



Natural Language Processing for Research Philosophies and Paradigms

Dissertation (DFINT91)

by

Ntombhimuni Mawila

Student No: 3689-857-0

submitted in accordance with the requirements for the degree of

MTECH: INFORMATION TECHNOLOGY (98802)

in the

College of Science, Engineering & Technology

School of Computing

Department of Computer Science

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor:

Prof Marcia Mkansi

Co-supervisor:

Prof Ernest Mnkandla

Date of submission: 28 February 2021

ABSTRACT

Research philosophies and paradigms (RPPs) reveal researchers' assumptions and provide a systematic way in which research can be carried out effectively and appropriately. Different studies highlight cognitive and comprehension challenges of RPPs concepts at the postgraduate level. This study develops a natural language processing (NLP) supervised classification application that guides students in identifying RPPs applicable to their study. By using algorithms rooted in a quantitative research approach, this study builds a corpus represented using the Bag of Words model to train the naïve Bayes, Logistic Regression, and Support Vector Machine algorithms. Computer experiments conducted to evaluate the performance of the algorithms reveal that the Naïve Bayes algorithm presents the highest accuracy and precision levels. In practice, user testing results show the varying impact of knowledge, performance, and effort expectancy. The findings contribute to the minimization of issues postgraduates encounter in identifying research philosophies and the underlying paradigms for their studies.

KEY TERMS:

Research; Philosophy; Paradigm; Corpus; Algorithm; Classification model; Classifier; Bag of Words; naïve Bayes; Researcher

ACKNOWLEDGEMENTS

First, I would like to thank the Creator for giving me life, strength, energy, and the perseverance to pursue this study. I would like to express my appreciation and gratitude to my supervisors, Prof. Marcia Mkansi and Prof Ernest Mnkandla, for their support, guidance, patience, contribution, and willingness to share invaluable insight and knowledge during this study. Your words of encouragement helped to see the study through. I would like to express my sincere gratitude to my sisters and brothers-in-law for being my biggest cheerleaders and stepping in when needed to help with mommy duties. To Mama, your words of encouragement, support and prayers have seen this project through. To Zanokuhle, the keeper of my smile, words fail me. Your patience and understanding during this period are beyond what any other 11-year old would have demonstrated; I owe you the world. To Donnaghue, no matter what the project demanded, you always came through. Your assistance cannot be quantified in any way; I appreciate all your help. To my husband, Zanethemba Jikijwa Mvaba, I am grateful for your support, your understanding and for affording me the opportunity to study, learn and realize my potential. Ndiyabulela!

DECLARATION

Name: NTOMBHIMUNI TLANGELANI MAWILA

Student No: 36898570

Degree: MTECH: INFORMATION TECHNOLOGY (98802)

Natural Language Processing for Research Paradigms and Philosophies

I declare that the above titled dissertation is my own work and all the sources that I have used or quoted have been duly acknowledged throughout the text and by means of a complete list of references. I further declare that I submitted the dissertation to originality checking software and that it falls within the acceptable requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at INISA for another qualification or at any other higher education institution

SIGNATURE : 

DATE: 2021/05/13

TABLE OF CONTENTS

LIST OF TABLES	x
GLOSSARY OF ACRONYMS	xi
1 CHAPTER 1: INTRODUCTION	12
1.1. Context of the Research	14
1.2. Problem Statement	15
1.3. Aim and Objectives of Research.....	16
1.4. Research Objectives.....	17
1.5. Research Questions	17
1.6. Significance of the Study	17
1.7. Scope and Limitations of the Study	18
1.8. Overview of Research Method.....	19
1.9. Chapter Summary.....	21
2 CHAPTER 2: LITERATURE REVIEW	22
2.1 Related Work.....	22
2.2. Theoretical Background.....	25
2.2.1. The theory of knowledge	25
2.2.2. The theory of classification in natural language processing	28
2.3. Overview of Corpora.....	36
2.4. Chapter Summary.....	37
3 CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY	38
3.1. Research design.....	40
3.2. Research Paradigm and Philosophy.....	41
3.3. Research Approach	42
3.3.1 Quantitative Research Approach	42
3.3.2 Qualitative research approach	43
3.3.3 Mixed research approach.....	43
3.4. Research Strategy	43
3.5. Research population and Sampling	46
3.6. Data collection	47
3.7. Data analysis	48
3.8. Reliability and Validity.....	48
3.9. Limitations and Constraints	49
3.10. Ethical Considerations.....	49

3.11.	Chapter Summary	50
4	CHAPTER 4: SYSTEM DESIGN AND IMPLEMENTATION.....	51
4.1	Systems Architecture.....	52
4.1.1	Web Browser/Client	52
4.1.2	Webserver.....	53
4.1.3	NLP application/Server-side.....	53
4.1.4	MySQL DB.....	53
4.2	System Design.....	53
4.2.1	Use case diagram	54
4.2.2	Class Diagrams	54
4.2.3	Entity Relationship Diagram	55
4.3	Implementation of the NLP Application.....	56
4.3.1.	Technical requirements.....	56
4.3.2.	Technologies Used	57
4.3.3.	User Interface.....	57
4.3.4.	Researcher's interaction with the RMI NLP application.....	59
4.3.5.	Administrator's interaction with the RMI NLP application	59
4.3.6.	Back-end.....	60
4.4	Input Data Creating the RPPs corpus.....	61
4.4.1	Data collection.....	61
4.5	Chapter Summary.....	68
5	CHAPTER 5: SYSTEM TESTING AND RESULTS ANALYSIS.....	69
5.1	SYSTEM TESTING	69
5.1.1	Evaluation of the trained models.....	69
5.2	Participants.....	76
5.3	Usability Test Results	77
5.3.1	Effort Expectancy	77
5.3.2	Performance Expectancy	84
5.3.3	Knowledge Expectancy.....	93
5.3.4	General Comments	102
5.4	Achievement of Objectives	107
5.4.1	Build a corpus of research philosophies and paradigms.....	107
5.4.2	Train a classifier and develop an NLP application to recommend RPPs.....	108
5.4.3	Test the application's ability to recommend a research philosophy and paradigm through user testing.....	108

5.5	Chapter Summary.....	109
6	CHAPTER 6: CONCLUSION AND RECOMMENDATIONS.....	110
6.1	Overview.....	110
6.2	Conclusion.....	111
6.3	Recommendations for Future Work.....	112
7	REFERENCES.....	114
8	APPENDICES.....	125
	APPENDIX A: DEFINITION OF KEY TERMS.....	125
	APPENDIX B: TEST SCRIPT.....	126
	APPENDIX C: SYSTEM USABILITY QUESTIONNAIRE.....	127
	APPENDIX D: USER MANUAL.....	131
	APPENDIX E: SAMPLE OF THE CORPUS.....	135
	APPENDIX F: RELATED STUDIES SOURCES.....	136
	APPENDIX G: ARCHITECTURE DIAGRAMS.....	138
	APPENDIX H: PARTICIPATION CONSENT FORM.....	143
	APPENDIX I: ETHICAL CLEARANCE CERTIFICATES.....	144
	APPENDIX J: TURNITIN RECEIPT.....	148
	APPENDIX K: SOURCE CODE.....	149
	APPENDIX L : ANNOTATION PROCESS.....	192
	APPENDIX M: BREAK DOWN OF SOURCES FOR THE CORPUS.....	193
	APPENDIX N: DATA STATEMENTS WORKSHEET.....	194

LIST OF FIGURES

Figure 1.1 The RPP classification app development process.....	19
Figure 2.1 Relationship between research philosophies and paradigms.....	26
Figure 2.2 Use of ML in NLP (Conversia, 2017).....	29
Figure 2.3 Process for training a classifier (Springboard, 2020).....	30
Figure 2.4 Machine Learning categories.....	31
Figure 2.5 Sigmoid function and graph for the SVM algorithm.....	35
Figure 3.1 CRSP-DM steps for data mining (Dsouza, 2018).....	40
Figure 3.2 Development of the RMI application.....	44
Figure 4.1 RMI flow diagram.....	59
Figure 4.2 Differences in the components of RPPs (Saunders et al., 2009).....	62
Figure 4.3 Chart showing the recommended RPPs.....	67
Figure 4.4 Named Entity Recognition.....	68
Figure 5.1 Steps followed in selecting the classifier.....	70
Figure 5.2 Data split for cross-validation.....	71
Figure 5.3 Comparison of the algorithms.....	71
Figure 5.4 The script and report for naïve Bayes classifier.....	72
Figure 5.5 The script and report for SVM classifier.....	74
Figure 5.6 The script and report for the Logistic Regression classifier.....	75
Figure 5.7 Analysis of participants.....	76
Figure 5.8 Analysis of participants' profession.....	76
Figure 5.9 Measurable items for Effort Expectancy.....	78
Figure 5.10 Analysis of N-technical users will find it easy to learn.....	78
Figure 5.11 Analysis of It is easy to get the system to do what I want it to do.....	79
Figure 5.12 Analysis of Interaction with system is clear and understandable.....	80
Figure 5.13 Analysis of The system is flexible to interact with.....	81
Figure 5.14 Analysis of the level of difficulty in mastering how to utilize the system.....	82
Figure 5.15 Analysis of the system is easy to use.....	83
Figure 5.16 Analysis of utilizing the research system will empower participants to accomplish research tasks more quickly.....	84
Figure 5.17 Analysis of using the system would improve my epistemological understanding.....	85
Figure 5.18 Analysis of Does the system perform well when there is the concurrent use of the system.....	86
Figure 5.19 Analysis of Acceptability of time required to generate the report.....	87
Figure 5.20 Analysis of Is the system functional and fit for purpose.....	88
Figure 5.21 Analysis of Attitude towards using the technology.....	89
Figure 5.22 Analysis of Support of the utilization of the system to enhance the learning.....	90
Figure 5.23 Analysis of I would like using technology to learn more about the subject matter the system addresses instead of the traditional way.....	91
Figure 5.24 Analysis of the level of anticipation of those aspects of research that will be supported by the use of the system.....	92
Figure 5.25 Analysis of I know and understand the concept of research philosophy.....	94
Figure 5.26 Analysis of I know and understand what research paradigm is.....	95
Figure 5.27 Analysis of I understand the value of research philosophies and paradigms in conducting research.....	96
Figure 5.28 Analysis of I know which research philosophy I espouse.....	97
Figure 5.29 Analysis of I understand the relevance of research philosophies and paradigms in conducting research.....	98
Figure 5.30 Analysis of how the concepts of research philosophies and paradigms are introduced is beneficial.....	99
Figure 5.31 Analysis of the ability to find out more information about research philosophies and paradigms.....	100
Figure 5.32 Crosstab analysis of participants who understand and know the value of research philosophies and paradigms.....	101
Figure 5.33 Analysis of I know and understand what a research paradigm is, and I support the idea of using the system to enhance the learning process.....	101

<i>Figure 5.34 Crosstab analysis of I know which research philosophy I espouse, and the system's introduction of the concepts of research philosophies and paradigms is beneficial</i>	102
<i>Figure 8.1 Research Methods Index NLP architecture diagram</i>	138
<i>Figure 8.2 Interaction of the components of the MVT pattern (Tutorialspoint, 2019)</i>	139
<i>Figure 8.3 User interaction with the NLP application</i>	139
<i>Figure 8.4 Administrator interaction with the NLP application</i>	140
<i>Figure 8.5 Class diagram of the NLP system</i>	140
<i>Figure 8.6 Entity relationship diagram of the NLP system</i>	141
<i>Figure 8.7 The NLP process flow diagram for text classification</i>	141
<i>Figure 8.8 The database schema</i>	142
<i>Figure 8.9 Create project applications</i>	149
<i>Figure 8.10 Installed apps</i>	149
<i>Figure 8.11 Create the data model</i>	150
<i>Figure 8.12 Update database schema</i>	150
<i>Figure 8.13 Adding url patterns for the nlp application</i>	150
<i>Figure 8.14 Defining custom views</i>	151
<i>Figure 8.15 Forms.py script</i>	151
<i>Figure 8.16 View of MySQL database schema</i>	152
<i>Figure 8.17 Creation of a sub directory of management and command</i>	152
<i>Figure 8.18 Use of the 'GET' and 'POST' methods</i>	152
<i>Figure 8.19 Python script for the classifying model</i>	185
<i>Figure 8.20 Script models.py to create the Natural Language Processing (NLP) database</i>	187
<i>Figure 8.21 Spreadsheet for the RPP corpus or dataset</i>	188
<i>Figure 8.22 Training sentences for the corpus</i>	188
<i>Figure 8.23 Words in the RPP corpus and the scores of occurrence in the corpus</i>	188
<i>Figure 8.24 Login page of the RMI application</i>	189
<i>Figure 8.25 Form for registering a user account</i>	189
<i>Figure 8.26 Landing page of the RMI NLP application</i>	190
<i>Figure 8.27 NLP questionnaire</i>	190
<i>Figure 8.28 Button to view the NLP report</i>	191
<i>Figure 8.29 RMI NLP report</i>	191

LIST OF TABLES

<i>Table 4.1 Development and deployment specifications</i>	<i>56</i>
<i>Table 4.2 BoW vector representation of corpus text data.....</i>	<i>65</i>
<i>Table 5.1 Comparison of the algorithm scores</i>	<i>71</i>
<i>Table 5.2 System tasks to be performed on the NLP system</i>	<i>77</i>
<i>Table 5.3 System Errors</i>	<i>103</i>
<i>Table 5.4 System Failure.....</i>	<i>104</i>
<i>Table 5.5 Comments</i>	<i>106</i>

GLOSSARY OF ACRONYMS

RMI	Research Methods Index
ERD	Entity Relationship Diagram
API	Application Programming Interface
MVT	Model View Template
MySQL	My Structured Query Language
HTML	Hypertext Transfer Markup Language
IDE	Integrated Development Editor
OS	Operating System
URL	Uniform Resource Locator
AI	Artificial Intelligence
ORM	Object-Relational-Mappers
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RPP	Research Paradigms and Philosophies
HTTP	Hypertext Text Transfer Protocol

CHAPTER 1: INTRODUCTION

Research philosophy refers to the development of knowledge together with the source and nature of that knowledge. It guides one in collecting, analysing, and using data (Saunders et al., 2012). Research paradigm refers to the system of common beliefs, agreements, and assumptions a community has about the world around them (Saunders et al., 2009; Kivunja and Kuyini, 2017). These assumptions and beliefs shape ones' understanding of a research question, provide guidance for methods to use when conducting the research, and on how to interpret the collected data (Saunders et al., 2009; Guba & Lincoln, 1994; Denzin & Lincoln, 2005). The literature reveals that many higher degree research students and some career researchers find it difficult to identify the research philosophies and paradigms guiding their research (Kivunja and Kuyini, 2017). This study uses Natural Language Processing (NLP) techniques to help researchers identify research philosophies and paradigms.

NLP has been used in classification, among other tasks, to determine the categories of given texts or documents. This has been spurred on by the large volumes of digital texts and the need to organize them (Sebastiani, 2002). NLP refers to the use of a computer in the manipulation of everyday human languages to produce meaningful responses. This study set out to recommend/determine the category of a set string of text (T) using a supervised classifier given a fixed set of research philosophy and paradigm categories ($C = C_1, C_2, C_3, \dots C_n$). These categories of research philosophies and paradigms (RPPs) were sourced from various philosophy publications such as PhilPapers, Stanford Encyclopedia of Philosophy, Google Scholar, IBSS, Philosophy Basics publications, journals, and theses. The recommendation of RPPs is achieved through manipulating responses provided by a system user and mapping them to target RPPs categories or labels most linked to a user's expressed ideas.

NLP relies on machine learning algorithms such as naïve Bayes, logic regression, support vector machine (SVM), and decision trees for classification. These algorithms have been used in various NLP problems such as spam detection and sentiment analysis (Romanov, 2014; Sebastiani, 2002). This study experimented

with the naïve Bayes, logistic regression, and SVM algorithms trained with predefined categories of RPPs texts gathered into a corpus. Based on Python's classification report function, the naïve Bayes outperformed the other algorithms and was thus selected as the best algorithm suited for this study.

The study also set out to collect text data, which included the epistemology, ontology, and axiology components of each of the RPPs categories into the corpus used to train the algorithms. The Bag of Words (BoW) model was used to represent the RPPs data utilized to train the classification algorithms. Chapter 5 discusses the results.

By recommending RPPs at the onset of a research project, the developed NLP application contributes significantly to establishing which methods researchers can use to collect, analyse, and interpret the data gathered through their research. This gives researchers a holistic view of the knowledge generation process and provides a general perspective on how they relate to the knowledge they are generating. It is envisaged that this will contribute to researchers' creativity and can also improve the research quality.

From the computer experiment results carried out to evaluate the performance of the algorithms, Naïve Bayes reveals the highest precision level of 85%, the accuracy of 70%, recall rate of 76%, and an f1-score of 76% for the study. Hence, it was used to create a web application to classify user input into research philosophies and paradigms categories. In practice, user testing results show the varying impact of knowledge, performance, and effort expectancy that yield moderate but significant improvement to cognitive and comprehension of research philosophies and paradigms at the postgraduate level. The remainder of this chapter provides a general review of the context of research philosophies and paradigms. The chapter further presents the problem statement, purpose, objectives, research questions, significance, scope, limitations of the study, and the research methodology. The chapter concludes with an outline of subsequent chapters and a chapter summary.

1.1. Context of the Research

The knowledge generation and production process in academe requires a lot of understanding and lies at the core of research philosophies and paradigms. The skyrocketing investment in learning, teaching, and induction of researchers to increase technical, conceptual, and professional human capacity at higher education institutions is a testament to the need to understand these concepts (Lamont 2014; Polster & Newson, 2007).

Researchers' approach to investigating a chosen study topic is dependent on how they think about the problem and how it can be studied to achieve results. Researchers have their own views about their topic of choice. These views guide their beliefs, assumptions, and how they think about society and help in forming a structure to establish how they view the world around them (Chilisa & Kawulich, 2012). According to Makombe (2017), a research process must be guided by a particular paradigm. However, he further notes that many researchers never mention their guiding paradigm due to the confusion that arises about the relationship between paradigms, research methods, and research designs.

To avoid the sometimes narrow definitions of concepts and to prevent confusion, this study defines research paradigm as typical ways in which a group of people has common beliefs, share the same values and assumptions about how research design in a particular field of study (Mittwede, 2012). According to Chilisa and Kawulich (2012), the following attributes characterize research paradigms:

- Ontology – A belief system that represents an individual's understanding of what constitutes reality
- Epistemology – The analysis of the nature of knowledge, the logic of and reason for belief.
- Methodology – The systematic methods followed in enquiring about a phenomenon in a particular area of research

According to Flowers (2009), any research undertaking should consider different research paradigms, ontologies, and epistemologies because they influence how a study is conducted. He further notes that a lack of understanding of these aspects of

research will invariably lead to a researcher opting for methods that are not compatible with a researcher's stance, therefore leading to incoherent study results. Saunders et al. (2012) assert that having an appropriate research strategy is important because the strategy ensures that:

- the research issues are addressed in a manner that is relevant and consistent with the overall research topic;
- research questions are addressed holistically;
- the research is in line with its intended purpose; and
- it outlines the general standards of hypothetical reasoning, a technique for comprehension, point of view, and mindfulness, which are all used to get information on the real world

This study leveraged the power of NLP classification algorithms in developing an application that focused on the multi-label classification of user input into RPPs categories. Most importantly, it introduces the concepts of research philosophies and paradigms to researchers in a comprehensive, objective way by analysing the ideas a researcher provides, and mapping them into relevant philosophies and paradigms. The following section details the problem statement.

1.2. Problem Statement

According to Kivunja and Kuyini (2017), the articulation of research paradigms as a concept is quite elusive to most scholars and equally challenging to apply in research proposals. Researchers have limited exposure to the research paradigm and philosophical stance at their disposal and suitable for their chosen research questions (Mertens, 2014). Consequently, researchers have insufficient knowledge about the value that an appropriate research paradigm or philosophy can have on their knowledge production process (Mertens, 2014). The lack of knowledge about research philosophies and paradigms makes it difficult for researchers to ascertain the kind of methodology to use to formulate a research problem and determine an effective and appropriate way of collecting, processing, and analysing data (Žukauskas et al., 2018). Mertens (2014) points out that not having an understanding of one's paradigm or its associated philosophical assumption does not necessarily

mean that a researcher does not have any assumptions; instead, it reveals a bane that impacts the quality of knowledge production.

This study is designed to solve this problem of unexamined and unrecognized assumptions by predicting and exploring various appropriate models through a set of typologies or parameters responded to by users (researchers). Furthermore, the study intends to enhance teaching and learning by using natural language processing (NLP) to develop an annotated dataset or corpus of research philosophies and paradigms (RPPs). The annotated data in the corpus is used simultaneously with supervised machine learning – the naïve Bayes, Support Vector Machines, and Logistic Regression algorithms – to classify input, thereby recommending or predicting the philosophical stance and the underlying paradigm adopted by researchers for their knowledge production process. Thus, the information will be extracted and transformed from various sources on existing concepts and definitions of research philosophies and paradigms. The study used the RPPs corpus to train the machine learning supervised classification algorithms to simplify the complexity embedded in identifying research paradigms and philosophies by classifying input text into RPPs categories. This will enhance the knowledge production and development process, thereby improving teaching and learning.

1.3. Aim and Objectives of Research

The study aimed to improve the learning and teaching process by developing a specialised corpus of research philosophies and paradigms and using a model that will classify input variables into research philosophies and paradigms (RPPs) categories. This was done through a natural language processing interface and machine learning algorithms that access the RPPs categories. The result was to determine the research philosophies and underlying paradigms that researchers can adopt for their knowledge production process.

1.4. Research Objectives

To simplify the complexity entailed in understanding or establishing the philosophies and underlying paradigms necessary to augment the teaching and learning process, the study endeavoured to meet the following objectives:

- a. Extract and transform information of existing definitions and concepts constituting research philosophies and paradigms from various sources and build a corpus or knowledge base of philosophies and paradigms that will be used for algorithm training purposes
- b. Train and test the performance of a supervised classification algorithm with the created RPPs corpus and use the algorithm to build an application to determine researchers' philosophical stance
- c. Evaluate the application's ability to recommend a research philosophy and paradigm through user testing

1.5. Research Questions

This research project seeks to answer the following questions:

- a) What components makeup research philosophy and paradigm, and which key texts will be stored in a corpus as features to be used in the RPPs categories?
- b) Which supervised classification algorithm is suitable for determining a researcher's philosophical stance?
- c) What are the effects of the application on the different technology acceptance model constructs?

1.6. Significance of the Study

The study intends to make theoretical, practical, and methodological contributions in the knowledge generation process. In practice, the study intends to contribute to the minimization of issues researchers encounter in identifying research philosophies and underlying paradigms in line with their research as outlined by Chilisa and Kawulich (2012); Makombe (2017); Mkansi and Acheampong (2012).

From theory, the problem statement being pursued by this study will provide a great contribution in enhancing knowledge production in teaching and learning. A literature review has not yielded any positive results in establishing a corpus/dataset of research philosophies and paradigms, nor an application that uses NLP technologies in classifying user input into RPPs categories. Therefore, this study makes a significant contribution by developing a natural language processing (NLP) application and a corpus that will be used for the classification or prediction of RPPs for researchers. This project can add value to the study of NLP in line with what other researchers in similar fields have done. For example, Crowston et al. (2012) explored NLP techniques in social research. The paper was written to demonstrate how NLP can be used, for qualitative data analysis to provide advanced analytic capabilities to help reduce the amount of textual data analysis a human would need to perform. This study intends to elevate the discoveries similar to those of Crowston et al. (2012).

1.7. Scope and Limitations of the Study

As this study was primarily designed to help researchers identify research philosophies and underlying paradigms for their studies, a classification model was trained only with the RPPs categories. The output from the application only shows a researcher's view of the world. It does not recommend nor provide further insight into the type of methods or techniques used for collecting, analyzing, and interpreting research data based on each of the research philosophies and paradigms produced as output. This inability to recommend the methods to be used constitutes a limitation of this study that needs further exploration.

Further, the only available sources of information on research philosophies and paradigms are in the form of unstructured content such as encyclopaedia (e.g., the Stanford Encyclopaedia of Philosophy), dictionaries (e.g., The Oxford Dictionary of Philosophy), libraries (e.g., The Philosophical Library) and indexes (e.g., Philosopher's Index) (Szarko, 2017). Although the limitation was addressed by developing a corpus containing structured information on as many paradigms and philosophies as possible, the process was laborious. The process required the

identification of components and attributes that make up a paradigm and philosophy. The algorithms' performance was also identified as another limitation of the study but improved with the increase of the dataset.

1.8. Overview of Research Method

Text classification refers to the task of NLP that facilitates the process of allocating texts to predefined labels. This study aims to develop an application that will classify user input into research philosophy and paradigm categories postgraduate students' cognitive and comprehension dimensions. This is achieved by training machine learning classification algorithms using text data relevant to RPPs and deploying the best performing algorithm to create the RPP application. In this section, the process followed in producing the RPPs classification system is described as shown in the workflow in Figure 1.1.

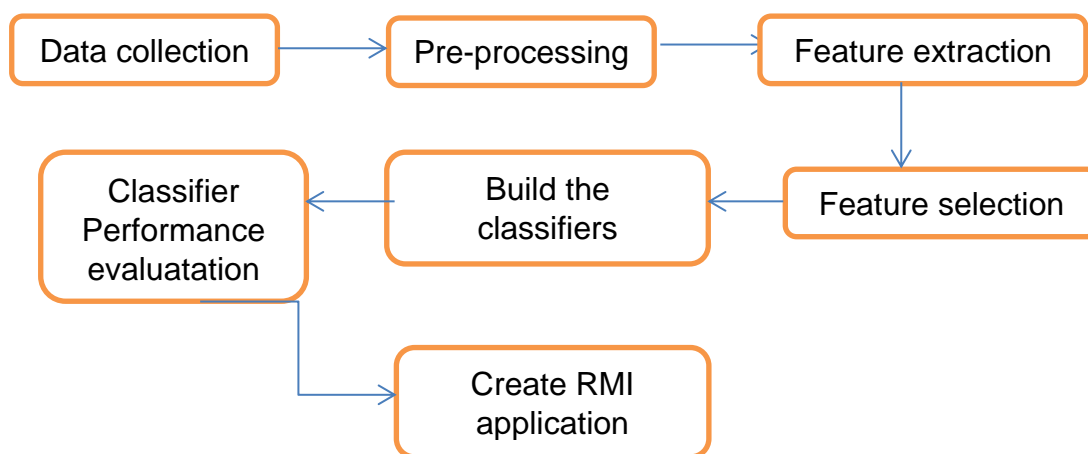


Figure 1.1 The RPP classification app development process.

The following steps are followed in developing the RPPs classification application:

- a.) Text data on RPPs were collected from various online sources, including Google Scholar, IBSS, Encyclopedia, PhilPapers, theses, and journals to create the corpus. The study set out to find texts about research philosophies and paradigms from the various sources, and the search yielded a total of 323, of which only 180 were used for this study. The framework established by Saunders et al. (2009) was followed in identifying the RPPs components to be used, which focuses on how people get to know what they know, how people

view the world around them, and the nature of values in relation to research as prescribed by each RPP.

- b.) Pre-processing was done on the collected text to standardize and normalize it by checking for typing errors and removing capitalization in preparation to train the classifying algorithm. Pre-processing of text included tokenizing, lemmatizing, data annotation, removing stop words, punctuation, and non-alphabetic characters. Section 4.4 further describes the processes used in creating and preparing the RPPs corpus to learn the ML classification algorithms.
- c.) Features were then extracted from the standardized and normalized text to represent the text as numerical vectors creating a vocabulary for each RPP category in the corpus. The BoW model was used to represent the pre-processed text in vectors of equivalent length.
- d.) A subset of the extracted features was selected from the original text data to improve classifier scalability, accuracy, and efficiency by constructing vector space. This is achieved by retaining words with the highest score based on how important the words are in an RPPs category (Kaur & Saini, 2015).
- d.) The study then experimented with classification models to select one. The process started with the splitting of the created RPPs corpus into a train and a test set. The train set was then used to learn the classifiers, after which the classifiers were tested using the test set to validate their performance.
- e.) The performance of the classifiers given the same train and test data was compared for accuracy, precision, recall, and F1-measure.
- f.) The naïve Bayes classifier was selected to create the RPPs classification application based on its performance compared to the other classifiers. The technology acceptance model (TAM) was used to evaluate the usability of the application by system users.

The collection of RPPs text data gathered in this study enabled the training and validation of the naïve Bayes, logistic regression, and the SVM classification algorithms. While each algorithm was evaluated for performance using F1-measure, precision, recall, and accuracy, the naïve Bayes classifier performed better than the rest (the results are discussed in Chapter 5) and was thus used for the RPPs application.

This section provided an overview of the research methods and strategies applied in the study. A detailed discussion of the methods and strategy will be outlined in Chapter 3. The study used a mixed approach as it is deemed an appropriate methodology. According to Oberiri (2017), quantitative research is aimed at the analysis and the quantifying of phenomena to acquire results through the application of specific statistical techniques. He further argues that quantitative methods are divided into the following categories: experimental research, correlational research, causal-comparative research, and survey research.

The distinct benefit of using quantitative research methods is that the data gathered can be tested, its reliability can be ascertained, its subjective interpretation avoided, and errors mitigated (Devault, 2019). This study relies on experimentation and observation as part of the quantitative approach and therefore uses positivism as its research paradigm. The suitability of this research paradigm is confirmed by Saunders et al. (2009) when they argue that the positivist paradigm guides the assumption of a quantitative approach.

1.9. Chapter Summary

This study successfully trained and evaluated a classification model using the collected corpus data. The corpus data consisted of data that best described the epistemology, ontology, and axiology aspects of the RPPs. The best performing classifier, the naïve Bayes with an 85% precision, was then used to create a web application to classify user input into research philosophies and paradigms categories. User testing of the system following the TAM was done to assess whether the system achieved its objective of helping researchers establish their RPP, thereby improving teaching and learning.

This chapter provided an introduction to the study by stating the purpose of the research and the research problem. Furthermore, the study presented the context of the study, the research questions, its significance, scope, and limitations. Ultimately, the chapter presented a brief overview of the research approach adopted and concluded with the chapter summary.

CHAPTER 2: LITERATURE REVIEW

The large volumes of digital texts and the need to have them organized has seen an exponential increase in automated text classification technologies in the last couple of years in the industry and academic communities (Sebatiani, 2002; Lei et al., 2010; Kaur & Saini, 2015). Text classification is usually designed as a learning task. A classifier is trained on how to differentiate between categories or labels based on features extracted from corpus data. The process of classifying text typically involves the use of machine learning algorithms whose accuracy is dependent on the amount of available annotated data (Lei et al., 2010). Classification algorithms include support vector machines (SVM), the Rocchio classifiers, Naive Bayes, Maximum Entropy (Sebatiani, 2002; Lei et al., 2010).

This chapter presents text classification research, followed by the theoretical concepts involved in natural language processing and the application of machine learning algorithms in classifying text. This is then followed by a review of the literature on corpora; the chapter also shows the importance of research philosophies and paradigms in conducting research and concludes with a summary.

2.1 Related Work

Natural language processing (NLP) technologies have become more sophisticated in recent years (Hirschberg & Manning, 2015). However, no research has been published about the application of NLP or machine learning for the classification of text into research philosophy and paradigm categories. Similarly, there are no corpora for research philosophies and paradigms. The extant literature refers to the use of specialized corpora for the classification or prediction of results in such fields as medicine, politics, and customer relationship management due to the unavailability of relevant corpora (Connor & Upton, 2004). These corpora have been created using multiple resources, including the internet, databases, medical journals, and inaugural speeches (Ogren et al., 2006). This section discusses work done in relation to this research study (see APPENDIX F for journal information). It focuses particularly on the use of classifiers in assigning predefined categories to text (topic

modelling), provides related studies on using the Bag of Words model, research philosophies and paradigms studies, and an overview on corpora.

2.1.1. Text classification or topic modelling

Literature does not reveal any work done in relation to the use of NLP for the prediction or classification of user ideas and concepts into research philosophies and paradigms; however, NLP has been used in marketing research (Leeson et al., 2019; Yu & Kwok, 2011), aviation (Kumar & Zymbler, 2019; Xu & Kumar, 2015; Tanguy et al., 2016; Abdebin et al., 2010), and education (Bhatnagar et al., 2016; Waters et al., 2017; Ananiadou, 2010). Leeson et al. (2019) explored the use of NLP (topic modelling and Word2Vec) in analysing qualitative data from interview transcripts, the results of which were terms that accurately defined the subjects of interviewees' ideas and concepts. Yu and Kwok (2011) found that an SVM classifier with term frequency-inverse document frequency (tf-idf) weighted part of speech features performed better than one trained with word features in classifying direct marketing and communication messages into predefined topics. In aviation, feedback obtained from Twitter where features were extracted using Glove dictionary word embedding and n-grams to classify customer experience were classified using SVM (support vector machine), convolutional neural network (CNN), and several ANN (artificial neural network). The results showed that CNN outperforms the other models in identifying associations that could help airlines improve customer experience (Kumar & Zymbler, 2019). At the centre of text classification is the bag of words model (BoW) used for representing text data in machine learning. It represents data in a numerical form that makes it possible for computers to understand. According to Brownlee (2017) and Kowsari et al. (2019), the BoW has been successfully used by machine learning in applications such as computer vision, Bayesian spam filters, NLP, information retrieval, document classification, and machine learning. Section 4.4 discusses in detail the use of the BoW model for this study. While the studies provide a great platform for the classification of data, they equally reveal a gap in research philosophies and paradigms.

2.1.2. Research philosophies and paradigms

The literature reveals studies that have been conducted to explain the concepts of research philosophies and paradigms which researchers need to understand in order to apply them in their studies (Creswell, 2013; Khaldi, 2017; Makombe, 2017; Kivunja & Kuyini, 2017; Scotland, 2012; Saunders 2009). These studies speak to the objective of assisting researchers in designing effective research proposals and applying methodologies best suited to a researcher's choice of research paradigm by exploring the philosophical underpinnings of their research (Scotland, 2012).

Saunders et al. (2009) go as far as creating a framework, referred to as the research onion, which shows the process involved in knowledge development. The research onion framework posits that the research process should be based on a philosophy that will inform the method and strategy to be adopted (Saunders et al., 2009). Scotland (2012) shows how philosophy supports research by linking the methodology and methods of research to the epistemology and ontology of research philosophy. Khaldi (2017) shows the techniques to be used in some of the research paradigms, while Kivunja and Kuyini (2017) provide a rationale for a researcher's choice of paradigm and also suggest how they can "locate their research into a paradigm".

Previous studies tend to focus on a maximum of five research philosophies, with the most prevalent ones being interpretivism, pragmatism, positivism, constructivism and realism (Bajpai, 2011; Saunders et al., 2012; Dudoskiy, 2018; Zakauskas, 2018). Awareness of just a few research philosophies and paradigms limits researchers' methods. So far, the literature does not reveal any studies that have ventured into NLP and ML's use in recommending RPPs to researchers as this study has done. The following section gives a brief overview of corpora and their use in NLP.

2.1.3. Corpora

The literature shows that there is currently no available corpus suitable for training a classification model for research philosophies and paradigms. Hence, this study

intends to create a specialised corpus of research philosophies and paradigms. Most NLP tasks have used data accessible from search engines, and their performance improves in relation to the increase of data in a corpus (Liu & Curran, 2006; Xiao, 2010; Wallgrün et al., 2014; Wagner et al., 2018). Wagner et al. (2018) developed a corpus from Twitter messages and extracted term features. They then applied text processing, including term frequencies, feature selection, and stop word removal. Further detail about the development of the RPPs corpus is discussed in Section 4.4.

The related studies discussed in this section show the application of text classification in various industries for target marketing and improving user experience. They also provided an overview of the Bag of Words (BoW) model and its use in text classification. The section also provided a review of studies based on research philosophies and paradigms related to the availability of a corpus, NLP, and machine learning. The following section provides a theoretical background on text classification.

2.2. Theoretical Background

This study is rooted in two theoretical frameworks: the theory of knowledge, the governing theory behind the research philosophies and paradigms concept, and the NLP theory that provides the tools or techniques through which the knowledge theory can better be understood using technology. The following sections discuss both the theories of knowledge and NLP.

2.2.1. The theory of knowledge

Knowledge is described in Merriam-Webster (n.d.) as the scope of understanding or acquaintance with concepts and can be created or acquired through practice, collaboration, interaction, and education. The theory of knowledge posits that to create reliable knowledge about research phenomena, the philosophical approach and paradigm that form the basis for the researcher need to be identified (Žukauskas et al., 2018). Research philosophy provides a systematic insight into a researcher's thoughts. It forms the basis for the research because it helps with the formulation of the research question, the choice of a research strategy, the collection of data, the

processing, and analysis of the collected data in research (Žukauskas et al., 2018; Saunders et al., 2009; Guba & Lincoln, 1994). Research paradigms are related to the philosophical stance and provide a conceptual framework or a set of assumptions that guide how research is carried out (as shown in Figure 2.1). According to Saunders et al. (2009), these are the epistemological, ontological, and axiological assumptions about the phenomena being researched as discussed in the next subsection.

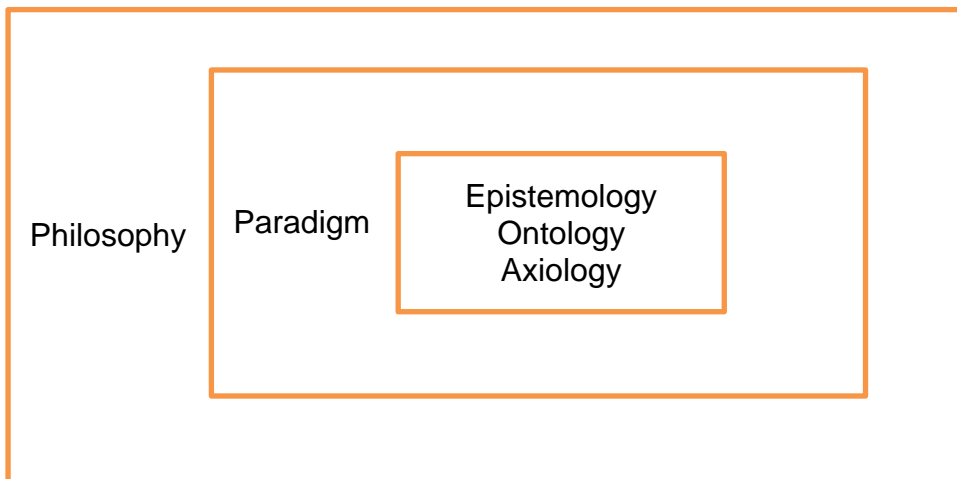


Figure 2.1 Relationship between research philosophies and paradigms

2.2.1.1. Epistemology

According to Lawson (2000), the theory of knowledge posits that in order to acquire reliable knowledge, one must be aware of their epistemic stance as it determines the kind of knowledge that is created, gathered, and presented. The creation, gathering, and presentation of this knowledge is based on a particular epistemic stance's systematic view on reality, knowledge of the reality, and the meaning that can be ascribed to that reality (Mittwede, 2012; Saunders et al., 2009; Guba & Lincoln, 1994). According to Brix (2014, 2017) and Lyles (2014), the epistemic stance concerns itself with the individual and collective creation of different kinds of knowledge, which can be achieved through various social and cognitive processes of action and interaction. Therefore, it follows that to conduct research, one needs to develop assumptions about the research, its nature, and knowledge (Žukauskas, 2018; Brix, 2014). These assumptions are based on how one views the world around

them, the nature of reality, and the impact of values on the research described in the next subsections (Mittwede, 2012).

2.2.1.2. Ontology

Ontology refers to how researchers see the world around them and formulate reality about the structure and nature of things, often independently of their existence (Merriam-Webster, n.d.; Guarino et al., 2009). The ontology describes the nature and structure of things based on their shared conceptualization or how they are generally categorized and their relations (Harispe et al., 2015; Guarino et al., 2009). Ontology aims at answering questions such as the ones that follow:

- What things exist?
- What categories do things belong to?
- Does objective reality exist?
- What is the meaning of the verb “to be”?

Ontology is closely linked to a researcher’s understanding and view of their world. This provides them with philosophical, theoretical, and methodical foundation to base their research (Žukauskas, 2018; Harispe et al., 2015; Guarino et al., 2009).

2.2.1.3. Axiology

Axiology refers to the philosophical study of goodness and value (Harispe et al., 2015; Guarino et al., 2009). These values are further categorized into ethics and aesthetics. Ethics are linked to daily human actions and the questioning of morals. Therefore, they are focused on the classification of actions or things as good or bad and the degree to which they belong to such classification (Schroeder, 2016; Mubeshera, 2012). Aesthetics involve the scrutiny of how things appear and establishing the subjective value of beauty in things.

The theory of knowledge discussed in this section provides essential concepts that this study considered in developing the application for recommending research philosophies and paradigms using natural language processing. The next section

discusses the theoretical concept of text classification using natural language processing.

The theory of knowledge discussed in this section provides essential concepts that this study considered in developing the application for recommending research philosophies and paradigms using natural language processing. The next section discusses the theoretical concept of text classification in NLP.

2.2.2. The theory of classification in natural language processing

NLP, an artificial intelligence application, is considered the bridge between computational linguistics and computer science (Frankhauser, 2015; Novoseltseva, 2017). NLP introduces computational techniques that have been theoretically motivated to analyze and represent texts that occur naturally at some or other level of linguistic analysis. It is meant to grant applications processing capabilities like humans (Hirschberg & Manning, 2015; Kumar 2012). NLP deals with language aspects such as phonology, morphology, syntax, semantics, and pragmatics (Couto, 2015; Kumar, 2012; Kurdi, 2016).

According to Reynoso (2019), NLP facilitates getting computer systems to understand the complex and diverse human text and speech to interpret and establish their intent. NLP further enables computer systems to resolve ambiguous and confusing human language by adding structure to unstructured input data (Reynoso, 2019, Filannino, n.d).

NLP has been applied in sentiment analysis, topic labelling and text or document classification (Brownlee, 2018). The accomplishment of these tasks relies heavily on machine learning algorithms to help machines understand nuances in human language (Marr, 2016). The next section provides a theoretical discussion of text classification as applied in this study.

2.2.2.1. Text classification

Text classification refers to the task of NLP that enables the classification of text into one category or another based on its content. Text classifiers have been used to organise, structure, and categorise various texts from a wide range of sources such

as files, the web, journals, documents, and user input. NLP text classification is divided into the following categories:

2.2.2.1.1. Machine learning-based systems

Machine learning-based systems use statistical techniques (or algorithms) to enable computers to parse historical pre-labelled data and learn from it for text classification or predictive analysis (Bonaccorso, 2018; Dangeti, 2017; Genzel, 2016). The experience gained from the data is used to improve or make predictions without being expressly programmed (induction); for instance, computers learn by themselves through exposure to data (Bonaccorso, 2018; Marr, 2016). Figure 2.2 depicts the involvement of machine learning algorithms in Natural Language Processing (NLU and NLG).

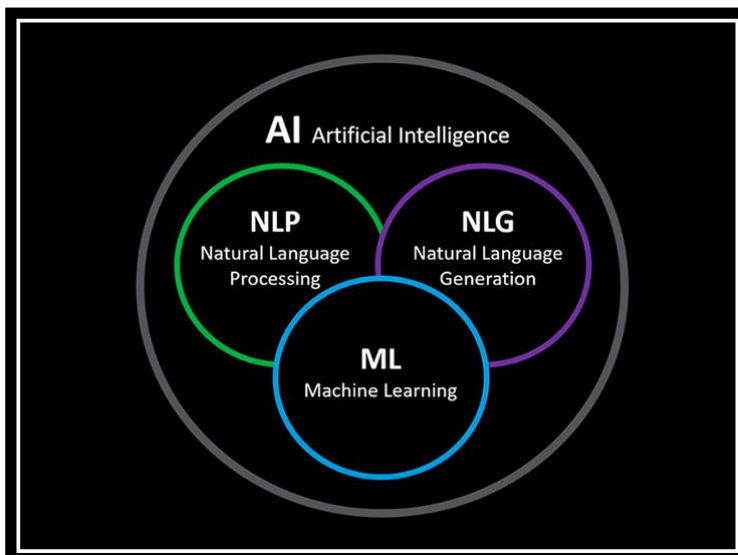


Figure 2.2 Use of ML in NLP (Conversia, 2017)

ML algorithms can learn how to associate text; for a particular input, they can link it to an expected output. ML algorithms cannot work on the raw text; therefore text needs to undergo a process called feature extraction or engineering that enables the conversion of the raw text into numerical vectors. The process of transforming the data into vectors is done through the use of the BoW, Term Frequency–Inverse Document Frequency (TF-IDF), One-Hot Encoding, or distributed representation methods (Le & Mikolov, 2014). Following the feature extraction process, algorithms are trained on how to identify data using the created vectors. The process involves splitting historical data into a train and test set; the latter is

used to validate the algorithm. Figure 2.3 shows the process followed in training a classifier.

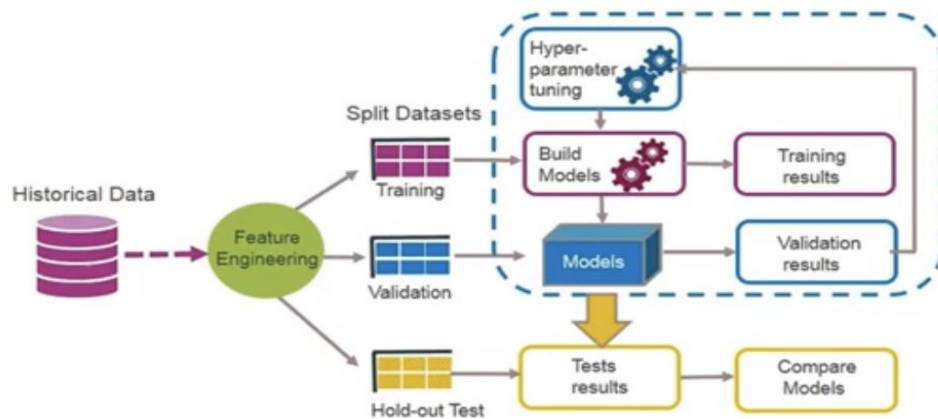


Figure 2.3 Process for training a classifier (Springboard, 2020)

ML algorithms are categorized into three types: supervised, semi-supervised and unsupervised learning (Brownlee, 2018), as depicted in Figure 2.4 and discussed in the sections that follow.

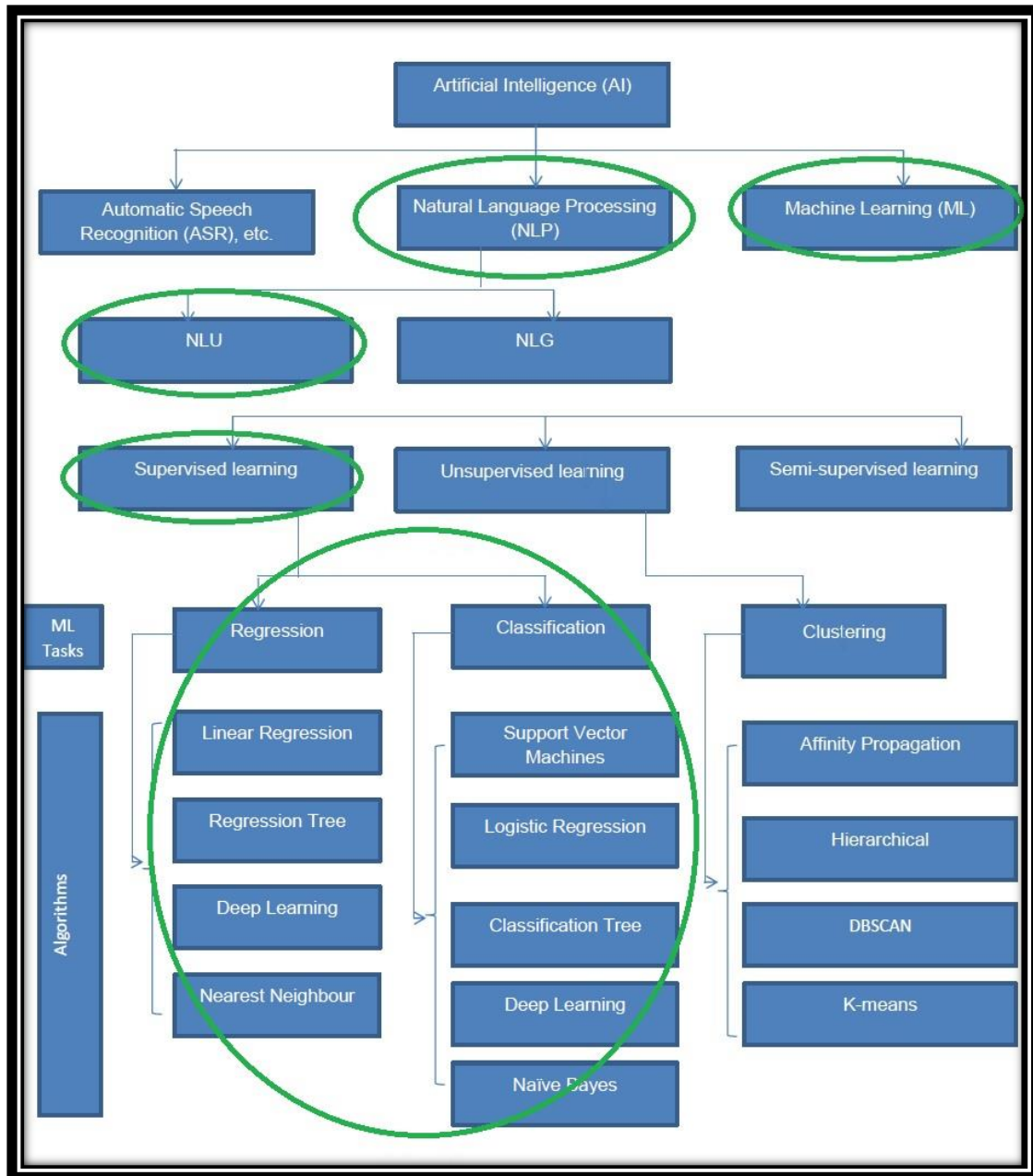


Figure 2.4 Machine Learning categories

2.2.2.1.1.1. Unsupervised machine learning

These are machine learning techniques where the models require no supervision because they can discover information independently (Brownlee, 2019; Kaur & Saini, 2015). The models use and analyse unlabelled data for the discovery of patterns and features used for classification. Unsupervised learning models typically use clustering and association algorithms.

2.2.2.1.1.2. Supervised machine learning

Supervised machine learning algorithms are those algorithms whose mapping function is provided with a set of training data from which it can learn to identify patterns used for predictions (Brownlee, 2015; Garbade, 2018). The training data gets split into a training and testing set. The training set is used to learn how to classify using certain characteristics in the set. The test set validates whether the algorithm can classify and is performing as expected. Supervised machine learning algorithms are typically grouped into the following categories (Brownlee, 2019; Garbade, 2018; Kaur & Saini, 2015):-

- **Regression:**– in regression algorithms, learning is based on the identification of patterns in input where continuous outcomes are calculated and predicted, resulting in a numerical output the regression (continuous numerical output).
- **Classification:**– Classification, one of the basic functions in data-mining, is used to structure natural language texts by automatically assigning predefined labels or categories in a process called supervised text classification (Bird et al., 2009). This task of classifying text is achieved by using classification algorithms that analyse and identify elements in the content of input texts and assign them pre-defined categories or class labels (Eisenstein, 2018; Vyatkina, 2014). According to Lei et al. (2010) and Vyatkina (2014), text classification algorithms have many practical applications in NLP. They have been used in industry and academic networks for sentiment analysis, topic modelling, cybercrime prevention, knowledge management, intent discovery, targeted marketing, and automation of CRM tasks.

Supervised text classification is regarded as a learning task due to the requirement to learn or train a classifier on using text features to distinguish between categories available in a given set of texts (Lei et al., 2010). The features used to train classifiers are automatically extracted from a myriad of labelled texts and/or documents. The automatic learning process requires

the deployment of statistical and machine learning techniques such as the Bayesian classifier, BoW model, maximum entropy, decision tree, SVM, nearest-neighbour classifiers, and logistic regression models (Pérez-Ortiz et al., 2016). The next section briefly overviews the BoW classification and naïve Bayes classifier models used in this study. The following section provides a description of the algorithms in this study.

2.2.2.1.1.3. naïve Bayes classifier

Naïve Bayes classifiers are statistical algorithms used in text classification. They are modelled around the Bayes Theorem, which assumes that features used for the category or class prediction are independent of other features in the same class (Dai et al., 2007; Peng et al., 2004; Ray, 2015). This assumption is based on the use of each individual event's occurrence based on the computation of the conditional probabilities of the occurrence of two events (Agarwal, 2012). This means that each feature in a class contributes to the probability of an item belonging to that class, with an algorithm computing the likelihood of an event belonging to a specific category.

Owing to their acquisitive learning character and the ability to predict across multiple classes, the naïve Bayes classifiers yield far better results and have a high success rate compared to other classifying algorithms (Dai et al., 2007; Ray, 2015). The naïve Bayes classifiers need less training data and perform better in cases where feature independence is assumed and when confronted with categorical input variables based on text instead of numerical values (Ray, 2015). They are mostly used in text classification, spam filtering, sentiment analysis, multi-class prediction, real-time prediction, and recommendation systems.

The family of naïve Bayes classifiers is composed of the Bernoulli, which can be applied in the presence or absence of features. Bernoulli is a binary algorithm, the multinomial algorithm that considers feature vector frequency for multi-label classification, and the Gaussian algorithm that is only applied with a continuous distribution of features (Ray, 2015; Brownlee, 2016).

Naïve Bayes classifiers have a wide application and success in NLP; however, their limitation lies in the assumption that predictors are independent and cannot be classified in cases where they do not have pre-trained categories (Ray, 2015).

2.2.2.1.1.4. Support Vector Machine (SVM) classifiers

Another supervised classification algorithm experimented within this study is the SVM. According to Agarwal and Xhiao (2012), SVM classifiers achieve their classification by first determining optimal boundaries inherent in different classes using linear or non-linear delineations between different classes to partition the data space. SVMs use a small subset of data to separate it across a decision boundary, making them probabilistic binary linear classifiers (Bridgelall, 2017). SVMs also do not require a lot of data to train WITH to provide accurate classification results; however, as they are binary classifiers, they do not work well with multi-label classification problems. Although the SVM classifiers perform very well with limited data, in cases where there are overlapping classes, their performance is below average as they are not tuned to explain classifications based on probabilities.

2.2.2.1.1.5. Logistic Regression (LR)

The logistic regression (LR) is a predictive analysis ML learning algorithm that used the probability concept to solve classification problems. This algorithm uses the sigmoid function (as shown in figure 2.5) to determine the probability of data belonging to either one of two classes, denoted by 0 and 1, as it is a binary classifier. To enable the multiclass classification experiment required for this study, the one-vs-rest scheme or the cross-entropy loss is applied by passing values to the multi-class argument in Python. The default one-vs-rest (OvR) scheme was used in training the LR algorithm to classify RPPs.

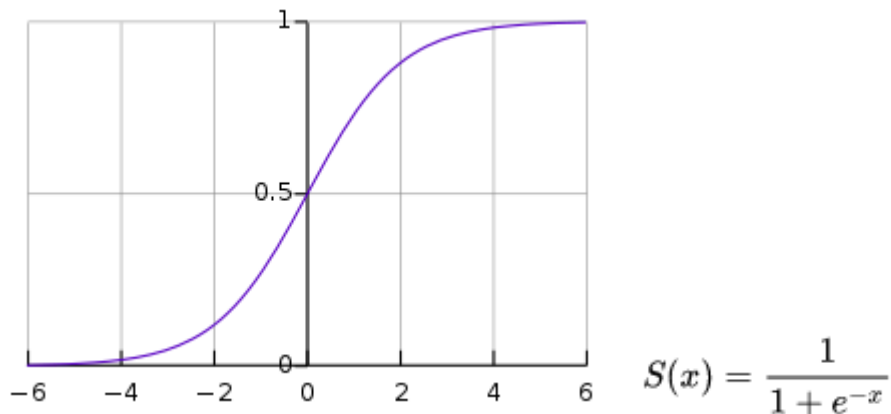


Figure 2.5 Sigmoid function and graph for the SVM algorithm

2.2.2.1.2. Rule-based systems

This kind of classification system uses a set of handcrafted linguistic rules for the prediction of text categories. The rules are based on the context of text elements that are semantically relevant to identify target categories. Rule-based classification algorithms are typically divided into:

Rule Induction Algorithms follow the IF_THEN tree structure extracted from data through association rule mining, data mining methods or sequential covering algorithms Tung (2016).

Rule Ranking Measures algorithms use a ranking system to rules that are used to predict categories or classes. They base their prediction on measuring the usefulness of a rule using predetermined values. Rule ranking measures are used in both rule induction algorithms to trim out rules that are not necessary to improve their efficiency (Tung, 2016).

Class prediction algorithms learn from training sets that have known classes and validated using test data sets for validation to predict new cases (Tung, 2016).

2.2.2.1.3. Hybrid algorithms

These kinds of algorithms are a combination of rule-based and machine learning algorithms.

Machine learning approaches to text classification rely on a collection of data referred to as a corpus to learn from. A corpus provides data for machine learning approaches and real language for the evaluation of the algorithms. The following section provides an overview of corpora used with machine learning.

2.3. Overview of Corpora

Automatic text classification of natural language depends on the availability of a collection of texts, or a corpus, to train a machine learning algorithm on how to classify texts. Corpora are important in NLP because they provide data for machine learning approaches and provide real language to evaluate algorithms. Corpora contain a collection of annotated texts that have been specifically sampled to be utilized in NLP tasks (Liu & Curran, 2006). They are generally constructed for a specific purpose and are usually representative of the genre they are constructed for. An example of some of the different kinds of corpora follows:

General corpora - represent a full range of language use varieties. They contain massive volumes of text from various domains of both spoken and written language (Leech, 1995).

Historical corpora - These kinds of corpora contain language texts from different periods. They are used to study the evolution of language over time. The ARCHER (A Representative Corpus of Historical English Registers) is one such corpus.

Specialised corpora - These kinds of corpora contain texts belonging to a particular context or genre. They usually have a finite number of texts, the size of which is manageable. Such corpora include the International Corpus of Learner English, the Michigan Corpus of Spoken English, the Reuters Newswire Topic Classification, and the Nottingham Health Communication Corpus.

The following corpora have been used for NLP and machine learning tasks, including information retrieval, classification, text summarization, machine translation, and teaching and learning.

- The DIALOG mathematical proof dataset (Wolska et al., 2004) - A Wizard-of-Oz dataset involving an automated tutoring system that attempts to advise students on proving mathematical theorems
- British National Corpus (BNC) (Leech, 1995) - contains approximately 10 million words of dialogue used for thesauri, grammar books, teaching material, and usage guides, among others (Aston, 2000).
- The BROWN corpus - acclaimed as the first million-word English corpus to be published electronically, is used for part of speech tagging.
- Gutenberg Corpus - contains free text derived from electronic books by Project Gutenberg.

Research shows that there is currently no available corpus suitable for training a classification model for research philosophies and paradigms. For this reason, this study motivates the creation of a specialised corpus of research philosophies and paradigms. According to Liu and Curran (2006), most NLP tasks have used data accessible from search engines, and their performance improves with the increase in data in a corpus. Such tasks include machine translation, text summarization, and document classification.

2.4. Chapter Summary

This chapter presented related studies conducted in other fields and the theoretical background of this study. It also discussed the theories of knowledge and NLP, specifically on machine learning, and classification with more detail on supervised classifiers and corpora.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

Research methodology refers to the process involved in systematically solving a research problem (Saunders, 2009). According to Saunders (2009), research methodology involves learning techniques used in conducting experiments, surveys, and tests. Research purposes are commonly categorized as, information gathering (discovering, uncovering and/or exploring relationships amongst variables) and theory testing (testing and understanding causal relations between variables) (Salkind, 2010; Thyer, 2010). This study follows guidelines in the Cross-Industry Standard Process for Data Mining (CRSP-DM) methodology (see Figure 3.1), which encompass both qualitative and quantitative methods, in developing the RMI web application for the following reasons:

- It provides a framework for guidelines and experience documentation
- The inclusion of industry established processes that help with the data mining tasks
- The methodology is not domain-specific and encourages best practices
- The provision of a framework for planning and managing projects

In achieving the study's objectives, the guidelines provided in the CRSP-DM were followed as detailed below:

Qualitative phase

- The methodology helped define the research philosophies and paradigms problems experienced by the researcher and, as such, enabled the definition of the study's objectives. The following analysis steps were followed in defining the problem and establishing the objectives of the study:
 - The study determined the study's objectives through a review of the literature to find out the common issues related to the identification of a researcher's philosophical stance and the teaching of research philosophies and paradigms to research students.
 - The literature review identified that there were inherent difficulties and or uncertainties when it comes to identifying individual research philosophies and guiding paradigms for many novice researchers.
 - The study determined to find out documented research philosophies and paradigms and identified the key components that make up each of

the RPPs such as the ontological, epistemological and axiological stance of each of the RPPs.

- With the RPPs identified, the study set out to train and test machine learning algorithms to determine which is best suited for developing the RMI Web application.
- The study was able to identify the kind of data that will be suitable to address the problems researchers face in identifying their relevant philosophies and paradigms
- Following the above processes, the study went on to pre-process the collected data and extracted features that can be used to train the machine learning algorithms.

Quantitative phase

- The methodology provided a systematic way to train and evaluate the machine learning algorithms for the classification of RPPs and the deployment of the trained model. The following processes were followed:
 - The study determined the algorithms to be used based on the size of the dataset or corpus
 - The corpus was split into a train and test set at 70-30%
 - The selected algorithms were trained and tested on the split corpus. Using the classification report function in Python, the outcome of the classification for the algorithms was compared to select one best suited for the study.

In addressing the objective of this study, computer experiments were conducted to evaluate classification algorithms to select the one best suited for natural language processing classification of RPPs. Developing a software system and the processes entailed in selecting the appropriate algorithm to be used in the system is firmly guided by the approaches found in quantitative methods. The positivist paradigm informs the quantitative approach as its underlying epistemological stance. Accordingly, this study is based on mixed research methods, which will be explained in detail in the sections that follow.

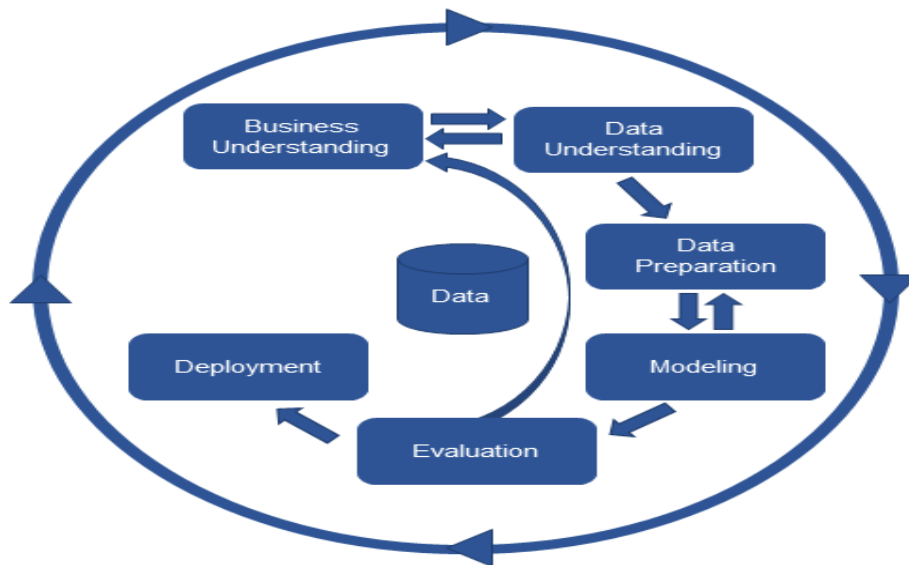


Figure 3.1 CRISP-DM steps for data mining (Dsouza, 2018)

3.1. Research design

According to Creswell (2013) and Leavy (2017), research design provides an overview of the steps through which data will be produced, collected, scrutinized, and have its meaning explained in a study (Creswell, 2013; Leavy, 2017; Peri & Bellami, 2011). In short, a research design is a framework or structure in which an inquiry can be carried out (Sileyew, 2019). Therefore, research design helps in minimizing the chances of reaching incorrect conclusions from data by answering research questions unambiguously (Inaam, 2016). This framework includes selecting one or a combination of qualitative, quantitative, and mixed methods, which are the primary ways or methods of research design (Creswell, 2013; Leavy, 2017).

In this study, mixed research methods were employed. The classification of research philosophies in natural language processing is modelled around the supervised machine learning algorithms. This requires the training and experimental testing or validation of machine learning algorithms. In selecting a machine learning classification model best suited to classify the input into philosophy and paradigm categories predefined in the RPPs corpus, the study followed the quantitative approach. Computer experiments were conducted to evaluate the classification models to determine which of the models (naïve Bayes, logistic regression, and support vector machines) yields the best results when classifying. The RPPs corpus developed, as explained in Section 2.2.2.1, was split into a test and validation set to

train the classification algorithms. Evaluation and adaption of the study's algorithm were based on the recall, precision, accuracy, and f1-measure scores of Python's classification report function.

Quantitative methods were also employed in creating feature vectors in the RPPs corpus used to train the ML classification models. To this effect, the study was able to select a suitable classification model.

The study used the qualitative approach to collect data that was used to create the corpus. A review of the literature to establish documented research philosophies and paradigms was conducted. The epistemological, ontological, and axiological components of the RPPs were selected as features for the corpus.

To evaluate the system for usability and technology acceptance, the study used a quantitative survey questionnaire to gather data from participants. This information helped to establish if the system had an impact on the improvement of teaching and learning. The sections that follow will explore the purpose of research paradigms with a brief outline of a few of them and the one adopted for this study.

3.2. Research Paradigm and Philosophy

Paradigm refers to the pattern of thinking peculiar to a study field or established research traditions. It can be viewed as a framework that includes accepted theories, models, approaches, methodologies, the frame of reference and traditions, for observing and understanding a particular field of study (Perri & Bellamy, 2012). Paradigms show us how to develop descriptions, explanations, and interpretations within the confines of the discipline we do our research (Makombe, 2017). Although there are many research paradigms such as Positivism, Relativism, Interpretivism, and Pragmatism, this section will dedicate more time to a detailed discussion of the positivist paradigm, which this study has adopted as its framework of research.

Positivism advocates for experimental observation and explanation of events that can be correlated with information in line with one's senses (Caldwell, 2010). Accordingly, positivism establishes the cause and effect relationship of variables and how these influence a particular outcome. It presumes that generalizable theoretical

models can be developed to predict outcomes that can explain the causal relationship of predictors (Flowers, 2009; Scotland, 2012). The positivist paradigm is prescriptive and purports that objective truth can only be derived or discovered through strict adherence to methodological rules in science. Positivism relies on experience as a source of knowledge that is gained by understanding human behaviour through observation and reason. The purpose of this research is to use NLP processes (tokenizing, lemmatizing, using WordNet, and named entity recognition) and ML algorithms to develop a system and create a corpus that will be used to train a machine learning model for the prediction or classification of research philosophies and paradigms. The positivist paradigm supports the systematic way of developing both the NLP classification model and a corpus used to train the classification model. For this reason, the positivist paradigm is the most suitable paradigm that will be applied in this study. To achieve this, the study will follow mixed research methods, as discussed in the section that follows.

3.3. Research Approach

Research methods are categorized into qualitative research (inquiry relies on non-numeric data), quantitative research (deal with numerical data as a way of investigating phenomena) and mixed research methods (combine at least one of the techniques in available in qualitative and quantitative research methods) (Kumar, 2014). The section that follows will discuss the quantitative, qualitative, and concludes with the mixed research design, which is suited for training a classification model and building a corpus.

3.3.1 Quantitative Research Approach

According to Creswell (2013), quantitative research involves the gathering of quantifiable data on which statistical, mathematical, and computational techniques are applied to systematically investigate a phenomenon. The quantitative research design data are collected and then transformed into numerical values and graphs for interpretation (Maxwell, 2013). These data are collected through the use of surveys, polls, questionnaires, etc. Research in software development is primarily concerned with the acceptance or appropriateness of a system developed for a specific purpose.

3.3.2 Qualitative research approach

Qualitative research is research in which non numerical data are collected and analysed so that concepts, experiences, and opinions can be understood from such data (Saunders et al., 2009). This kind of research approach is used for the gathering of in-depth understanding of a research topic. This kind of research effectively obtains information such as values, behaviours, opinions, and social contexts specific to a population with a common culture.

3.3.3 Mixed research approach

The mixed research approach which this study has adopted combined both the qualitative and quantitative research approaches in answering the research questions, thereby meeting its objectives.

The qualitative research design was suitable for this study since the qualitative tools helped determine whether the developed system and the corpus are appropriate for the enhancement of learning through the introduction of research philosophies and paradigms. The data required for this study were collected from various studies on research philosophies and paradigms and research participants.

The quantitative research approach was suitable for evaluating the machine learning algorithms and usability of the system. A post usability questionnaire using the Likert scale of 1-5 was used to gather data about the RMI NLP application's usability. The participants were purposefully selected to use the system and answer the questionnaire. The classification report function in Python was used to evaluate the machine learning algorithms to determine their precision, accuracy, and f1-scores in order to select the best algorithm to use with the RMI Web application.

3.4. Research Strategy

According to McLeod (2012), an experiment is a scientific procedure used as the primary investigation method to test a hypothesis, show or explain a known fact, or make a discovery. Experiments are used to manipulate causal variables and for the

measurement of the dependent ones to produce quantitative data used in the prediction of phenomena (McLeod, 2012). In this study, computer experiments were carried out by training and testing some of the classification algorithms to evaluate, which will have a high accuracy rate of classifying input into RPPs categories. The experiment process involved creating a corpus with RPPs categories and descriptions. A Spacy entity ruler was also created in Python for the purpose of identifying and extracting RPPs components (ontology, epistemology, and axiology) from the input. The detailed process is elaborated upon in Chapter 4.

Building a training corpus for the classification or predictive model forms part of the experimental design method. This further provides a rationale for the adoption of the mixed research method. The RPPs corpus was made available for learning and testing of the machine learning algorithms used in the experiment. Figure 3.2 displays the steps that were followed.

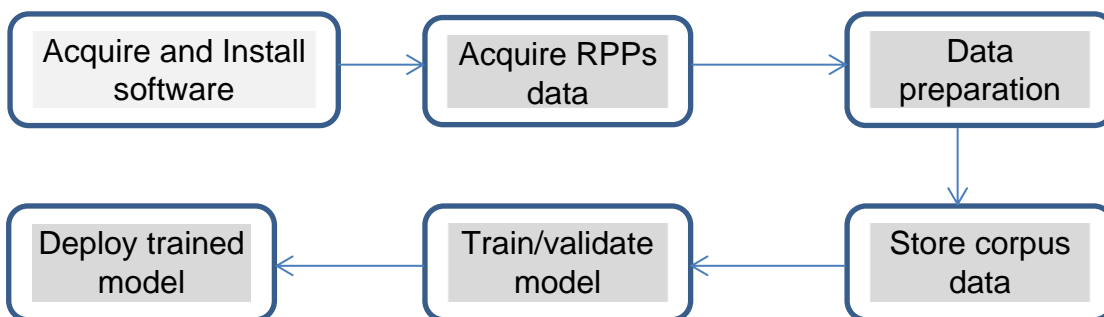


Figure 3.2 Development of the RMI application

Python NLP libraries, available through the NLTK, were installed to process text that was used to build a corpus. The NLTK includes the tools necessary for the data analysis part of the development phase.

Step 2. Acquire RPPs data

The study was able to identify the kind of data that will be suitable to address the problems researchers face in identifying their relevant philosophies and paradigms. The study further identified the lack of structured RPPs corpus; therefore data were collected from various sources (PhilPapers, Stanford Encyclopedia of Philosophy,

Google Scholar, IBSS, Philosophy Basics, Encyclopaedia) and converted into a machine-readable format. Further detail is discussed in Chapter 4, Section 4.4.

Step 3. Data preparation

- Cleaning up the collected data to correct spelling and standardize character representation and the editing of text formats. This process allows for the effective use of the collected data and the elimination of unwanted text. Having clean and standardized data allows for the learning of meaningful features by the models and helps avoid overfitting them with irrelevant noise. Cleaning and standardizing were done in this phase through the following pre-processing steps of NLP:
- **Tokenizing** (breaking down sentences into tokens):
- **Part of speech tagging:** Assigning morpho-syntactical features to words, based on their context, to enable simple syntactic searches;
- **Parsing:** present grammatical entities of sentences and the relations between them in an abstract form by producing a dependency tree (Zeroual & Lakhouaja, 2018);
- **Annotation:** Individual words in the corpus text data were annotated to identify the components of the RPPs (epistemology, ontology, and axiology). This annotation is part of feature extraction in preparation for the classification task and the named entity recognition for the components for the training of the algorithm.
- **Lemmatization:** Reducing surface words to their canonical form etc.

Step 4. Store structured data in corpus

The prepared data were stored in the corpus to train a classification model. As machine learning algorithms work only with numerical data, this process involved transforming corpus texts into vectors. The study used the Bag of Words model to represent the corpus data. Further detail is discussed in Chapter 4, Section 4.4.

Step 5. Train and select a classification model

The training was done using the natural language toolkit (NLTK) and Python's Sci-kit Learn's libraries. The developed corpus is split into a train and test set and used to learn and test the machine learning algorithms. The machine learning algorithms experimented within this study are the naïve Bayes, support vector machine, and logistic regression algorithms. Once the ML algorithms were trained and tested, the classification report was generated to show the scores relating to each algorithm's performance given the same the RPPs corpus data. The results of the evaluation are presented in Section 5.1.1.

Step 5: Deploy the model

Once the algorithms are evaluated, the one that performed well and was deemed fit for the study was used as a model for the web application. In this process, the trained model was deployed as a web application using the Django web framework as detailed in Chapter 4, section 4.3.

Following the steps detailed above, the application could classify user input into research philosophies, and paradigm categories once deployed. A series of user tests ensued to test the usability of the deployed syweb application. Section 5.3 covers the details of the usability tests. The following section discusses the sampling techniques engaged in this study.

3.5. Research population and Sampling

Research population refers to a collection of objects or individuals who are the focus of a research investigation for a particular criterion, and the definition of the objects of a particular research population is dependent on the nature of the research (Alvi, 2016; Miles et al., 2018; Thompson, 2012). The descriptions of the research philosophies and paradigms, the classification machine learning algorithms and the participants were purposely selected to constitute a population for this study. According to Thompson (2012) and Daniel (2011), sampling involves selecting a small part of a larger population group (sample) to observe to estimate or determine characteristics or parameters about the whole population without perturbing or disturbing that population. Due to notable constraints in resources (time, financial

and human resources), the sample should be big enough to warrant statistical analysis (Miles et al., 2018). Research sampling is divided into probability and non-probability sampling. It is distinguished by whether everyone in a sampling frame has an opportunity of being nominated for the study or not.

This study used purposive sampling, a form of non-probability sampling, to collect both research philosophies and paradigm data and test the NLP application for appropriateness and acceptability. The purposive sampling was appropriate for this study as researchers can choose participants by relying on their own judgement. Purposive sampling was used to collect RPP sample data and sample participants: graduates, post-graduates, and academics to test the system. This sampling technique provided the study with necessary insight from the targeted population of both RPPs and the system's target users.

3.6. Data collection

The process through which information on a topic of interest is systematically gathered and measured to answer research questions, testing hypothesis, and the evaluation of outcomes is referred to as data collection (Boslaugh, 2010; Hafiz et al., 2014; Ohlsen, 2012;). Data collection is classified into primary and secondary data (available data that has already been published) collection methods and must be informed by the research design in line with the sampling choice (Daniel, 2011; Hafiz et al., 2014; Phillips & Stawarski, 2016).

Publications containing research philosophies and paradigms were manually obtained from various sources such as books, journals, dictionaries and records for inclusion in the RPP corpus. According to Zeroual and Lakhouaja (2018), corpus data are mainly collected manually, automatically or through crowdsourcing using the help of experts. A questionnaire was used to gather data from participants to determine whether the developed system and the corpus are appropriate for this study. The following section describes the necessary analysis and processing of the collected data.

3.7. Data analysis

Data analysis refers to the process wherein data collected for a research study is interpreted and summarized to determine patterns, trends, and relationships; and involves applying deductive and inductive logic (Zeroual & Lakhouaja, 2018). Data analysis is categorized into qualitative and quantitative techniques. In creating the RPPs corpus, the collected data was cleaned and standardized using the pre-processing tasks for natural language processing. The RPPs data were manually annotated, taking into account the provided annotation guidelines for the various components of the RPPs limited only to the following:-

- **Research philosophy and paradigm names:** RPP names were annotated according to the components below:-
- **Ontology:** Phrases relating to the nature of reality, being, or existence will be annotated as such;
- **Epistemology:** The annotation of texts relating to the methods of acquisition of knowledge, the justification of belief; and
- **Axiology:** Information or texts relating to values, ethics, or aesthetics will be extracted and annotated. The annotation process will engage processes such as lemmatizing, speech tagging, parsing, stemming, and tokenization (see APPENDIX L). Analysis of text will include the following (Zeroual & Lakhouaja, 2018);

The data gathered from participants were analysed using the IBM SPSS Statistics software, a tool for manipulating and deciphering survey data. The variables to be tested were coded into the SPSS variable view, where the coding of the scores per variable was done. Responses from the participants' questionnaires were captured into the data view section. Once all the input was captured, the data were analysed using the descriptive statistics algorithms for frequencies and cross-tabulation. The results from this process are discussed in Chapter 5, Testing and Results.

3.8. Reliability and Validity

Since the classification or prediction of RPPs relied heavily on the RPPs corpus, it was therefore imperative that the annotated texts are validated. Validation in this

study included comparing the same annotations from different independent coders, as this study is part of a bigger project involving two other coders. According to Artstein (2007), reliability can be measured by assessing the following:

Stability: the same annotations at different points should not give significantly different results;

Reproducibility: different coders should be in agreement and produce nearly the same annotations; and

Accuracy: an expert's and coder's annotations should not be far different.

The measures above were observed to produce reliable and valid corpus. This study achieved reliability by comparing annotations with those of the other coders in the project.

3.9. Limitations and Constraints

Some of the challenges in constructing the RPPs corpus were ensuring that the corpus data represents the selected theme (i.e., the findings can be generalized) and that appropriate data are identified (Sealey & Pak, 2018). Of the 323 identified RPPs, only 180 of them could be used for the study. Most of them had no readily available text due to the majority of research focusing on the four major RPPS: positivism, pragmatism, interpretivism and postpositivism for their studies. Another reason was that the RPPs are not popular in academia and research as many researchers are only exposed to the four most common ones. Zeroual and Lakhouaja (2018) maintain that collecting sufficient data on each text category of a corpus remains a challenging task. According to Sealey and Pak (2018), the compilation of a thematic corpus is not yet established. It is not advisable as it presents challenges when it comes to identifying linguistic items of the entities to be denoted.

3.10. Ethical Considerations

Researchers are bound by the ethical code of the institutions they serve and the country's constitution in which they conduct their research. There are ethical and psychological implications for conducting research. Similarly, there are consequences for research participants about which research must reflect, anticipate, mitigate, and address. Researchers must recognise, acknowledge, and

respect the rights of their research participants. Researchers may not infringe the rights and dignity of participants in their research. As such, the participants in this study were informed and consented to be part of the study. They were also made aware that there is no monetary gain to participating in the study. The information provided by the participants was not divulged to any other person in any form, except to the participant concerned. Equally, researchers may not cause harm to their employers through their research. They shall uphold the professional standard and conduct their investigation with the highest level of integrity. They must be transparent and honest (Shamoo & Resnik, 2015). The participants in this study were made aware of the intention of the study and were asked to sign a consent form before participating in the study (the consent form is attached as APPENDIX H). To uphold the above-mentioned moral, ethical, and statutory considerations, the researcher scrupulously acted per the Software Engineering Code of Ethics and Professional Practices, which provides guidelines for the professional, responsible and ethical way in which software engineers should behave (Gotterbarn et al., 2001) and also obtained the ethical clearance certificate from UNISA (APPENDIX I).

3.11. Chapter Summary

This chapter outlined the research methodology for the study, motivating for the use of quantitative methods. The data collection and analysis process were also discussed. The limitation and constraints and the ethical considerations of the study were discussed. The next chapter addresses the implementation of the NLP system.

CHAPTER 4: SYSTEM DESIGN AND IMPLEMENTATION

The literature shows that research philosophies and paradigms are important in knowledge generation. They provide a meaningful and credible way in which a researcher can collect, analyse, and interpret data for their study. It further reveals that many novice researchers do not know their guiding philosophy and paradigm at the onset of their research and that establishing these is quite confusing. Research philosophies and paradigms have been widely documented in various journals, encyclopaedia, scholarly articles, books, and search engines. In contrast, in academia, researchers are only exposed to just a few of them. Gathering the information into a corpus will make it easy for researchers to discover and learn about different research philosophies and paradigms and identify one that resonates with their studies. In this study, the RPPs corpus is used for training algorithms in classifying text data into RPPs categories. Machine learning embodies several classifying algorithms that can be deployed together with natural language processing techniques for this purpose. A combination of the ML algorithms, natural language processing techniques, and Django web framework are used in this study to make the following contributions: 1) create a research philosophies and paradigms corpus that will be used to train machine learning algorithms, 2) train and test machine learning algorithms and select the best performing one to develop a web-based application for classifying user input, and 3) evaluate the performance of the web-based application. This will help researchers identify research philosophies and paradigms in line with their studies, thereby improving the knowledge generation process. Chapter 2 presented the literature related to this work area while Chapter 3 presented the methodology.

This chapter presents the approach used to develop the proposed Research Methods Index NLP application for classifying text. The chapter outlines the process followed by discussing the system architecture in 4.1, the system design in 4.2, the implementation of the system in 4.3, and the input data required for the system in 4.4, followed by a summary in 4.5. See APPENDIX G (Figure 8.1) for the diagram relating to the system architecture.

4.1 Systems Architecture

As the RMI application is meant to be used for academic purposes, the study saw it fit to create a web-based application that can be accessed from multiple locations. This meant that the system's architecture should cater to multiple users at different locations by employing technologies that will make that kind of access possible.

The NLP application is hosted on the Microsoft Azure public cloud computing platform to ease access from multiple locations. The study adhered to the architectural style referred to as a representative state transfer (RESTful) client-server design in deploying the NLP application. Python was used to write an NLP code that is used for the classification of input. The Django Web framework was used to construct the NLP web API and used for passing data between the user interface and the back-end.

The user interface was developed using the Django Web framework. The Django Web framework supports the Model View Template (MVT) pattern, where the developer provides the model, and Django uses the template and views to map the model to the Uniform Resource Locator (URL). The Django Web framework then renders the URL to the user interface. APPENDIX G (Figure 8.2) shows how the components of the MVT pattern interact in response to a user's request.

The MySQL database stores information on research philosophies and paradigms, user account information, user input, system roles, and reports. A user will be provided with a link to the system, allowing them to register an account, answer the questionnaire, and then view their report.

The four primary components of the NLP Research Methods Index (RMI) application considered for this study are discussed below:

4.1.1 Web Browser/Client

The web browser resides in a personal computer that is connected to the internet through an ISP or other means. A researcher uses the web browser to

connect to the RMI application. The web browser is linked to the webserver through a POST/GET method;

4.1.2 Webserver

The communication between the client and the webserver is through the Hypertext Transfer Protocol (HTTP). The function of the server is mainly to store, process, and deliver web content (HTML pages) as requested by the client;

4.1.3 NLP application/Server-side

The server-side refers to the NLP application that processes input, classifies it into RPPs categories, and then issues a participant's report. This report is based on the information request from the MySQL database; and

4.1.4 MySQL DB

The MySQL database stores all RPPs data, participant data, information and responses, which are then used to generate a report that recommends research philosophy and paradigm categories to participants.

4.2 System Design

This study's output is a software application that will accept user input and classify it into research philosophies and paradigms using machine learning and natural language processing. Part of the software development lifecycle includes the design phase to set a specific standard to be followed during the development phase to minimise flaws to the system. In designing the RMI NLP application, the study followed the Do not Repeat Yourself (DRY) software design principles. This design principle was opted for because it guards against the duplication of code through abstraction, thereby avoiding code complications. It also makes the code easy to debug in cases where there are bugs in the code. According to Rodger (2011), applying the DRY principle in software development improves the performance of applications being developed and also the effort required because of the reduced number of lines of code needed. In developing the NLP RMI application, the code

used to calculate the class score commonality for training a model has also been used to calculate the input variable's score.

The following section provides an understanding of the behaviour of the system and how it does the classification of a researcher's sentiment into research philosophy and paradigm categories. It employs the use case-, class- and entity-relationship diagrams (ERD) to provide a detailed systems design of the RMI (NLP) application. The diagrams and their explanations will follow in the next subsections:

4.2.1 Use case diagram

The NLP system's detailed system design is presented in this section through use case-, class- and entity-relationship diagrams (ERD). There are two role players or actors in this system: the researcher, who interacts with the system by undertaking a questionnaire, and the administrator responsible for maintaining the system. The researcher must register on the system, sign-in, answer the questionnaire, and view their report. Figure 8.3 shows the user interaction with the NLP application. The administrator will sign in to the system to maintain tables and user accounts as shown in APPENDIX G, Figure 8.4. The sub-section that follows, sub-section 4.2.2, presents the class diagram of the NLP application.

4.2.2 Class Diagrams

This section intends to show the NLP application structure by showing the classes represented in the application, together with their attributes and methods.

The NLP RMI system has four class diagrams representing the classes used to achieve the goal of classifying user input into the RPPs. These classes are created using the Django model script. The class diagrams are shown in APPENDIX G (Figure 8.5). The `auth_user` superclass has eleven attributes (ID, Password, Last_login, Is_superuser, Username, First name, Last_name, Email, Is_staff, Is_active, Date_Joined). These attributes are captured by users (administrators and researchers) when they register to access and use the system.

The login method inherited by both the subclasses researcher and administrator is used to sign-in on the system. This method passes username and password that are used to verify the credentials of a user and grants them access once verification is successful. Both researcher and administrator use the other methods as in the functions that follow: `take_questionnaire()`, for answering the questionnaire provided, `generate_report()` to create and view a report from the captured questionnaire answers, `tokenize()` to create tokens from the answers, `lemmatize()` to lemmatize some of the tokens.

As the two role players on the system, the researcher and administrator inherit data from the user superclass. The administrator has additional methods as follows; `modify_user()` change user credentials, `view_user_reports()` view other reports, `delete_user_accounts()` delete inactive user accounts. The `nlp_paradigm` class contains the names and descriptions of research philosophies and is linked to the `nlp_user_answer` class with an 'is-a' relationship.

The `nlp_user_answer` class is used to store the answers a researcher provides and is associated with the user superclass through a 'has-a' relationship. The `nlp_user_answer` is used to generate a report for a user and present results using the `nlp_cluster` and `nlp_paradigm` classes. The `nlp_cluster` class contains the paradigm IDs, names, and their components (ontology, axiology, epistemology) and is associated with the `nlp_paradigm` class through a 'belong-to; relationship.

4.2.3 Entity Relationship Diagram

The RPPI prototype system database consists of four key tables and Figure 8.6 describes the entity relationship diagram (ERD) of how the tables relate to each other. MySQL relational database management system is used to develop the NLP system as it is open-source. Three tables (`nlp_paradigm`, `nlp_cluster`, and `nlp_user_answers`) were created by running the `models.py` script in Django to define fields and behaviours of the MRI-NLP application that will be storing the data. The `auth_user` table is created using MySQL script, and it stores user credentials (both administrator and researchers). The entity-relationship diagram of the tables is shown in APPENDIX G (Figure 4.6).

The primary key for the `auth_user` table is the ID field and it links with the `nlp_user_answer` table for authentication purposes through a one-to-one relationship, as shown in the Class diagram in Figure 8.5. The `nlp_user_answer` table stores user answers and will be used for classification purposes. The `nlp_paradigm` table, whose primary key is ID, links to the `nlp_cluster` table, with a one-to-many clustering of research philosophies into respective research paradigms.

4.3 Implementation of the NLP Application

The NLP system involves the user (front-end) and the functionality (backend) components. The NLP system's implementation, which includes technical requirements, technologies used, and the interaction of both the back- and front-end components, are discussed in this section.

The process of text classification as seen from the point of view of automatic text classification systems can be clearly separated into two main phases, namely, (1) information retrieval phase when numerical data is being extracted from the text; and (2) main classification phase when an algorithm processes this data to decide as to which category should the text belong.

4.3.1. Technical requirements

The minimum technical requirements for developing, running, and deploying the NLP system are detailed in this section. The systems specifications, implementation environment, and functional testing, will be discussed. The NLP system is web-based and runs on any of the available internet browsers but requires internet connectivity to be accessed by anyone who intends to use it. Table 4.1 shows the development specifications.

Table 4.1 Development and deployment specifications

	Development Environment	Development Environment
Operation System	Windows 10	Linux
Memory		
Storage instance	500G RAM	500G RAM

Platform	64-bit	32- or 64-bit
Programming Language	Python 2.7.x or 3.4.x	Python 2.7.x or 3.4.x
Database	SQLite, MySQL	SQLite, MySQL
Web Framework	Django	Django
Package manager	Pip/Anaconda	Pip/ Anaconda
Integrated Development Environment(IDE)	Spyder/Jupyter/Python IDE	Spyder/Jupyter/Python IDE

4.3.2. Technologies Used

The front-end (user interface) and back-end solutions of the NLP system are implemented using open source technologies and additional hardware technologies at hand. This classification depends on the RPP categories corpus, which contains the descriptions of research philosophies and paradigms. The RPP corpus uses the tokenizer, stemmer, and WordNet lemmatizer in preparation for storage as a corpus with NLTK (see APPENDIX G, Figure 8.7 The NLP process flow diagram for text classification). This is then used to train a classification model that will be used to classify user input. A classifying model is the algorithm or a procedural process a computer follows in accomplishing a task that assigns input text categories depending on its content (Miller and Ranum, 2013). The sub-sections that follow discuss the details of each interface;

4.3.3. User Interface

The Django web framework, written in Python on the Anaconda IDE, was used to develop the user interface. The web framework Django is suitable for both front-end and back-end solutions as it has a complete collection of suitable libraries. Some of these libraries include the URL management templating language, the authentication mechanism, and navigation tools. These libraries facilitate data selection, formatting, and displaying of texts. Django runs on an activated Python virtual environment as follows;

- Create project applications by running the command: *python manage.py startapp nlp_project* and updated the setting.py file with the installed applications as shown in APPENDIX K (Figure 8.10)

- The database tables were created by observing all installed apps in the settings.py file using command: *python manage.py migrate*
- The administrator user for the nlp_project was created using the command: *python manage.py createsuperuser*
- Create an application for the project to host project settings and the WSGI using command: *python manage.py startapp core* WSGI_APPLICATION = 'nlp_project.wsgi.application'
- Adjust database settings on the settings.py file to reflect the database used, in this case it is MySQL, and to set up the URL templates as represented by (APPENDIX K, Figure 8.9)

Once the database was created the following stages ensued:

Creating the data model

The nlp_project data model was defined by adding the code in APPENDIX K (Figure 8.11) to the models.py file in the nlp_project folder. After defining the models, the database schema was updated by running the commands in APPENDIX K (Figure 8.12) to migrate all changes and resulted in the schema in Figure 8.8

Designing the URLs

The URLs for the nlp_project were designed by editing the urls.py file and adding URL patterns for the NLP application as in APPENDIX K (Figure 8.13).

Creating custom views

Custom views represent the data that is being passed between the web interface and the backend of the RMI NLP application. The custom views were defined in the view.py file through the code in APPENDIX K (Figure 8.14).

Creating the RMI NLP Application's templates and forms

The following web templates for the nlp_project were created in HTML for user interaction;

Login.html- used every time a user wants to use the system

Nlp_questions.html – for the questionnaire and answers

Nlp_results.html – for displaying the graph showing the percentages per RPP

Nlp_report.html – for showing the report after submitting the questionnaire answers

Register.html – for user registration

The SignUpForm and the nlp_form for user answers forms are automatically created through the forms.py script as in APPENDIX K (Figure 8.15).

Once all these forms and templates are created, invoking the command `python manage.py makemigrations` updates the database schema with all the changes as shown in APPENDIX K (Figure 8.16). In APPENDIX D, a user manual of the NLP system is attached, showing all the templates that were created in this process.

4.3.4. Researcher's interaction with the RMI NLP application

A researcher is provided with the link to the RMI application, deployed through the Azure cloud service. The researcher interacts with the system by initially registering a user account for them to be able to sign-in. The user account is activated when the researcher signs in after the successful account registration process. Once signed-in, the user can proceed to take a questionnaire and answer the questions posed. At the end of the questionnaire, the researcher can submit their answers, view their report, and sign-out. Figure 4.1 shows the flow diagram of this interaction; the full procedure is illustrated in Figures 8.24-8.29 with screenshots of the web application.

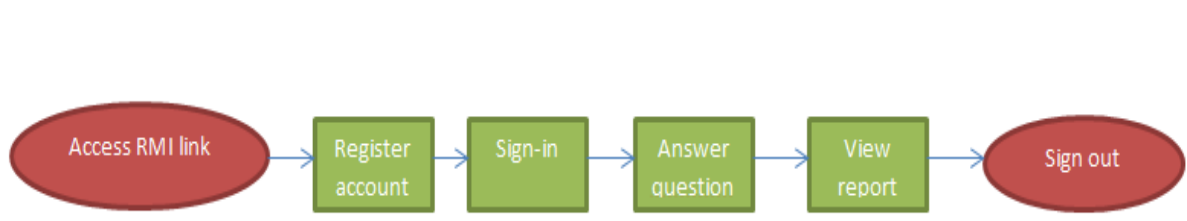


Figure 4.1 RMI flow diagram

4.3.5. Administrator's interaction with the RMI NLP application

The administrator is assigned the super-user role, enabling them access to the database to maintain tables (modify, insert and/or delete table entries) and views.

4.3.6. Back-end

To setup, a development environment for the NLP application, the open-source Anaconda platform for Python data science was acquired installed by downloading the files from <https://www.anaconda.com/distribution/>. Anaconda runs on platforms including Linux, Mac OS and Windows. It is regarded as the industry standard for developing, testing, and training machine learning models and data science. The Anaconda platform has a host of integrated development environments (IDE), including Eclipse and PyDev, IDLE, Spyder, and Jupyter. This study used the Scientific Python Development Environment (Spyder) IDE to develop scripts to train and test the classification model.

The Django Python Web framework is also used for the backend for manipulating data sources by creating a subdirectory of management and command scripts, as shown in APPENDIX K, Figure 8.17.

The `init__.py` script was used to mark directories in the `nlp_project` directory as Python package directories. The Python model or script for classifying user input was also saved on the NLP project directory and the RPPs corpus used to train the model. The `views.py` script acts as a controller by linking the HTML templates to the `RMITextclassify.py` script. The script passes a string parameter, obtained and concatenated from the questionnaire answers, to the `RMITextclassify.py` script. The `classify(sentence)` function of the `RMITextclassify.py` processes the parameter and returns the respective RPPs categories to the `views.py` script. The script then passes the categories to the `nlp_report.html` script for viewing on the web using the 'GET' and 'POST' methods as shown in APPENDIX K, Figure 8.18.

The Django web framework uses the object-relational mapping layer to interact between the application and the MySQL relational database. The backend solution for training and testing the classification model was developed in the Anaconda platform through the Spyder IDE. Figure 8.19 shows the `classify(sentence)` function of the `RMITextclassifier` model. The `RMITextclassifier` model, in APPENDIX K page 154, is a set of instructions written in the Python programming language. The model accepts user input and then process the input for the recommendation of research

philosophy and paradigm categories. This function returns the recommended RPPs categories to the views.py script in APPENDIX K, page 175 .

The open-source relational database system, MySQL, is used to store user profiles, the questionnaire, user answers, paradigm, and clusters, as shown above. The database tables were created by executing the models.py script of the Django web framework. The models.py script works as an Object-relational mapping tool by creating a virtual object database or python representation of entities in a relational database management system (RDBMS), MySQL, by using python classes. The code for the models.py script is shown on APPENDIX K page 153 and the resulting database schema in APPENDIX G (Figure 8.16).

4.4 Input Data Creating the RPPs corpus

In this study a corpus of research philosophies and paradigms was created from various philosophy publications including but not limited to; PhilPapers, Stanford Encyclopedia of Philosophy, Google Scholar, IBSS, Philosophy Basics, etc. The data collected for the corpus was mainly focused on the three components that represent specific features of each RPPs category: epistemology, ontology and axiology. A total of 323 RPPs were collected from the various sources, but the study used only 180 as not enough data could be gathered due to the unavailability of data for the excluded RPPs. APPENDIX N shows the data statements worksheet for the RPPs corpus. The bag-of-words model, or BoW, is used for the RMI application to represent text data for the RPPs corpus. The BoW involves using a vocabulary of known words and measuring known words in a context. Following is a detailed explanation of the steps undergone in creating the corpus:

4.4.1 Data collection

The classification model in this study required a specific corpus or dataset of RPPs to be used for learning. The corpus used for this study was created with data obtained from various sources such as PhilPapers, Stanford Encyclopedia of Philosophy, Google Scholar, IBSS, Philosophy Basics, Encyclopaedia, etc. (see APPENDIX M). These sources yielded 323 research philosophies, although most of

them did not have enough content (data and/or reviews) to be used for the study. This study considered this collection of data representative of the available RPPs. The final corpus consists of 180 research philosophies and paradigms with information relating to their epistemology, ontology, and axiology, excluding the ones with little or no data. The study focused on collecting data on the three components because, although each of the RPPs serves the same purpose of generating knowledge, how that purpose is achieved differs for each one of them depending on the nature of reality, what constitutes knowledge, and the role of values in a study. Figure 4.2 shows an example of the differences in the components of four RPPs as adopted from Saunders et al. (2009). A spreadsheet sample of the corpus is attached in APPENDIX E and Figure 8.22 shows an example of how the algorithms are trained using the spreadsheet.

Ontology (nature of reality or being)	Epistemology (what constitutes acceptable knowledge)	Axiology (role of values)
Positivism		
Real, external, independent One true reality (universalism) Granular (things) Ordered	Scientific method Observable and measurable facts Law-like generalisations Numbers Causal explanation and prediction as contribution	Value-free research Researcher is detached, neutral and independent of what is researched Researcher maintains objective stance
Critical realism		
Stratified/layered (the empirical, the actual and the real) External, independent Intransient Objective structures Causal mechanisms	Epistemological relativism Knowledge historically situated and transient Facts are social constructions Historical causal explanation as contribution	Value-laden research Researcher acknowledges bias by world views, cultural experience and upbringing Researcher tries to minimise bias and errors Researcher is as objective as possible
Interpretivism		
Complex, rich Socially constructed through culture and language Multiple meanings, interpretations, realities Flux of processes, experiences, practices	Theories and concepts too simplistic Focus on narratives, stories, perceptions and interpretations New understandings and worldviews as contribution	Value-bound research Researchers are part of what is researched, subjective Researcher interpretations key to contribution Researcher reflexive

Figure 4.2 Differences in the components of RPPs (Saunders et al., 2009)

Once the relevant data were collected, the study proceeded in pre-processing the data for feature selection for each of the RPPs class labels or categories. According to Agarwal (2016), the use of class labels to supervise the process of selecting features is necessary as it ensures that features highly skewed towards a certain category are selected for the learning process. The following section details processes that were followed in creating the RPPs corpus in preparation for the learning and validation of the classification algorithms.

4.4.1.1. Pre-processing phase

In the pre-processing phase, the collected texts are cleaned up, removing special characters and formats to present the texts in clear word order. The following processes are involved in the pre-processing phase;

- **Tokenization:-** Tokenization is useful in NLP as the tokens can be used to, amongst other things, count the number of words present and word frequencies in text. Tokenization involves using a lexer (lexical analysis) in identifying instances of sequence of characters or words, referred to as tokens, in a given sentence (Nardkarni et al., 2011; Thanaki, 2016) through invoking the *nltk.sent_tokenize(sentence)* function in python as a precursor to stemming and lemmatization. The function produces the tokens by considering the beginning and ending of words as boundaries.
- **Part of speech tagging:** Based on their context, the tokens are then assigned morpho-syntactical features that will make syntactic searches of words possible (Zeroual & Lakhouaja, 2018).
- **Stemming and Lemmatizing:** Lemmatization reduces words to their dictionary form, considering the meaning of words in sentences or nearby sentences, whereas stemming establishes relationships between words by reducing them to their basic or root form (NSS, 2017; Thanaki, 2016; Pedrycz & Chen, 2016). In different lemmatizing forms of the same word were converted from verbs, singular and plural forms, and tenses to inflected forms.
- **Removing stopwords:** In this process, common words like 'the', 'a', 'this', 'for', etc., are removed from the collection of words in a sentence. These

words bear no weight in the categorization of RPPs due to their use in regular expressions. Common words not specific to any RPPs categories were identified and removed using the `set(stopwords.words('english'))` function available in Python. Such words include words such as; 'the' 'by' 'such' 'when' 'at', etc.

4.4.1.2. Feature selection

This step is done to mitigate the high feature space dimensionality problem experienced in text classification by determining and measuring the importance of words and retaining only the highly scored words as relevant features (Kaur & Saini, 2015; Tang et al., 2014). The features this study considered relevant are texts relating to how knowledge is acquired, how researchers view the world around them, and what impact their values have in conducting research based on the research philosophies framework Saunders et al. (2000) and Guba and Lincoln (1994). According to Tang et al. (2014), the feature selection process ensures that the time it takes to train a classification algorithm is reduced. Once the features are selected they are extracted and vectorized as explained in the next subsection;

4.4.1.3. Feature Extraction

Owing to the high dimensionality of text features and noisy features in text data, feature extraction is fundamental in classification tasks (Agarwal, 2016). This process involves creating vectors from the available text by scoring each word or token that appears in a category or class and representing the result as a numeric feature of that class. Although the text can be represented in different ways, this study uses the Bag of Words (BoW) model. In text classification, this model has been deployed widely due to its ease of use (Aggarwal & Zhai, 2012). The BoW model is explained in the next subsection;

4.4.1.4. The Bag of Words model

BoW text modelling technique is used in NLP to convert text into numbers (vectors) for classification or use with any other ML algorithm (Soumya & Shibily, 2014). This

model converts text into BoW for text categories and stores the total count of occurrences of frequently used words in a category. The words are used as the basis for representing a specific category. Traditionally the BoW has been used for topic modelling and text classification (Sebastiani, 2002). With further technology developments, the BoW model is used for feature extraction for training machine learning algorithms. Soumya and Shibily (2014) augmented the BoW by incorporating the co-occurrence (word terms that frequently appear next to each other) of word sets as a feature for classification.

The Bag of Words model represents texts belonging to a category as tokens placed in a container regardless of their structure, grammar, syntax, or word order (Kowsari et al., 2019). The model uses word frequencies as features to characterize texts, which are represented as numerical vectors.

Creating vectors involves scoring each word or token that appears in the description of the RPPs. Owing to the undetermined number of words in the RPP vocabulary, the study used an indefinite-length document representation to score the words and mark them as a Boolean value to show their presence or absence (0 or 1, respectively). All the different words for the RPPs are represented as a BoW for the corpus. The respective frequencies in each category are regarded as features (as shown in Table 4.2) used as equivalent vectors of numbers for the words with a fixed length. Figure 8.23 shows the scores of occurrence in the corpus.

Table 4.2 BoW vector representation of corpus text data

- *RPP1 : The world is composed of two fundamental substances*
- *RPP2 : External natural kind terms independent of language*
- *RPP3 : Objects in the world exist and function only as relational entities in the world*

	W or ld	Co mp ose	Fund amen tal	sub stan ce	Ext ern al	na tur e	ki nd	indep ende nt	te m	Lan gua ge	O bj ec t	e xi st n t	fun ctio n	rel at e	en tit y
RPP1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
RPP2	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0
RPP3	2	0	0	0	0	0	0	0	0	0	1	1	1	1	1

After assigning the vocabulary for each RPP, the occurrence of words in each RPP is scored by calculating the frequency at which each feature appears in each category of an RPP. These features include common words like 'the' 'this,' etc., which yield a high number of frequencies in any given set of text (Brownlee, 2017; Kowsari et al., 2019). To avoid this, the BoW model uses the term frequency-inverse of document frequency (TF-IDF) to normalize them for training a classifier and determining the input text category (Brownlee, 2017). This is achieved by rescaling the word frequencies by comparing how often the words appear in each of the RPPs, with the most frequently appearing words such as 'there' and 'that' being ignored. This then produces a final score, which is used as a final weight or feature for each RPP. The extracted RPPs features or vectors are then used to train classifiers because ML algorithms cannot handle text directly but work with numbers (Aggarwal & Zhai, 2012). Although the BoW model has seen successes in text classification, it also has its inadequacies. These are due to the model's requirements for a carefully designed vocabulary, difficulty in modelling sparse representations of words, and the disregard of context and meaning (Kowsari, 2019).

Once the user has completed the questionnaire on the RMI application, the input is concatenated into a string used for the next process of calculating scores. A comparison of the user input's score against the corpus category scores yields the three topmost RPPs closely linked to the input, which represents a researcher's worldview. The results are presented in a chart showing the degree to which a researcher is aligned to a particular RPP, as in Figure 4.3:

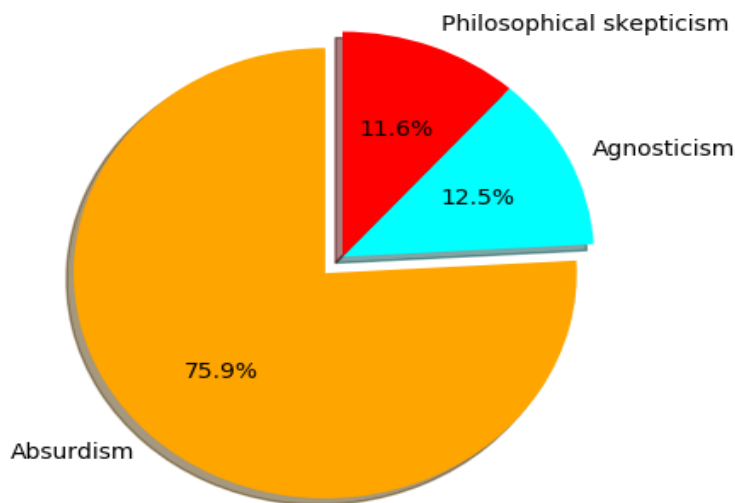


Figure 4.3 Chart showing the recommended RPPs

4.4.1.5. Named Entity Recognition

The process of recognizing named entities is used in the NLP system to identify RPPs components (epistemology, axiology, and ontology) in input text captured on the questionnaire. This is done through the use of the NLTK Spacy library's rule-based entity recognition model. The entity ruler, a pipeline component, is used to add RPP components based on pattern dictionaries. APPENDIX K shows the script used for the recognition of research philosophies and paradigms named entities. The phrase pattern of the entity ruler was used to add new RRP entities by labelling text with 'label' and 'pattern' keys as follows;

```

{"label": "Epistemology", "pattern": "interaction"},
{"label": "Ontology", "pattern": "Universal"},
{"label": "Axiology", "pattern": "social"},

```

Once the entity ruler is created, it is saved and used to recognize named entities in a user's answer to the questionnaire. Figure 4.4 shows the results of the named entity recognizer.

Named Entities

[one version | Ontology', 'independent | Ontology', 'individual | Ontology', 'fact | Ontology', 'values | Axiology', 'independent | Ontology', 'reason | Epistemology', 'reason |', 'reason | Epistemology', 'individual | Ontology']

Figure 4.4 Named Entity Recognition

4.5 Chapter Summary

The system design and architecture of the proposed RMI NLP system was discussed in this chapter. The chapter also presented an easy to use application that will guide prospective researchers to establish their research philosophies and underlying paradigms. The system requirements and the implementation environment were discussed using UML diagrams to show the system flow. A representation of the NLP application architecture was also shown, followed by the system design with data flow diagrams. The chapter also used UML diagrams to show how a user will interact with the system. The following chapter, Chapter 5, will cover the usability and the functional tests of the RMI NLP system.

CHAPTER 5: SYSTEM TESTING AND RESULTS ANALYSIS

This chapter presents the functional and usability test results from different users to assess the overall technology acceptance and usability of the NLP application. System testing and results analysis were done to answer the research questions as outlined in Chapter 1. The researcher used data that were gathered using a post-usability questionnaire for the analysis. Section 5.1 presents the system test results and findings. Section 5.2 presents an overview of the study participants, followed by the analysis of the usability test results in Section 5.3. The achievement of the study's objectives is discussed in Section 5.4, and Section 5.5 summarises the chapter.

5.1 SYSTEM TESTING

The pre-processing of corpus data yielded vectors represented as a Bag of Words (BoW) that can be used to train classifiers to label the text. The corpus data were split into 70-30% train-test sets as shown in Figure 5.2. For the classification of input into RPPs categories, three classifiers were trained and tested to evaluate and select the one that best fits the study's purpose (Figure 5.1). The naïve Bayes, support vector machine (SVM), and logistic regression classifiers were used for this study because they have a high bias/low variance and do not need large training sets to learn from (Brownlee, 2016). Bias in algorithms refers to the assumptions an algorithm makes for it to learn, whereas variance refers to the level to which a target function will change given a different set of training data (Brownlee, 2016). The classifier's evaluation and selection were made by measuring the F1-score, the precision, and recall results of classifiers tested (Table 5.1). The naïve Bayes, SVM, and the logistic regression were tested using the Sci-Kit Learn library's classification report. All classifiers were trained with a sample of seven (7) RPPs categories from the corpus using the Scikit-Learn pipeline class. The following section shows the evaluation of the trained models.

5.1.1 Evaluation of the trained models

According to Berrar (2018) cross-validation is used to estimate a model's ability to produce a true prediction and to prevent it from overfitting. The evaluation phase in

this study reviewed the trained and tested algorithms to determine how accurate they are and whether they meet project's objective and goals using a randomised 2-fold cross-validation. The algorithms were trained iteratively (repeat=10) with different subsets of the training data split into a train and test set at 70-30 as shown in Figure 5.2. Table 5.1 shows the mean classification scores as a result of the cross-validation. To select an appropriate classifier that is fit for the purpose, the following items were measured (Kaur & Saini, 2015):

- **Performance evaluation:** this is the process where the performance of the classifier is being evaluated for;
- **Accuracy:** measures the level of correctness in classifying texts;
- **Precision:** measures how well does the classifier match input against the predefined categories, measured by the percentage of texts that are accurately classified;
- **Recall:** measures the completeness of a classifier; and
- **F1 measure:** measures the combination of both precision and recall.

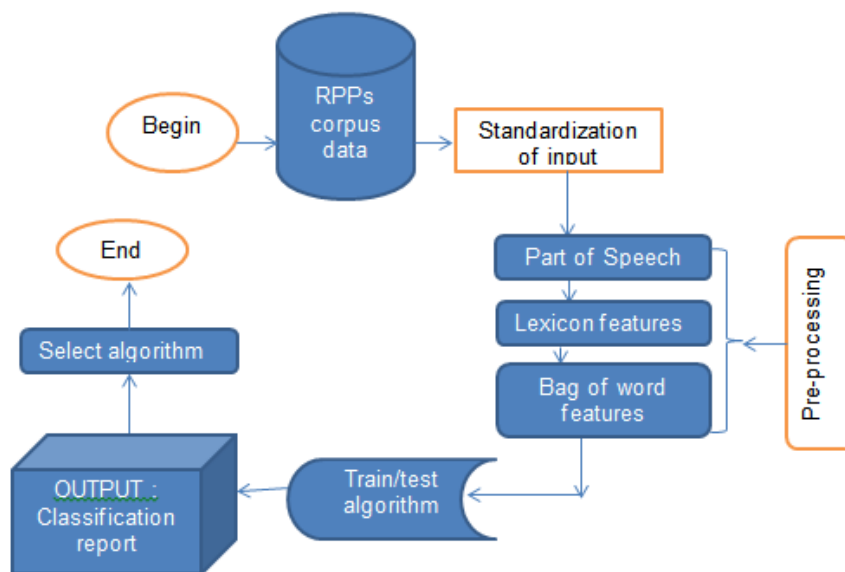


Figure 5.1 Steps followed in selecting the classifier

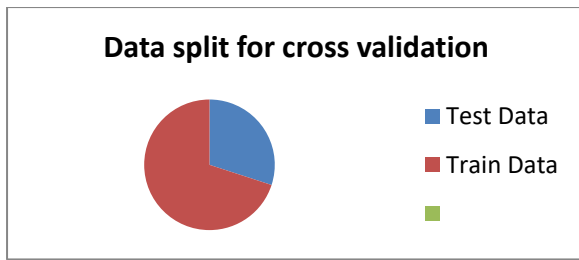


Figure 5.2 Data split for cross-validation

Table 5.1 Comparison of the algorithm scores

	%			
	precision	Accuracy	recall	f1-score
SVM	78	76	76	75
naïve Bayes	85	70	76	76
LR	80	77	77	77

	%			
	precision	accuracy	recall	f1-score
SVM	78	76	76	75
naïve Bayes	85	70	76	76
LR	80	77	77	77

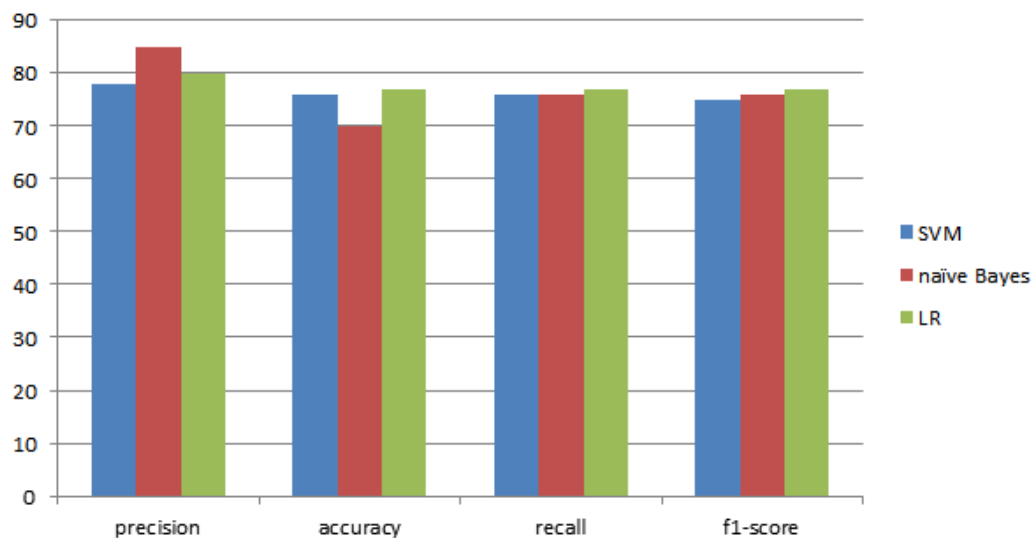


Figure 5.3 Comparison of the algorithms

Initial results obtained from training and evaluating the algorithms with minimal corpus data resulted in an inferior performance for all algorithms. An increase in training data, however, saw a significant increase in the performance of all algorithms. Figure 5.3 compares the performance of the algorithms. The following sections provide detail on the performance of each of the algorithms.

5.1.1.1 Naive Bayes (NB) Classifier

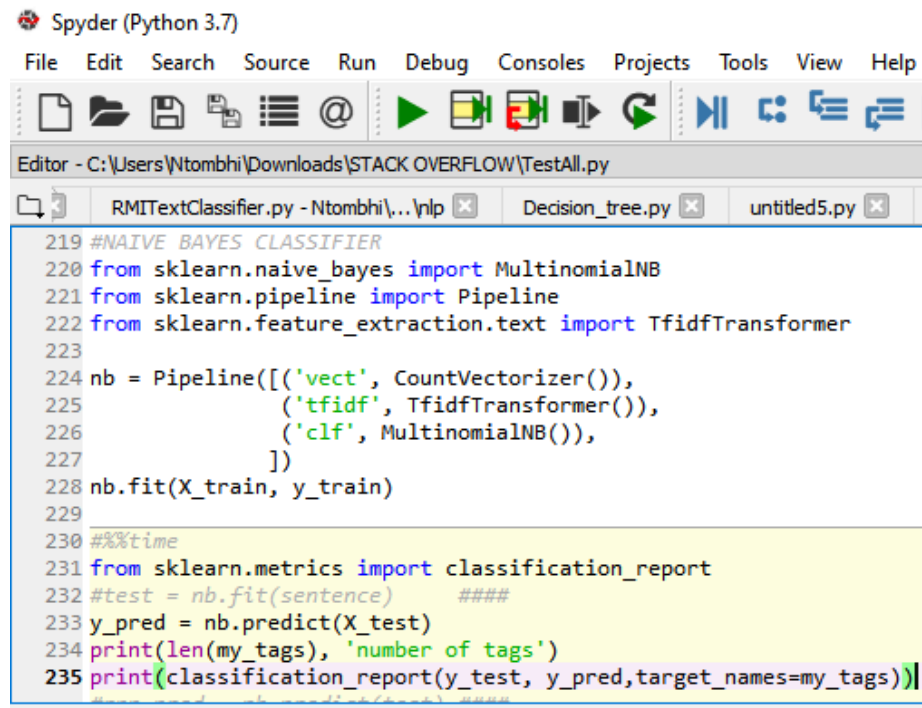
The NB classifier yielded an 85% precision, as shown in Figure 5.4. The accuracy rate of classification for the highest RPP category was at 70%. Both recall and f1-scores were at 76%.

```
Accuracy of NAIVE BAYES CLASSIFIER 0.7067498581962564
      precision    recall  f1-score   support

PostPositivism    0.54     0.96     0.69     488
   Realism        0.85     0.83     0.84     563
  Scepticism      0.96     0.71     0.82     223
 Subjectivism     0.86     0.57     0.68     270
   Positivism     0.00     0.00     0.00     89
 Interpretivism   1.00     0.01     0.02     130

   micro avg     0.71     0.71     0.71    1763
   macro avg     0.70     0.51     0.51    1763
  weighted avg     0.75     0.71     0.67    1763

NB accuracy 0.7067498581962564
Precision score W NB: 0.8542571152628757
Precision score NB: [0.95901639 0.8259325 0.71300448 0.56666667
RECALL score NB: [0.53855006 0.84699454 0.95783133 0.85955056 0.
F1 score F1 NB: 0.7463651621112375
```



```
Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Ntombhi\Downloads\STACK OVERFLOW\TestAll.py
RMITextClassifier.py - Ntombhi\...\nlp x Decision_tree.py x untyped5.py x
219 #NAIVE BAYES CLASSIFIER
220 from sklearn.naive_bayes import MultinomialNB
221 from sklearn.pipeline import Pipeline
222 from sklearn.feature_extraction.text import TfidfTransformer
223
224 nb = Pipeline([('vect', CountVectorizer()),
225               ('tfidf', TfidfTransformer()),
226               ('clf', MultinomialNB()),
227               ])
228 nb.fit(X_train, y_train)
229
230 ###time
231 from sklearn.metrics import classification_report
232 #test = nb.fit(sentence) #####
233 y_pred = nb.predict(X_test)
234 print(len(my_tags), 'number of tags')
235 print(classification_report(y_test, y_pred, target_names=my_tags))
```

Figure 5.4 The script and report for naïve Bayes classifier

5.1.1.2 The Support Vector Machines (SVM)

Although SVM is considered one of the best classification algorithms, it yielded a 78% precision in this instance. The accuracy rate of classification for the highest RPP category was at 76%. Both recall and f1-scores were at 76% and 75%, respectively. The training function and classification report for the SVM is shown in Figure 5.5.

```

Accuracy of SUPPORT VECTOR MMACHINES 0.7612024957458877
      precision    recall  f1-score   support

PostPositivism      0.73      0.84      0.78       488
   Realism          0.86      0.89      0.87       563
   Scepticism       0.89      0.87      0.88       223
   Subjectivism     0.71      0.62      0.66       270
   Positivism       0.35      0.44      0.39        89
Interpretivism     0.64      0.25      0.36       130

   micro avg       0.76      0.76      0.76      1763
   macro avg       0.69      0.65      0.66      1763
   weighted avg    0.76      0.76      0.75      1763

SVM accuracy 0.7612024957458877
Precision score W SVM: 0.7871377868793393
RECALL score: [0.72775801 0.85763293 0.88990826 0.70886076 0.34513274 0.64      ]
Precision score SVM: [0.83811475 0.88809947 0.86995516 0.62222222 0.43820225 0.24615385]

from sklearn.linear_model import SGDClassifier

sgd = Pipeline([('vect', CountVectorizer()),
                ('tfidf', TfidfTransformer()),
                ('clf', SGDClassifier(loss='hinge', penalty='l2', alpha=1e-3, random_state=1, max_iter=5, tol=None)),
                ])
sgd.fit(X_train, y_train)

#%%time

y_pred = sgd.predict(X_test)

print('number of RPPs CATEGORIES ', len(my_tags) )
print('Accuracy of SUPPORT VECTOR MMACHINES %s' % accuracy_score(y_pred, y_test))
print(classification_report(y_test, y_pred, target_names=my_tags))

```

```

number of RPPs CATEGORIES 11
Accuracy of SUPPORT VECTOR MMACHINES 0.15384615384615385

```

	precision	recall	f1-score	support
Realism_Absolute idealism	0.27	0.60	0.37	5
Scepticism_Absurdism	0.00	0.00	0.00	1
Realism_Accidentalism	0.12	0.40	0.19	5
Realism_Action theory	0.00	0.00	0.00	0
Realism_Actualism	0.12	0.14	0.13	7
Scepticism_Agnostic atheism	0.00	0.00	0.00	5
Interpretivism_Agnostic theism	0.00	0.00	0.00	2
Scepticism_Agnosticism	0.33	0.33	0.33	6
Realism_Anti Foundationalism	0.00	0.00	0.00	10
Interpretivism_Anti-realism	0.00	0.00	0.00	6
Positivism_Atomism	0.00	0.00	0.00	5
micro avg	0.15	0.15	0.15	52
macro avg	0.08	0.13	0.09	52
weighted avg	0.09	0.15	0.11	52

Figure 5.5 The script and report for SVM classifier

SVMs perform well with limited training datasets; however, according to Dhiraj (2019), in instances where the target classes overlap, their performance is poor. The study's findings correspond to those by Deepika et al. (2019).

5.1.1.3 The Logistic Regression Classifier

The precision rate of classification for the highest RPP category was 80% for the logistic regression classifier. It showed a 77% classification accuracy, which was slightly higher than the naïve Bayes and the SVM classifiers. Both recall and f1-scores were at 77%. The training function and classification report for the logistic regression classifier are shown in Figure 5.6.

```

Accuracy of LOGISTIC REGRESSION 0.7725467952353943
Precision score W LR: 0.804764056556409
Precision score LR: [0.83401639 0.88809947 0.92376682 0.6555
RECALL score LR: [0.7440585 0.81967213 0.90350877 0.7662337
F1 score F1 LR: 0.784256501905808

```

	precision	recall	f1-score	support
PostPositivism	0.74	0.83	0.79	488
Realism	0.82	0.89	0.85	563
Scepticism	0.90	0.92	0.91	223
Subjectivism	0.77	0.66	0.71	270
Positivism	0.43	0.48	0.46	89
Interpretivism	0.62	0.22	0.33	130
micro avg	0.77	0.77	0.77	1763
macro avg	0.71	0.67	0.67	1763
weighted avg	0.77	0.77	0.76	1763

```

269
270 #LOGISTIC REGRESSION
271
272 from sklearn.linear_model import LogisticRegression
273
274 logreg = Pipeline([('vect', CountVectorizer()),
275                   ('tfidf', TfidfTransformer()),
276                   ('clf', LogisticRegression(n_jobs=1, C=1e5)),
277                   ])
278 logreg.fit(X_train, y_train)
279
280 ###time!
281
282 y_pred = logreg.predict(X_test)
283 print('number of RPPs CATEGORIES ', len(my_tags) )
284 print('Accuracy of LOGISTIC REGRESSION %s' % accuracy_score(y_pred, y_test))
285 #print('Precision score: ', precision_score(y_pred, y_test, average='weighted'))
286 #print('Precision score: ', precision_score(y_pred, y_test, average=None))
287 #print(classification_report(y_test, y_pred, target_names = my_tags, labels=rc
288
289 print(classification_report(y_test, y_pred, target_names=my_tags))

```

Figure 5.6 The script and report for the Logistic Regression classifier

The logistic regression classifiers are known for being vulnerable to overfitting and their inability to solve non-linear problems because they operate on a linear surface (Kumar, 2018). This means they can only extend to classification problems with distinctly separable multiple classes.

The combination of the BoW data representation model and the naïve Bayes classifier is used in this study based on the overall performance of the algorithm on a

limited amount of data. Another advantage of using the naïve Bayes algorithm is its scalability and inherent ability to perform multiclass classifications. Given categorical input variables, the performance of the classifier is good.

5.2 Participants

The participants were a combination of thirty graduates, postgraduates, undergraduates, and academic staff. The participants were purposely chosen because they have been, to some extent, exposed to the concept of research philosophies and paradigms. The participants each received an email with the link to the system and a user manual for the application. Also attached were the participant consent form and a post-usability questionnaire, which they needed to complete and return. All participants were informed that the purpose of the exercise was to test the acceptability and functionality of the system in introducing research philosophies and paradigms. The Statistical Package for the Social Sciences (SPSS) software was used to analyse the data. Figure 5.7 shows the total number of participants who agreed to participate and participated in the study.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	12	40.0	40.0	40.0
	Female	16	53.3	53.3	93.3
	Did not answer	2	6.7	6.7	100.0
	Total	30	100.0	100.0	

Figure 5.7 Analysis of participants

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Academic(PhD)	4	13.3	13.3	13.3
	Graduate(BA, MA, etc.)	16	53.3	53.3	66.7
	Post-graduate	2	6.7	6.7	73.3
	Undergraduate	1	3.3	3.3	76.7
	Did not answer	6	20.0	20.0	96.7
	Not applicable	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

Figure 5.8 Analysis of participants' profession

The participants comprised 40% male, 53.3% female, and 6.7% other, as shown in Figure 5.6. The demographic dimensions in Figure 5.8 reveal that the findings reflect

the views of research students with master’s degrees (53%), followed by academic experts (13%), far much more than undergraduate students’ (3%). This means that most of the participants have been previously exposed to and are at a level where they should understand the concepts of research philosophies and paradigms.

5.3 Usability Test Results

After implementing the system, a usability test was conducted to obtain feedback from participants who used the system. This was achieved by using a modified version of the technology acceptance model (TAM) to establish and ensure whether the system was performing as intended. The participants were required to complete a list of tasks detailed in the user manual (APPENDIX D) and the system Usability questionnaire (APPENDIX C). Table 5.2 presents the list of tasks participants had to perform on the system.

Table 5.2 System tasks to be performed on the NLP system

TASKS
Register account
Login
Take the Questionnaire
View report
Logout

The system usability questionnaire used the Likert rating score of between 1 and 5, ranging between strongly disagree and strongly agree. It was divided into the following sub-sections: effort expectancy, performance expectancy, attitude towards using the technology, knowledge expectancy, and general comments, all outlined in the sections that follow.

5.3.1 Effort Expectancy

Participants were given a manual to follow in completing the system tasks in Table 5.1. These tasks included registering a user account, signing in the system, taking a questionnaire, and viewing the generated report. This section provides an evaluation

report of the ability and ease at which the participants could accomplish the tasks. Figure 5.9 shows the Effort Expectancy variable and the measurable items thereof.

		Statistics					
		EE (PEU) System easy to learn for non-technical users	EE (PEU) It is easy to get the system to do what I want it to do	EE (PEU) Interaction with system is clear and understandable	EE (PEU) The system is flexible to interact with	EE (PEU) It is easy to become skillful in using the system	EE (PEU) The system is easy to use
N	Valid	30	30	30	30	30	30
	Missing	0	0	0	0	0	0

Figure 5.9 Measurable items for Effort Expectancy

The following section discusses the results of each of the Effort Expectancy measurable items:

5.3.1.1 Non-technical users will find it easy to learn

EE (PEU) System easy to learn for non-technical users

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	2	6.7	6.7	10.0
	3 Neutral	9	30.0	30.0	40.0
	4 Agree	6	20.0	20.0	60.0
	5 Strongly agree	12	40.0	40.0	100.0
Total		30	100.0	100.0	

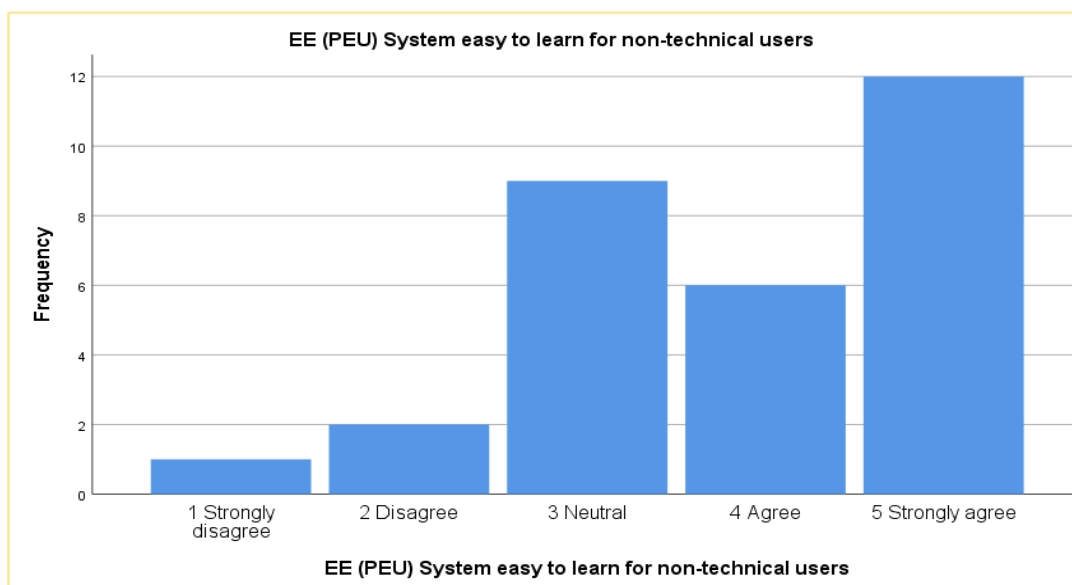


Figure 5.10 Analysis of N-technical users will find it easy to learn

Figure 5.10 shows that most participants (60%) expressed that one does not need to be technically inclined to figure out how to utilize the system. Only 10% of the participants expressed that it will be difficult for non-technical users without an NLP background to figure out how to utilize the system and 30% of all participants expressed no difficulties nor enthusiasm about the ease of use.

5.3.1.2 It is easy to get the system to do what I want it to do

EE (PEU) It is easy to get the system to do what I want it to do

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	4	13.3	13.3	13.3
	3 Neutral	10	33.3	33.3	46.7
	4 Agree	11	36.7	36.7	83.3
	5 Strongly agree	5	16.7	16.7	100.0
	Total	30	100.0	100.0	

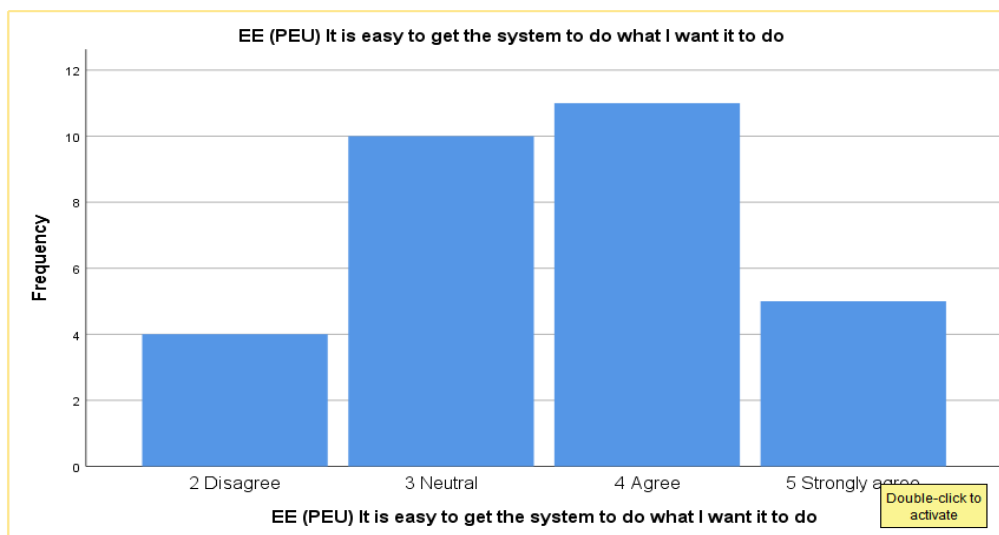


Figure 5.11 Analysis of It is easy to get the system to do what I want it to do

Figure 5.11 shows that most participants (53.4%) reported that it was easy to accomplish what they wanted to be done on the system, while 13.0% of the participants disagreed.

5.3.1.3 Interaction with the system is clear and understandable

EE (PEU) Interaction with system is clear and understandable

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	1	3.3	3.3	3.3
	3 Neutral	6	20.0	20.0	23.3
	4 Agree	17	56.7	56.7	80.0
	5 Strongly agree	5	16.7	16.7	96.7
	Did not answer	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

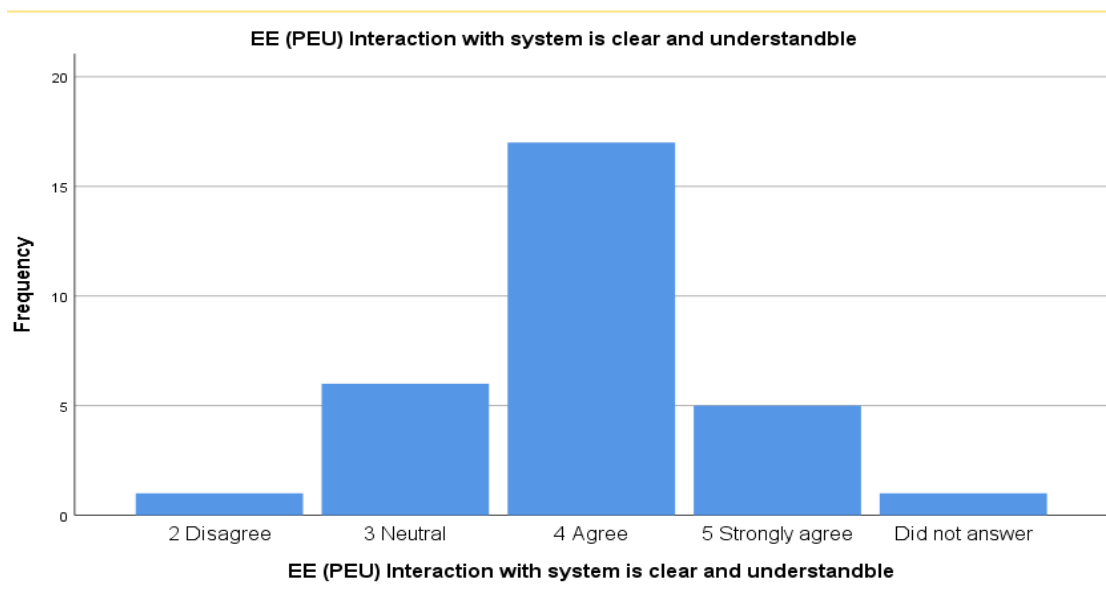


Figure 5.12 Analysis of Interaction with system is clear and understandable.

Twenty-two of the participants (73%) found the system to be understandable and straightforward, as shown in Figure 5.12. In comparison, only one participant (3%) did not understand how to use the system. Of the remaining participants, 6 or 20% of all participants remained neutral, while one did not answer the question.

5.3.1.4 The system is flexible to interact with

EE (PEU)The system is flexible to interact with

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	1	3.3	3.3	6.7
	3 Neutral	8	26.7	26.7	33.3
	4 Agree	14	46.7	46.7	80.0
	5 Strongly agree	6	20.0	20.0	100.0
	Total	30	100.0	100.0	

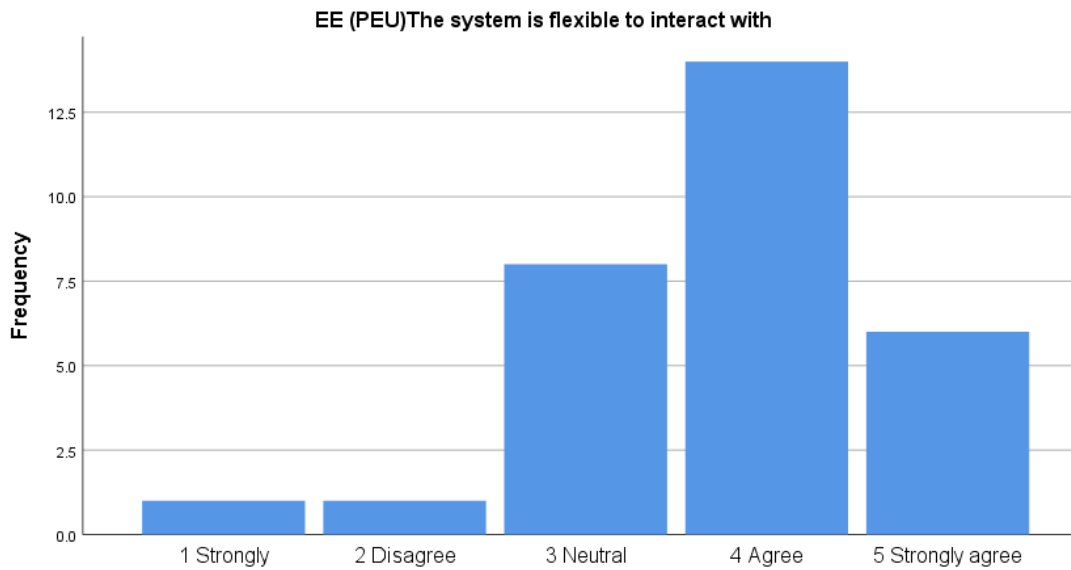


Figure 5.13 Analysis of The system is flexible to interact with

Figure 5.13 shows that twenty participants (66.6%) found that the system was flexible to interact with, with only 2 participants (6.6%) not agreeing. The remaining 8 participants (26.7%) remained neutral.

5.3.1.5 Level of difficulty in utilizing the system

EE (PEU) It is easy to become skillful in using the system

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	1	3.3	3.3	3.3
	3 Neutral	10	33.3	33.3	36.7
	4 Agree	12	40.0	40.0	76.7
	5 Strongly agree	7	23.3	23.3	100.0
	Total	30	100.0	100.0	

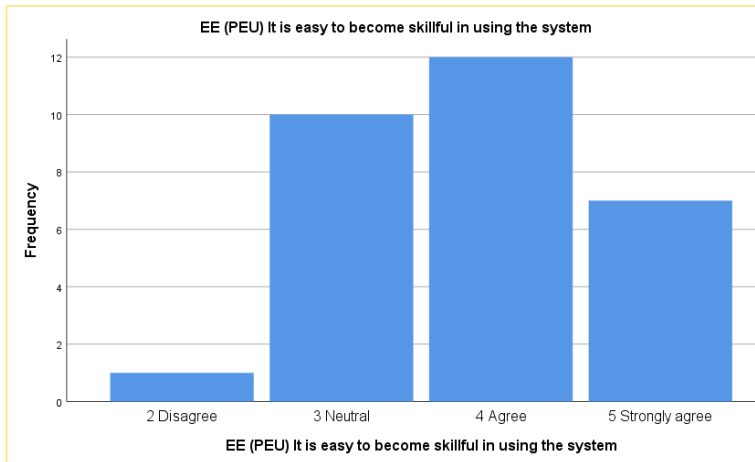


Figure 5.14 Analysis of the level of difficulty in mastering how to utilize the system

Figure 5.14 shows that while nineteen participants (63.3%) reported that they found it easy to master using the NLP application, 10 participants (33.3%) remained neutral, and 1 participant (3.3%) disagreed.

5.3.1.6 The level of difficulty in utilizing the system

EE (PEU) The system is easy to use

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	1	3.3	3.3	3.3
	3 Neutral	10	33.3	33.3	36.7
	4 Agree	8	26.7	26.7	63.3
	5 Strongly agree	11	36.7	36.7	100.0
	Total	30	100.0	100.0	

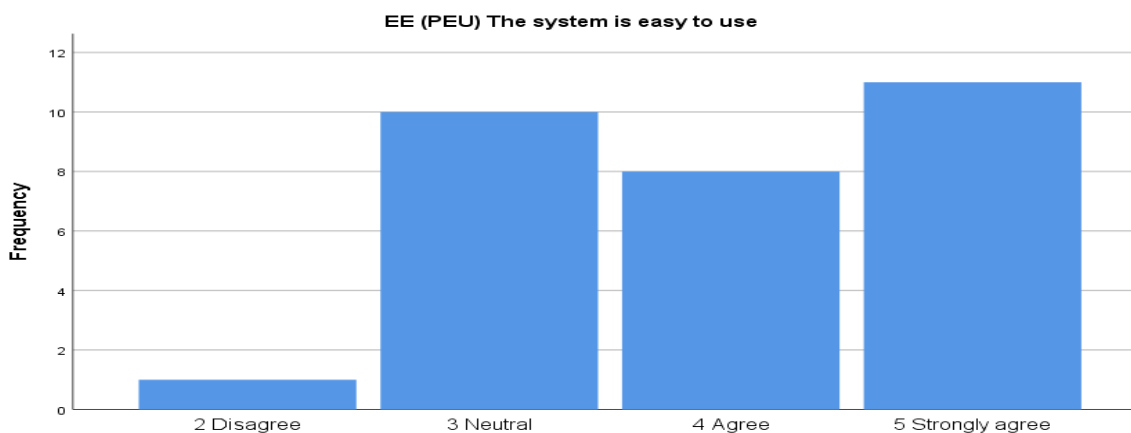


Figure 5.15 Analysis of the system is easy to use

Figure 5.15 shows that nineteen participants constituting a majority of 63.3% of the participating population reported that they could use the NLP application without much effort as opposed to 1 participant (3.3%) who disagreed. Ten participants (33.3%) remained neutral.

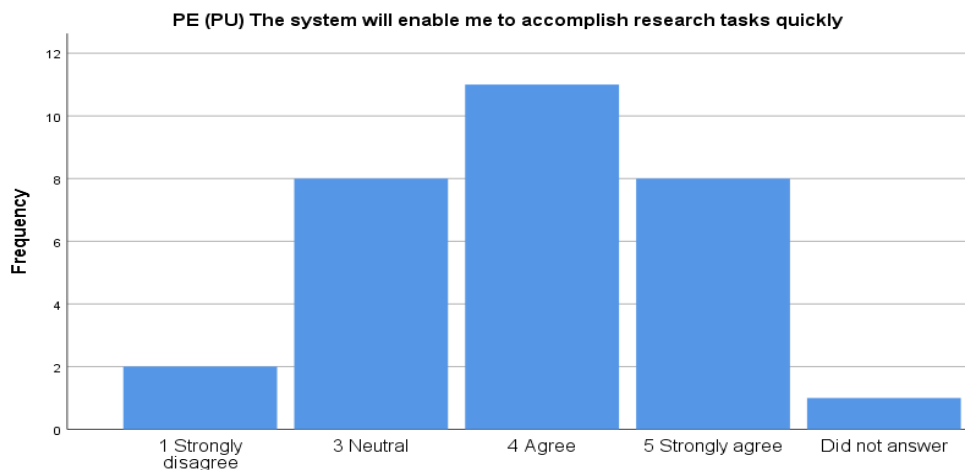
The interpretation of the data for the Effort Expectancy variable shows that all participants could register an account on the system with little to no effort with the sole feedback that participants need a clear instruction that they need to register an account before signing-on. The login process went through successfully for all participants. They were also able to navigate easily to get to the questionnaire to answer the questions, view the generated report with no assistance, and sign-out. Most participants showed the system was easy to navigate and that it would be easy for non-technical persons to use. In the researcher's view, the study's finding regarding the usability of the system proved the ease of using it. In comparison,

findings regarding acceptability prove that researchers will use the system as they embark on their research endeavours.

5.3.2 Performance Expectancy

This section aimed to establish the degree to which the participants believe that using the system will further improve their performance and understanding of concepts when doing research. It also evaluated whether participants would understand and know the effective and appropriate ways of collecting, processing, and analyzing research data after interacting with the system. The results of the items measured are discussed below.

5.3.2.1 Utilizing the system in research will empower participants to accomplish research tasks more quickly



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	2	6.7	6.7	6.7
	3 Neutral	8	26.7	26.7	33.3
	4 Agree	11	36.7	36.7	70.0
	5 Strongly agree	8	26.7	26.7	96.7
	Did not answer	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

Figure 5.16 Analysis of utilizing the research system will empower participants to accomplish research tasks more quickly

Figure 5.16 shows that nineteen participants (63.3%) say that utilizing the research system will empower them to complete research tasks more quickly. Eight participants (26.7%) remain neutral, 6.7% of all participants disagree that using the system in research will enable them to accomplish research tasks more quickly, while 3.3% did not answer the question.

5.3.2.2 Using the system would improve my epistemological understanding

PE (PU) The system will improve my epistemological understanding					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	3	10.0	10.0	10.0
	3 Neutral	9	30.0	30.0	40.0
	4 Agree	8	26.7	26.7	66.7
	5 Strongly agree	9	30.0	30.0	96.7
	Did not answer	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

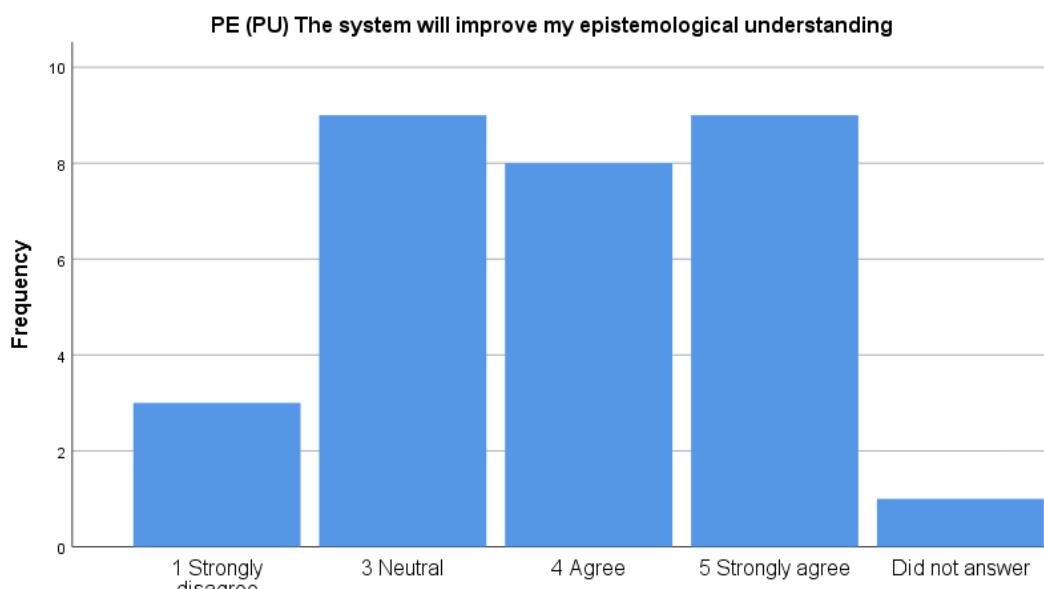


Figure 5.17 Analysis of using the system would improve my epistemological understanding

Figure 5.17 shows that seventeen participants constituting 56.7% of the participating population agree that utilizing the system will improve their performance and

understanding of concepts when doing research. Three participants constituting 10% of the participating population disagreed. Nine participants constituting 30% of the participating population remained neutral, whereas one participant did not answer the question.

5.3.2.3 Does the system perform well when there is the concurrent use of the system

PE (PU) The system performs well when used concurrently with other users

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	2	6.7	6.7	6.7
	3 Neutral	16	53.3	53.3	60.0
	4 Agree	5	16.7	16.7	76.7
	5 Strongly agree	7	23.3	23.3	100.0
	Total	30	100.0	100.0	

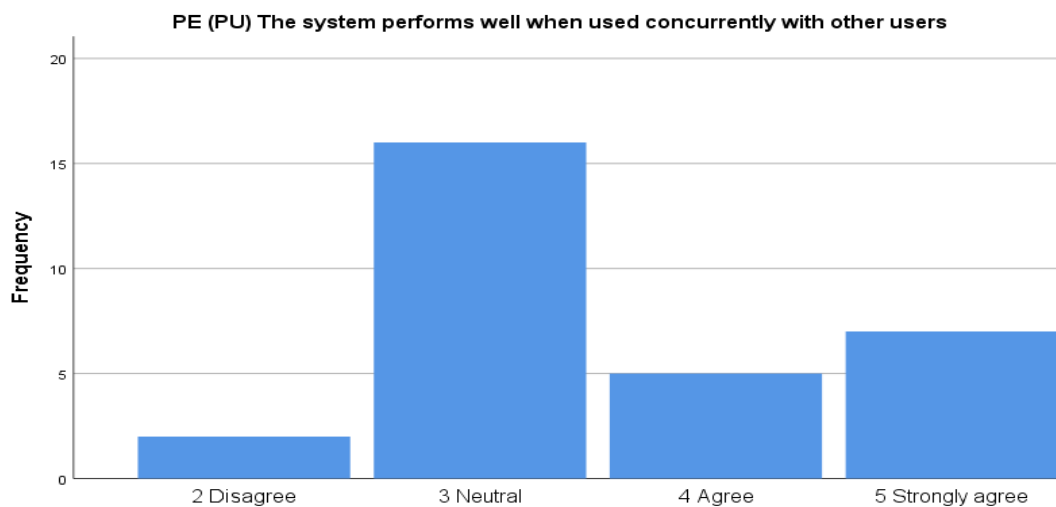


Figure 5.18 Analysis of Does the system perform well when there is the concurrent use of the system

Figure 5.18 shows that sixteen participants (53.3%) responded in a neutral way when asked about the system's performance given the concurrent usage. Eleven participants constituting (40%) agree that the system does perform well with concurrent usage. The remaining 6.7% assert that the system does not perform well when there are concurrent users on the system.

5.3.2.4 Acceptability of time taken to generate a report the database is acceptable

PE (PU) The time required to fetch results from database is acceptable

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	1	3.3	3.3	6.7
	3 Neutral	3	10.0	10.0	16.7
	4 Agree	8	26.7	26.7	43.3
	5 Strongly agree	17	56.7	56.7	100.0
	Total	30	100.0	100.0	

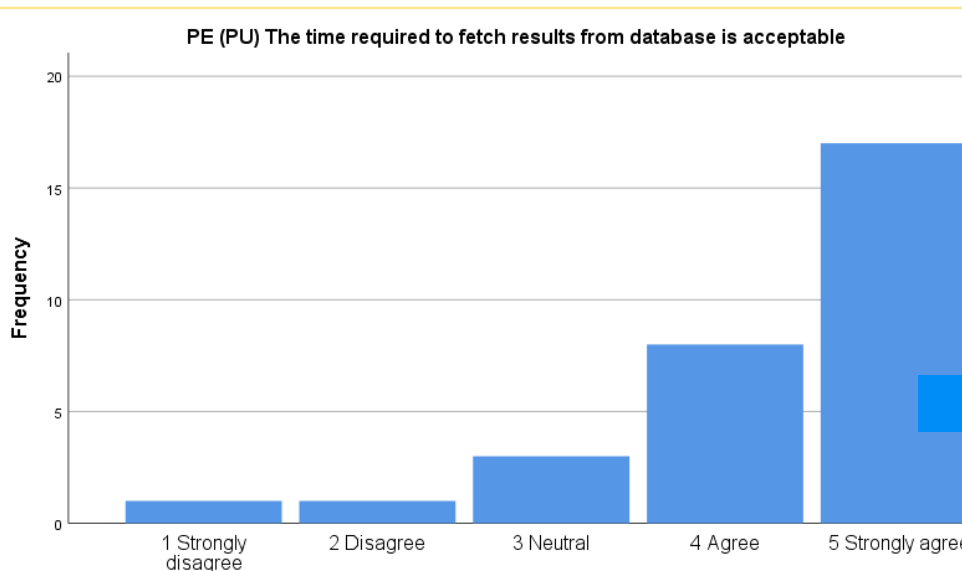


Figure 5.19 Analysis of Acceptability of time required to generate the report

Figure 5.19 shows that twenty-five participants constituting (82.6%) are satisfied with the time required to generate a report. Three participants constituting (10%) were neutral, whereas 2 participants (6.6%) disagreed.

5.3.2.5 Is the system functional and fit for purpose

PE (PU) The system is useful

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Disagree	1	3.3	3.3	3.3
	3 Neutral	10	33.3	33.3	36.7
	4 Agree	7	23.3	23.3	60.0
	5 Strongly agree	12	40.0	40.0	100.0
	Total	30	100.0	100.0	

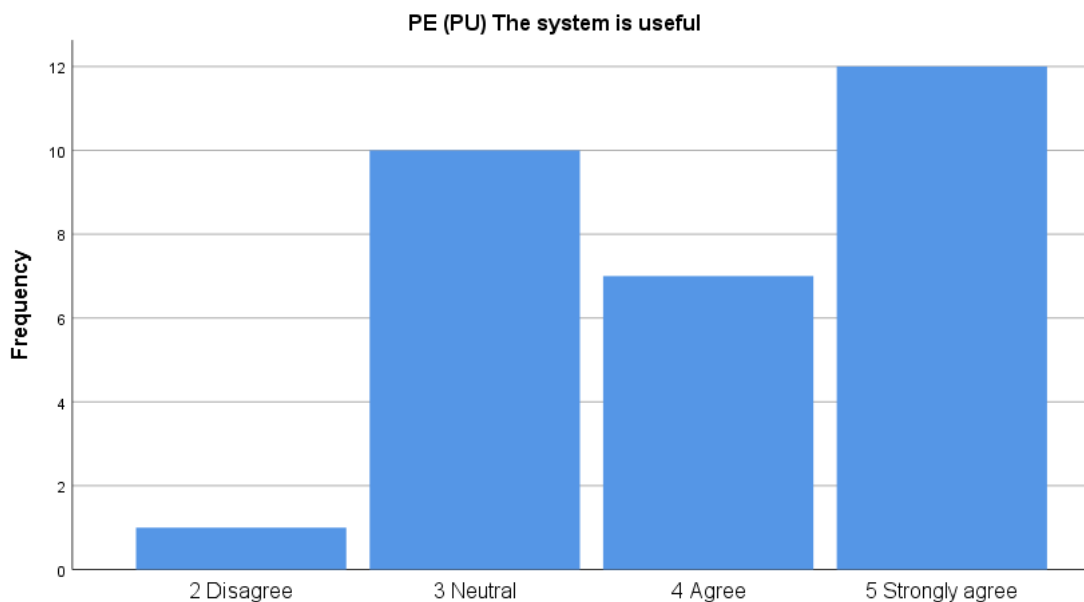


Figure 5.20 Analysis of Is the system functional and fit for purpose

The results in Figure 5.20 show that nineteen participants (63%) said the system was useful for research. Ten participants were neutral about the system's usefulness, while one participant disagreed. These participants represented 33.3% and 3.3% of the participating population, respectively.

Interpretively the Performance Expectancy variable shows that most participants believe that there will be some significant enhancement in their research performance. Using the system will help them achieve research tasks quicker. Only a minority of the participants did not find the system useful in their research as the

system does not help improve their understanding of research philosophies and paradigms.

5.3.2.6 Using the system requires an understanding of Natural Language Processing

This variable tested whether participants support using the system to assist in conducting research and whether they are likely to use the system in the future. The section also assessed whether using the system required an understanding of natural language processing concepts. The measured items are discussed below.

AE (ATB) Using the system requires an understanding of NLP concepts

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	4	13.3	13.3	13.3
	2 Disagree	7	23.3	23.3	36.7
	3 Neutral	3	10.0	10.0	46.7
	4 Agree	8	26.7	26.7	73.3
	5 Strongly agree	8	26.7	26.7	100.0
	Total	30	100.0	100.0	

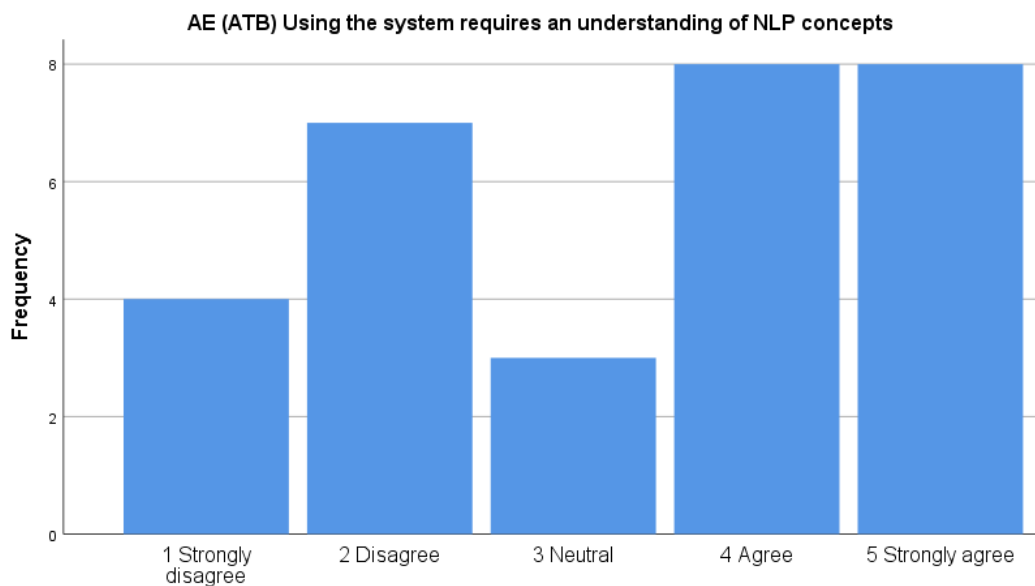


Figure 5.21 Analysis of Attitude towards using the technology

Figure 5.21 above shows that sixteen participants constituting 53.4% of all participants assert that using the system requires understanding natural language processing concepts. Three participants constituting of all participants were neutral,

whereas eleven participants constituting 36.6% of all participants disagreed. This finding is quite surprising and contrary to the finding on effort expectancy. As seen in section 5.3.1 participants were able to use the system without challenges. This finding can be attributed to the lack of understanding of what the natural language process means.

5.3.2.7 Support of the utilization of the system to enhance the learning process

AE (ATB) I support the idea of using the system to enhance the learning process

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	4	13.3	13.3	16.7
	3 Neutral	2	6.7	6.7	23.3
	4 Agree	10	33.3	33.3	56.7
	5 Strongly agree	12	40.0	40.0	96.7
	Did not answer	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

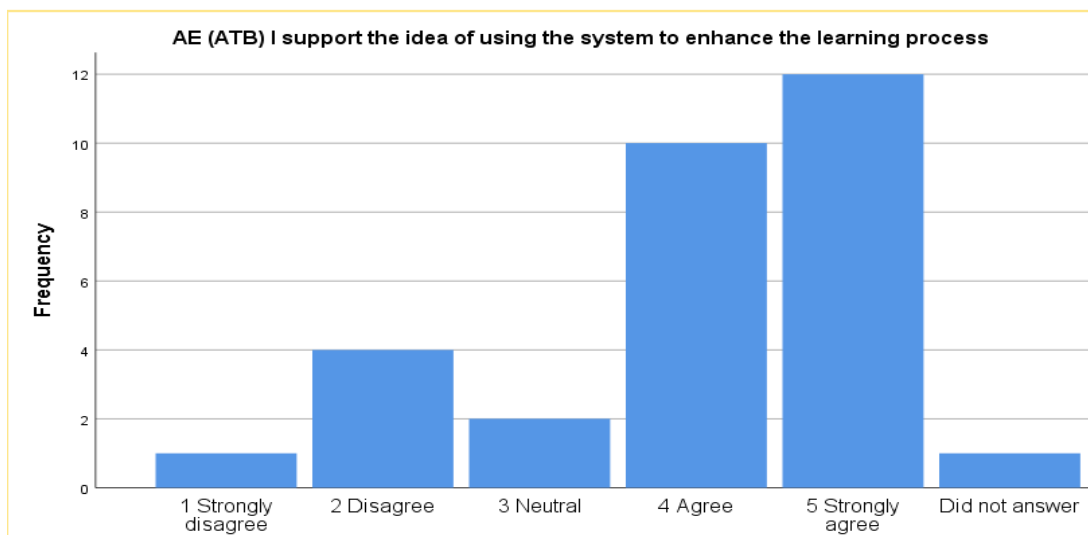


Figure 5.22 Analysis of Support of the utilization of the system to enhance the learning

Twenty-two of the participants (73.3%) support using the system to enhance their learning process, as shown in Figure 5.22. Two of the participants representing 6.7%

of all participants remained neutral, while five participants representing 16.6% of all participants did not support using the system to enhance their learning process. Only one participant representing 3.3% of all participants did not respond.

5.3.2.8 I would like using technology to learn more about the subject matter the system addresses instead of the traditional way

AE (ATB) I support the idea of using the system to enhance the learning process * AE (AF) I will like using technology to learn more about the subject matter the system addresses instead of the traditional way
Crosstabulation

Count

		AE (AF) I will like using technology to learn more about the subject matter the system addresses instead of the traditional way					Total
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree	
AE (ATB) I support the idea of using the system to enhance the learning process	1 Strongly disagree	1	0	0	0	0	1
	2 Disagree	1	1	2	0	0	4
	3 Neutral	0	0	1	0	1	2
	4 Agree	0	1	1	6	2	10
	5 Strongly agree	0	0	4	2	6	12
	Did not answer	0	0	1	0	0	1
Total		2	2	9	8	9	30

Figure 5.23 Analysis of I would like using technology to learn more about the subject matter the system addresses instead of the traditional way

Figure 5.23 shows a cross-tabulation of the participants who expressed interest in using the system for the learning process over traditional methods. The findings show that twenty-two of the participants representing 63.3% of all participants support using the system to enhance their learning process and would like to use technology to learn more about research philosophies and paradigms. Five participants (16.3%) do not support enhancing the learning process through technology. They will not like using it to learn more about research philosophies and paradigms. Of the remaining participants, 6.7% of all the participating population was neutral about the support and use of technology, whereas 3.3% of the participating population did not respond.

5.3.2.9 Level of anticipation of aspects of research that will be supported by the use of the system

AE (ATB) I support the idea of using the system to enhance the learning process * AE (AF) I look forward to those aspects of research that will be supported by the use of the system Crosstabulation							
Count	AE (AF) I look forward to those aspects of research that will be supported by the use of the system					Total	
	1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree		
AE (ATB) I support the idea of using the system to enhance the learning process	1 Strongly disagree	1	0	0	0	0	1
	2 Disagree	1	2	1	0	0	4
	3 Neutral	0	1	0	0	1	2
	4 Agree	0	0	1	8	1	10
	5 Strongly agree	0	0	2	2	8	12
	Did not answer	0	0	1	0	0	1
Total	2	3	5	10	10	30	

Figure 5.24 Analysis of the level of anticipation of those aspects of research that will be supported by the use of the system

Figure 5.24 shows a cross-tabulation of participants who support the system during the study against future intentions to use the system. The findings reveal that twenty-two of the participants representing 63.3% of all participants who support the use of the system to enhance their learning process, look forward to those aspects of research that will be supported by the system's use. Five participants representing 16.3% of the participating population do not support enhancing the learning process through technology. They are not looking forward to those aspects of research that will be supported by the system's use. This could be attributed to the fact that the participants believe the introduction of technology will have a detrimental effect on the production of quality knowledge, as observed in Gambini's (2019) study. Such effects include the lack of in-depth understanding or research philosophies and paradigms and how to apply various research methods. Of the remaining participants, 6.7% of all the participating population was neutral about their outlook of those aspects of research that will be supported by the system's use, whereas 3.3% of the participating population did not respond.

The interpretation of data in the Attitude Expectancy variable displays that the attitude towards using the technology to establish research philosophies and underlying paradigms rated above average for most participants. More than half of the participants support the utilization of the system to improve their learning process. They are looking forward to using the system to learn more about research

philosophies and paradigms. This can be attributed to the participants getting real-time feedback about their guiding research philosophy and paradigm and exploring literature on more RPPs. Other participants are not certain whether the system will be able to improve or enhance their learning process. Other participants still prefer the traditional way of establishing what their research philosophies and paradigms are. According to Creswell (2013), the traditional way means researchers start their career by assuming their general philosophical orientation about the world based on previous research experiences, teacher's philosophical disposition, and the structure of the subject's curriculum. According to Saunders et al. (2009), this is based on beliefs and assumptions about the world.

5.3.3 Knowledge Expectancy

This variable aimed to establish whether participants understood the concepts of research philosophies and paradigms after engaging with the system, whether the system recommended RPP resonates with their beliefs and worldview, and if the way the concepts were introduced is beneficial to their research enterprise. It also assessed whether participants understood the value and relevance of research philosophies and paradigms in research. Participants were also asked if they knew what their research philosophy was. The Knowledge Expectancy variable had seven measurable items, and the results of the items are discussed next.

5.3.3.1 I Know and understand the concept of research philosophy

KE (CO) I know and understand the concept of research philosophy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	3	10.0	10.0	13.3
	3 Neutral	5	16.7	16.7	30.0
	4 Agree	19	63.3	63.3	93.3
	5 Strongly agree	2	6.7	6.7	100.0
	Total	30	100.0	100.0	

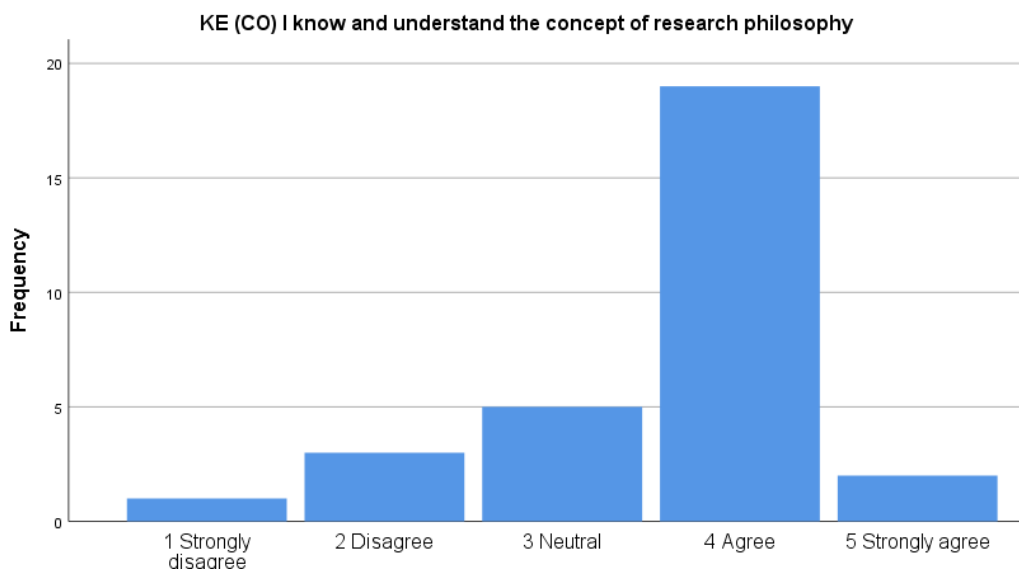


Figure 5.25 Analysis of I know and understand the concept of research philosophy

Figure 5.25 shows that the system helped people understand the concept of research philosophies and paradigms. The findings show that twenty-one of the participants representing 70% of all participants knew and understood the research philosophy concept. Five participants representing 16.7% of all participants remained neutral, while four participants representing 13.3% of all participants did not know and understand the concept.

5.3.3.2 I Know and understand what research paradigm is

KE (CO) I know and understand what a research paradigm is

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	2	6.7	6.7	10.0
	3 Neutral	8	26.7	26.7	36.7
	4 Agree	14	46.7	46.7	83.3
	5 Strongly agree	5	16.7	16.7	100.0
	Total	30	100.0	100.0	

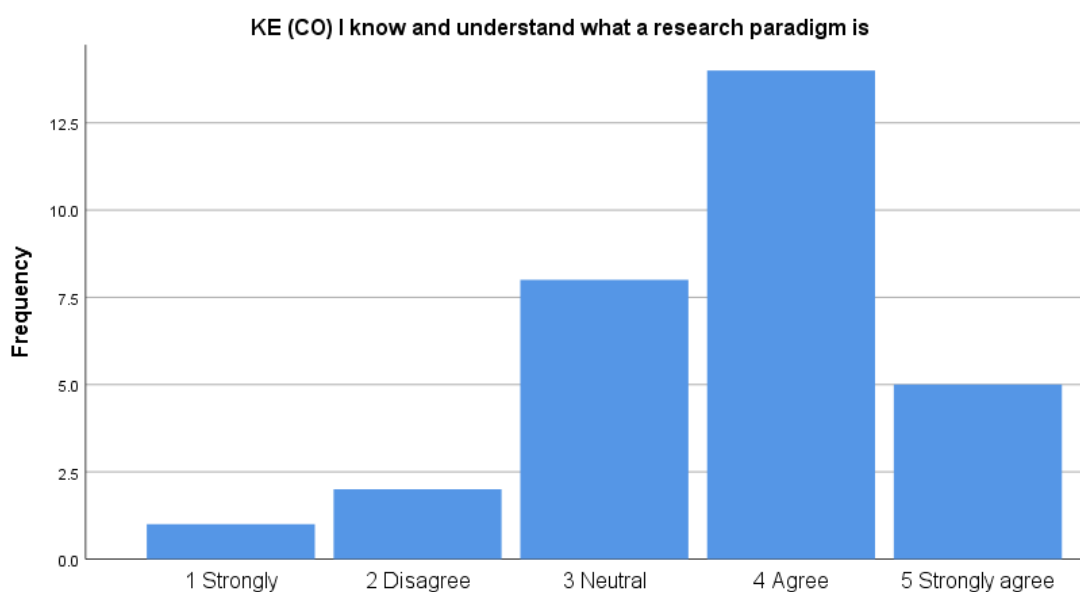


Figure 5.26 Analysis of I know and understand what research paradigm is

Nineteen of the participants representing 63.4% of all participants attest knowledge and understanding of the research paradigm as guided by the system and shown in Figure 5.26. Eight participants representing 26.7% of all participants remained neutral, while three participants representing 10% of all participants did not know and understand what a research paradigm is.

5.3.3.3 I understand the value of research philosophies and paradigms in conducting research

KE (CO) I understand the value of research philosophies and paradigms in conducting research

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	1	3.3	3.3	6.7
	3 Neutral	8	26.7	26.7	33.3
	4 Agree	16	53.3	53.3	86.7
	5 Strongly agree	4	13.3	13.3	100.0
	Total	30	100.0	100.0	

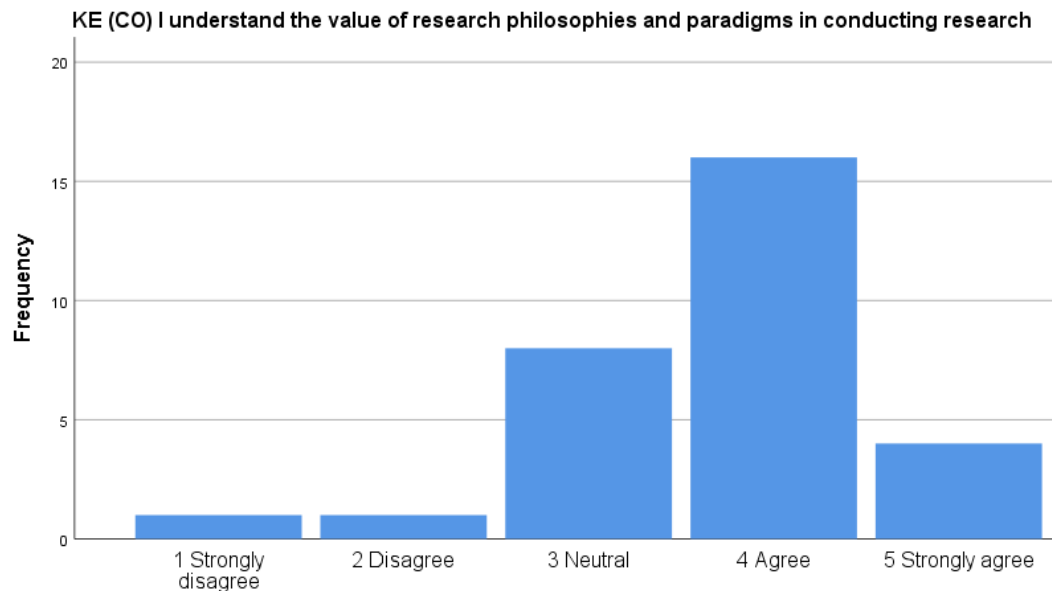


Figure 5.27 Analysis of I understand the value of research philosophies and paradigms in conducting research

As shown in Figure 5.27, twenty of the participants representing 66.6% of all participants express the system's role in enhancing teaching and learning. Their findings show that the system helps enhance understanding of the value of research philosophies and paradigms in conducting research. Eight of the participants representing 26.7% of all participants remained neutral. In comparison, two participants (6.6%) did not understand the value of research philosophies and paradigms in conducting research.

5.3.3.4 I know which research philosophy I espouse

KE (CO) I know which research philosophy I espouse					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	6	20.0	20.0	20.0
	2 Disagree	1	3.3	3.3	23.3
	3 Neutral	7	23.3	23.3	46.7
	4 Agree	10	33.3	33.3	80.0
	5 Strongly agree	6	20.0	20.0	100.0
Total		30	100.0	100.0	

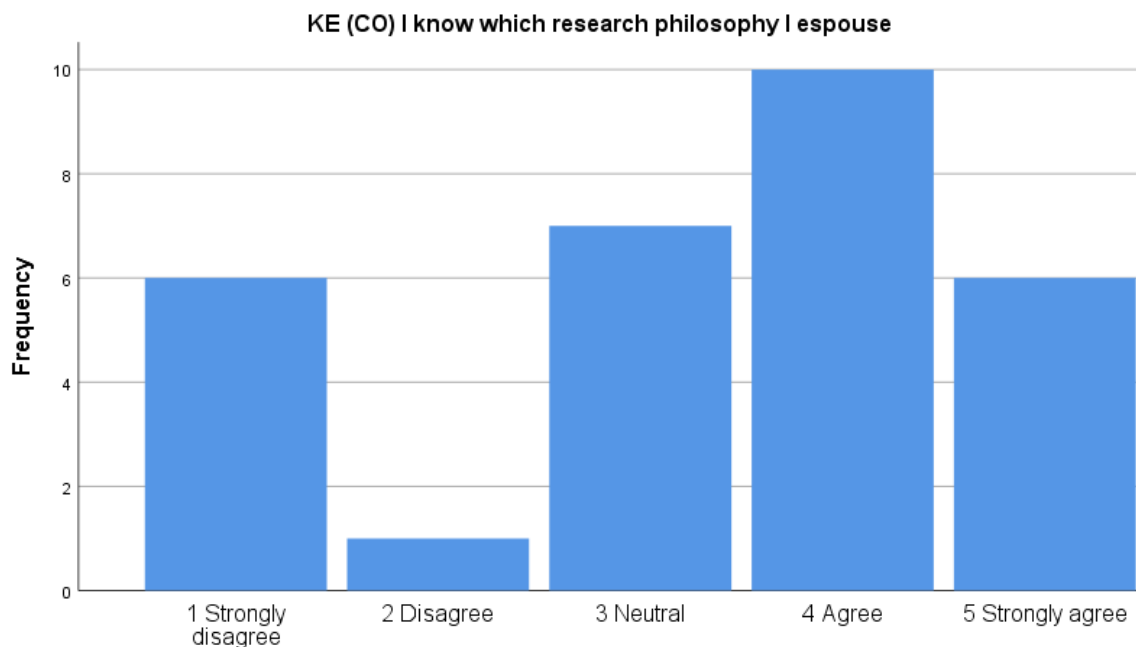


Figure 5.28 Analysis of I know which research philosophy I espouse

Figure 5.28 above shows the number of participants who had a better understanding of their research philosophy due to the system. Accordingly, 16 participants representing 53.3% of all participants knew, 7 participants representing 23.3% of all participants remained neutral. In comparison, the remaining 7 participants representing the other 23.3% of all participants did not know which research philosophy they espoused.

5.3.3.5 I Understanding how relevant research philosophies and paradigms are to conducting research.

KE (CO) I understand the relevance of research philosophies and paradigms in conducting research

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	2	6.7	6.7	6.7
	2 Disagree	3	10.0	10.0	16.7
	3 Neutral	6	20.0	20.0	36.7
	4 Agree	10	33.3	33.3	70.0
	5 Strongly agree	9	30.0	30.0	100.0
	Total	30	100.0	100.0	

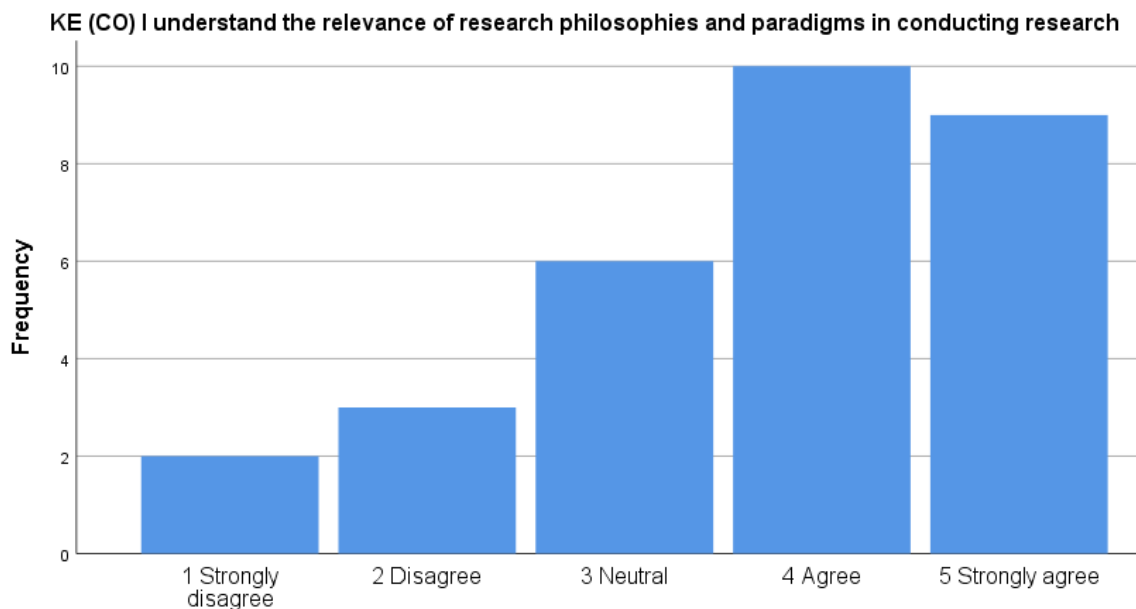


Figure 5.29 Analysis of I understand the relevance of research philosophies and paradigms in conducting research.

Figure 5.29 shows that the system can heighten the awareness and relevance of research philosophies to scholars. Accordingly, nineteen participants constituting 63.3% of all participants understood. Six participants constituting 20% of all participants remained neutral, while the remaining five participants constituting the other 16.7% of all participants did not understand the relevance of research philosophies and paradigms conducting research.

5.3.3.6 The way in which the concepts of research philosophies and paradigms are introduced is beneficial

KE (IN) The system's introduction of the concepts of research philosophies and paradigms is beneficial

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3 Neutral	11	36.7	36.7	36.7
	4 Agree	11	36.7	36.7	73.3
	5 Strongly agree	7	23.3	23.3	96.7
	Did not answer	1	3.3	3.3	100.0
	Total	30	100.0	100.0	



Figure 5.30 Analysis of how the concepts of research philosophies and paradigms are introduced is beneficial

Figure 5.30 above shows that eighteen participants constituting 60% of all participants agree that system's introduction of the concepts of research philosophies and paradigms is beneficial. Eleven participants constituting 53.3% of all participants were not certain and remained neutral, while one participant constituting 3.3% of all participants did not answer.

5.3.3.7 Ability to find out more information about research philosophies and paradigms

KE (IN) I am able to find out more information about research philosophies and paradigms

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Strongly disagree	1	3.3	3.3	3.3
	2 Disagree	3	10.0	10.0	13.3
	3 Neutral	6	20.0	20.0	33.3
	4 Agree	14	46.7	46.7	80.0
	5 Strongly agree	6	20.0	20.0	100.0
	Total	30	100.0	100.0	

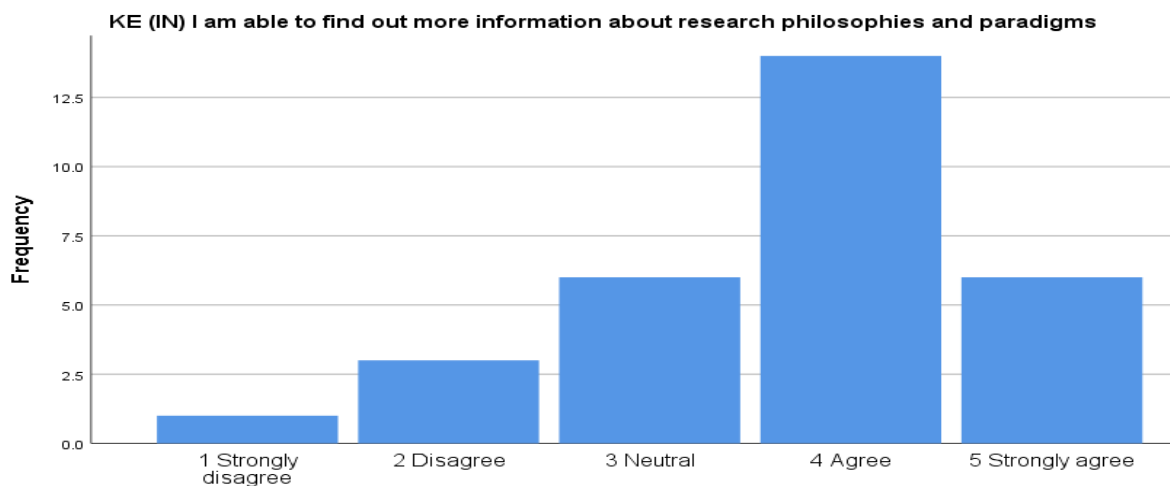


Figure 5.31 Analysis of the ability to find out more information about research philosophies and paradigms

Figure 5.31 above shows that twenty participants constituting 66.7% of all participants could find out more information about research philosophies and paradigms in using the system. Six participants, constituting 20% of all participants remained neutral, while four participants constituting 13.3% of all participants could not find out more information.

5.3.3.8 Cross-tabulation analysis of Knowledge Expectancy items

KE (CO) I know and understand what a research paradigm is * KE (CO) I understand the value of research philosophies and paradigms in conducting research Crosstabulation

Count

		KE (CO) I understand the value of research philosophies and paradigms in conducting research					Total
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree	
KE (CO) I know and understand what a research paradigm is	1 Strongly disagree	1	0	0	0	0	1
	2 Disagree	0	0	2	0	0	2
	3 Neutral	0	1	3	4	0	8
	4 Agree	0	0	2	12	0	14
	5 Strongly agree	0	0	1	0	4	5
Total		1	1	8	16	4	30

Figure 5.32 Crosstab analysis of participants who understand and know the value of research philosophies and paradigms

Figure 5.32 above shows that of the nineteen of the participants who know and understand what research paradigms are, sixteen also understand the value of research philosophies and paradigms in conducting research, while three of them remain neutral. Of the eleven remaining participants, eight were neutral regarding knowing and understanding research paradigms, with four participants understanding the value of research philosophies and paradigms in conducting research. At the same time, three remained neutral in that regard and one not knowing what value research philosophies and paradigms add to research. Of the three participants who did not know or understand research paradigms, two were neutral about understanding the value of research philosophies and paradigms in conducting research. In contrast, one did not understand the value at all.

KE (CO) I know and understand what a research paradigm is * AE (ATB) I support the idea of using the system to enhance the learning process Crosstabulation

Count

		AE (ATB) I support the idea of using the system to enhance the learning process					Total	
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree		Did not answer
KE (CO) I know and understand what a research paradigm is	1 Strongly disagree	0	0	1	0	0	0	1
	2 Disagree	0	1	0	1	0	0	2
	3 Neutral	0	0	0	4	4	0	8
	4 Agree	0	3	1	4	6	0	14
	5 Strongly agree	1	0	0	1	2	1	5
Total		1	4	2	10	12	1	30

Figure 5.33 Analysis of I know and understand what a research paradigm is, and I support the idea of using the system to enhance the learning process

Of those nineteen participants who know and understand what research paradigms are, fourteen of them support the idea of using the system to enhance the learning process, four disagree, while one remains neutral, as shown in Figure 5.33 above.

		KE (IN) The system's introduction of the concepts of research philosophies and paradigms is beneficial				
		3 Neutral	4 Agree	5 Strongly agree	Did not answer	Total
KE (CO) I know which research philosophy I espouse	1 Strongly disagree	5	0	0	1	6
	2 Disagree	1	0	0	0	1
	3 Neutral	3	3	1	0	7
	4 Agree	1	8	1	0	10
	5 Strongly agree	1	0	5	0	6
Total		11	11	7	1	30

Figure 5.34 Crosstab analysis of I know which research philosophy I espouse, and the system's introduction of the concepts of research philosophies and paradigms is beneficial

Figure 5.34 shows that of the sixteen participants who knew which research philosophy they espoused, only two were neutral about whether the system's introduction of research philosophies and paradigms concepts would be beneficial. The remaining fourteen participants agreed that the system's introduction of research philosophies and paradigms concepts would be beneficial.

The interpretation of the data of the Knowledge Expectancy variable displays that a majority of the participants support the idea of using the system to improve the learning process and will benefit from the way the system introduces the concepts of research philosophies and paradigms. Most of the participants were able to gather more information about research philosophies and paradigms by using the system, and only a few were not able to find more information. The results also show that participants who knew and understood the research paradigm concept also know and understand the value of research paradigms in conducting research. The finding supports the assertion by Abubakar (2016) that to understand the importance of research it is imperative to understand the value and importance that research paradigms have in research inquiries.

5.3.4 General Comments

In this section, participants were expected to give feedback on issues not addressed on the preceding sections. The issues addressed were system errors and failure, general comments, and whether the participants would like to participate in future research of this kind. These are discussed in the sections that follow;

5.3.4.1 System errors

Table 5.3 System Errors

List System Errors

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	None	10	33.3	33.3	33.3
	I didn't experience any	4	13.3	13.3	46.7
	Grammar apps and GoGo Survey	2	6.7	6.7	53.3
	None picked up	1	3.3	3.3	56.7
	No glitches were experienced	2	6.7	6.7	63.3
	No errors experienced	1	3.3	3.3	66.7
	The system was very slow, probably network issues	1	3.3	3.3	70.0
	No glitches, the app has spell check so my spelling errors were highlighted	1	3.3	3.3	73.3
	Did not answer	6	20.0	20.0	93.3
	Not applicable	2	6.7	6.7	100.0
	Total	30	100.0	100.0	

Table 5.3 shows the participants' responses when asked to report any glitches, spelling or formatting errors and/or parts of this app that were inaccessible. Nineteen participants constituting 63.3% of all participants reported none; they were not aware of, did not experience any, and did not pick up any glitches. One participant, 3.3%,

commented that the app has spell check, which highlighted their spelling errors. One participant constituting 3.3% of all participants reported that the system was very slow and attributed that to network issues. Six participants constituting 20% of all participants did not answer, while one participant, 3.3%, noted grammar apps and GoGo survey. Two participants, 6.7% of all participants, reported that the item did not apply to them.

The interpretation of the system errors data displays that most participants did not experience any errors while interacting with the system. A few participants commended the availability of a spelling and grammar check, which means they could answer the questionnaire with ease.

Table 5.4 System Failure

System Failure

		Frequen cy	Percent	Valid Percent	Cumulative Percent
Valid	None	6	20.0	20.0	20.0
	I didn't experience any	3	10.0	10.0	30.0
	I do not understand the question	1	3.3	3.3	33.3
	When we try to extract the data we need to subscribe first	1	3.3	3.3	36.7
	It did not fail	3	10.0	10.0	46.7
	None this system only accepts responses no output or any other form of interaction	1	3.3	3.3	50.0
	I am not aware of any failures	1	3.3	3.3	53.3
	No problems; however, with more training in how to use the system I could have been much quicker	1	3.3	3.3	56.7

	Frequen cy	Percent	Valid Percent	Cumulative Percent
I don't know	1	3.3	3.3	60.0
At no point during the survey	1	3.3	3.3	63.3
Only when it is not well maintained	1	3.3	3.3	66.7
If the software is unable to solve the intended problems	1	3.3	3.3	70.0
Did not answer	8	26.7	26.7	96.7
Not applicable	1	3.3	3.3	100.0
Total	30	100.0	100.0	

Table 5.4 shows that sixteen participants constituting 5 3.3% of all participants reported none. They were not aware of and did not experience any failure. At no point in the survey did the system fail. One of the sixteen participants noted that the system only accepts responses, no output or any other form of interaction, while another said that with more training in how to use the system, they could have been much quicker. Eight participants constituting 26.7% of all participants did not answer. One participant, 3.3%, reported that they did not understand the question. Another participant representing 3.3% reported that they did not know if the system failed. The remaining number of participants were split between one who noted that when they try to extract the data, they need to subscribe first, another one who said the system would fail only when it is not well maintained, another one said the system would fail if the software is unable to solve the intended problems. Only one participant noted that the question did not apply to them.

The system failure data interpretation displays that most participants did not experience any failures while interacting with the system. Also noted in the data is the need to train participants on how to use the system.

5.3.4.2 Comments

Table 5.5 Comments

Comment

		Frequenc		Valid	Cumulative
		y	Percent	Percent	Percent
Valid	None	5	16.7	16.7	16.7
	It is an interesting research topic allowing the user to move out of their own research comfort zone	1	3.3	3.3	20.0
	No comment	1	3.3	3.3	23.3
	it was quick and easy to use with challenging questions	1	3.3	3.3	26.7
	N/A	1	3.3	3.3	30.0
	It is very useful and accurate as well	2	6.7	6.7	36.7
	I am looking forward to learning more skills necessary on how to be effective in using the system	1	3.3	3.3	40.0
	This was interesting and informative	1	3.3	3.3	43.3
	Fun to use and introduce these concepts	1	3.3	3.3	46.7
	No further comment	1	3.3	3.3	50.0
	The system was very clear and understandable	1	3.3	3.3	53.3
	Did not answer	13	43.3	43.3	96.7
	Not applicable	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

- Table 5.5 shows the general comments made by the participants. Twenty-three participants constituting 76% of all participants did not answer, had no comments or found the section did not apply to them. The participants who commented made up the remaining 24% of all participants, each with the comments below:-

- “The system was very clear and understandable”;
- “Fun to use and introduce these concepts”;
- “This was interesting and informative”;
- “It is very useful and accurate as well”;
- “it was quick and easy to use with challenging questions”;
- “I am looking forward to learning more skills necessary on how to be effective in using the system”; and
- “It is an interesting research topic allowing the user to move out of their own research comfort zone”.

The interpretation of the data in the comments section displays that most of the participants found the system interesting, easy to understand, and a fun way of being introduced to research philosophies and paradigms concepts. In contrast, the other participants had no further input about the system.

5.4 Achievement of Objectives

The study aimed to improve the learning and teaching process by developing a specialised corpus of research philosophies and paradigms and using a model that will classify input variables into research philosophies and paradigms (RPPs) categories. The study set out to address the following objectives: 1) building a corpus of research philosophies and paradigms that will be used to train a classification model, 2) conduct computer experiments to identify a classifier suitable for this study and use it to develop an application and 3) assess the efficacy of the application through user testing. This section discusses how the study’s objectives were met.

5.4.1 Build a corpus of research philosophies and paradigms

The first objective of building a corpus of research philosophies and paradigms was achieved by collecting data on research philosophies and paradigms from sources such as PhilPapers, Stanford Encyclopedia of Philosophy, Google Scholar, IBSS, Philosophy Basics publications, journals, theses, etc. A total of 323 research philosophies and paradigms were discovered, out of which only 180 were used for the corpus. All collected data was labelled with relevant research philosophies and

paradigms categories. The Bag of Words model was used to represent the data in a machine-readable format by creating vectors and creating features for each of the RPPs to enable the training of the classification algorithms.

5.4.2 Train a classifier and develop an NLP application to recommend RPPs

The second objective of training a classifier and developing an NLP application to recommend RPPs was achieved by successfully implementing and training a classification model using the RPP corpus. The training corpus was split into a train and test set at a ratio of 70-30%. Three machine learning algorithms (Logistic Regression, Support Vector Machine and naïve Bayes) learnt how to classify using 70% of the training corpus data. The remaining 30% of the corpus data was used to validate the performance of the algorithms. Classification reports for the algorithms were generated with the naïve Bayes classifier performing better than the others with a precision rate of 85%, an accuracy rate of 70%, and a recall rate of 76%.

The research methods index application was developed using Python programming language, the naïve Bayes algorithm, and the RPPs corpus. The application was then deployed on the Microsoft Azure cloud through the Django Web framework. The developed NLP Research Methods Index is available through the following link <http://unisa-rppi.westeurope.cloudapp.azure.com:8082/nlp>.

5.4.3 Test the application's ability to recommend a research philosophy and paradigm through user testing

The third objective of testing the application's ability to recommend a research philosophy and paradigm through user testing is meant to assess whether the systems achieved its aim of improving the knowledge production process by recommending research philosophies and paradigms to system users. Users logged in to the system and participated in answering the questionnaire relating to their worldviews, the nature of reality, and the relevance or importance of research values. User input was passed to the Python script created for the trained classification for processing. After that the input data were used to determine which research philosophy and paradigm to classify input text into. The process was

achieved by transforming the input text into vectors and using the trained model to classify.

5.5 Chapter Summary

This chapter presented the rationale for selecting the naïve Bayes algorithm for classification by comparing performance results of the three classification algorithms and the rationale behind selecting the Bag of Words model for the representation of the text corpus. The usability and acceptability testing results were also evaluated and presented in this chapter, together with an analysis of whether and how the objectives set out at the beginning of the study were met. Analysis of the various measurable variables of the usability and technology acceptance test was done, which revealed a general feeling of acceptance for the NLP Research Methods Index application. Chapter 6 will discuss future recommendations of the study and also provide the conclusion of the dissertation.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

This chapter summarises the study by discussing the recommendations, conclusions, and limitations of the study. An overview of the research problem is presented first, followed by a discussion of the research questions. Recommendations are then made based on the findings obtained in Chapter 5, followed by limitations of the study. Recommendations for future research will be presented, followed by a conclusion of the whole study.

6.1 Overview

The objective of the study is to enhance learning and teaching by developing a natural language processing (NLP) classification model or application and an RPP corpus that will be used to train the classification model. The classification model had to be trained on the RPP corpus to recommend research philosophies and paradigms when provided with the input text. Users had to test the system to evaluate if it adds value to the learning enterprise. To this end, the study accomplished the following:

- Built an RPP corpus using NLP to be used to train a machine learning algorithm on how to classify input text into research philosophies and paradigms.
- Successfully trained and tested the naïve Bayes classifier and used it to develop the research methods index application.
- Recommended research philosophies and paradigms to researchers to help improve the knowledge generation process

The study proposed to develop an NLP classification model and build an RPP corpus that will be used to classify a participant's input into research philosophies and paradigms in the corpus to recommend a research philosophy and paradigm to the participant. As detailed in Chapter 4, the development of the system was done using the Spyder IDE and partitioned into the back-end and front-end. The Django Web Framework was used to develop the front-end (user interface), whereas the back-end was developed using Python and relied on MySQL to create the database. This process is shown in the system architecture section. Once the system was

working, a usability test was conducted to assess the overall usability, appropriateness, and acceptability of the technology. Participants were provided with a test script (APPENDIX B) and a post usability questionnaire (APPENDIX C) to evaluate their system experience. The results indicate that participants were able to complete the expected tasks without any assistance once they could register a user account and log into it. The detailed results of the testing were discussed in Chapter 5. The study achieved its objectives through the design and development of a classification model that was trained on an RPP corpus built for the same intention.

6.2 Conclusion

This study is an academic endeavour that has navigated through a minefield of sources and conducted research activities with the ultimate objective of developing a research methods index (RMI) system based on NLP technologies. The RMI system was developed to discern concepts of research philosophies and paradigms. This was achieved by developing an RPPs corpus, which was then represented with the Bag of Words model in numerical vectors. This representation was used to train and evaluate the performance of the naïve Bayes, support vector machine, and logistic regression algorithms for the classification of text into research philosophy and paradigm categories. The naïve Bayes performed better than the logistic regression and support vector machine algorithms, with a precision rate of 85%. Thus it was used to create the RMI application.

The system's development also included preparing a questionnaire for participants. The questionnaire was prepared to solicit responses about the axiological, epistemological, and ontological stances of participants. Based on the responses, the system provided a report, which recommended relevant philosophies and paradigms to participants.

With these activities accomplished, the study has achieved its objectives: 1) building a research philosophies and paradigms corpus, 2) training and selecting a classifier that used to develop an NLP application for classifying text into research philosophy and paradigm categories, and 3) evaluating the application's ability to determine the research philosophy and paradigm a user's input belongs to. The corpus is of great

value to the study as it was used for training the algorithm used for recommending relevant RPPs based on the inputs of the participants. During this study, several limitations and research challenges were also encountered, mainly due to unavailability of corpus on RPPs. Consequently, many sources of information had to be consulted to ensure that the system has in its corpus known RPPs to cater to researchers across a wide range of fields of study.

As stated in the study, there has been no published research about applying NLP or ML in the classification of text into RPPs categories. Therefore, this study achieved its objective of creating a corpus whose utility has been tested and confirmed through the feedback of research participants. It makes a modest contribution to the enhancement of knowledge production. It also makes a unique contribution, not just to learning and teaching but also to establishing a guiding philosophy and paradigm in a research endeavour. This study's achievements could be enhanced through future developments as outlined in the recommendations for the future work section above.

6.3 Recommendations for Future Work

Feedback from study participants and the study's findings identified some gaps from which the developed RMI application would benefit through further research. These future developments will greatly improve the functional and practical input to learning and teaching. The following enhancements and functionalities are recommendations for future work:

a. Discovery and Identification of new and emerging RPPs

A literature review in this study uncovered more research philosophies and paradigms that are not covered in a single book, lectures, encyclopaedia, and publications. Therefore, the study assists researchers in discovering research philosophies and paradigms beyond those covered in most publications. However, the study does not go further to establish if there are new and emerging RPPs. It would be beneficial to the knowledge enterprise to extend the study's current landscape to include the discovery of new and emerging RPPs.

b. Demonstrate how philosophy shapes methods and helps generate knowledge

As the study focuses only on helping researchers discover their research philosophy and underlying paradigms, it does not guide the researcher into how their philosophy shapes methods and helps them generate knowledge. This inadequacy means that researchers still have to undergo further study or research to establish which methods to use to generate knowledge. Further research that will see the study extended to guiding researchers into how their philosophy shapes methods and how it helps them generate knowledge will add more value to the current enterprise.

c. Use the RMI application to find more resources on RPPs

In printing out a report of recommended RPPs for a researcher, the study provides further detail about each of the recommended RPPs; however, the detail is not exhaustive. A compilation of more sources or resource locators for each of the RPPs would benefit the study. It will allow researchers to further interrogate their recommended RPPs to determine methods and tools used to conduct research. The extension of the RMI application through the implementation of a web-crawler, searching the internet for more journals and other resources on the recommended paradigms and philosophies, will enhance the study substantially, thereby improving the learning and knowledge generation enterprise.

d. Identify the research philosophy and paradigm of individuals in a particular group

Another area of future research in relation to this study is the need to enhance the system to a level where it can identify if there is a common guiding paradigm in a particular group of individuals such as social sciences, health sciences, and education. This will help in solidifying the view that to research in a particular field such as science, there are specific guiding philosophies and underlying paradigms as attested to by Saunders et al. (2012), Guba and Lincoln (1994), and Denzin and Lincoln (2005).

REFERENCES

- Kumar, S. and Zymbler, M. 2019. A machine learning approach to analyze customer satisfaction from airline tweets. *J Big Data* **6**, 62. <https://doi.org/10.1186/s40537-019-0224-1>
- Reynoso, R. 2019. What is NLP (Natural Language Processing)
<https://learn.g2.com/natural-language-processing>
- Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.
- Springboard India. 2020. Hands-on Training with Machine Learning Algorithms: Decision Tree and Random Forest
<https://in.springboard.com/blog/machine-learning-algorithms-decision-tree-random-forest/>
- Xiao, R.. (2010). Corpus creation.
- Dsouza, V. (2018). An Analysis of Housing Rental Sector in Ireland.
- Filannino, M. (n.d.) Dbworld e-mail classification using a very small corpus', project of machine learning course, university of manchester.
- Kadhim, Ammar. (2018). An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security*, **16**. 22-32.
- Rao, S. (2018). The Philosophical Paradigm of Financial Market Contagion Research
SSRN: <https://ssrn.com/abstract=3183951> or <http://dx.doi.org/10.2139/ssrn.3183951>
- Ohno-Machado, L., Nadkarni, P., & Johnson, K. (2013). Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature. *Journal of the American Medical Informatics Association : JAMIA*, **20**(5), 805. <https://doi.org/10.1136/amiajnl-2013-002214>
- Levers, M.-J. D. (2013). Philosophical Paradigms, Grounded Theory, and Perspectives on Emergence. SAGE Open. <https://doi.org/10.1177/2158244013517243>
- Guba, E. G. & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In Denzin, N.K. & Lincoln, Y.S. *Handbook of qualitative research*, 3rd Edn. (pp. 105 – 117). California: Sage.
- Denzin, N. K., Lincoln, Y. S. (2005). Introduction: The discipline and practice of qualitative research. In Denzin, N., Lincoln, Y. (Eds.), *The SAGE handbook of qualitative research* (3rd ed., pp. 1-32). Thousand Oaks, CA: Sage.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**:1–47
- Tang, J., Alelyani, S. and Huan Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications* (2014),

- Agarwal, B., Mittal, N.: Prominent Feature Extraction for Sentiment Analysis. Springer, Cham (2016)
- Koeva, Svetla & Stoyanova, Ivelina & Leseva, Svetlozara & Dekova, Rositsa & Dimitrova, Tsvetana & Tarpomanova, Ekaterina. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*. 65. 10.15398/jlm.v0i1.33.
- Wagner, Jorge & Wilkens, Rodrigo & Idiart, Marco & Villavicencio, Aline. (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese.
- Evans, D. (2007). Corpus building and investigation for the Humanities. University of Birmingham. <http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit1.pdf>
- Ludovic Tanguy, Nikola Tulechki, Assaf Urieli, Eric Hermann, Céline Raynal. Natural Language Processing for aviation safety reports: from classification to interactive analysis. *Computers in Industry*, Elsevier, 2016, 78, pp.80-95. [ff10.1016/j.compind.2015.09.005](https://doi.org/10.1016/j.compind.2015.09.005). [ffhalshs-01322238f](https://doi.org/10.1016/j.compind.2015.09.005)
- Zakauskas, P. (2018). Philosophy and Paradigm of Scientific Research. IntechOpen
- Xu, B. & Kumar, S. (2015). A Text Mining Classification Framework and its Experiments Using Aviation Datasets.
- Zubrinic, Krunoslav & Milicevic, Mario & Zakarija, Ivona. (2013). Classification of Concept Maps Using Bag of Words Model. 118-123.
- Dudoskiy, J. (2018). The Ultimate Guide to Writing a dissertation in Business Studies: A Step-by-Step Assistance.
- Bajpai, N. (2011) "Business Research Methods". Pearson education India
- Abedin, M. A. U., Ng, V. and Khan. L. (2010). Cause identification from aviation safety incident reports via weakly supervised semantic lexicon construction. *J. Artif. Int. Res.*, 38:569–631
- Soumya, G.K. and Shibily, J. (2014) Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 16, Issue 1, Ver. V (Jan. 2014), PP 34-38 www.iosrjournals.org
- Romanov, A., Lomotin, K., & Kozlova, E. (2019). Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts. *Data Science Journal*, 18(1), 37. DOI: <http://doi.org/10.5334/dsj-2019-037>
- Abubakar, A. (2016). Understanding the use of research paradigm and theory in the discipline of library and information science research: Reflection on Qualitative and Quantitative Approach.

Alvi, M.H. (2016). A Manual for Selecting Sampling Techniques in Research. University of Karachi. Iqra University. Retrieved from https://mpr.aub.uni-muenchen.de/70218/1/MPRA_paper_70218.pdf

Artstein, R. (2007). Quality Control of Corpus Annotation Through Reliability Measures. Retrieved from <http://ufal.mff.cuni.cz/acl2007/tutorials/index.php/t5/>

Asiri, S. (2018) Machine Learning Classifiers. Accessed from <https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623>

Ohno-Machado, L., Nadkarni, P., & Johnson, K. (2013). Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature. *Journal of the American Medical Informatics Association : JAMIA*, 20(5), 805. <https://doi.org/10.1136/amiajnl-2013-002214>

Aston, G. (2000). Learning English with the British National Corpus.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python:(Annotated ed.). Beijing: OReilly.

Bonaccorso, G. (2017). Machine learning algorithms: Reference guide for popular algorithms for data science and machine learning. Birmingham, UK: Packt.

Boslaugh, S.E. (2010). Secondary data Source. Encyclopedia of Research Design, Volume 1 edited by Salkind , N.J. SAGE

Brownlee, J. (2018). Supervised and Unsupervised Machine Learning Algorithms. Retrieved from <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>

Brownlee, J. (2015). A Gentle Introduction to the Bag of_Words Model, Machine Learning Mastery. Retrieved from <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>.

Brownlee, J. (2016, October 25). Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning. <https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/>

Caldwell, B. (2010). Beyond Positivism. Routledge. <http://dx.doi.org/10.2139/ssrn>.

Cambridge University Press. (2019). Cambridge online dictionary, Cambridge Dictionary online. Retrieved: Open Educational Resources (OER) Portal at <http://temoa.tec.mx/node/324>

Castle, N. (2017). Supervised vs. Unsupervised Machine Learning. Retrieved from <https://blogs.oracle.com/datascience/supervised-vs-unsupervised-machine-learning>

- Chilisa, B., & Kawulich, B. (2012). Selecting a research approach: Paradigm, methodology and methods. 51-61. University of West Georgia. Retrieved from https://www.researchgate.net/publication/257944787_Selecting_a_research_approach_Paradigm_methodology_and_methods
- Connor, U. and Upton, T.A. (2004) *Discourse in the Professions: Perspectives from Corpus Linguistics*. ISBN 9027222878, 9789027222879. John Benjamins Publishing Conversia. (2017) <https://medium.com/@conversica/ai-machine-learning-and-nlp-your-basic-guide-to-ai-technology-and-what-it-can-do-bd211a32e4c9>
- Couto, J. (2015). The definitive guide Natural Language Processing. [Blog] Monkey Learn. Available at: <https://monkeylearn.com/blog/definitive-guide-natural-language-processing/> [Accessed 25 Sep. 2018].
- Cox, C. (2011). Corpus linguistics and language documentation: challenges for collaboration. In: Newman, J., Harald Baayen, R., Rice, S. (eds.) *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, pp. 239–264. Rodopi, Amsterdam
- Creswell, J. W. (2013). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*.
- Crowston, K., Allen, E. E., & Heckman, R. (2012). Using Natural Language Processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, 15(6), 523–543. [Google Scholar](#), [Crossref](#)
- Dai, W., Xue, G. , Yang, Q. and Yu, Y. 2007. "Transferring Naive Bayes Classifiers for Text Classification", *Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence*, pp. 540-545
- Dangeti, P. (2017). *STATISTICS FOR MACHINE LEARNING -: Essential statistical concepts for exploring predictive... analytics and machine learning using python and r* (4th ed.). S.I.: PACKT PUBLISHING LIMITED.
- Daniel, J. (2011). *Sampling essentials: practical guidelines for making sampling choices*, SAGE Publications, Inc., Thousand Oaks, California, [Accessed 16 September 2018], doi: 10.4135/9781452272047.
- Devault, G. (2019) Advantages and Disadvantages of Quantitative Research <https://www.thebalancesmb.com/quantitative-research-advantages-and-disadvantages-2296728>

Deepika, K., Kancherla, J., Devi, B., & Veeranjanyulu N. (2019). Effect of Different Kernels on the Performance of an SVM Based Classification. *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-7, Issue-5S4, February 2019

Flowers, P. 2009. Research Philosophies- Importance and relevance

Frankhauser, W. (2015). Artificial Intelligence Applications: Natural Language Processing. [ebook] CreateSpace Independent Publishing Platform.

Garbade, M. (2018, October 15). A Simple Introduction to Natural Language Processing [Blog post]. Retrieved from <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>

Genzel, D. (2016, September 23). What Are The Differences Between AI, Machine Learning, NLP, And Deep Learning? Retrieved from <https://www.forbes.com/sites/quora/2016/09/23/what-are-the-differences-between-ai-machine-learning-nlp-and-deep-learning/#20133b49274f>

Gotterbarn, D., Miller, K., Rogerson, S., Barber, S., Barnes, P., Burnstein, I., ... Werth, L. H. (2001). Software Engineering Code of Ethics and Professional Practice. *Science and Engineering Ethics*, 7(2), 231-238

Kumar, R. (2014) Research Methodology: A Step-by-Step Guide for Beginners. 4th Edition, SAGE Publications Ltd., London.

Julia Hirschberg, J. and D. Manning, D.M. (2015) Advances in natural language processing. Retrieved from <https://nlp.stanford.edu/~manning/xyzzzy/Hirschberg-Manning-Science-2015.pdf>

Inaam, D.M.I. A. (2016). Research Design. Retrieved from SSRN: <https://ssrn.com/abstract=2862445> or <http://dx.doi.org/10.2139/ssrn.2862445>

Kadriu, A., Abazi, L. and Abazi, H. (2019), "Albanian Text Classification: Bag of Words Model and Word Analogies," *Business Systems Research*, Vol. 10 No. 1, pp. 74–87. DOI: <https://doi.org/10.2478/bsrj-2019-0006>

Kisser, M. (2016). Introduction to Natural Language Processing. [Blog post] Algorithmia.

Kivunja, C., & Kuyini, A.B. (2017). Understanding and Applying Research Paradigms in Educational . *International Journal of Higher Education*
DOI: <https://doi.org/10.5430/ijhe.v6n5p26>

Kowsari, K., Meimandi, K.J. Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D. 2019. Classification Algorithms: A Survey

- Kumar, E. (2012). Natural Language Processing. New Delhi: I. K. International Publishing House Pvt.
- Kumar, S. (2014). Research methodology: A step by step guide for beginners (4th ed.). Jaipur, India: Yking Books.
- Kaur, J. & Saini, J. (2015). A Study of Text Classification Natural Language Processing Algorithms for Indian Languages. VNSGU Journal of Science and Technology. 4. 162-167.
- Kumar, A. (2018, September 12) Multivariate Multilabel Classification with Logistic Regression._Retrieved from <https://acadgild.com/blog/logistic-regression-multiclass-classification>
- Kurdi, M. Z. (2016). Natural Language Processing and computational linguisticsn1. London: ISTE.
- Lamont, M. (2014). How do University, Higher Education and Research Contribute to Societal Well-Being? Higher Education in Societies,9-16. doi:10.1007/978-94-6209-746-9_2
- Leavy, P. (2017). Research Design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches. Guilford Publications.
- Leech, G.N. (1995) "Corpora." *The Linguistics Encyclopedia*, ed. by Kirsten Malmkjaer. Routledge, 1995)
- Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. International Journal of Qualitative Methods. <https://doi.org/10.1177/1609406919887021>
- Lei, S., Rada, M. and Mingjun, T. (2010). Cross Language Text Classification by Model Translation and Semi-Supervised Learning
- Liu, V. and Curran, J.R. (2006). Web Text Corpus for Natural Language Processing. Proceeding of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy
- Loog, M. (2018) [Machine Learning Techniques for Space Weather](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/supervised-classification)
<https://www.sciencedirect.com/topics/earth-and-planetary-sciences/supervised-classification>
- Ma, E. (2018). 3 Basic approaches in Bag of Words which are better than Word Embeddings. Retrieved from

<https://towardsdatascience.com/3-basic-approaches-in-bag-of-words-which-are-better-than-word-embeddings-c2cbc7398016>

Maini, V. (2017, August 19). Machine Learning for Humans, Part 3: Unsupervised Learning. Retrieved from

<https://medium.com/machine-learning-for-humans/unsupervised-learning-f45587588294>

Makombe, G. (2017). An Expose of the Relationship between Paradigm, Method and Design in Research. *The Qualitative Report*, 22(12), 3363-3382. Retrieved from <https://nsuworks.nova.edu/tqr/vol22/iss12/18>

Marr, B. (2016). What Is The Difference Between Artificial Intelligence And Machine Learning? Retrieved from

<https://www.google.co.za/amp/s/www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/amp/>

Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach*.

McLeod, S. (2012). Experimental Method. Retrieved from <https://www.simplypsychology.org/experimental-method.html>

Mertens, D. M. (2014). Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods. Retrieved from https://www.sagepub.com/sites/default/files/upm-binaries/29985_Chapter1.pdf

Merriam-Webster. (n.d.). Algorithm. In Merriam-Webster.com dictionary. Retrieved August 8, 2020, from <https://www.merriam-webster.com/dictionary/algorithm>

Miles, R. H., Williams, L. A., Burke, J., & Gojak, L. (2018). *Your mathematics standards companion, grades 6-8: What they mean and how to teach them*.

Miller, B. & Ranum, D., (2013). *Problem Solving with Algorithms and Data Structures Release 3.0*

Mittwede, S. K. (2012). Research Paradigms and Their Use and Importance in Theological Inquiry and Education. *Journal of Education and Christian Belief*, 16(1), 23–40. doi:10.1177/205699711201600104

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168328/>

Novoseltseva, K. (2017, September 27). Natural Language Processing companies & examples. Retrieved from <https://apiumhub.com/tech-blog-barcelona/natural-language-processing-projects/>

- Oberiri, A. (2017). Quantitative Research Methods : A Synopsis Approach. *Arabian Journal of Business and Management Review (kuwait Chapter)*. 6. 40-47. 10.12816/0040336.
- Ogren, P. V., Savova, G., Buntrock, J. D., & Chute, C. G. (2006). Building and evaluating annotated corpora for medical NLP systems. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2006*, 1050.
- O'Keeffe, A. & McCarthy, M. (2012) 'Historical perspective: What are corpora and how have they evolved?', in O'Keeffe, A. & McCarthy, M.(eds), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, p 3-13.
- Ovchinnikov, A., Gorelik, J., & Livschitz, V. (2016, November 18). Creating training and test data sets and preparing the data [Web log post]. Retrieved from <https://blog.griddynamics.com/creating-training-and-test-data-sets-and-preparing-the-data-for-twitter-stream-sentiment-analysis-of-social-movie-reviews>
- Pedrycz, W., & Chen, S. (2016). *Sentiment Analysis and Ontology Engineering An Environment of Computational Intelligence*. Springer International Publishing.
- Peng, F., Schuurmans, D. & Wang, S. *Information Retrieval* (2004) 7: 317. <https://doi.org/10.1023/B:INRT.0000011209.19643.e2>
- Perri, G., & Bellamy, C. (2012). *Principles of methodology: Research design in social science*. Los Angeles: SAGE.
- Pérez-Ortiz, M.; Jiménez-Fernández, S.; Gutiérrez, P.; Alexandre, E.; Hervás-Martínez, C.; Salcedo-Sanz, S. A review of classification problems and algorithms in renewable energy applications. *Energies* **2016**, *9*, 607. [[Google Scholar](#)] [[CrossRef](#)]
- Phillips, P. P., & Stawarski, C. A. (2016). *Data collection: Planning for and collecting all types of data*. Place of publication not identified: Pfeiffer.
- Polster, C., & Newson, J. A. (2015). A penny for your thoughts: How corporatization devalues teaching, research, and public service in Canadas universities. Retrieved from <https://link.springer.com/article/10.1007/s10734-005-1118-z>
- Ptaszynski, M., Rzepka, R., Araki, K. & Momouchi, Y. (2011). Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, Issue 1, pp. 24-36.
- Salkind, N. J. (2010). *Encyclopedia of Research Design, Volume 1*. SAGE Publishers

Rodger, R. (2011) *Beginning Mobile Application Development in the Cloud*. John Wiley & Sons. ISBN 9781118203354

Sarkar, D. (2018, June 19). *A Practitioner's Guide to Natural Language Processing (Part I) - Processing & Understanding Text*. Retrieved from <https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72?gi=913e6a3e665c>

Saunders, M., Lewis, P., & Thornhill, A. (2012). Multiple methods research design. In *Research methods for business students*.

Saunders, M., Lewis, P. and Thornhill, A. (2009). Understanding research philosophies and approaches. *Research Methods for Business Students*. 4. 106-135.

Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English Language Teaching*, 5(9), pp.9–16.

Sealey, A., & Pak, C. (2018). First catch your corpus: Methodological challenges in constructing a thematic corpus. *Corpora*, 13(2), 229-254. doi:10.3366/cor.2018.0145

Shamoo, A. E., & Resnik, D. B. (2015). *Responsible conduct of research*. Oxford: Oxford University Press.

Shams, R., Elsayed, A. and Akter, Q.M. (2010). "A corpus-based evaluation of a domain-specific text to knowledge mapping prototype", A special issue of *Journal of Computers*, Academy Publisher

Sileyew, K.J. (2019). *Research Design and Methodology* [Online First], IntechOpen, DOI: 10.5772/intechopen.85731. Available from: <https://www.intechopen.com/online-first/research-design-and-methodology>

Szarko, M. (2017). *Philosophy: Philosophy Databases*. Retrieved from <https://libguides.mit.edu/philosophy>

Thompson, S. K. (2012). *Sampling*. Hoboken: J. Wiley.

Thyer, B. A. (2010). *The handbook of social work research methods*. Los Angeles: SAGE.

Tutorialspoint, 2019 Retrieved from https://www.tutorialspoint.com/django/django_overview.htm

Gambini, B. (2019, October 22). Rethinking the role of technology in the classroom: Study finds added access can lead to decrease in students' academic motivation. *ScienceDaily*. Retrieved from

<https://phys.org/news/2019-10-rethinking-role-technology-classroom.html>

Vyatkina, N. (2014) *Approaching Language Transfer Through Text Classification: Explorations in the Detection-Based Approach*. Bristol, UK: Multilingual Matters, 2012. Pp. viii, 189. ISBN 978–1–84769–697–7.

Williams, C. (2007). Research methods. *Journal of Business & Economic Research*, 5(3), 65-72

Wolska, M., Bao Quoc Vo, B., Tsovaltzi, D., Kruijff-Korbayová, I., Karagjosova, E., Horacek, H., Fiedler, A. & Benz Müller, C. (2004). An Annotated Corpus of Tutorial Dialogs on Mathematical Theorem Proving. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

<https://arxiv.org/ftp/arxiv/papers/1204/1204.6364.pdf> Zeroual, I., & Lakhouaja, A. (2018). Data science in light of Natural Language Processing: An overview. *Procedia Computer Science*, 127, 82-91.

doi:10.1016/j.procs.2018.01.101

Žukauskas, P., Jolita Vveinhardt, J. & Regina Andriukaitienė, R. (2018) Philosophy & Paradigm of Scientific Research. DOI: 10.5772/intechopen.70628

Lawson, A.E. How Do Humans Acquire Knowledge? And What Does That Imply About the Nature of Knowledge?. *Science & Education* 9, 577–598 (2000). <https://doi.org/10.1023/A:1008756715517>

Merriam-Webster. (n.d.). Knowledge. In *Merriam-Webster.com dictionary*. Retrieved December 8, 2020, from

<https://www.merriam-webster.com/dictionary/knowledge>

Brix, J. 2014. Improving individual knowledge construction and re-construction in the context of radical innovation. *International Journal of Innovation and Learning*, 15 (2) (2014), pp. 192-209

Lyles, M.A. 2014. Organizational Learning, knowledge creation, problem formulation and innovation in messy problems. *European Management Journal*, 32 (1) (2014), pp. 132-136

Brix, J. 2017. Exploring knowledge creation processes as a source of organizational learning: A longitudinal case study of a public innovation project, *Scandinavian Journal of Management*, Volume 33, Issue 2, Pages 113-127

Žukauskas, P., Vveinhardt, J and Andriukaitienė, R. 2018. Philosophy and Paradigm of Scientific Research, Management Culture and Corporate Social Responsibility, Pranas Zukauskas, Jolita Vveinhardt and Regina Andriukaitienė?, IntechOpen, DOI: 10.5772/intechopen.70628. Available from:

<https://www.intechopen.com/books/management-culture-and-corporate-social-responsibility/philosophy-and-paradigm-of-scientific-research>

Mittwede, S. K. (2012). Research Paradigms and Their Use and Importance in Theological Inquiry and Education. *Journal of Education and Christian Belief*, 16(1), 23–40. <https://doi.org/10.1177/205699711201600104>

Merriam-Webster. (n.d.). Ontology. In *Merriam-Webster.com dictionary*. Retrieved December 9, 2020, from <https://www.merriam-webster.com/dictionary/ontology>

Guarino, N., Oberle, D., and Staab, S. (2009). What is an Ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Tung A.K.H. (2016) Rule-Based Classification. In: Liu L., Özsu M. (eds) *Encyclopedia of Database Systems*. Springer, New York, NY. https://doi.org/10.1007/978-1-4899-7993-3_559-2

Harispe S, Ranwez S, Janaqi S, et al. Semantic similarity from natural language and ontology analysis. *Synth Lect Hum Lang Technol* 2015; 8: 1–254.

Schroeder, Mark, "Value Theory", *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2016/entries/value-theory/>.

Tufail, Mubeshera. (2012). Axiology. 10.13140/RG.2.1.2456.8408.

Edelheim, Johan R. , "Ontological, epistemological and axiological issues" , in *The Routledge Handbook of Tourism and Hospitality Education* ed. Dianne Dredge , David Airey and Michael J. Gross (Abingdon: Routledge, 10 Oct 2014), accessed 10 Dec 2020 , Routledge Handbooks Online.

Quoc V. Le and Tomas Mikolov, *Distributed Representations of Sentences and Documents*, (2014) <http://bit.ly/2GJBHjZ>

APPENDICES

APPENDIX A: DEFINITION OF KEY TERMS

In this section important key terms for the study are defined for ease of understanding.

Researcher

In the present study the term researcher refers to any person involved in the generation of knowledge. The term is used as an umbrella for students, lecturers and knowledge workers alike.

Research Philosophy

Research Philosophy refers to a set of convictions and presumptions about the generation and improvement of information (Saunders et al., 2009)

Research Paradigm

Research paradigm refers to the ideological direction of knowledge workers towards the social world they research (Saunders et al., 2009)

Ontology

Ontology refers to a belief system that represents an individual's understanding of what constitutes reality. Ontology shapes the way in which a researcher's views and study the object of their research (Saunders et al., 2009)

Corpus

A corpus is a database or collection of textual material used in natural language processing tasks (Shams, 2010)

Algorithm

Algorithm refers to computer program that provides a computer with procedural steps of accomplishing a task. (Merriam-Webster, n.d)

Classifier

A classifier is a tool that implements the functional mapping of features to the predefined categories in a supervised classification model (Loog, 2018).

Supervised classification model

Supervised classification models are classification algorithms that use training data to know and understand how to relate input text to predefined categories of classes (Asiri, 2018).

APPENDIX B: TEST SCRIPT

INTRODUCTION

The Research Methods Index (RMI) system is a straight-forward system that will automatically predict and assist in determining a participant's research philosophical stance and the underlying paradigm closely aligned to their beliefs. A participant will receive a link (<http://unisa-rppi.westeurope.cloudapp.azure.com:8082/nlp>) that they can launch to access the system. Once the participant launches the system, they should perform the tasks that are related to RMI prototype system only. Internet access and the registration of an account are required before performing any activities in the system. A proper email address is required for registration purpose and activation of the user account. The testing activities to be completed by participants are as follows:

1. Register to use the system

The participants must register an account by clicking on the **'REGISTER HERE'** button to get the registration form. They must then fill in their details to register an account; once the registration is successful, the user will then be able to login and use the system.

2. Login after registration

Once the registration is successful, the participant can login using the registered username and password. Upon successful login, the system will launch the home page with the option to **'TAKE QUESTIONNAIRE'**. The system will not allow a participant to login until an account is registered.

3. Take a questionnaire

The participants are expected to take the questionnaire. At the end of the questionnaire, they must click on the **'SUBMIT'** button to store their responses in the system. An option to view the report will display.

4. View report

Once the questionnaire is completed and submitted, participants must click on the **'VIEW REPORT'** button, the system will automatically generate a report with the three topmost recommended research philosophies and paradigms based on their response.

APPENDIX C: SYSTEM USABILITY QUESTIONNAIRE

Research Methods Index (RMI)-Natural Language Processing (NLP) Software Evaluation

Thank you for providing feedback about the RMI-NLP Software.

Before we begin, please provide some important background information.

Section A: Biographical data								
<p>1. Gender</p> <p><input type="checkbox"/> Male <input type="checkbox"/> Female</p>								
<p>2. Profession (Please tick where applicable)</p> <p><input type="checkbox"/> NLP Expert</p> <p><input type="checkbox"/> Academic (PhD)</p> <p><input type="checkbox"/> Graduate (BA, MA, etc)</p> <p><input type="checkbox"/> Student</p> <p><input type="checkbox"/> Other (Please specify) _____</p>								
<p>If Student please complete follow-up section below:</p> <p><input type="checkbox"/> Postgraduate <input type="checkbox"/> Undergraduate</p> <p>Year of study : _____</p> <p>Qualification under study</p> <p><input type="checkbox"/> Bachelors' Degree <input type="checkbox"/> Master's Degree <input type="checkbox"/> PhD</p>								
Section B: Variables, root constructs, definitions, and scales				1	2	3	4	5
Please complete the questionnaire below by ticking an option from 1-5 on the rating scale (5 being strongly agree/ 1 being strongly disagree), in response to the following statements								
Variable	Construct	Definitions	Items					
Effort	Perceiv	The degree to	1. Learning to operate the					

Expectancy	ed ease of use	which the user believes that using a system would be free of effort	system would be easy for a non-technical user					
			2. I would find it easy to get the system to do what I want it to do					
			3. My interaction with the system would be clear and understandable					
			4. I would find the system to be flexible to interact with					
			5. It would be easy for me to become skilful at using the system					
			6. I would find the system easy to use					
Performance Expectancy	Perceived usefulness	The degree to which the user believes that using the system would enhance his or her research performance and understanding	1. Using the system in my research would enable me to accomplish research tasks more quickly					
			2. Using the system would improve my epistemological understanding					
			3. Does the system perform well when there is concurrent use of the system					
			4. The time required to fetch results from the database is acceptable					
			5. I would find the system useful					
Attitude towards using the technology	Attitude toward behaviour	User's positive or negative feelings about performing the target behaviour	1. Using the system requires an understanding NLP concepts					
			2. I support the idea of using the system to enhance the learning process					
	Affect	User's liking of	3. I would like using					

		the behaviour	technology to learn more about the subject matter the system addresses instead of the traditional way					
			4. I look forward to those aspects of research that will be supported by the use of the system					
Knowledge Expectancy (pre questionnaire)	Concepts	Understanding of concepts	1. I know and understand the concept of research philosophy					
			2. I know and understand what a research paradigm is					
			3. I understand the value of research philosophies and paradigms in conducting research.					
			4. I know which research philosophy I espouse					
			5. I understand the relevance of research philosophies and paradigms in conducting research.					
	Introduction	Introduction to concepts	6. The way in which the concepts of research philosophies and paradigms are introduced is beneficial.					
			7. I am able to find out more information about research philosophies and paradigms.					
Open ended questions (Please answer all questions below)								
1. Please list any glitches, spelling/formatting errors, or parts of this app that were inaccessible?								
<hr/>								

2. At what point does software fail given more users/transactions?

3. Comment

Future research (Please answer where applicable)

Would you be willing to participate in a similar survey in the near future?

Yes

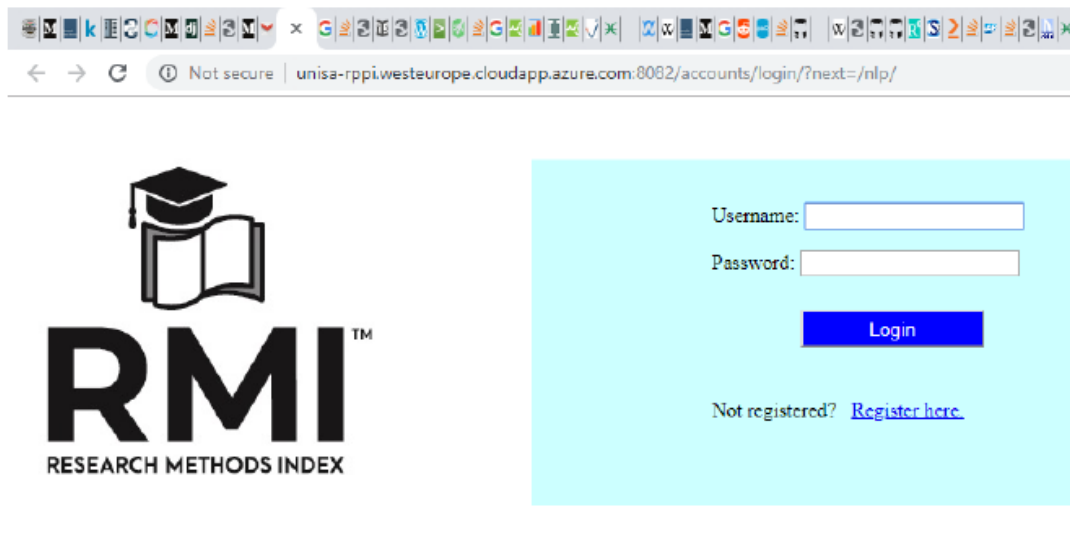
No

If Yes, please provide email address where you can be reached at:

APPENDIX D: USER MANUAL

A researcher will be provided with the link to the RMI application, deployed through the Azure cloud service. The researcher will then register a user account for them to be able to sign-in. Once signed-in, the user can proceed to taking a questionnaire and answer the questions posed. At the end of the questionnaire the researcher can submit their answers, view their report and sign-out. The procedure is illustrated in Figures 4.7, 4.8, 4.9, 4.10 and 4.11;

i. User registration (Researcher)



Copyright © 2018 - University of South Africa.

Login page of the RMI application

- a) Click on 'register here'



Sign up

Username: Required. 150 characters or fewer. Letters, digits and @/./+/_ only.

First name: Optional.

Last name: Optional.

Email: Required. Inform a valid email address.

Password:

Your password can't be too similar to your other personal information.
Your password must contain at least 8 characters.
Your password can't be a commonly used password.
Your password can't be entirely numeric.

Password confirmation: Enter the same password as before, for verification.

Form for registering user account

- b) Fill in the registration form noting the naming conventions and password rules, on completion click on sign-up

[Logout](#) [Report Home](#)



Welcome Test1



Landing page of the RMI NLP application

- c) Click on Take Questionnaire and answer questions that follow

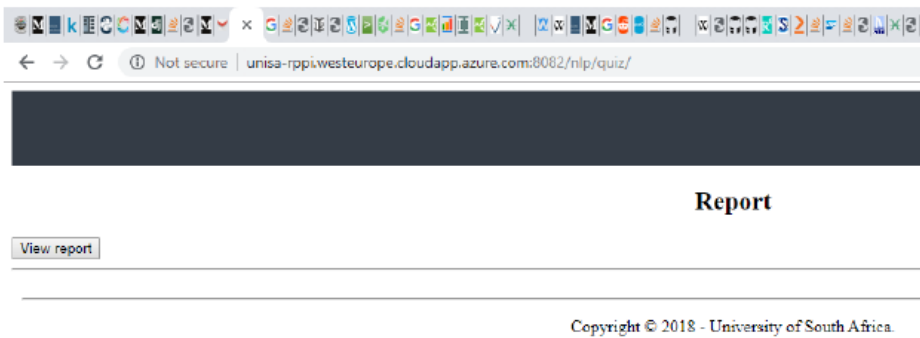
Please answer the following questions to the best of your understanding

Question 1: How many versions of the truth can there be in a given situation?	There is only one version of the truth
Question 2: How can the truth be influenced?	The truth is universal
Question 3: How can truth be justified?	The truth is independent of the individual
Question 4: How is knowledge acquired i.e., how do we know what we know?	Reality is based on knowledge or perception of a situation or fact
Question 5: What influences what we know?	Ascribes the motion and changes of the world to some external force
Question 6: How many sources of knowledge are there?	accessible to human reason and accessible in a non-sophisticated way
Question 7: How can knowledge be advanced?	Moral beliefs can be justified non-inferentially or inferentially
Question 8: What is the importance of values and ethics?	Independent of individual opinion and free from bias
Question 9: How can personal values influence the truth?	Knowledge about moral principles solely acquired through reasoning
Question 10: What determines our values and ethics?	Good and evil are objects of moral truths of an universal validity

Submit

NLP questionnaire

d) Submit questionnaire



Button to view the NLP report



RMI NLP report

View report and Sign out

APPENDIX E: SAMPLE OF THE CORPUS

Natural Language Processing								
1	Variables	Nature	Meaning	Tokenizing	PoS Tagging	Word Embeddings	Stemming	NER
2	Absolute idealism	Single	Truth exists independent of us, whether we know it or not Physical world is only	'Truth', 'exists', 'independent', 'of', 'us,', 'whether', 'we',	Truth/NNP exists/NNS independent/JJ of/IN us/PRP whether/IN we/PRP know/VBP it/PRP or/CC	knowledge understanding skills practice information scientific teaching influence development process communication	Truth exist independ	Truth (exists independent/RL) of us, whether we know it or not Physical world is only an (appearance/BL) to our expression of mind/ (Only one/RL) reality or world view in a (well balanced manner/TR)
3	Absolute idealism		Knowledge can be seen as mental or spiritual in nature Knowledge can be	'Knowledge', 'can', 'be', 'seen', 'as', 'mental', 'or', 'spiritual',	Knowledge/NNP can/MD be/VB seen/VBN as/IN mental/JJ or/CC spiritual/JJ in/IN nature/NN		Knowledge can be obtain through pure uniform	Knowledge can be seen as (mental or spiritual/VL) in nature Knowledge can be obtained through (pure uniform spiritual consciousness/MT) All views come together in a state of harmony
4	Absolute idealism		Values and morals are representation of the truth, not the truth itself	'Values', 'and', 'morals', 'are', 'representation', 'of', 'the', 'truth',	Values/NNS and/CC morals/NNS are/VBP representation/NN of/IN the/DT truth/NN not/RB		Valu and moral are represent of the truth	Values and morals are (representation /AE)of the truth, not the truth itself
5	Absurdism	Unknown	The existence or non-existence of anything is meaningless A world of contradictions where nothing is significant	'The', 'existence', 'or', 'non-existence', 'of', 'anything', 'is', 'meaningless.', 'A', 'world', 'of'	NN The/DT existence/NN or/CC non/NN -/: existence/NN of/IN anything/NN is/VBZ meaningless/NN A/DT world/NN of/IN	meaningless however thus being therefore cause often both that whole sometimes	A world of contradict where noth is signific	The existence or non-existence of anything is (meaningless/RL) A world of contradictions where (nothing is significant/RL) Individuals and their (existence are not important/BL)
6	Absurdism		The world is full of contradictions and not easy to understand Impossible to obtain	'The', 'world', 'is', 'full', 'of', 'contradictions', 'and', 'not',	NN The/DT world/NN is/VBZ full/JJ of/IN contradictions/NNS and/CC not/RB easy/JJ to/TO		The world is full of contradict and not easi to	The world is full of contradictions and not easy to understand (Impossible to obtain full knowledge/SC) in an unreasonable world Better to rely on one's (experience/MT) than phenomena
7	Absurdism		No hereafter, or life after death Nothing needs to be taken seriously	'No', 'hereafter', 'or', 'life', 'after', 'death.', 'Nothing',	NN No/DT hereafter/NN or/CC life/NN after/IN death/NN Nothing/NN needs/VBZ to/TO be/VB		No hereaft , or life after death	(No hereafter/ET), or life after death (Nothing needs to be taken seriously/AE)
8	Accidentalism (philosophy)	Multiple	Things have only accidental properties, no essential properties, or no common nature.	'Things', 'have', 'only', 'accidental', 'properties',	NN Things/NNS have/VBP only/RB accidental/JJ properties/NNS no/DT essential/JJ properties/NNS	properties values current system specific levels	All event and properti are accident	Things have only (accidental properties/TR), (no essential properties/TR), or no common nature. All events and properties are (accidental/BL). The occurrence of some events is either (not necessitated/RL) or (not
9								

APPENDIX F: RELATED STUDIES SOURCES

Field	Year	Title	Author	DOI	Journal
TEXT CLASSIFICATION REFERENCES					
Marketing Research	2018	Text Classification for Organizational Researchers: A Tutorial. Organizational Research	Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N.	https://doi.org/10.1177/1094428117719322	Methods, 21(3), 766–799.
	2011	Classifying Business Marketing Messages on Facebook.	Yu, Bei & Kwok, Linchi. (2011).		
Aviation	2019	A machine learning approach to analyze customer satisfaction from airline tweets	Kumar, S., Zymbler, M., 62 .	https://doi.org/10.1186/s40537-019-0224-1	
	2015	A Text Mining Classification Framework and its Experiments Using Aviation Datasets.	Xu, Brian & Kumar, Sathish.		
	2016	Natural Language Processing for aviation safety reports: from classification to interactive analysis.	Ludovic Tanguy, Nikola Tulechki, Assaf Urieli, Eric Hermann, Céline Raynal.	ff10.1016/j.compind.2015.09.005ff. ffhalshs-01322238f	Computers in Industry,

Field	Year	Title	Author	DOI	Journal
TEXT CLASSIFICATION REFERENCES					
	2010	Cause identification from aviation safety incident reports via weakly supervised semantic lexicon construction.	Abedin, M. A. U., Ng, V. and Khan. L.		J. Artif. Int. Res., 38:569–631
Educati on	2016	Text classification of student self-explanations in college physics questions.	S. Bhatnagar, M. Desmarais, N. Lasry, and E. S. Charles.		In Proc. 9th Intl. Conf. Educ. Data Min ., pages 571–572, July 2016.
	2017	Short-answer responses to stem exercises: Measuring response validity and its impact on learning	A. Waters, P. Grimaldi, A. S. Lan, and R. G. Baraniuk.		In Proc. Conf. Edu. Data Mining, pages 374–375,
	2010	Supporting the education evidence portal via text mining. Philosophical transactions.	Ananiadou, S., Thompson, P., Thomas, J., Mu, T., Oliver, S., Rickinson, M., Sasaki, Y., Weissenbacher, D., & McNaught, J. (2010).	https://doi.org/10.1098/rsta.2010.0152	Series A, Mathematical, physical, and engineering sciences, 368(1925), 3829–3844.

APPENDIX G: ARCHITECTURE DIAGRAMS

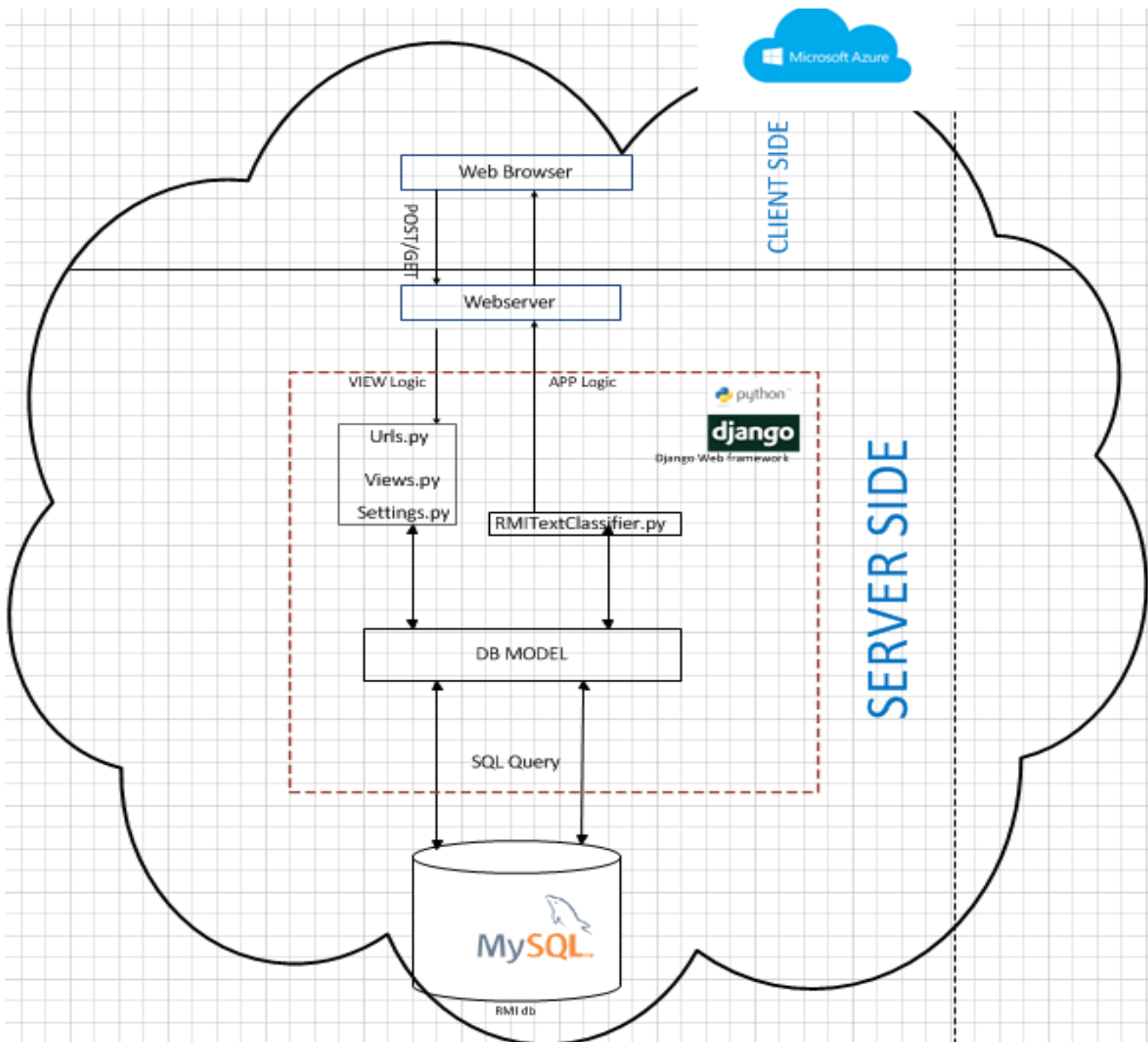


Figure 8.1 Research Methods Index NLP architecture diagram

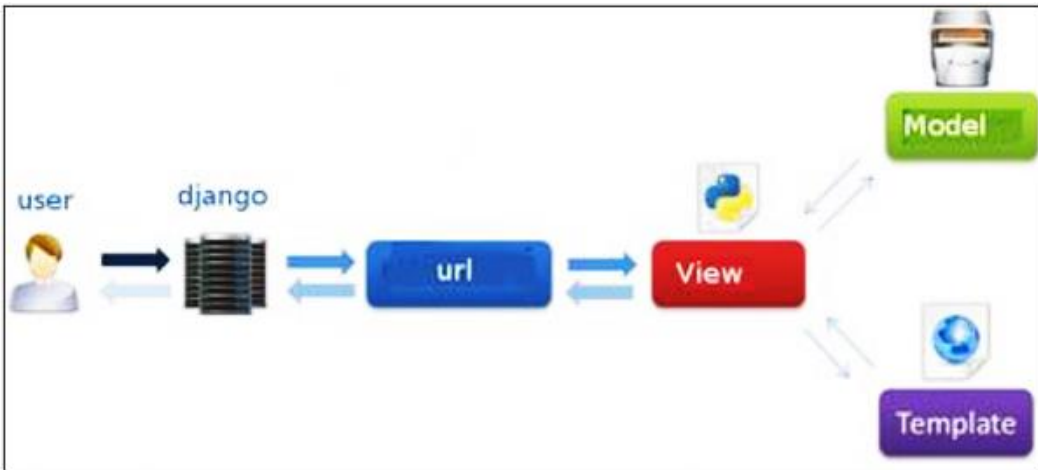


Figure 8.2 Interaction of the components of the MVT pattern (Tutorialspoint, 2019)

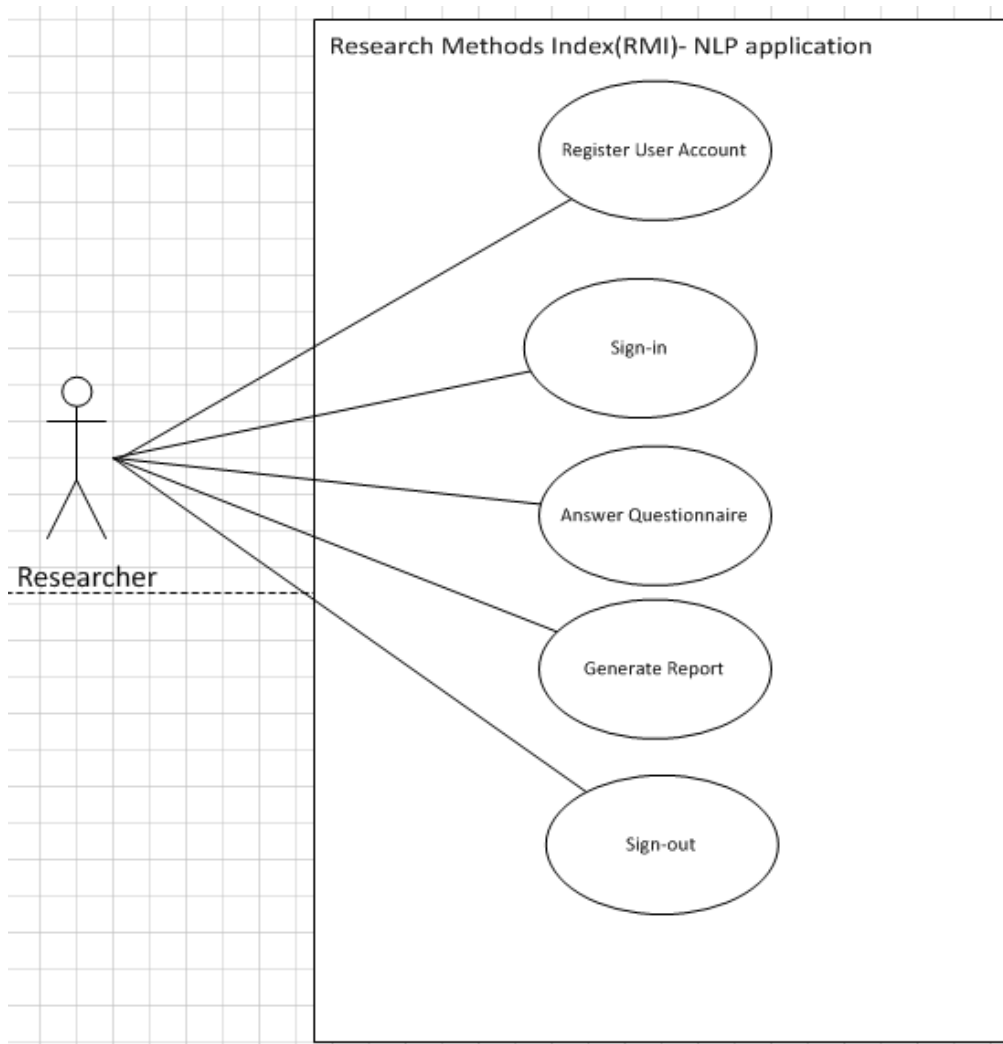


Figure 8.3 User interaction with the NLP application

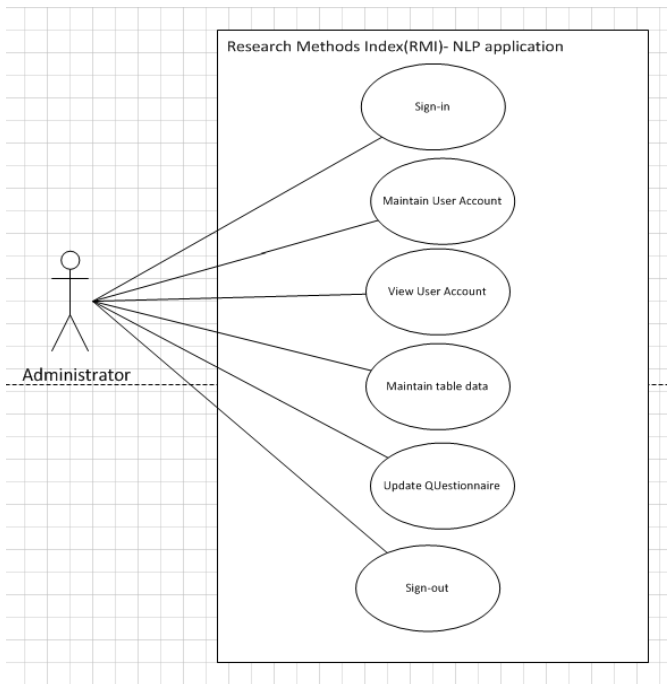


Figure 8.4 Administrator interaction with the NLP application

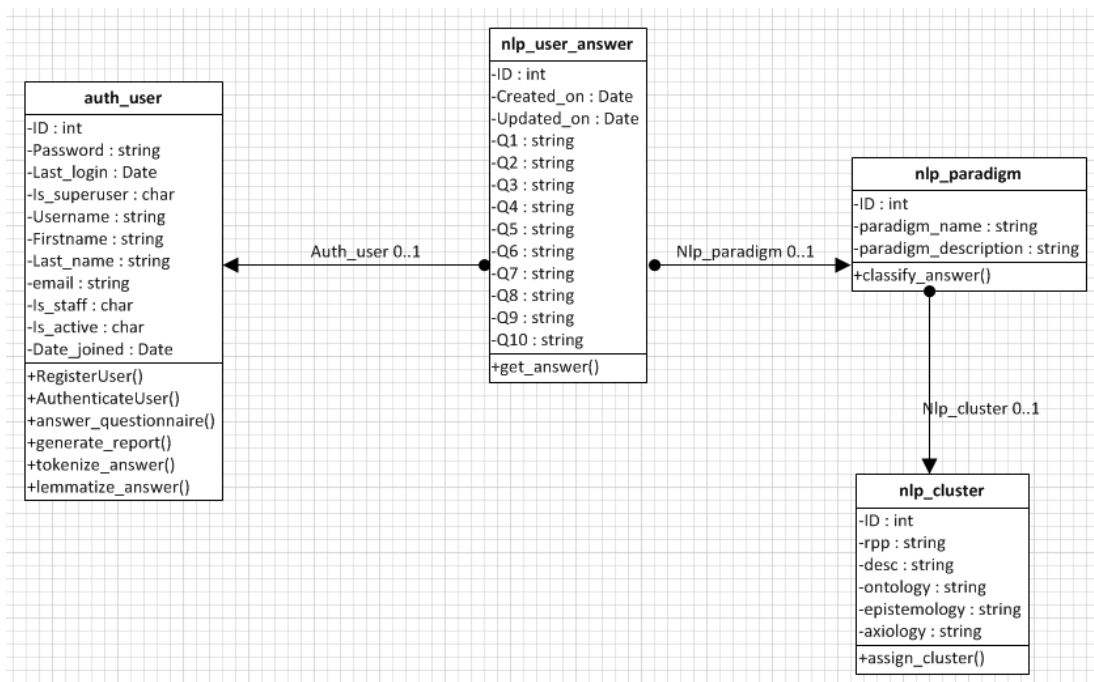


Figure 8.5 Class diagram of the NLP system

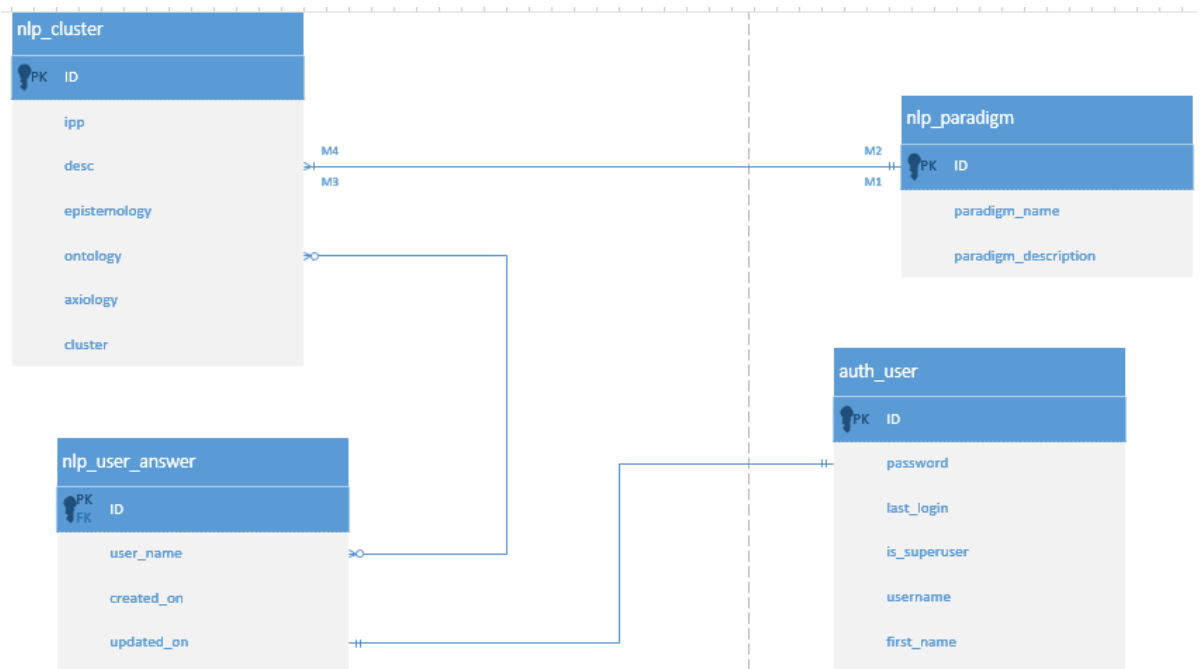


Figure 8.6 Entity relationship diagram of the NLP system

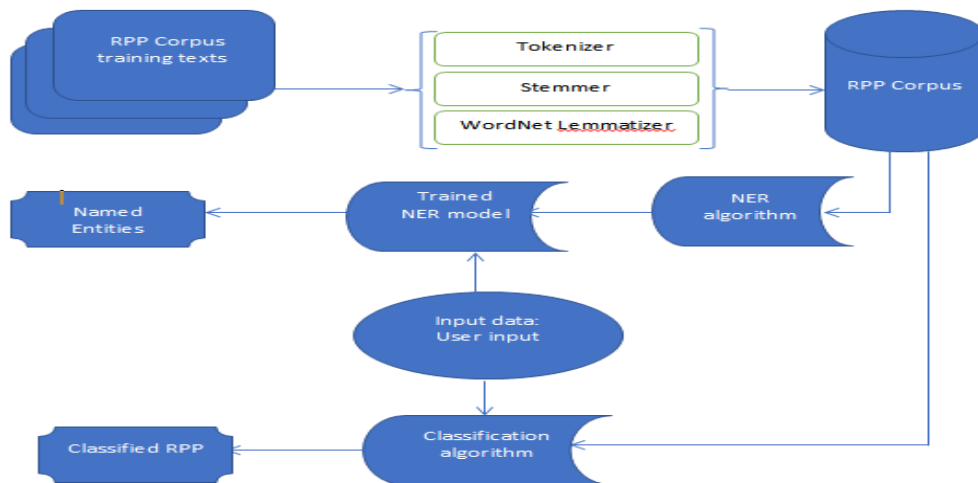


Figure 8.7 The NLP process flow diagram for text classification

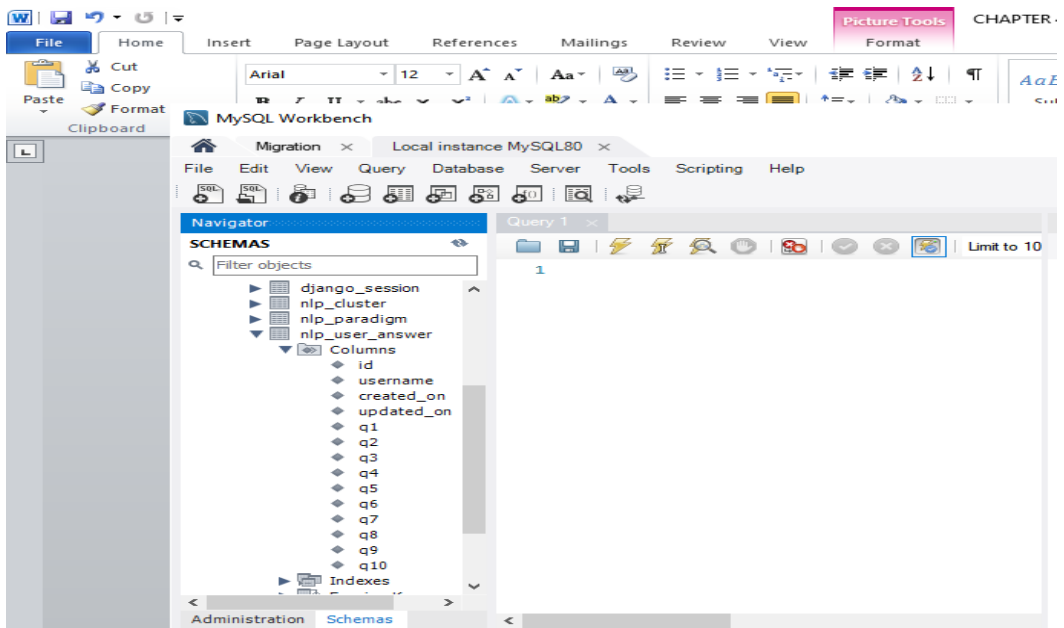


Figure 8.8 The database schema

APPENDIX H: PARTICIPATION CONSENT FORM



26 October 2019

Title: Natural Language Processing for Research Philosophies and Paradigms

Dear Participant

I am Ntombhi Mkalipi, a Masters' student in the Department of Information Technology at Unisa, under the supervision of Prof M Mkansi, with Prof E Mnkandla as the co-supervisor. I am currently registered for MTech Computer Science and am busy with the development of research software, using Natural Language Processing, to enhance researchers' understanding of some complex research phenomenon such as research paradigms and philosophies. I request your participation in order to evaluate and conduct:

- **Software evaluation (Design expectancy, effort expectancy, performance expectancy, graphic user interface and acceptance)**

You are invited to participate in this survey because of your expertise and interest in research. While there is no direct benefit in participating, this is part of my quest to developing and delivering an efficiently working Natural Language Processing (NLP) system that will help in enhancing teaching and learning at tertiary levels.

Please complete the attached questionnaire by answering all the questions asked. There are no right or wrong answers, but PLEASE ANSWER ALL QUESTIONS as honestly as possible. The software evaluation survey should take about 10-20 minutes to complete.

After the completion of the questionnaire, a report with three topmost recommended research philosophies and their underlying paradigms that are closely aligned with your world view will be generated. The report is based on the responses that you will have provided on the questionnaire. You are most welcome to save and research them further for yourself – it should be an interesting discovery.

The University of South Africa complies with the relevant data protection legislation and your responses and personal details will not be divulged to anyone else. No source, individual or institution, will be identified or comment attributed without written permission of the originator and you may withdraw at any time without consequence of any kind.

The data will be kept and used for academic purposes such as writing an article for publication in an academic journal. Such an article will only refer to post-graduate students as a whole and no names whatsoever (no names are ever requested on the questionnaire) will be mentioned. My computer is password protected and I am the only one who has access to it.

Feedback is available upon request. You may contact me anytime. There is no penalty or loss of benefit for non-participation.



University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

APPENDIX I: ETHICAL CLEARANCE CERTIFICATES



COLLEGE OF ECONOMIC AND MANAGEMENT SCIENCES RESEARCH ETHICS REVIEW COMMITTEE

10 May 2016

Dear Prof Mkansi

Ref #: 2016_CRERC_009(FA)

Name of applicant: Prof Marcia Mkansi

Student number #: 90215028

Decision: Ethics Approval

Name: Prof Marcia Mkansi, mkansm@unisa.ac.za, 012 429-2339 or 084 901 0362

Proposal: Knowledge product for knowledge development: a theory of constraints perspective

Qualification: n.a.

Thank you for the application for research ethics clearance by the College of Economic and Management Sciences Research Ethics Review Committee for the above mentioned research. Final approval is granted from 10 May 2016 to 11 May 2018.

For full approval: The application was expedited reviewed in compliance with the Unisa Policy on Research Ethics by members the CRERC.

The proposed research may now commence with the proviso that:

- 1) The researcher/s will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.
- 2) Any adverse circumstance arising in the undertaking of the research project that



University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

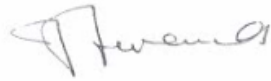
is relevant to the ethicality of the study, as well as changes in the methodology, should be communicated in writing to the CRERC.

- 3) *An amended application could be requested if there are substantial changes from the existing proposal, especially if those changes affect any of the study-related risks for the research participants.*
- 4) *The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study.*

Note:

The reference number **2016_CRERC_009(FA)** should be clearly indicated on all forms of communication [e.g. Webmail, E-mail messages, letters] with the intended research participants, as well as with the CRERC.

Kind regards,



Prof JS Wessels

Chairperson of the CRERC, CEMS, UNISA
012 429-6099 or wessejs@unisa.ac.za



Prof M.T. Mogale

Executive Dean: CEMS
mogalmt@unisa.ac.za



University of South Africa
Pretter Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

Request for:

Extensions of Ethical clearance number: 2016_Crec_009FA, which was granted on the 10 of May 2016. The project commenced in March 17, a year after the certificate which impacted on the anticipated completion time. Please extend it up until May 2021.

UNIVERSITY OF SOUTH AFRICA CRERC RESEARCH ETHICS REVIEW COMMITTEE

RESEARCH ETHICS PROGRESS REPORT FOR HUMAN RESEARCH ETHICS ¹

2017

If you have any questions about or require assistance with the completion of this form, please contact your supervisor (master's or doctoral students), or the Research Ethics and Integrity Advisor: engelm1@unisa.ac.za

IMPORTANT:

GUIDELINES FOR COMPLETING THE PROGRESS REPORT

1. **Ethics approval is valid for the time period stipulated on the research ethics approval certificate** in accordance with the risk category of the study (**non-health negligible risk studies** = between 3 and 5 years; **low risk PhD studies** = maximum of 5 years; **low risk Master and non-degree studies** = maximum of 3 years; **Medium and High risk studies** = validity period is risk dependent and annual renewal could be required; **health research** = annual renewal)
2. A **progress report** is an application for (a) **renewal of ethics approval**, a (b) **request for amendment** to a current application or (c) **notification of a completed/terminated research project**. It must be submitted well before the ethics approval expiry date, so that the progress report can be reviewed and the project re-approved prior to the expiry date.
3. **No research may continue without a valid research ethics certificate and re-approval.**
4. The **progress report** should contain **sufficient information** to allow the ethics review committee (ERC) to conduct a substantive and meaningful review of the progress of the project, including any challenges or problems encountered.
5. **Requests for amendments** must be accompanied by supporting documents essential for review purposes, i.e. updated Informed Consent Leaflets, Data Collection Instruments, Risk Assessment, Measures to Ensure Data Security, etc.

For applicant use <i>*This section is needed for record keeping and reporting.</i>	
DATE OF REPORT (when submitted to the EFC)	10 May 2016
ETHICS CERTIFICATE REFERENCE NUMBER	2016_CRERC_009(FA)

¹ This document is based on the content of the following reports: 1) Research Ethics Progress Report Human Research Ethics, Tshwane University of Technology & 2) Progress Report, Health Research Ethics Committee, Faculty of Health Sciences, Stellenbosch University and Form FHS)16, University of Cape Town.

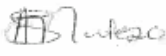


Approved by UERC V1
27 July 2017

University of South Africa
Pretorius Street, Midrand, City of Tshwane
PO Box 394 UNISA 0001 South Africa
Telephone: +27 12 429 3011 Facsimile: +27 12 429 4150
www.unisa.ac.za

CURRENT ETHICS APPROVAL WAS GRANTED UNTIL	11 May 2018
EXPECTED DATA OF RESEARCH COMPLETION (MONTH & YEAR)	
TYPE OF REPORT a) Report of ongoing project to renew ethics approval	31 May 2021
b) Request for amendments	
c) Report of completed/terminated project	


**This section is for office use only.*

APPLICATION NUMBER	
DATE PROCESSED (submitted to reviewers)	
RISK LEVEL (<i>low, medium or high</i>)	
TYPE OF REVIEW (<i>expedited or full committee review</i>)	
AGENDA DATE <i>(For expedited transactions, the agenda date is the date the expedited approval gets reported or ratified at the convened ERC)</i>	
DECISION OF ERC (<i>approved, referred back, disapproved</i>)	
DATE OF ISSUING APPROVAL CERTIFICATE OR FEEDBACK LETTER	
Period for which approval is valid (Approved until/next renewal date)	
Signature Chairperson of the ERC	
Date signed: 12/03/2019	
<u>Comments to principal researcher from the ERC</u>	

PRIVACY INFORMATION:

The information you provide on this form is collected for the primary purpose of assessing your progress up to date (including completion of a research project) or to approve amendments to the current research. This information will be entered into a database to assist with future

APPENDIX J: TURNITIN RECEIPT

 Turnitin Originality Report

Natural Language Processing for Research
Philosophies and Paradigms by
Ntombhimuni Mawila

From Revised Language Edited
Thesis/Dissertation (CSET M&D Students)

Similarity Index 14%	Similarity by Source	
	Internet Sources:	8%
	Publications:	3%
	Student Papers:	11%

Processed on 18-Dec-2019 10:18 SAST
ID: 1236438583

Word Count: 21557

sources:

- 1** 1% match (Internet from 11-Dec-2019)
http://uir.unisa.ac.za/bitstream/handle/10500/26158/dissertation_mphahlele_sm.pdf?isAllowed=y&sequence=1
- 2** < 1% match (student papers from 20-Dec-2015)
[Submitted to Universiti Teknologi MARA on 2015-12-20](#)
- 3** < 1% match (Internet from 28-Aug-2019)
<https://simpleisbetterthancomplex.com/tutorial/2017/02/18/how-to-create-user-sign-up-view.html>
- 4** < 1% match (Internet from 07-Dec-2018)
<https://core.ac.uk/download/pdf/140678.pdf>
- 5** < 1% match (student papers from 17-Oct-2016)
[Submitted to University of Hong Kong on 2016-10-17](#)
- 6** < 1% match (Internet from 02-Jun-2019)
<https://torina.top/detail/347/>
- 7** < 1% match (student papers from 15-Aug-2018)
[Submitted to TechKnowledge on 2018-08-15](#)
- 8** < 1% match (Internet from 09-Dec-2017)
<https://arxiv.org/pdf/1512.05742.pdf>
- 9** < 1% match (student papers from 29-Sep-2017)
[Submitted to University College London on 2017-09-29](#)
- 10** < 1% match ()
http://ir.uitm.edu.my/id/eprint/16393/2/TM_NORHAFIZAH%20CHE%20MAT%20AC%2013_5.pdf
- 11** < 1% match (Internet from 11-Oct-2012)
<http://tyovoima.tracon.fi/sources/3358212263>

APPENDIX K: SOURCE CODE

```
DATABASES = {
    'default': {

        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'nlp',
        'USER': 'root',
        'PASSWORD': '****',
        'HOST': '127.1.0.0',
        'PORT': '3306',
    }
}

ROOT_URLCONF = 'nlp_project.urls'

TEMPLATES = [
    {
        'BACKEND':
'django.template.backends.django.DjangoTemplates',
        'DIRS': [],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',
                'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]
```

Figure 8.9 Create project applications

```
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'nlp',
]
```

Figure 8.10 Installed apps

```

from django.db import models
from django.contrib.auth.models import User

class user_answer(models.Model):
    username = models.CharField(max_length=50, null=True)
class cluster(models.Model):
    rpp = models.CharField(max_length=100, null=True)
class paradigm_answer(models.Model):
    answer = models.CharField(max_length=1000)
    question = models.ForeignKey(paradigm_question, on_delete=models.CASCADE)
    score = models.IntegerField(null = True)
    def __str__(self):
        return self.answer

```

Figure 8.11 Create the data model

```

python manage.py makemigrations nlp_project
python manage.py migrate

```

Figure 8.12 Update database schema

```

from django.contrib import admin
from django.conf.urls import include, url
urlpatterns = [
    url('admin/', admin.site.urls),
    url('', include(('nlp.urls', 'nlp'), namespace= 'nlp')),
]

```

Figure 8.13 Adding url patterns for the nlp application

```

from django.shortcuts import render
from django.http import HttpResponseRedirect
from django.shortcuts import get_object_or_404, render, redirect,
render_to_response
from django.template import RequestContext
from django.urls import reverse
from django.views.generic import TemplateView, FormView,
ListView
from . models import user_answer, paradigm_question, cluster
from django import forms
from . RMITextClassifier import classify
from . RuleBased_NER import named_entity
from nlp.forms import SignUpForm, nlp_form
from django.contrib.auth.decorators import login_required
from django.contrib.auth.mixins import LoginRequiredMixin
from django.utils.decorators import method_decorator
from django.contrib.auth.models import User
from django.contrib.auth import login, authenticate

```

Figure 8.14 Defining custom views

```

from django import forms
from . models import user_answer
from django.contrib.auth.forms import UserCreationForm
from django.contrib.auth.models import User
from django.shortcuts import get_object_or_404, render, redirect,
render_to_response
from django.forms import modelformset_factory, TextInput
from django.utils.safestring import mark_safe

class SignUpForm(UserCreationForm):
    first_name = forms.CharField(max_length=30, required=False,
help_text='Optional.')
    last_name = forms.CharField(max_length=30, required=False,
help_text='Optional.')
    email = forms.EmailField(max_length=254, help_text='Required.
Inform a valid email address.')
    class Meta:
        model = User
        fields = ('username', 'first_name', 'last_name', 'email',
'password1', 'password2', )

class nlp_form(forms.ModelForm):
    class Meta:
        model = user_answer
        fields = ('q1', 'q2', 'q3', 'q4', 'q5', 'q6', 'q7', 'q8', 'q9', 'q10',)

```

Figure 8.15 Forms.py script

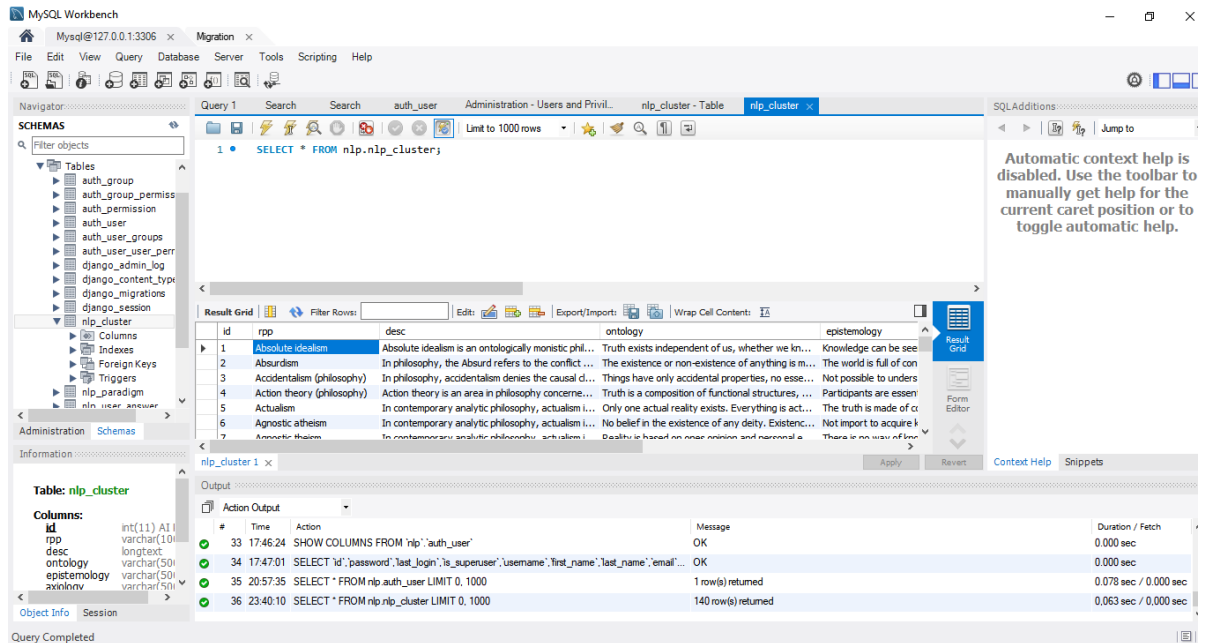


Figure 8.16 View of MySQL database schema

```

Nlp_project/
  __init__.py
  Models.py
  management/
    __init__.py
    commands/
      __init__.py
      my_command.py
  tests.py
  views.py

```

Figure 8.17 Creation of a sub directory of management and command

```

from .RMITextClassifier import classify
def clusterView(request):
    template_name = 'nlp/results.html'
    if request.method == ('GET') or request.method == ('POST'):
        try:
            dirname = 'nlp/static/media/' + str(request.user)
            nlp_query_data = []
            nlpUserAnswers =

            nlp_predict = classify(nlp_data, dirname)

```

Figure 8.18 Use of the 'GET' and 'POST' methods

MODELS.PY

```
from django.db import models
```

```
from django.contrib.auth.models import User
```

```
# Create your models here.
```

```
class user_answer(models.Model):
```

```
    username = models.CharField(max_length=50, null=True)
```

```
    created_on = models.DateTimeField(auto_now_add=True)
```

```
    updated_on = models.DateTimeField(auto_now=True)
```

```
    q1 = models.TextField(max_length=5000, null=True)
```

```
    q2 = models.TextField(max_length=5000, null=True)
```

```
    q3 = models.TextField(max_length=5000, null=True)
```

```
    q4 = models.TextField(max_length=5000, null=True)
```

```
    q5 = models.TextField(max_length=5000, null=True)
```

```
    q6 = models.TextField(max_length=5000, null=True)
```

```
    q7 = models.TextField(max_length=5000, null=True)
```

```
    q8 = models.TextField(max_length=5000, null=True)
```

```
    q9 = models.TextField(max_length=5000, null=True)
```

```
    q10 = models.TextField(max_length=5000, null=True)
```

```
class cluster(models.Model):
```

```
    rpp = models.CharField(max_length=100, null=True)
```

```
    desc = models.TextField(max_length=5000, null=True)
```

```
    ontology = models.CharField(max_length=5000, null=True)
```

```
    epistemology = models.CharField(max_length=5000, null=True)
```

```
    axiology = models.CharField(max_length=5000, null=True)
```

```
    cluster = models.CharField(max_length=100, null=True)
```

```
class paradigm(models.Model):
```

```
    paradigm_name = models.CharField(max_length=200)
```

```
    paradigm_description = models.CharField(max_length=1000)
```

```

def __str__(self):
    return self.paradigm_name

class paradigm_question(models.Model):
    question = models.TextField(max_length=1000)
    # paradigm_component =
    models.ForeignKey(paradigm_component,on_delete=models.CASCADE,
    default='1')
    def __str__(self):
        return self.question
class paradigm_answer(models.Model):
    answer = models.CharField(max_length=1000)
    question = models.ForeignKey(paradigm_question,on_delete=models.CASCADE)
    score = models.IntegerField(null = True)
    def __str__(self):
        return self.answer

```

RMITEXTCLASSIFIER.PY

```

# use natural language toolkit
import nltk
from nltk.stem.lancaster import LancasterStemmer
from nltk.stem import WordNetLemmatizer
import numpy as np
import matplotlib.pyplot as plt
from pathlib import Path
from django.shortcuts import render
from nltk.corpus import wordnet
import spacy
import random
from spacy import displacy
from spacy.util import get_lang_class
#from django.shortcuts import render

```

```

# word stemmer
stemmer = LancasterStemmer()
lemmatizer = WordNetLemmatizer()

## classes of training data
#training_data =
open("C:/Users/Ntombhi/Anaconda3/nlp_project/nlp/CORPUS_DATA.txt", "r")
#training_data = training_data.read()
##if training_data.mode == 'r':
## print('success')
chart_data = []
training_data = []
#
training_data.append({"class":"Realism_Absolute idealism","sentence":"Truth exists
independent of us, whether we know it or not. Physical world is only an appearance
to our expression of mind. Only one reality or world view in a well balanced manner.
Knowledge can be seen as mental or spiritual in nature. Knowledge can be obtained
through pure uniform spiritual consciousness. All views come together in a state of
harmony. Values and morals are representation of the truth, not the truth itself"})
training_data.append({"class":"Interpretivism_Voluntarism ", "sentence":"A person's
will dictates their reality or views The truth is personal to an individual based on their
own will Evidence is not required to knowing the truth, only beliefs Values determine
personal acts and goodness results due to acts of goodwill"})

#print ("%s sentences of training data" % len(training_data))
#print(training_data)

# In[16]:

# capture unique stemmed words in the training corpus
corpus_words = {}
class_words = {}

```

```

#####LEMMATIZER WordNet Link
classes = list(set([a['class'] for a in training_data]))
for c in classes:
    # prepare a list of words within each class
    class_words[c] = []

# loop through each sentence in our training data
for data in training_data:
    # tokenize each sentence into words
    for word in nltk.word_tokenize(data['sentence']):
        # discard special characters
        if word not in ["?", "s"]:
            # stem and lowercase each word
            lemmatized_word = lemmatizer.lemmatize(word, pos="a") #.lower()

            # Add new lemmatized word to corpus
            if lemmatized_word not in corpus_words:
                corpus_words[lemmatized_word] = 1
            else:
                corpus_words[lemmatized_word] += 1

            # add the word to words in class list
            class_words[data['class']].extend([lemmatized_word])
#####END LEMMATIZER

```

```

# calculate a score for a given class
def calculate_class_score(sentence, class_name, show_details=True):
    score = 0
    # tokenize each word in our new sentence
    for word in nltk.word_tokenize(sentence):

```

```

# check to see if the stem of the word is in any of our classes
if stemmer.stem(word.lower()) in class_words[class_name]:
    # treat each word with same weight
    score += 1

    if show_details:
        print (" match: %s" % stemmer.stem(word.lower() ))
return score

## Find the class with the highest score
#for c in class_words.keys():
# print ("Class: %s Score: %s \n" % (c, calculate_class_score(sentence, c)))

# calculate a score for a given class taking into account word commonality
def calculate_class_score_commonality(sentence, class_name, show_details=True):
    score = 0

    # tokenize each word in our new sentence
    for word in nltk.word_tokenize(sentence):
        # check to see if the stem of the word is in any of our classes
        if stemmer.stem(word.lower()) in class_words[class_name]:
            # treat each word with relative weight
            score += (1 / corpus_words[stemmer.stem(word.lower())])

            if show_details:
                print (" match: %s (%s)" % (stemmer.stem(word.lower()), 1 /
                corpus_words[stemmer.stem(word.lower())]))

    return score

```

```

# return the class with highest score for sentence: topmost three classes
def classify(sentence, *args, **kwargs):

    #dirname = Path.cwd().joinpath("algorithms").joinpath("media").joinpath(str(args[0]))
    token = nltk.word_tokenize(sentence)
    # print('TOKENIZED', token)
    lwrds = ( [lemmatizer.lemmatize(w, get_wordnet_pos(w)) for w in
nltk.word_tokenize(sentence)])
    # print('wordwrds')
    dirname = Path.cwd() / str(args[0])

    if not Path(dirname).is_dir():
        dirname.mkdir()

    high_class = None
    high_score = 0
    mid_class = None
    mid_score = 0

    low_class = None
    low_score = 0

    other_class = None
    other_score = 0
    other1_class = None
    other1_score = 0
    prdgm6_class = None
    prdgm6_score = 0
    prdgm7_class = None
    prdgm7_score = 0

    # print(sentence)

```

```

# loop through our classes
for c in class_words.keys():
    # calculate score of sentence for each class
    score = calculate_class_score_commonality(sentence, c, show_details=False)
    chart_data.append([c.split("_")[1],score])
    # keep track of highest score
    if score > high_score:
        high_class = c
        high_score = score
#     print(high_class)
    if score < high_score and score > mid_score:
        mid_class = c
        mid_score = score
#     print(mid_class)
    if score < mid_score and score > low_score:
        low_class = c
        low_score = score
#     print(low_class)

#####
    if score < low_score and score > other_score:
        other_class = c
        other_score = score
#     print(other_class)
    if score < other_score and score > other1_score:
        other1_class = c
        other1_score = score
#     print(other_class)

    if score < other1_score and score > prdgm6_score:
        prdgm6_class = c
        prdgm6_score = score

```

```
if score < prdgm6_score and score > prdgm7_score:
    prdgm7_class = c
    prdgm7_score = score
#####
```

```
if high_class is not None:
    if high_class.find("_") == -1:
        cluster1 = high_class
        paradigm1 = high_class
    else:
        word = high_class
        cluster1 = word.split("_")[0]
        paradigm1 = word.split("_")[1]
```

```
if mid_class is not None:
    if mid_class.find("_") == -1:
        cluster2 = mid_class
        paradigm2 = mid_class
    else:
        wm = mid_class
        paradigm2 = wm.split("_")[1]
        cluster2 = wm.split("_")[0]
```

```
if low_class is not None:
    if low_class.find("_") == -1:
        cluster3 = low_class
        paradigm3 = low_class
    else:
        wl = low_class
        paradigm3 = wl.split("_")[1]
        cluster3 = wl.split("_")[0]
```



```
if other_class is not None:
    if other_class.find("_") == -1:
        cluster4 = other_class
        paradigm4 = other_class
    else:
        wl = other_class
        paradigm4 = wl.split("_")[1]
        cluster4 = wl.split("_")[0]
```

```
#
```

```
if other1_class is not None:
    if other1_class.find("_") == -1:
        cluster5 = other1_class
        paradigm5 = other1_class
    else:
        wl = other1_class
        paradigm5 = wl.split("_")[1]
        cluster5 = wl.split("_")[0]
```

```
if prdgm6_class is not None:
    if prdgm6_class.find("_") == -1:
        cluster6 = other1_class
        paradigm6 = other1_class
    else:
        wl = prdgm6_class
        paradigm6 = wl.split("_")[1]
        cluster6 = wl.split("_")[0]
```

```
if prdgm7_class is not None:
    if prdgm7_class.find("_") == -1:
        cluster6 = other1_class
        paradigm6 = other1_class
```

```

else:
    wl = prdgm7_class
    paradigm7 = wl.split("_")[1]
    cluster7 = wl.split("_")[0]

```

Plot the graph with three top RMIs:

```
#####BAR CHART
```

```

print(chart_data)
objects = (paradigm1,paradigm2,paradigm3)    #, paradigm4)
y_pos = np.arange(len(objects))
performance = [high_score, mid_score,low_score]    #, other_score]

plt.bar(y_pos, performance, align='center', alpha=0.5, color=['purple', 'blue',
'cyan'])    #, 'yellow'])
plt.xticks(y_pos, objects, rotation='vertical')
plt.subplots_adjust(bottom=0.5)
plt.margins(0.3)
plt.ylabel('% Alignment')
plt.title('Recommended RPPs')

```

```
#####
```

```
#####PIE CHART
```

```

#     height = [high_score, mid_score,low_score, other_score, other1_score,
prdgm6_score, prdgm7_score]
#     bars = (paradigm1,paradigm2,paradigm3, paradigm4, paradigm5,paradigm6,
paradigm7)    #(word2,wm1, wl1)
#     y_pos = np.arange(len(bars))

```

```

# explode = (0.1,0,0,0,0,0,0)
# labels = str( paradigm1 ), str(paradigm2),str( paradigm3),str( paradigm4), str(
paradigm5), str( paradigm6), str( paradigm7) #str( word2 ), str(wm1),str( wl1)

# str(high_class), str(mid_class),str(low_class)
# Create bars

# plt.bar(y_pos, height, color=['purple', 'blue', 'cyan'])

# colors=['orange', 'cyan', 'red','blue','pink']
# plt.pie(height, explode=explode, colors=colors,labels=labels, autopct='%1.1f%%',
shadow=True,
#         startangle=90)

#####END CHARTS
# Save graphic
# plt.show()
plt.savefig(str(dirname) + '/pie.PNG')
plt.clf()

print( high_class, high_score)
print( mid_class, mid_score)
print( low_class, low_score )
print(other_class ,other_score)
#         print(other_class ,other_score, high_class, high_score, '\n', mid_class,
mid_score, '\n', low_class, low_score )
#return results
return {"cluster1": cluster1, #high_class,

```

```

        "cluster2": cluster2, #mid_class,

        "cluster3": cluster3, #low_class,

#         "FOR PIE CHART
        "paradigm1":paradigm1, #LABEL
        "high_score": high_score, #FIGURE
        "paradigm2": paradigm2,
        "mid_score": mid_score,
        "paradigm3": paradigm3,
        "low_score": low_score,
#         FOR PIE CHART
        "token": token,
        "lemmatized":lwrđ,
#####
        "other1": other_class,
        "other2":other_score,
        "paradigm4":paradigm4,
        "cluster4": cluster4,
        "paradigm5":paradigm5,
        "cluster5": cluster5,
        "paradigm6":paradigm6,
        "cluster6": cluster6,
        "paradigm7":paradigm7,
        "cluster7": cluster7,

#####
    }

def get_wordnet_pos(word):
    """Map POS tag to first character lemmatize() accepts"""

```

```

tag = nltk.pos_tag([word])[0][1][0].upper()
tag_dict = {"J": wordnet.ADJ,
            "N": wordnet.NOUN,
            "V": wordnet.VERB,
            "R": wordnet.ADV}

return tag_dict.get(tag, wordnet.NOUN)

```

RESULTS.HTML

```
{%extends "nlp/header.html" %}
```

```
{% block content %}
```

```
<p class="courier">
```

```
<p class="courier"><h1> Research philosophy / paradigm report for {{
user.first_name }} {{ user.last_name }}</h1></p>
```

```
<p class="courier"></p>
```

```
<section>
```

```
<div class="courier">
```

```
<h3>Research philosophies and paradigms</h3>
```

```
<p class="courier"> This report displays recommended philosophy / paradigm that
can be employed in your future research and knowledge creation. Research
philosophy can be thought of as underlying and guiding principles or roadmaps that
a research is based upon. Philosophy is a multi-dimensional concept that is linked to
personal ideas about the world, entities, how they interact and exchange knowledge
with each other. </p>
```

```
<p class="courier">Research philosophies are ideologies or stance that a researcher
takes during research undertaking. These guide the research in choosing a strategy,
roadmap, research sources and methods of obtaining the required knowledge from
the sources. The report can assist you with recommending a philosophy or more that
that is closer to your ideologies based on answers, which you provided during the
questionnaire.
```

```
</p><p class="courier">testing fonts</p>
```

```
</div>
```

<h3> The report is based on a consolidated framework of the work of Denzin and Lincoln (2005) and Saunders et al. (2015) and derived using Natural Language Processing (NLP) and will show recommended research philosophies and their respective paradigms, with the following components:</h3>

<div style="overflow-x:auto;">

<p class="courier"> </p>

<table class="table">

<tr><td> Ontology – ideas about what exists and can be known in the world and even whether it is important to know about this existence.</td> </tr>

<tr><td> Epistemology – the feasibility and extent to which knowledge can be acquired, for example can we know anything for certain. And if it possible to obtain knowledge what means of acquisition can be used and how can we justify this knowledge. </td></tr>

<tr><td> Axiology – the influence that a researcher’s personal values may have on the outcomes of research. This also includes ethical behaviour of researchers during knowledge creation.</td></tr>

</t>

</div>

</section>

<p class="courier"></p>

<div style="overflow-x:auto;">

<table class="table">{% for item in main_paradigm %}

<tr class="even">

{{item.rpp}}

<ul class="b">[More Info]

Click here for {{item.rpp}} Info 1

```
        <li><a
href="../../../static/rmi_documents/output/Research_Methods_for_Business_Students_
C.pdf" target="_blank">Click here for {{item.rpp}} Info 2</a></li>
```

```
        <li><a
href="../../../static/rmi_documents/output/Vosloo_JJ_Chapter_5.pdf"
target="_blank">Click here for {{item.rpp}} Info 3</a></li>
```

```
    </ul>
```

```
</ul>
```

```
</tr>
```

```
<tr class="odd">
```

```
<td>Description</td><td>{{item.desc}}</td>
```

```
</tr>
```

```
<tr class="even">
```

```
<td>Ontology</td><td>{{item.ontology}}</td>
```

```
</tr>
```

```
<tr class="odd">
```

```
<td>Epistemology</td><td>{{item.epistemology}}</td>
```

```
</tr>
```

```
<tr class="even">
```

```
<td>Axiology</td><td>{{item.axiology}}</td>
```

```
</tr>
```

```
{% endfor %}
```

```
</table>
```

```
</div>
```

```
<!--<p><h4> Predicted Paradigm & Philosophy cluster for {{ user }} is <u>{{ cluster
}}</u></h4></p>
```

```
<div style="overflow-x:auto;">
```

```
<!--<table>{% for item in main_paradigm %}
```

```
<tr class="even">
```

```
<td>Paradigm</td><td>{{item.rpp}}</td>
```

```
</tr>
```

```

<tr class="odd">
<td>Description</td><td>{{item.desc}}</td>
</tr>
<tr class="even">
<td>Ontology</td><td>{{item.ontology}}</td>
</tr>
<tr class="odd">
<td>Epistemology</td><td>{{item.epistemology}}</td>
</tr>
<tr class="even">
<td>Axiology</td><td>{{item.axiology}}</td>
</tr>
{% endfor %}
</table-->
</div>

```

<p><h2>Natural Language Processing </h2> </p>

```

<table >
<td class="courier"> <b>Natural Language Processing (NLP)</b> - This report uses
Natural Language Processing algorithm on user input captured in the questionnaire,
for classification into research philosophies and paradigms. This classification is
based on the created Research Paradigm and Philosophies (RPP) categories
corpus. The RPP corpus gets tokenized, stemmed, lemmatized and then used to
train the classification algorithm.

```

```

The Bag of Words (BoW) model is used to calculate a score for each given RPPs
category in the corpus. The same model is also used to calculate a score for and
classsify the user input. A comparison of the user input's score against the corpus
category scores yields the three topmost RPPs that are closely linked to a
researcher's worldview. The results are presented in a pie chart showing the degree
to which a researcher is aligned to a particular RPP as below:

```

```

</td>
</table>

```

<p><h3>Tokenizing</h3></p>

```

<table >

```


<td class="courier">Tokenizing refers to the process where the input string is broken down into individual words, phrases or even sentences, referred to as tokens, separated by a whitespace. Special characters, especially punctuation marks, and other symbols are ignored in this process. The tokens are used as input for the Stemming and Lemmatization processes of NLP. The tokens are checked for the number of occurrences within the corpus and then score of each word noted. The score are used to tally up the vectors for each class or category of the corpus, to be used later when comparing values between input data and corpus, which assists in classification.

The section that follows shows the tokenized user response; </td>

</table>

<p></p>

<p></p>

<div style="overflow-x:auto;">

<table border = "5" bordercolor="#b9c7cb">

<tr class="even">

<th>Tokenized Response</th>

</tr>

<tr class="odd">

<td>{{nlp_predict.token}}{{nlp_predict.tokenized}}</td>

</tr>

</table>

<p> </p>

<p> </p>

<p> </p>

<h3>Lemmatization</h3>

<table >

<td class="courier">Lemmatization refers to the processes of changing a word back to it's base form in relation to the context in which the word appears. Inflected forms of a word are grouped together and treated as a single item for analysis purposes. The WordNet lexical database is used to lemmatize the tokens using the WordNetLemmatizer algorithm. These lemmatized words, or tokens, are used to compare with words, or tokens, in the RPP corpus and each word found gets scored for that particular category. At the end of the process the scores are tallied, with the highest score representing the topmost RPP that is recommended for a user.

The section that follows shows the lemmatized user response;</td></table>

<p></p>

<div style="overflow-x:auto;">

<table border = "5" bordercolor="#b9c7cb">

<tr class="even">

<th>Lemmatized Response</th>

</tr>

<tr class="odd">

<td> {{nlp_predict.lemmatized}}</td>

</tr>

</table>

<p><h3>Named Entity Recognition(NER)</h3></p>

<table >

<td class="courier">Named Entity Recognition refers to the process in information extraction that seeks to locate and classify named entities in text into pre-defined categories.

NER is used in Natural Language Processing (NLP) of a user's input to identify the components that make up the research philosophies and paradigms. The section that follows shows the identified entities in a user's response together with the respective labels; </td>

</table>

<p></p>

<p></p>

<div style="overflow-x:auto;">

<table border = "5" bordercolor="#b9c7cb">

<tr class="even">

<th>Named Entities</th>

</tr>

<tr class="odd">

<td>{{entities}}</td>

</tr>

</table>

<h3>NLP Text classification</h3>

<table >

<td class="courier">Natural Language Processing(NLP) text classification in this context aims to automatically classify user input into one or more of the research paradigms and philosophies(RPPs) a user is most aligned to. The table that follows shows the topmost three(3) RPPs that closely resemble the user's research philosophy and paradigm. </td></table>

<p></p>

<div style="overflow-x:auto;">

<p class="courier"> </p>

<table class="table" border = "5" bordercolor="#b9c7cb"> <tr>

<th>Research Philosophy</th>

<th>Description</th>

<th><a href="" atl="" title="Ontology is the study of being. It focuses on several related questions:

What things exist? (stars yes, unicorns no, numbers . . . yes?)

What categories do they belong to? (are numbers physical properties or just ideas?)">Ontology</th>

<th><a href="" atl="" title="Epistemology is the study of knowledge and justified belief

concerned with the following questions: What are the necessary and sufficient conditions of knowledge? What are its sources? What is its

structure, and what are its limits? As the study of justified belief, epistemology aims to answer questions such as: How we are to understand

the concept of justification? What makes justified beliefs justified? Is justification internal or external to one's own mind?">Epistemology</th>

<th><a href="" atl="" title="Axiology is a branch of philosophy that studies judgements about the value. Specifically, axiology is engaged with assessment of the role of researcher's own value

on all stages of the research process.">Axiology</th>

<th>Paradigm</th>

</tr>

{% for item in other_paradigms %}

<tr class="{% cycle 'even' 'odd' %}">

<td>{{item.rpp}}</td>

<td>{{item.desc}}</td>

<td>{{item.ontology}}</td>

<td>{{item.epistemology}}</td>

<td>{{item.axiology}}</td>

<td>{{item.cluster}}</td>

</tr>

{% endfor %}

</table>

<p> <h4>This table shows the three topmost recommended philosophies and the underlying paradigms based on the user {{User}}'s response </h4></p>

</div>

<p class="courier"> </p>

<!--

<p> {{nlp_predict.cluster1}}</p>

>

<p>{{nlp_predict.high_score}}</p>
<p>{{nlp_predict.mid_class}}</p>
<p>{{nlp_predict.mid_score}}</p>
<p>{{nlp_predict.low_class}}</p>
<p>{{nlp_predict.low_score}}</p>
<p>{{nlp_predict.nlp_cluster}}</p>
<p>{{nlp_predict.paradigm_one}}</p>
<p>{{nlp_predict.paradigm_two}}</p>
<p>{{nlp_predict.paradigm_three}}</p>
<p>{{nlp_predict.paradigm_four}}{{User}}</p>

<div style="overflow-x:auto;">

<p> </p>

</div>

<!--

{% for item in nlp_predict %}

<p> {{ item }}</p>

{% endfor %} -->

<p> </p>

<p> </p>

<p> </p>

<p> </p>

<p><H4> The graph illustrates the order of predicted Philosophies </H4></p>

</p>

{% for item in userAnswers %}

 {{ item.answer_name }}


```
{% endfor %}  
{% endblock %}
```

RULEBASED_NER.PY

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Tue Oct 1 22:56:15 2019
```

```
@author: Ntombi
```

```
"""
```

```
import spacy
```

```
from spacy.lang.en import English
```

```
from spacy.pipeline import EntityRuler
```

```
from spacy import displacy
```

```
def named_entity(sentence, *args, **kwargs):
```

```
    nlp = spacy.blank('en')      #spacy.load("en_core_web_sm")
```

```
    ruler = EntityRuler(nlp)
```

```
    nlp = English()
```

```
    ruler = EntityRuler(nlp)
```

```
# Create labels and patterns(specific words) for RPP components
```

```
    patterns = [{"label": "Ontology", "pattern": "individual"},
```

```
                {"label": "Axiology", "pattern": "social"},
```

```
                {"label": "Epistemology", "pattern": "language"},
```

```
                {"label": "Ontology", "pattern": "contradictory"}]
```

```
    ruler.add_patterns(patterns)
```

```
    nlp.add_pipe(ruler)
```

```

entities =[]

#ruler1 = EntityRuler(nlp)
    ruler.from_disk("C:/Users/Ntombhi/Anaconda3/lib/site-
packages/en_rmi1_patterns.jsonl") # loads patterns only
    ruler.from_disk("C:/Users/Ntombhi/Anaconda3/lib/site-
packages/en_rmi1_entity_ruler")

    doc = nlp(sentence)
#
# print([(ent.text, ent.label_) for ent in doc.ents])
# displacy.serve(doc, style="ent") # print entities using Spacy's entity visualizer

#Save entities and lables for the results template
for ent in doc.ents:
    text = ent.text
    label = ent.label_
    entities.append(text+' | '+label)
print(entities)

return(entities)

```

VIEWS.PY

```

from django.shortcuts import render
from django.http import HttpResponseRedirect
from django.shortcuts import get_object_or_404, render, redirect,
render_to_response
from django.template import RequestContext
from django.urls import reverse
from django.views.generic import TemplateView, FormView, ListView
#from django.forms.formsets import formset_factory
from . models import user_answer,paradigm_question, cluster

```

```

from django import forms
from . RMITextClassifier import classify
#from . MRI_NERecognizer import named_entities
from . RuleBased_NER import named_entity
# Create your views here.
from nlp.forms import SignUpForm, nlp_form #,FeatureFormSet #worldViewForm
from django.contrib.auth.decorators import login_required
from django.contrib.auth.mixins import LoginRequiredMixin
from django.utils.decorators import method_decorator
from django.contrib.auth.models import User
from django.contrib.auth import login, authenticate

#register view
def register(request):
    if request.method == 'POST':
        form = SignUpForm(request.POST)
        if form.is_valid():
            form.save()
            username = form.cleaned_data.get('username')
            raw_password = form.cleaned_data.get('password1')
            user = authenticate(username=username, password=raw_password)
            login(request, user)
            return redirect('..')
    else:
        form = SignUpForm()
    return render(request, 'nlp/register.html', {'form': form})

#Home page view
@method_decorator(login_required, name='dispatch')
class IndexView(FormView):
    template_name = 'nlp/index.html'

```



```

def get (self,request):
    return render(request, self.template_name, {'User': request.user})

#Cluster report view
@login_required
def clusterView(request):

    template_name = 'nlp/results.html'
    if request.method == ('GET') or request.method == ('POST'):

        try:
            ##NLP text classification
            dirname = 'nlp/static/media/' + str(request.user)
            nlp_query_data = []
            nlpUserAnswers = user_answer.objects.filter(username=request.user).values_list('q1','q2','q3','q4','q5','q6','q7','q8','q9','q10')

            if nlpUserAnswers.count() > 0:

                for item in nlpUserAnswers:
                    nlp_query_data.append(item)
                    sentence = nlp_query_data[0]

                nlp_data = ' '.join(sentence)

                nlp_predict = classify(nlp_data, dirname)          #RPP classification
                # nlp_entities = named_entities(nlp_data, dirname)
                nlp_entities1 = named_entity(nlp_data, dirname)  #named entities

            if nlp_predict["paradigm1"] is not None:

```

```

print(nlp_predict["paradigm1"])

rpp_one =cluster.objects.filter(rpp=nlp_predict["paradigm1"])
rpp_two =cluster.objects.filter(rpp=nlp_predict["paradigm2"])
rpp_three =cluster.objects.filter(rpp=nlp_predict["paradigm3"])
other_paradigms = rpp_one.union(rpp_two,rpp_three)
else:
#         other_paradigms = cluster.objects.filter(cluster=user_cluster)[:2]
#     else:
        return render(request, 'nlp/')
#####
#
return render(request, template_name,
               { 'other_paradigms':other_paradigms,
                 'User': request.user,
                 'entities':nlp_entities1,
                 #'cluster': user_cluster,
#                 'nlp_cluster': nlp_cluster,
                 'nlp_predict':nlp_predict})
# 'userAnswers': userAnswers

except:
    return render(request, 'nlp/')

#nlp view
@login_required
def nlpView(request):
    template_name = 'nlp/nlp_questions.html'

```

```

if request.method == 'POST':
    #getting values from post
    form = nlp_form(request.POST)
    completed = user_answer.objects.filter(username=request.user).count()

    if form.is_valid:
        if completed > 0:
            user_answer.objects.filter(username=request.user).delete()

        form.save(commit=False)
        form.instance.q1 = form.cleaned_data.get('q1')
        form.instance.q2 = form.cleaned_data.get('q2')
        form.instance.q3 = form.cleaned_data.get('q3')
        form.instance.q4 = form.cleaned_data.get('q4')
        form.instance.q5 = form.cleaned_data.get('q5')
        form.instance.q6 = form.cleaned_data.get('q6')
        form.instance.q7 = form.cleaned_data.get('q7')
        form.instance.q8 = form.cleaned_data.get('q8')
        form.instance.q9 = form.cleaned_data.get('q9')
        form.instance.q10 = form.cleaned_data.get('q10')
        form.instance.username = request.user
        print(form.instance.q2)
        form.save(commit=True)

    return render(request,'nlp/report.html')

else:
    return render(request, template_name, {'User': request.user})

def reportView(request):
    template_name='nlp/report.html'

```

```
return render(request, template_name, {'User': request.user})
```

```
def tokenizeView(request):
```

```
    template_name='nlp/tokenize.html'
```

```
    return render(request, template_name, {'User': request.user})
```

```

def classify(sentence, *args, **kwargs):

    #dirname = Path.cwd().joinpath("algorithms").joinpath("media").joinpath(str(args[0]))
    token = nltk.word_tokenize(sentence)
    # print('TOKENIZED', token)
    lwrds = ([lemmatizer.lemmatize(w, get_wordnet_pos(w)) for w in nltk.word_tokenize(sentence)])
    # print('wordwrds')
    dirname = Path.cwd() / str(args[0])

    if not Path(dirname).is_dir():
        dirname.mkdir()

    high_class = None
    high_score = 0
    mid_class = None
    mid_score = 0

    low_class = None
    low_score = 0

    other_class = None
    other_score = 0
    other1_class = None
    other1_score = 0
    prdgm6_class = None
    prdgm6_score = 0
    prdgm7_class = None
    prdgm7_score = 0

    # print(sentence)

```

```

# loop through the RPP classes
for c in class_words.keys():
    # calculate score of sentence for each class
    score = calculate_class_score_commonality(sentence, c, show_details=False)
    chart_data.append([c.split("_")[1],score])
    # keep track of highest score
    if score > high_score:
        high_class = c
        high_score = score
    if score < high_score and score > mid_score:
        mid_class = c
        mid_score = score
    if score < mid_score and score > low_score:
        low_class = c
        low_score = score
    if score < low_score and score > other_score:
        other_class = c
        other_score = score
    if score < other_score and score > other1_score:
        other1_class = c
        other1_score = score

    if score < other1_score and score > prdgm6_score:
        prdgm6_class = c
        prdgm6_score = score

    if score < prdgm6_score and score > prdgm7_score:
        prdgm7_class = c
        prdgm7_score = score

```

```
if high_class is not None:
    if high_class.find("_") == -1:
        cluster1 = high_class
        paradigm1 = high_class
    else:
        word = high_class
        cluster1 = word.split("_")[0]
        paradigm1 = word.split("_")[1]

if mid_class is not None:
    if mid_class.find("_") == -1:
        cluster2 = mid_class
        paradigm2 = mid_class
    else:
        wm = mid_class
        paradigm2 = wm.split("_")[1]
        cluster2 = wm.split("_")[0]

if low_class is not None:
    if low_class.find("_") == -1:
        cluster3 = low_class
        paradigm3 = low_class
    else:
        wl = low_class
        paradigm3 = wl.split("_")[1]
        cluster3 = wl.split("_")[0]

if other_class is not None:
    if other_class.find("_") == -1:
```

```
cluster4 = other_class
paradigm4 = other_class
else:
    wl = other_class
    paradigm4 = wl.split("_")[1]
    cluster4 = wl.split("_")[0]
#
if other1_class is not None:
    if other1_class.find("_") == -1:
        cluster5 = other1_class
        paradigm5 = other1_class
    else:
        wl = other1_class
        paradigm5 = wl.split("_")[1]
        cluster5 = wl.split("_")[0]

if prdgm6_class is not None:
    if prdgm6_class.find("_") == -1:
        cluster6 = other1_class
        paradigm6 = other1_class
    else:
        wl = prdgm6_class
        paradigm6 = wl.split("_")[1]
        cluster6 = wl.split("_")[0]

if prdgm7_class is not None:
    if prdgm7_class.find("_") == -1:
        cluster6 = other1_class
        paradigm6 = other1_class
```



```

wl = prdgm7_class
paradigm7 = wl.split("_")[1]
cluster7 = wl.split("_")[0]

# Plot the graph with three top RMIs:

print(chart_data)
objects = (paradigm1,paradigm2,paradigm3) #, paradigm4)
y_pos = np.arange(len(objects))
performance = [high_score, mid_score,low_score] #, other_score]

plt.bar(y_pos, performance, align='center', alpha=0.5, color=['purple', 'blue', 'cyan'])
plt.xticks(y_pos, objects, rotation='vertical')
plt.subplots_adjust(bottom=0.5)
plt.margins(0.3)
plt.ylabel('% Alignment')
plt.title('Recommended RPPs')

```

Figure 8.19 Python script for the classifying model

```

from django.db import models

from django.contrib.auth.models import User

# Create the models here.

class user_answer(models.Model):

    username = models.CharField(max_length=50, null=True)

    created_on = models.DateTimeField(auto_now_add=True)

    updated_on = models.DateTimeField(auto_now=True)

    q1 = models.TextField(max_length=5000, null=True)

    q2 = models.TextField(max_length=5000, null=True)

    q3 = models.TextField(max_length=5000, null=True)

    q4 = models.TextField(max_length=5000, null=True)

    q5 = models.TextField(max_length=5000, null=True)

    q6 = models.TextField(max_length=5000, null=True)

    q7 = models.TextField(max_length=5000, null=True)

    q8 = models.TextField(max_length=5000, null=True)

    q9 = models.TextField(max_length=5000, null=True)

    q10 = models.TextField(max_length=5000, null=True)

class cluster(models.Model):

    rpp = models.CharField(max_length=100,null=True)

    desc = models.TextField(max_length=5000,null=True)

    ontology = models.CharField(max_length=5000, null=True)

    epistemology = models.CharField(max_length=5000, null=True)

    axiology = models.CharField(max_length=5000, null=True)

    cluster = models.CharField(max_length=100, null=True)

```

```

class paradigm(models.Model):

    paradigm_name = models.CharField(max_length=200)

    paradigm_description = models.CharField(max_length=1000)

    def __str__(self):

        return self.paradigm_name

class paradigm_question(models.Model):

    question = models.TextField(max_length=1000)

# paradigm_component = models.ForeignKey(paradigm_component,on_delete=models.CASCADE,
default='1')

    def __str__(self):

        return self.question

class paradigm_answer(models.Model):

    answer = models.CharField(max_length=1000)

    question = models.ForeignKey(paradigm_question,on_delete=models.CASCADE)

    score = models.IntegerField(null = True)

    def __str__(self):

        return self.answer

```

Figure 8.20 Script models.py to create the Natural Language Processing (NLP) database

Paradigm/Philosophy	Characteristics		Meaning
Accidentalism (philosophy)	(Some or all) things have only accidental properties, no essential properties, or no common nature. All properties are accidental.	Things have only accidental properties, no essential properties, or no common nature. All events and properties are accidental.	Things have only accidental properties, no essential properties, or no common nature.
Atomism	All Matter is composed of atoms, element of everything composite, principle and seed of everything in existence	All matter is composed of atoms Elements are composite or seeds of everything in existence	All matter is composed of atoms and elements are composite or seeds of everything in existence
Dualism	Mental events and properties are not physical. Relation between mental events	Mental events do not have physical existence Relation between mental events	Mental events do not have physical existence
Disjunctivism	Multiple and subjective reality that's isolated from the world around it.	Subjective multiple reality that is influence by one's thinking processes	Multiple reality that's isolated from the world around
Positivism	Objective reality.	One version of the truth, real existing substance, certainty and	One version of the truth. Universal, independent tru

Figure 8.21 Spreadsheet for the RPP corpus or dataset

```

32ss:"Realism Absolute idealism","sentence":"Truth exists independent of us, whether we know it or not. Physical world is only an appearance to our expression of mind. Only one reality o
33ss:"Scepticism Absurdism","sentence":"The existence or non-existence of anything is meaningless A world of contradictions where nothing is significant Individuals and their existence a
34ss:"Accidentalism","sentence":"Things have only accidental properties, no essential properties or no common nature. All events and properties are accidental. The occurrence of some eve
35ss:"Realism Action theory ","sentence":"Truth is a composition of functional structures, objects and properties Only the whole or composition of a system matters")
36ss:"Realism Action theory ","sentence":"Reality is everywhere and in all things Reality is everywhere and in all things Participants are essential in understanding the whole Participan
37ss:"Realism Actualism","sentence":"Only one actual reality exists Everything is actual and actually exists. Reality is constructed out of valuations. Collective and complimentary prope
38ss:"Scepticism Agnostic atheism","sentence":"There is not sufficient evidence to prove the existence of a god Logic is not sufficient to overcome the unknowability of the existence of
39ss:"Interpretivism Agnostic theism","sentence":"One believes in god Not certain about the characteristics of god God created the universe but is detached from how the universe works C
40ss:"Scepticism Agnosticism","sentence":"Humans cannot know of the existence of anything beyond the phenomena of their experience Cannot claim with conviction to know about the existenc
41ss:"Realism Anti Foundationalism","sentence":"Reality is not permanent Reality is based on experience The truth is dependent on an individual 's experience Reality is influenced by o
42ss:"Interpretivism Anti-realism","sentence":"The truth is dependent on human theoretical activities More than one truth exist Existence of objects and structures depends on individual
43ss:"Interpretivism Anti-Realism","sentence":"There is no objective reality Statements can be true or false independent of one's knowledge There is enough evidence to support the truth
44ss:"Positivism Atomism","sentence":"All matter is composed of atoms and elements are composite or seeds of everything in existence Atoms are indivisible, impenetrable, continuous, and
45ss:"Interpretivism Biocentric universe","sentence":"Nothing could exist ")
46ss:"Realism Realism","sentence":" sometimes called naturalism, in the arts is generally the attempt to represent subject matter truthfully, without artificiality and avoiding artistic
47ss:"Realism Realism","sentence":"In the visual arts, illusionistic realism is the accurate depiction of lifefoms, perspective, and the details of light and colour. But realist or natu
48ss:"Interpretivism Biocentric universe","sentence":"without consciousness. Energy can't be created or destroyed, only change forms The content of the consciousness is an ultimate real
49ss:"Scepticisism B-theory of time","sentence":"Tensed sentences do not exist and tense is not a fundamental feature of the world. Nothing that realistically distinguishes the present fro
50ss:"Positivism Bundle theory","sentence":"Material objects are bundles of properties and concrete particulars are bundles of universals Property possession is understood in terms of pa
51ss:"Positivism Bundle theory","sentence":"constitute material objects Knowledge obtained through analytical theory and philosophical views")
52ss:"Positivism Cartesian Doubt","sentence":"Reality is not able to be relied on conclusively The reality of the external world of things cannot be demonstrate to be true Our perceptio
53ss:"Positivism Categories of the understanding","sentence":"Events come as a surprise to the observer and have a major impact Events can only be explained after they have happened Real
54ss:"Realism Coherence theory of the truth","sentence":"Every true statement, insofar as it is true, describes its subject in the totality of its relationship with all other things. Eve

```

Figure 8.22 Training sentences for the corpus

```

Corpus words and counts: {'truth': 135, 'exists': 32, 'independent': 27, 'of': 315, 'u': 4, ',': 139, 'whether': 1, 'we': 28, 'know': 17, 'it': 38, 'or': 101, 'not': 93, '.': 138, 'physical': 25, 'world': 45, 'is': 314, 'only': 67, 'an': 36, 'appearance': 3, 'to': 154, 'our': 29, 'expression': 1, 'mind': 32, 'one': 76, 'reality': 107, 'view': 17, 'in': 90, 'a': 161, 'well': 1, 'balanced': 1, 'manner': 2, 'knowledge': 115, 'can': 93, 'be': 98, 'seen': 2, 'mental': 17, 'spiritual': 5, 'nature': 21, 'obtained': 8, 'through': 80, 'pure': 1, 'uniform': 1, 'consciousness': 10, 'all': 38, 'come': 8, 'together': 5, 'state': 6, 'harmony': 1, 'value': 31, 'and': 241, 'moral': 14, 'are': 101, 'representation': 3, 'the': 341, 'itself': 4, 'existence': 22, 'non-existence':

```

Figure 8.23 Words in the RPP corpus and the scores of occurrence in the corpus

i. User registration (Researcher)

Username:

Password:

Not registered? [Register here.](#)

Copyright © 2018 - University of South Africa.

Figure 8.24 Login page of the RMI application

a) Click on 'register here'

Sign up

Username: Required. 150 characters or fewer. Letters, digits and @/./+/_ only.

First name: Optional.

Last name: Optional.

Email: Required. Inform a valid email address.

Password:

Your password can't be too similar to your other personal information.
Your password must contain at least 8 characters.
Your password can't be a commonly used password.
Your password can't be entirely numeric.

Password confirmation: Enter the same password as before, for verification.

Figure 8.25 Form for registering a user account

b) Fill in the registration form noting naming conventions and password rules; on completion click on sign-up

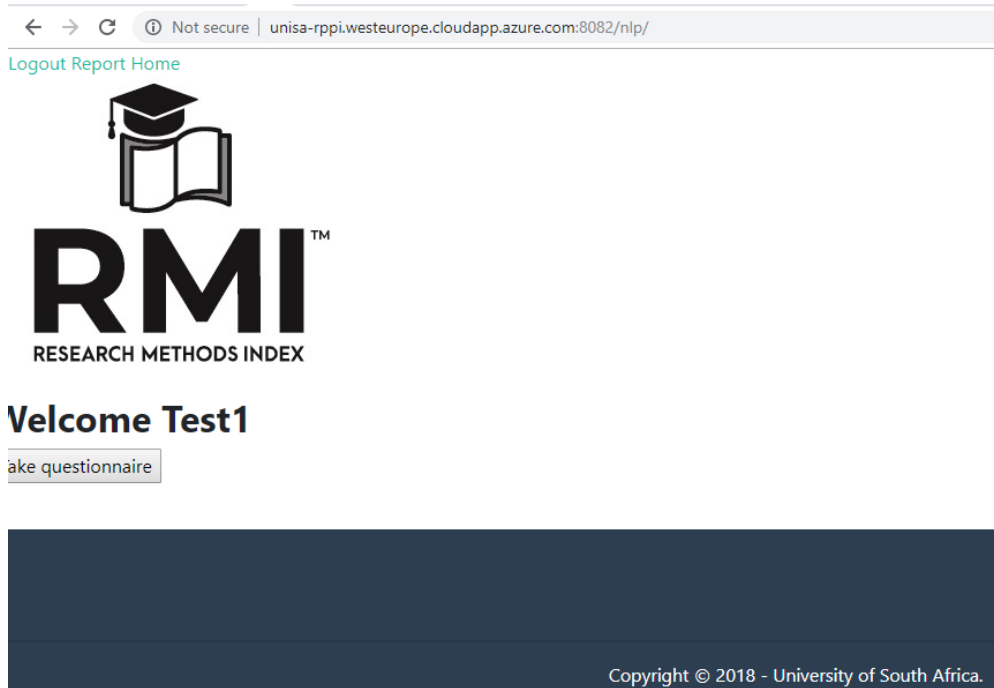


Figure 8.26 Landing page of the RMI NLP application

c) Click on Take Questionnaire and answer questions that follow

Please answer the following questions to the best of your understanding

Question 1: How many versions of the truth can there be in a given situation?	There is only one version of the truth
Question 2: How can the truth be influenced?	The truth is universal
Question 3: How can truth be justified?	The truth is independent of the individual
Question 4: How is knowledge acquired i.e., how do we know what we know?	Reality is based on knowledge or perception of a situation or fact
Question 5: What influences what we know?	Ascribes the motion and changes of the world to some external force
Question 6: How many sources of knowledge are there?	accessible to human reason and accessible in a non-sophisticated way
Question 7: How can knowledge be advanced?	Moral beliefs can be justified non-inferentially or inferentially
Question 8: What is the importance of values and ethics?	Independent of individual opinion and free from bias
Question 9: How can personal values influence the truth?	Knowledge about moral principles solely acquired through reasoning
Question 10: What determines our values and ethics?	Good and evil are objects of moral truths of an universal validity

Submit

Figure 8.27 NLP questionnaire

d) Submit questionnaire

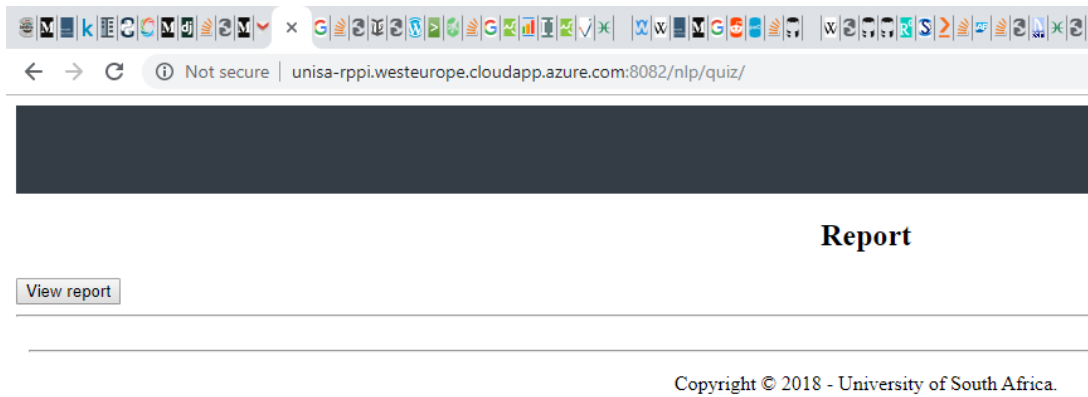


Figure 8.28 Button to view the NLP report

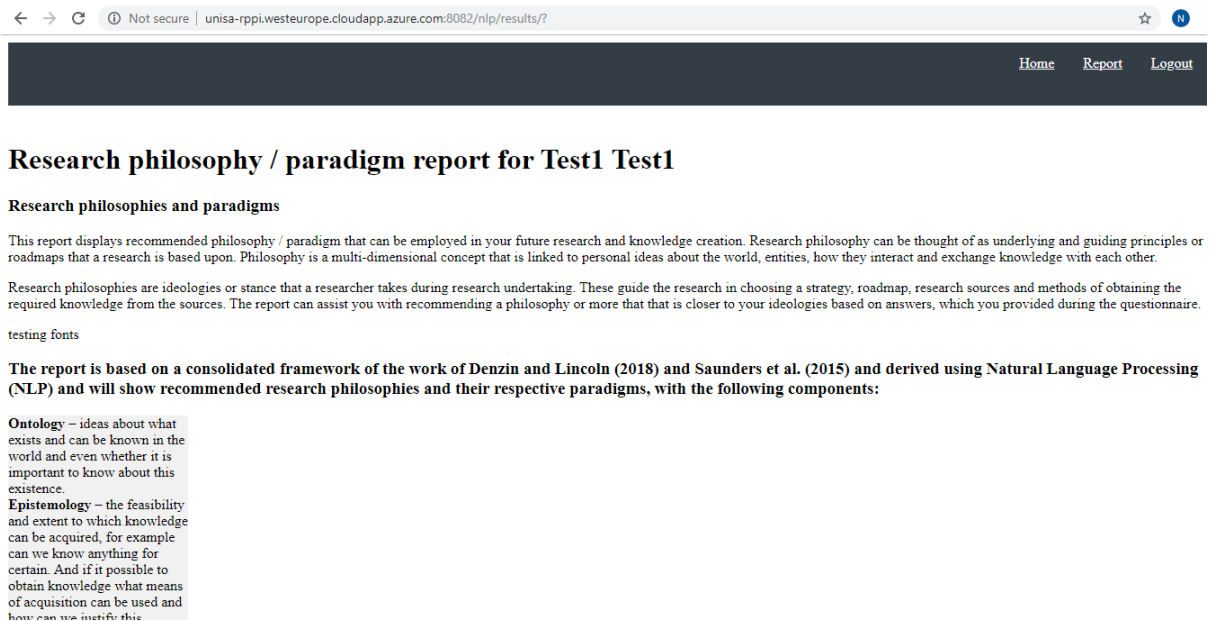


Figure 8.29 RMI NLP report

e) View report and Sign out

ii. Administrator

The administrator has access to sign-in onto the database, maintain tables (modify, insert and/or delete table entries) and view them.

APPENDIX L : ANNOTATION PROCESS

Tokenization

```
'train': 1, 'clos': 1, 'fulfil': 1, 'due': 1, 'goodwil': 1}

Class words: {'Scepticism_B-theory of time': ['tens', 'sent', 'do', 'not', 'ex', 'and', 'tens', 'is', 'not',
', 'feat', 'of', 'the', 'world', '.', 'noth', 'that', 'real', 'distinct', 'the', 'pres', 'from', 'the', 'pa
ut', '.', 'incomplet', 'sent', 'ar', 'transl', 'into', 'tenseless', 'sent', 'to', 'complet', 'them', 'witho
'of', 'mean', 'ev', 'in', 'tim', 'stand', 'in', 'rel', 'to', 'their', 'past', 'pres', 'fut'], 'Positivism_N
istemology': ['to', 'investig', 'knowledg', ',', 'inform', 'about', 'how', 'hum', 'acquir', 'knowledg', 'an
ontext', 'of', 'the', 'world', 'they', 'liv', 'in', 'is', 'imp', 'the', 'the', 'of', 'knowledg', 'should',
', 'as', 'continu', 'with', 'nat', 'sci', 'believ', 'ar', 'the', 'found', 'of', 'knowledg', 'the', 'conceiv
', 'larg', 'world', 'that', 'we', 'get', 'from', 'common', 'sens', 'and', 'sci', 'is', 'the', 'start', 'poi
nowledg', 'hum', 'being', 'develop', 'a', 'the', 'of', 'the', 'nat', 'world', 'on', 'the', 'bas', 'of', 'th
'input', 'sens', 'stim', 'lead', 'to', 'the', 'form', 'of', 'believ', 'about', 'the', 'world', 'found', 'b
'from', 'which', 'oth', 'believ', 'ar', 'infer', ',', 'just', 'on', 'believ', 'independ', 'of', 'individ',
'believ', 'and', 'influ', ',', 'fre', 'from', 'bia', '.'], 'Positivism_Eternalism (philosophy of time)': ['
past', 'object', 'ex', 'in', 'the', 'sam', 'way', 'pres', 'object', 'do', ',', 'no', 'diff', 'between', 't
', 'tru', 'independ', 'to', 'individ', 'etern', 'real', 'fact', 'obtain', 'through', 'funda', 'langu', 'and
', '.', 'stabl', 'rel', 'property', 'at', 'al', 'tim', '.', 'independ', 'observ', 'quant', 'by', 'tens', 'se
u', 'bas', 'valu', 'is', 'etern'], 'Interpretivism_Evolutionary epistemology': ['multipl', 'vert', 'of', 'r
'is', 'influ', 'by', 'the', 'extern', 'environ', 'real', 'is', 'influ', 'by', 'study', 'the', 'lif', 'form
iv', 'in', 'an', 'environ', 'the', 'process', 'of', 'acquir', 'knowledg', 'and', 'understand', 'is', 'throu
', 'knowledg', 'is', 'acquir', 'by', 'rel', 'a', 'subject', 'to', 'the', 'supposit', 'of', 'grad', 'develop
', 'knowledg', 'about', 'the', 'environ', 'is', 'gain', 'by', 'study', 'the', 'org', 'that', 'liv', 'in', '
al realism_Founderetism': ['ther', 'can', 'be', 'truth', 'that', 'ar', 'just', 'that', 'ar', 'not', 'suppo
ny', 'oth', 'truth', 'at', 'al', 'a', 'der', 'believ', 'ow', 'just', 'to', 'anoth', 'der', 'believ', 'or',
', ('', 'ii', ')', 'a', 'bas', 'believ', 'ow', 'just', 'to', 'anoth', 'bas', 'believ', 'or', 'believ', ',', ,
```

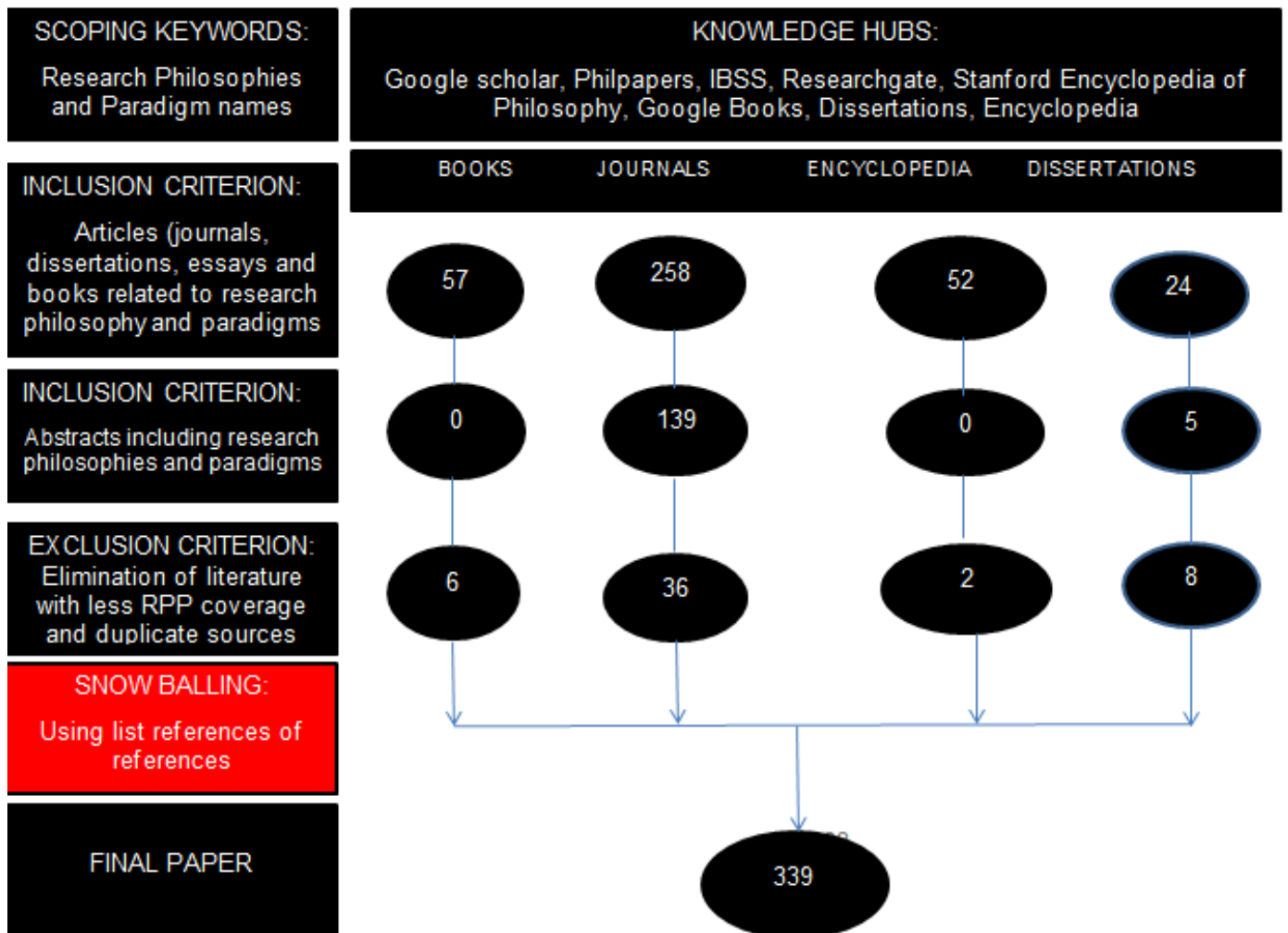
Part of Speech Tagging

```
([ reality_NN ])
<: is_VBZ :>
external_JJ and_CC independent_JJ of_IN ([ our_PRPS senses_NNS ])
```

Stemming and Lemmatization

```
independ knowledg fixabl meet
```


APPENDIX M: BREAK DOWN OF SOURCES FOR THE CORPUS



APPENDIX N: DATA STATEMENTS WORKSHEET

Data set name: Research Philosophies and Paradigms

Citation (if available): N/A

Data set developer(s): Marcia Mkansi, NT Mawila, T Catlyn and SM Mphahlele

Data statement author(s): NT Mawila

Others who contributed to this document:

May we draw on your notes and feedback in a report or publication we might write about data statements and how to develop them?

A. CURATION RATIONALE

This dataset includes texts relating to the epistemology, ontology and axiology of research philosophies and paradigms. The texts have been selected because they identify key concepts about the generation of knowledge based on each research philosophy and underlying paradigm. The dataset text is obtained from various sources such as PhilPapers, Stanford Encyclopaedia of Philosophy, Google Scholar, IBSS, Philosophy Basics, Encyclopaedia, etc. It consists of texts about 180 research philosophies and paradigms. The corpus was created with the intention to train classification models on RPPs categories which will be used to classify user input into these predetermined categories. For the creation of the corpus the study engaged in the following activities;

1. Identify and obtain texts on available research philosophies and paradigms
 2. The study proceeded in pre-processing the data for feature selection for each of the RPPs class labels or categories. This was done through the identification and extraction of information about their epistemology, ontology and axiology components
 3. The BoW text modelling technique was then used to convert the identified texts into numbers (vectors) for use with any machine learning algorithm
-

B. LANGUAGE VARIETY/VARIETIES

The language used for the RPPs-dataset is representative of the language used for conducting research and in line with the pedagogy philosophy of knowledge generation. This language is widely used in academia when mention of research methodology is made.

C. SPEAKER DEMOGRAPHIC

Being researchers, the compilers of the texts contained herein are conversant with the language used in knowledge generation. The texts are compiled by three

students, two of whom are completing their Master's and the other completing their PhD degrees.

D. ANNOTATOR DEMOGRAPHIC

The initial database of research philosophies and paradigms was first compiled by Professor Marcia Mkansi, This was later expanded by the team of three students working under her supervision (Ms NT Mawila, Ms T Catlyn and Mr SM Mphahlele).The annotators are these three students who are conversant with the concepts of research philosophies and paradigms at the university of South Africa (UNISA).

E. SPEECH SITUATION

N/A

F. TEXT CHARACTERISTICS

The dataset text is obtained from various sources such as PhilPapers, Stanford Encyclopaedia of Philosophy, Google Scholar, IBSS, Philosophy Basics, Encyclopaedia, Dissertations, etc. These sources provide further information about research philosophies and paradigms and how they have been used in conducting research.

G. RECORDING QUALITY

N/A. This dataset only includes text data.

H. OTHER

N/A

I. PROVENANCE APPENDIX

N/A. No datasets exist for research philosophies and paradigms.