	0.2884	Place of Birth	-0.2073	Work telephone number	-0.4194	Home telephone number	-0.4681	Work address	-0.5601
	Middle name	-0.0387	Home Unique Property Reference Number (House Number)	-0.4641	Work email address	-0.4194	Work telephone number	-0.4681	Expiry date	-1.5666
Namibia	Bank account details	2.7832	ID Number	1.9499	Last name	2.9146	Date of Birth	2.3110	Last name	8.2703
	Last name	1.5191	Last name	1.8532	Place of Birth	1.5894	Last name	1.9833	Date of Birth	6.5313
	Date of Birth	1.2382	Date of Birth	1.4663	Date of Birth	1.5157	Home address	1.3280	Home address	3.6567

	Home address	0.8871	Home address	1.0793	First name	0.8286	Home telephone number	1.0003	ID Number	2.7009
	Gender	0.6764	Place of Birth	0.9826	Home address	0.3623	ID Number	0.8365	Place of Birth	1.4632
	ID Number	0.3253	Expiry date	0.5957	Bank account details	0.2764	First name	0.0172	Bank account details	1.4320
	Expiry date	0.1848	Middle name	0.4022	Gender	-0.0058	Work email address	0.0172	Gender	0.2491
	Work address	-0.0961	Work address	0.0153	ID Number	-0.4107	Gender	-0.1466	First name	0.1674
	First name	-0.3068	Gender	-0.2749	Work telephone number	-0.5089	Expiry date	-0.1466	Expiry date	0.0268
	Middle name	-0.3068	First name	-0.3717	Home telephone number	-0.5212	Home email address	-0.1466	Middle name	-0.8712
Banks		**Government**		**Insurance**		**Universities & Schools**		**Overall**		Total
New Zealand	Home address	2.3007	Last name	2.2120	Date of Birth	3.1021	Date of Birth	2.3970	Date of Birth	9.8715
	Date of Birth	2.1604	Date of Birth	2.2120	Last name	1.2674	First name	1.9681	Last name	5.9005
	Bank account details	1.0377	Bank account details	1.8299	Bank account details	0.6941	Last name	1.4535	First name	2.7548
	Last name	0.9676	Home email address	0.6837	Work telephone number	0.5794	Gender	1.1105	Bank account details	2.6995
	Home telephone number	0.5115	First name	0.3016	First name	0.4647	Home telephone number	0.3386	Home address	2.5470
	Gender	0.1957	Home address	0.3016	Work email address	0.3500	ID Number	0.2528	Home telephone number	0.6229
	Middle name	0.1606	Gender	-0.2715	Home telephone number	0.2354	Home Unique Property Reference Number (House Number)	0.0813	Gender	0.4674
	First name	0.0203	Home Unique Property Reference Number (House Number)	-0.2715	Home address	0.1207	Expiry date	0.0813	Home email address	-0.7586
	Work telephone number	0.0203	ID Number	-0.2715	Expiry date	0.1207	Home address	-0.1760	Work telephone number	-0.9355
	Race	-0.0849	Expiry date	-0.2715	Home email address	0.1207	Middle name	-0.5191	Expiry date	-1.0315
Australia	Last name	2.2453	Date of Birth	2.4182	Home address	1.6592	Last name	2.5046	Last name	8.1047
	Home address	2.1168	Last name	2.3261	Date of Birth	1.5016	Date of Birth	2.5046	Date of Birth	7.5132
	Bank account details	1.2173	Home address	1.0831	Middle name	1.0287	Expiry date	1.2493	Home address	4.3966
	Date of Birth	1.0888	Middle name	0.8069	Last name	1.0287	Gender	0.4505	Gender	1.5809
	Home telephone number	0.4464	ID Number	0.7148	Gender	0.7135	Work telephone number	0.4505	Middle name	1.5768

	Middle name	0.3179	Gender	0.4846	Home telephone number	0.7135	ID Number	0.1081	Work telephone number	0.2263
	ID Number	0.1894	Expiry date	-0.1599	Work address	0.7135	First name	-0.1201	Home telephone number	0.0991
	Gender	-0.0676	Place of Birth	-0.3441	Work telephone number	0.5558	Home Unique Property Reference Number (House Number)	-0.3484	ID Number	-0.0081
	Home Unique Property Reference Number (House Number)	-0.0676	Work address	-0.3901	First name	0.3982	Home telephone number	-0.3484	Bank account details	-0.1198
	Work telephone number	-0.0676	Bank account details	-0.5282	Home Unique Property Reference Number (House Number)	0.3982	Home address	-0.4625	Work address	-0.6920
USA	Bank account details	2.1681	Date of Birth	2.9875	ID Number	2.3258	Date of Birth	2.4295	Date of Birth	8.8117
	Date of Birth	1.4699	First name	1.3768	Date of Birth	1.9248	First name	1.3541	ID Number	2.5972
	ID Number	1.4699	Last name	1.2102	First name	0.4010	Last name	1.3541	First name	2.3327
	Work telephone number	1.0335	Middle name	1.0436	Middle name	0.4010	Race	1.1886	Last name	2.2533
	Work email address	0.8590	Work telephone number	0.7659	Last name	0.4010	Gender	1.0232	Gender	1.3431
	Gender	0.5971	Work address	0.0994	Gender	0.4010	Home address	0.2787	Middle name	1.1031
	Home address	0.5971	Work email address	-0.1783	Home Unique Property Reference Number (House Number)	0.1604	Home telephone number	0.2787	Bank account details	0.6906
	Home telephone number	0.1608	Expiry date	-0.2894	Home telephone number	0.1604	Middle name	0.1959	Work telephone number	0.6168
	Home email address	-0.1011	Place of Birth	-0.3449	Home email address	0.1604	Place of Birth	-0.3004	Home address	0.5555
	Middle name	-0.5374	Home address	-0.4005	Work email address	0.1604	Work telephone number	-0.3004	Home telephone number	0.1993

Appendix C: Research Questionnaire

The University of Zambia

School of Natural Sciences

Development of identity attribute metrics model based on distance metrics

Felix Musama Lameck Kabwe (Student No: 2017014609)
MSc Computer Science

For more information or any queries, kindly get in touch on 0974285177

Dear Respondent,

I am a student at the University of Zambia in my final stage pursuing an MSc in Computer Science. As partial fulfillment for the award of a Master's degree, I am conducting a baseline study on: ***"Development of identity attribute metrics model based on distance metrics."***

You have been purposefully sampled to provide information for the topic indicated above. The information being collected is purely for academic purposes as such, it will be treated with maximum confidentiality. Subsequently, you are not supposed to indicate your name or any personal information that can lead to revealing of your identity.

Your co-operation will be greatly appreciated.
For more information or any queries, kindly get in touch with the following:

Project Supervisor: Dr. Jackson Phiri (0966 693 731) or
Head of Department: Mrs. Monica M. K. Kabemba (0211-293901)

Name of rganisation_____

Part 1: Demographic information (Please tick [√])

1. Gender: Male [ ] Female [ ]

2. Marital Status: Single [ ] Married [ ] Divorced [ ] Other [ ]

3. Age: 20 or under [ ] 21-30 [ ] 31-40 [ ] 41-50 [ ] 51-60 [ ] 61+ [ ]

4. Highest level of education:
   Grade 12 Certificate or below [ ] Diploma [ ] First degree [ ] Masters [ ] Ph.D. [ ]

5. Type of employment:
   Not working [ ] Salaried worker [ ] Self-employed [ ] Pensioner [ ]

6. Occupation (Please specify, e.g. “University lecturer in Computer Science”)

Part 2

We have a number of documents or sources of identification that we use in our daily life for the purpose of identifying entities/individuals, either physically or electronically. This research would like to find out from the respondent the documents/sources of identification that are used, in order of importance as is perceived by the respondent. A list of usually used documents has been given below, a respondent is requested to grade each document/source of identification. The grading ranges from 1 to 5, where five is the mostly important or valued document and 1 is the least important or valued document.

EXPLANATION OF IMPORTANT WORDS

i. ***Usefulness***
   The degree to which a person believes that using the particular document would help his or her job in identifying an individual/thing
ii. ***Ease of use***
   "The degree to which the document is perceived as being difficult to use"
iii. ***Image(status)***
   "The degree to which use of the document is perceived to enhance one's image or status in one's social system"
iv. ***Trust***
   How would the attributes on the document of the individual/thing being identified enhance trust?
v. ***User satisfaction***
   How satisfied with the use of the document

SD = strongly disagree | D = Disagree | N = Neutral | A = Agree | SA = Strongly Agree | NA= Not Applicable

SOURCE OF IDENTITIFICATION	Usefulness						Trust (Secure)						Ease of use						Image (Your status)						User satisfaction					
	SD	D	N	A	SA	NA	SD	D	N	A	SA	NA	SD	D	N	A	SA	NA	SD	D	N	A	SA	NA	SD	D	N	A	SA	NA
	1	2	3	4	5	-	1	2	3	4	5	-	1	2	3	4	5	-	1	2	3	4	5	-	1	2	3	4	5	-
BANKS																														
Mortgage application form																														
Bank loan application form																														
Credit card/Debit card application form																														
Student loan application form																														
Bank account application form and card																														

Employee ID																														
Club membership cards																														
INSURANCE																														
Life Insurance application form and card																														
Car Insurance application form and disc																														
Medical insurance application form and card																														
GOVERNMENT																														
National pension application form																														
National Visa application form and card																														
Building society account application form and card																														
Title Deeds																														
Marriage certificate																														
Divorce certificate																														
Driving license																														
National tax application form																														
National vehicle registration certificate																														
National birth certificate																														
National Passport																														
National ID card																														
Employee ID card																														
Club membership card																														
UNIVERSITIES AND SCHOOLS																														
Student ID card																														

University application form																														
School certificate (or Diploma/Degree/Masters / PhD)																														
Club membership card																														
Employee ID card																														
HOSPITAL																														
Medical insurance application form and card																														
Patient Registration Card																														
Club membership card																														
Employee ID card																														
MOBILE PHONES																														
Sim registration application form and sim card																														
Internet services registration application form and service																														
UTILITY BILLS																														
Water bill																														
Land rates																														
Electricity																														
Telephone																														
OTHERS																														
Baptism certificate																														
Club membership card																														
ANY OTHER																														

PLEASE NOTE

The respondent can add any five more documents that could not be listed above and grade them

Part 3

The identification documents mentioned above, may usually request for particular items of identification for a particular service. These items of identification are referred to as attributes. The following are some attributes that a service provider may request from you for registration of a service. The grading ranges from 1 to 5, where five is the mostly important or valued document and 1 is the least important or valued document.

ATTRIBUTE	**GRADING IN PERCIEVED IMPORTANCE OR VALUE (WHERE "1" is the LOWEST AND "5" IS THE MOST IMPORTANT)**				
	1	2	3	4	5
First name					
Middle name					
Last name					
Date of Birth					
Place of Birth					
Race					
Gender					
Home address					
Home Unique Property Reference Number (House Number)					
Home telephone number					
ID Number					
issuing authority					
Expiry date					
Home email address					
Work address					
Work telephone number					
Work email address					
Bank account details					
Height					

PLEASE NOTE

The respondent can add any five more documents that could not be listed above and grade them

Part 4: Actual use of Identity documents (Please tick [√])

1. How long have you been using Identity documents?
Under 1year [ ] 1-2 years [ ] 3- 4 years [ ] more than 4 years [ ]

2. On a weekly basis, how many times do you use Identity documents?
Not at all [ ] once a week [ ] 2-3 times [ ] more than 3 times [ ]