



**UNIVERSITY OF BOTSWANA
DEPARTMENT OF COMPUTER SCIENCE**

**DISSERTATION FINAL REPORT
Masters of Science in Computer Information System**

Intelligent HIV/AIDS FAQ Information Retrieval System Using Neural Networks

Student Name: Godfrey Mlambo

Student ID: 200908152

Date: September 2015

Supervisors: Professor Yirsaw Ayalew

Abstract:

HIV/AIDS has no cure to this date and it has been noted that the most effective means of mitigating the disease infectious rate is information sharing. HIV/AIDS Frequently Asked Questions (HIV/AIDS FAQ) is another approach for sharing information. The research proposes an automated FAQ information retrieval system for sharing HIV/AIDS FAQs. One of the challenges in FAQ retrieval is mapping of a user query in an FAQ retrieval system to appropriate FAQ question in the FAQ retrieval system repository. To address this challenge a number of approaches have been proposed, most of which are based on traditional information retrieval techniques.

The goal of this research is to design and implement an artificial neural network retrieval system to experiment mapping of arbitrary HIV/AIDS FAQ user question to similar in meaning or equivalent HIV/AIDS FAQ question stored in the FAQ retrieval system repository. Question to question similarity matching technique shall be used. System performance is benchmarked with traditional information retrieval (key word based) HIV/AIDS FAQ retrieval system. Golden standard approach is used to judge system efficiency using re-lection rate and recall rate metrics.

The study compiled an HIV/AIDS FAQ corpus. The Intelligent HIV/AIDS FAQ retrieval system (IHAFR) operational parameters were designed based on heuristics rules and experimental determinants. A portion of the HIV/AIDS FAQ corpus was used to train the IHAFR using MATLAB. Unknown HIV/AIDS FAQ questions were posed to the systems and performance benchmarked with traditional keyword based HIV/AIDS FAQ system. HIV/AIDS counselors, students participating in HIV/AIDS organized activities evaluated the performance of these systems.

The analysis results revealed that IHAFR had recall rate 79.17% and traditional keyword based FAQ system 55.83% for equivalent or similar HIV/AIDS FAQs. Traditional keyword based FAQ retrieval system attained a re-lection rate of 82.50%, compared to 61.67 % for neural network system. Based on these general results, the research concludes that neural network systems have a better ability to provide alternative FAQ questions which are semantically similar because of the neural network generalization trait. In contrast, key word based retrieval systems recall rate are poor because they do syntactical similarity matching, however this same trait gives them a better re-lection rate. Due to the generalization ability of the neural network approach, it could be an ideal technique for implementing HIV/AIDS FAQ retrieval as it semantically provides related FAQ question and therefore an answer.

Acknowledgement

The researcher pays his most profound gratitude and respect to the research's supervisor Professor Yirsaw Ayalew whose has shown me light from the darkness as he always says and recites the late president and former formidable president of South Africa "Education is the most powerful weapon which you can use to change the world - Nelson Mandela. I am sure this change also transcends into individuals as I am now an academic "light" bearer. I would like to thank all the academic staff in the department of computer of science at University of Botswana for their immeasurable support. Lastly I dedicate this research to my family in particular my sister T. Mpundu-Mlambo, my mother Mrs. D. Mlambo and my late father Mr. T. Mlambo for their profound support, may the Lord Bless them.

Table of Contents

Abstract:.....	i
Acknowledgement	ii
Table of Contents.....	iii
Table of Tables	vii
Tables of Figures.....	viii
CHAPTER 1	1
1.1. Introduction.....	1
1.2. Background of the Study	2
1.3. Question Answering System Types	4
1.3.1. Open Domain Question Answering System Types.....	5
1.3.2. Closed Domain Question Answering System Types	5
1.4. Concept of Frequently Asked Question (FAQ)	5
1.5. Question Similarity Matching.....	7
1.6. Problem Statement.....	9
1.7. Aim and Objectives.....	9
1.7.1. Research Objectives	9
1.8. Significance of the Research.....	10
1.9. Organization of the Dissertation	11
CHAPTER 2	12
2. Related Work.....	12
2.1. Similarity Matching Algorithms	12
2.1.1. Similarity Measurement Based on Strings:.....	12
2.1.2. Knowledge Based Similarity Measurement.....	16
2.1.3. Corpus Based Similarity Measurement (CBSM)	17
2.2. Approaches to FAQ Information Retrieval.....	19
2.2.1. Statistical Based FAQ Information Retrieval.....	20

2.2.2.	Template Based FAQ Information Retrieval	22
2.2.3.	Relevance Feedback Based FAQ Information Retrieval	22
2.2.4.	Probabilistic Based FAQ Information Retrieval	23
2.2.5.	Boolean Model Based FAQ Information Retrieval.....	25
2.2.6.	Fuzzy Model Based FAQ Information Retrieval	25
2.2.7.	Language Model Based FAQ Information Retrieval	26
2.2.8.	Latent Semantic Indexing Model (LSI) FAQ Information Retrieval.....	27
2.2.9.	Machine Learning Based FAQ Information Retrieval	27
2.3.	Approaches to FAQ Information Retrieval.....	28
CHAPTER 3		29
3.	Artificial Neural Network Approach.....	29
3.1.	Artificial Neuron.....	29
3.2.	Artificial Neuron Activation Functions	31
3.2.1.	Uni-Polar Sigmoid Function	31
3.2.2.	Bipolar Sigmoid Function.....	32
3.2.3.	Hyperbolic Tangent Function.....	33
3.2.4.	Radial Bases Function Neural Network (RBFNN).....	33
3.3.	Artificial Neural Network Architectures	34
3.3.1.	Artificial Neural Network Recurrent / Feedback Architecture	35
3.3.2.	Artificial Neural Network Feedforward Architecture	39
3.4.	Artificial Neural Network Training Algorithms	44
3.4.1.	Perceptron Learning Rule.....	46
3.4.2.	Backpropagation Training Rule	47
3.5.	FAQ Retrieval Using Neural Network	48
3.6.	Intelligent HIV and AID FAQ Retrieval Using Neural Networks (IHAFR) Architecture.....	51
CHAPTER 4		53

4. The Artificial Neural Network Development.....	53
4.1. HIV/AIDS FAQ Questions Source and Selection	54
4.2. HIV/AIDS Questions Processing (Document Reduction).....	55
4.2.1. Feature Extraction of HIV/AIDS FAQs: Experiment I.....	55
4.2.1.1. Tokenization.....	55
4.2.1.2. Stop Word Removal	56
4.2.1.3. Stemming:.....	56
4.2.2. Feature Selection of HIV and HIV FAQs: Experiment II.....	57
4.2.2.1. HIV/AIDS FAQ Key Word using Term Frequency and Inverse Document Frequency.....	57
4.2.2.2. HIV/AIDS FAQ Question Length Normalization.....	58
4.2.2.3. Vector Space Model	59
4.3. Dimensional Reduction Techniques with Principal Component Analysis: Experiment III ...	60
4.3.1. Principal Component Analysis via Eigenvector and Eigenvalue.....	61
4.3.2. Principal Component Analysis via Singular Value Decomposition	62
4.4. Selection of the Principal Analysis Factors: Experiment IV	64
4.4.1. Scree Plot Approach.....	64
4.4.2. Cumulative Percentage Variance technique.....	64
4.4.3. The Latent Root Criterion considers	65
4.5. Determining IHAFR Neural Network Parameters.....	66
4.5.1. Input and Output Neural Nodes: Experiment V	68
4.5.2. Hidden Neural Nodes: Experiment VI.....	70
4.5.3. Determining Neural Network Training Epochs:	73
4.5.4. Artificial Neural Network Activation Function and Rule: Experiment VII.....	74
4.5.4.1. Artificial Neural Node Activation Function.....	74
4.5.4.2. Artificial Neural Network Learning Rule.....	74
4.5.4.3. Activations Functions and Backpropagation Learning Algorithm Variants: Experiment VIII	77

4.5.5.	Similarity Matching of HIV/AIDS FAQ Questions: Experiment IX.....	79
4.6.	IHAFR Validation Performance	82
4.7.	IHAFR Generation of Responses to HIV/AIDS Queries	83
4.8.	Testing and Collection of Results for IHAFR System.....	85
4.9.	Evaluation of the IHAFR System	86
CHAPTER 5		89
5.	Experimental Evaluation	89
5.1.	IHAFRS Results.....	90
5.1.1.	Recall Rate Performance	90
5.1.2.	Re ction Rate Performance.....	90
5.2.	Main Issues	92
5.3.	Experiment Limitations	94
CHAPTER 6		95
6.1.	Conclusion	95
6.2.	Research Study Conclusion	95
6.3.	Contribution of this Research	97
6.4.	Future Work.....	98
References.....		99
APPENDIX 1: Backpropagation Training Algorithm Variants and Results.....		105
APPENDIX 2: Research Questionnaire for HIV/AIDS FAQ Questions.		111

Table of Tables

Table 1: Tokenized and Stop Word Removed AIDS FAQ Questions[1].....	56
Table 2 Examples of Stemmed Words for the HIV/AIDS FAQ questions[1].....	56
Table 3: PCA Component Determination and Specified Component Values	66
Table 4: Number of Input, Output and Hidden Neurons versus MSE.....	69
Table 5: Heuristic Thumb Rules:.....	71
Table 6: Performance Comparison with Different Transfer Activation Functions and Learning Rule	78
Table 7: IHAFR Neural Network Parameters.....	79
Table 8 Categorization of IHAFR Systems based on training cut-off points	86
Table 9 Categorization of IHAFR Systems based on training cut-off points	89

Tables of Figures

Figure 1: Typical HIV/AIDS FAQ Question and its Answer (Extract from MASA [23]).....	6
Figure 2: Representation of an FAQ Corpus as Term-Weights in Term Document Matrix	14
Figure 3: Spatial representation of question/documents and user query vectors.....	15
Figure 4: Artificial Neural Neuron Mathematical Model[67]	30
Figure 5: Uni-Polar Sigmoid Function Model	32
Figure 6: Bipolar Sigmoid Function Model.....	32
Figure 7: Hyperbolic Tangent Function Model	33
Figure 8: Single Layer Recurrent Neural Network Model	35
Figure 9: Multi Layer Recurrent Neural Network Model.....	35
Figure 10: Structural Layout of Self Organizing Neural Network Architecture	37
Figure 11: Architecture of Single Layered FeedForward.	39
Figure 12: Illustrate the Architectural Concept of Multi-Layered FeedForward Neural Network.....	40
Figure 13: Illustrate the Architectural Concept of Radial Basis Neural Network.....	41
Figure 14: Operational Concept of Gaussian RBF activated neuron in the Hidden Layer.....	42
Figure 15: Mapping of Input Vector to a Documents Using MLP Neural Network.	43
Figure 16 Unsupervised Learning Model	45
Figure 17: Reinforcement Learning Model	45
Figure 18: Supervised Learning Model	46
Figure 19: Intelligent HIV/AIDS FAQ Layout.....	51
Figure 20: Total number of HIV/AIDS FAQ questions and the Key words [1].....	54
Figure 21: Tokenized HIV/AIDS FAQ Questions and its Tokens.[1].....	55
Figure 22 Scree Plot of PCA Components Against Eigenvalues	65
Figure 23: Cumulative Percentage Plot Variance against Eigenvalues	65
Figure 24: The Latent Root Criterion Lot against Eigenvalues.....	66
Figure 25: Multilayered Feedforward Neural Network Structure.	68

Figure 26: Excel Plot for the Table 5a: Number of Input Neurons versus MSE.....	69
Figure 27: MATLAB Plot for the Table 5: Number of Input versus MSE.....	70
Figure 28: Determining the Hidden Neurons for the ANN Architecture using the Validation Data Set	72
Figure 29: Plot of Hidden Neurons versus Least Mean Square Error	72
Figure 30: Illustrations of range of Epoch for IHAFR	73
Figure 31: Training the Neural Network PCA dimensional reduced VMS for HIV/AIDS Matrix.....	76
Fig 32: Question <i>QM</i> to Question <i>FAQN</i> () Mapping using Neural Network.....	80
Figure 33: Backpropagation training rule for the multilayered feed forward ANN	81
Figure: 34 Linear Regression Analysis for the Classification	82
Figure 35: The Training Diagram for the IHAFR Based on the Experimental Determined Parameters.	83
Figure 36a: The query interface for entering the HIV/AIDS query.....	84
Figure 36b: Response Interface to Show all Possible Answers[1].	84
Figure 37: Plot of the Recall Rate Compared to the Responses of the Experimented Systems.	91
Figure 38: Plot of the Re action Rate Compared to the Responses of the Experimented Systems[1].	92

CHAPTER 1

1.1. Introduction

AIDS persist as a challenging disease to control its infection rate worldwide and more so in African countries including Botswana and this is mainly due to incidence risk taking behavior of the general populace. In most African countries educational campaigns and medical therapies are strategies used to manage and control the prevalence of HIV/AIDS because there is no cure.

HIV/AIDS Frequently Asked Questions (HIV/AIDS FAQ) constitute as one of the strategies for information sharing. Therefore, the research proposes an Intelligent HIV/AIDS FAQ Retrieval System (IHAFR) using Artificial Neural Network as a technique to implement sharing information about HIV/AIDS FAQs answers thorough posed arbitrary HIV/AIDS user queries. Question to question similarity matching using Artificial Neural Network (ANN) technique is proposed for selecting a similar or equivalent HIV/AIDS FAQ existing in the IHAFR repository[1].

HIV/AIDS FAQ questions from MASA, IPOLETSE booklets compiled by the Ministry of Health in government of Botswana and also other reliable and authentic HIV/AIDS FAQs were compiled to constitute a corpus. The HIV/AIDS FAQ corpus was pre-processed and sub ected to feature selection. Feature selection entails annotating text into numerical values using text properties like term frequency (*TF*) and inverse document frequency (*IDF*) to derive a numerical term - FAQ question matrix. This matrix defines an HIV/AIDS FAQ vector for each FAQ in the repository. Document reduction technique was used to filter out terms that do not significantly contribute to question identification and the final product deduced was a term - FAQ question matrix used to train the Neural Network system thus creating an inbuilt HIV/AIDS FAQ Knowledge Base (HIV/AIDS FAQ KB) within the IHAFR.

HIV/AIDS user query when submitted to the system shall be sub ected to the same process of pre-processing. The HIV/AIDS user query processed shall be matched with existing HIV/AIDS FAQs in the IHAFR FAQ repository. Relevant HIV/AIDS FAQs are extracted and ranked in order of relevancy to the submitted query. A traditional information FAQ retrieval using Vector Space Model technique was used as a baseline system to facilitate comparative analysis and evaluation of the IHAFR so that its response and accuracy could be determined.

1.2. Background of the Study

National AIDS Coordinating Agency (NACA) a Botswana government organization in its publication of July 2008 reported that adult HIV prevalence grew rapidly during the early 1990s, reached its peak of around 26% in 2000 before declining to 24% by the time of the Botswana AIDS Impact Survey in 2004 and to about 21% in 2007[2]. Joint United Nations Programme on HIV/AIDS [3] issued a survey report (HIV/AIDS ESTIMATES (2009))and estimated adults aged 15 to 49 years had a prevalence rate of 24.8%.”

USAID a popular nongovernmental organization operating in a number of Southern African countries, inclusive Botswana published a report (September 2010) on the country’s HIV prevalence rate revealing that 23.9 % of adults aged 15 to 49 years are HIV positive and the pace of new infections could be slowing[4].

NACA in its publication of March 2014 reported that adult HIV prevalence is declining and currently stands at 18.5% which shows the positive impact of educational campaigns and medical therapies despite the absence of a complete cure to the deadly disease [5]. Based on these observations this research is proposing to improve HIV/AIDS educational campaigns and awareness campaign tools by experimenting the feasibility of implementing an automated Intelligent HIV/AIDS FAQ retrieval system[1] using question to question mapping technique when searching for a relevant HIV/AIDS FAQ.

The government of Botswana in collaboration with other key stakeholders took a bold and stem measure to mitigate the ravaging and devastating disease. Key to these measures are information sharing programs, AIDS awareness campaign, education about the disease and other methods used to alert people about the illness. AVERT an international HIV/AIDS charity, working to avert HIV/AIDS worldwide, through education, treatment and care, commented that “Botswana’s long-term vision is to have no new HIV infections by 2016, when the nation will celebrate 50 years of independence. This will never be achieved without a massive and sustained HIV prevention campaign”[6].

Currently different HIV/AIDS campaigns include information dissemination strategies like public awareness and education comprising bill boards, banners and fliers, IPOLETSE a telephone call centre which provides answers to Frequently Asked Questions (FAQs) on HIV and AIDS, a website hosting and publications [7], Knowledge, Innovation & Training Shall Overcome AIDS (KITSO) a university of Harvard Research and Training program on HIV/AIDS [8] , Makgabanenga radio serial drama behavioral modeling people on afflicting HIV/AIDS issues[9].

Shortcomings noted from these adopted approaches are advertising on radio is considered as an inhibiting factor because of limited society who can access radio[10] and radio programme awareness campaigns must be initiated to let people know when health programmes will be on air [4]. Botswana has safe-sex billboards and posters everywhere, but it is unclear whether anyone pays attention [4]. Call centers use fixed landlines whilst the general populace has more access to mobile phones according to a Botswana Telecommunications Authority (BTA)[11] 2010 survey which revealed that there are 136 593 fixed telephone subscribers compared to 2 339 029 mobile users from a population of 1, 776,494. Call centers using fixed lines which do not support services like Short Message Services (SMS). Publications in general induce a sense of information load as people normally would want to get an answer for a specific question than trying to read the whole pamphlet or book in search of a simple answer.

Currently there are no appropriate, conducive and contemporary Information Communication Technology (ICT) techniques for accessing information resources on HIV/AIDS in Botswana as noted by Masizana-Katongo et al [1, 12]. Researches in Information Retrieval (IR) and Natural Language Processing (NLP) have lead to a Question Answering (QA) System, a service which is also capable of being interrogated, answering and disseminating information effectively and conveniently through use of ICT and Natural language (NL).

Automated Question Answering using Frequently Asked Questions (FAQ) is another category of QA systems. People sharing a common interest are able to get answers to recurring questions from this knowledge domain, using a FAQ information retrieval system. At the moment manual publications like IPOLETSE, MASA and reliable and authenticated HIV/AIDS sites have many HIV/AIDS FAQ question answer pair. These HIV/AIDS FAQs question answer pair could be converted to an electronic knowledge base embedded in an FAQ retrieval system hence provide easy access to answers on frequently asked questions on HIV and AIDS.

The problem that is common with using Natural language (NL) is the possibility of posing questions in diverse ways but meaning a similar thing is common. User questions are inherent of arbitrary words, word sense disambiguity and lexico grammatical content. For example what are symptoms of AIDS? How do you know signs of AIDS? What symptm of aids? Hw d u know sgns of aids? These are typical HIV/AIDS FAQ user questions that can be posed to an FAQ information retrieval system by users but portraying the same meaning, with different grammatical content and word sense.

Artificial Neural Networks (ANN) a branch of artificial intelligence has been proved to resolve this problem more efficiently than any conventional information retrieval method. ANN approach mimics the human brain ability to abstractly resolve incomplete data using acquired knowledge from examples and learning done. An intelligent and automated FAQ retrieval system implies that the system is sub ected to a knowledge discovery process through a learning process where some HIV/AIDS FAQ questions are presented to the system. The system in turn shall learn the patterns and trends confined within the HIV/AIDS FAQ questions in a way of relating certain terms in users queries to questions stored in the system knowledge base [13]. The research proposes that an Intelligent HIV/AIDS FAQ Retrieval ANN could be an ideal solution to resolve arbitrary HIV/AIDS FAQ question and provide similar or equivalent in meaning HIV/AIDS FAQ question which bears answers to the questions posed.

1.3. Question Answering System Types

Question Answering is a technology that has emanated from information retrieval concepts and it has improved the retrieval of information to a specific answer than providing a multitude of answers as characterized by Yahoo, Google search engines, etc [14]. USENET is a popular QA system that provides answers to frequently asked questions from a pool of a wide-ranging library of previously-answered questions[15]. Agrawal [16] defines Question Answering as the task whereby an automated machine answers, arbitrary questions formulated in natural language. Khiyal et at al[17] further elaborates question answering as a technology that provides spontaneous methods of information access compared to existing and famous information retrieval systems like Yahoo, Google, etc. Question Answering systems are largely classified as closed or open domain systems.

1.3.1. Open Domain Question Answering System Types

Mohammed et al [18] defines an open domain question answering system as a system that deals with open ended questions and relies on general world knowledge found on world wide web (WWW), text corpus, etc

1.3.2. Closed Domain Question Answering System Types

Closed-domain question answering relate to questions under a specific sub ect, supported with sub ect-specific knowledge formalized in a knowledge repository. FAQ information retrieval system is a category of closed domain QA systems. The FAQ information retrieval system retrieve existing FAQ questions that are already predefined and organized as question answers (QA pairs) stored in an FAQ knowledge base.[19]. Database and Ontology QA systems are yet another class of closed domain QA systems. They process user queries and retrieve predefined answers based on the user question which is transformed into a database query or template i.e. SQL or SPARQL etc.

The research adopts a question answering system belonging to the closed domain category in particular the FAQ information retrieval system. This model suite well with an information retrieval system the research intends to design and build for HIV/AIDS FAQ information retrieval system. This HIV/AIDS FAQ information retrieval model should handle domain information (HIV/AIDS) which has can be implemented as an HIV/AIDS FAQ question both from the processing and mapping context.

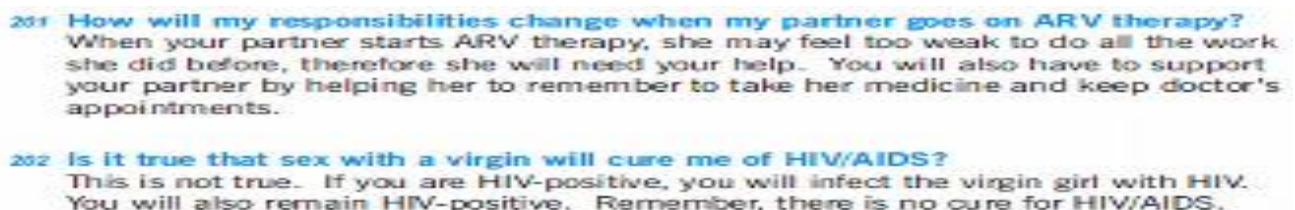
1.4. Concept of Frequently Asked Question (FAQ)

Isiaka and Salim [20] state that FAQ stands for “frequently asked questions” and further implies that there should be an access mechanism that provides access to the FAQ question, which is frequently asked by users, usually about different issues of concern or those about common interest and the answers are typically shown with the questions. The key concept of a collection of FAQs file is to keep a testimony pertaining to common perceptions in a given community about a particular sub ect and the questions being asked, then answers are made available, particularly to newcomers of the group who may otherwise ask the same questions again and again[21]. The same sentiments and reflections on the need and importance of constantly asked questions (FAQ), their availability is echoed by [22]

“...If many users asked the same questions (FAQ) in a certain period of time (a hot topic), then the answers to these questions will be checked manually and directly linked to the questions....this aspect is also found in many new generation search engines, such as Ask eeves...”

In summary the FAQs are created out of clear interest by people and their intention is to seek knowledge and have a common consensus about a problematic issue prevalent in a community. The FAQ’s answers are authored by domain experts since the general populace raises constant concerns over the same issue. In the research study context, HIV/AIDS is a pandemic disease which has caused untold suffering world over. Many repeated questions have been asked to seek knowledge about the disease with view to know and understand the nature of the disease.

In Botswana the IPOLETSE and MASA HIV/AIDS FAQ manuals are responses to the need for such knowledge on HIV/AIDS. The HIV/AIDS FAQs questions in these manuals are arranged as shown in figure 1 where each HIV/AIDS FAQ question is written in blue is provided with an answer. This set is known as an HIV/AIDS FAQ question answer pair. In this research we intended to store this HIV/AIDS FAQ in a FAQ system repository and every posed HIV/AIDS query shall be mapped to a corresponding or similar in meaning HIV/AIDS FAQ question so as to retrieve an answer.



201 How will my responsibilities change when my partner goes on ARV therapy?
When your partner starts ARV therapy, she may feel too weak to do all the work she did before, therefore she will need your help. You will also have to support your partner by helping her to remember to take her medicine and keep doctor's appointments.

202 Is it true that sex with a virgin will cure me of HIV/AIDS?
This is not true. If you are HIV-positive, you will infect the virgin girl with HIV. You will also remain HIV-positive. Remember, there is no cure for HIV/AIDS.

Figure 1: Typical HIV/AIDS FAQ Question and its Answer (Extract from MASA [23])

The research selects the FAQ information retrieval model as an appropriate question answering system model to be used for implementing an HIV/AIDS FAQ information retrieval system because it deals with FAQ questions. Furthermore, some questions on HIV/AIDS FAQ are found in IPOLETSE and MASA HIV/AIDS FAQ manuals. These FAQ questions can be converted to an electronic platform for easy access and utilization by people who might want to enquire and know on HIV/AIDS issues.

1.5. Question Similarity Matching

Similarity is an occurrence where resemblance is brought upon two distinct objects and certain common properties of these entities are used as features to infer the measure of resemblance. In the context of defining similarity between two textual objects [24] defines it as an evaluation of “... level of similarity between a subject document and the dataset documents”. A subject document in this case refers to a query to be compared against dataset documents stored in a system repository and equivalent or similar document(s) can be retrieved. Another definition of similarity measure in reference to textual similarity defines the process as “a function which computes the degree of similarity between a pair of vectors or documents since queries and documents are both vectors”[25]

So many text similarity measuring techniques or algorithms exist, approaches like Dice coefficient, Overlap coefficient, Jaccard, Cosine, Asymmetric, Dissimilarity and many more have been used to effect similarity measurement of textual objects as discussed in various literature reviews. The common consensus with regards to these conventional similarity measuring algorithms is their poor performance [26]. The main challenge articulated is the inability of the traditional techniques to effectively implement similarity measurement based on key concepts like semantics meaning, contextual meaning and conceptual meaning hence they lack the ‘intelligent aspect’ or the ‘reasoning dimension’. For instance when these approaches are used to determine the similarity between the key words car and automobile they infer them as poorly related words. A computation based on the Dice coefficient and Jaccard gives value 0.07693 respectively and cosine similarity gives 0.182574. These similarity measurement values factually show that traditionally similarity measures do not show any relationship at all but practically and logical these words do bear a strong relation and point to the same object.

A new order of textual object similarity measurement, which is gaining momentum and is being tried in different circles of information retrieval, is artificial neural networks techniques. The approach is based on the theory of artificial intelligence and bordering on the premises of machine learning where a system is taught using sample data of a particular domain to discover patterns, trends and relationships of key words that constitute documents within a corpus. On this basis, the system will be queried with generic data of the same type and the system should be able to respond with generalized or specific responses and in a way mimicking how humans reason and relate to provide

relevant and similar responses. This approach tends to combine the context of intelligent comparison, thus implying contextual, semantic and conceptual comparison of text objects using artificial neural networks.

A similarity measure in information retrieval system plays a very critical role as it facilitates the retrieval and ranking of relevant responses to a presented query. Furthermore similarity measure is used as a filter mechanism by defining a threshold value which would in turn determine the number of documents to be retrieved as relevant to the question. In some models of information retrieval the similarity measure or magnitude is used as a feedback measure to improve the magnitude and quality of retrieved documents

Thus, similarity measure is a key and critical mechanism in any information retrieval system. The efficiency of an information retrieval system is based on its ability to retrieve relevant documents based on a presented user query. This perception simply emphasizes the fact that a correct selection and implementation of similarity matching technique is very critical to the design of an information retrieval system as critically remarked by [27] and cites

“...commonly used techniques such the Cosine and accard...treat words as though they are independent of one another, which is unrealistic...words are not isolated but always relate to each other to form meaningful structures and to develop ideas... [27]”

Similarity matching techniques which are able to combine semantic and conceptual mechanism thus emulating human intelligence should be able to retrieve documents that are relevant to the query posed. In this regard, the research shall experiment using HIV/AIDS FAQs, design and train an information retrieval neural network based system with a sample of HIV/AIDS FAQs so that it would be able to learn, discover the trends and patterns and relationships of key words of HIV/AIDS FAQs that constitute the corpus. The system should be able to perform an intelligent similarity matching based on key words of posed user query and yield relevant and similar HIV/AIDS FAQ question hence the answer.

The research intends to determine the effectiveness of a question to question intelligent similarity measurement and compare with a conventional similarity measuring technique. If the results prove to be good the system could be recommended as a tool for sharing HIV/AIDS information based on frequently asked questions. As stated in the background section the intelligent information retrieval system could play a significant role in information dissemination hence mitigate the impact of HIV/AIDS prevalence in Botswana and Africa?

1.6. Problem Statement

The research problem statement is to develop an automated FAQ retrieval system which will automatically search the FAQ repository to see if the same or similar question exists in the repository. If the same or similar question is found, then the corresponding answer can be provided, however, determining the semantic similarity between a user question and questions in the FAQ repository is a difficult task. Praksher [28] comments that approaches to develop QA systems that require language understanding which is perfect seem doomed to failure because novel language, incomplete language, and error prone language are the norm, not the exception. The difficulty is due to the fact that the same question can be expressed using different words which have similar meanings through arbitrary or colloquial expressions which do not follow a language syntactic structure.

1.7. Aim and Objectives

The goal of the research proposal is to design and implement intelligent HIV/AIDS FAQ Retrieval System using neural networks for retrieval of relevant of HIV/AIDS FAQ question and its answer based on a user query.

1.7.1. Research Objectives

- 1. Survey and compile an HIV/AIDS FAQ corpus from authentic and reliable sources like IPOLETSE, MASA and United Nations World Health Organization (WHO) for HIV/AIDS FAQ questions.**
- 2. Pre-process the HIV/AIDS FAQ corpus to create a Vector Space Model and Principal Component Analysis Matrix to train the Intelligent HIV/AIDS FAQ Information Retrieval System using MATLAB and Java NETBEANS**
- 3. Design and parameterize neural network architecture for implementing the intelligent HIV/AIDS Information Retrieval System through heuristic rules.**

4. **Train the neural network using suitable training rule with representative knowledge on FAQ HIV/AIDS FAQ questions using MATLAB**
5. **Test the Intelligent HIV/AIDS FAQ Information Retrieval System by querying with HIV/AIDS queries and record Results for evaluation specialist in HIV and AIDS.**
6. **Evaluate the effectiveness of the system using appropriate golden standard/ground truth approach.**

1.8. Significance of the Research

As much as there are so many researches on information retrieval using neural networks, this research brings a different dimension in the research community as it attempts to test and verify the theory of machine learning a branch of artificial intelligence to design, train and implement an intelligent information sharing tool that can be used socially, to communicate on most frequently asked questions (FAQs) and their answers on the most devastating disease HIV and AIDS. Currently there is no cure to this deadly disease and the best and effective mode of mitigating the prevalence and spreading of the disease is information sharing.

Thus the purpose behind this research is to determine the efficacy of searching for a similar document or question in a given electronic corpus by using artificial intelligence notably through use of artificial neural networks techniques. A comparative analysis to indicate the effectiveness of the information retrieval using neural network techniques shall be done using the most accepted and standard information retrieval model the classical Vector Space Model, as acknowledge by [29].

If the performance of the system is good it is envisaged that the intelligent HIV/AIDS FAQ Retrieval System can be recommended to replace HIV/AIDS FAQ IPOLETSE and MASA manuals to share information about HIV/AIDS FAQ for Botswana community. These tools can also be used as a complement to the call centre on HIV/AIDS to answer HIV/AIDS FAQ questions 24/7 depending on the platform to be used.

1.9. Organization of the Dissertation

Chapter 1 will introduce the research problem in the context of conceptualizing an FAQ retrieval system that can use potential techniques like Artificial Intelligence in particular machine learning to retrieve FAQ answers based on arbitrary and colloquial FAQ user queries to facilitate information sharing on HIV/AIDS a devastating and disease without cure so far.

Chapter 2 reviews start of art in the domain of information retrieval by discussing various Information Retrieval models and those aligned to FAQ information retrieval. Key techniques applied especially on question to question similarity matching using incomplete and noisy queries and what the research could learn and adopt.

Chapter 3 brings to attention up to par with technical details of Neural Networks design and architectures and their approach in information retrieval, finally the research neural network architecture to implement.

Chapter 4 introduces tasks and procedures of compiling authentic HIV/AIDS FAQ s from reliable and authentic sources, experimental tasks for determining the optimal functional parameters of the artificial neural network in context of the compiled HIV/AIDS FAQ questions its implementations and results.

Chapter 5 presents evaluation experimental results, main issues of the research and experimental limitations in relation to effectiveness of the artificial neural networks as an approach to FAQ information retrieval technique for HIV and AIDS.

Chapter 6 gives concluding remarks on the research and also the research contribution to body of knowledge on information retrieval in particular for FAQs more inclined to HIV/AIDS using neural networks. The chapter further elaborates on thoughts of possible future work regarding information retrieval using neural network for HIV/AIDS FAQs.

CHAPTER 2

2. Related Work

In section 2.1, the paper shall evaluate the various approaches that have been used to implement similarity matching for textual objects. Section 2.2 provides an overview on types of information retrieval models and those used to implement FAQ Information retrieval system. The research shall relate to and adopt an intelligent approach and with any due modifications to implement a similarity measurement approach of textual objects for FAQ questions and in the context of HIV/AIDS FAQs which embrace semantic, conceptual and contextual analysis characteristics.

2.1. Similarity Matching Algorithms

Measuring similarity of textual objects like word to word, sentence to sentence, question to question and document to document in a given data corpus facilitates a critical role of deciding which textual objects are similar and also the degree of similarity. In their article [30], they mention three key approaches to text object measurement as string based measurement which is categorized as Character Based Similarity Measurement (CBSM) and Term Based Similarity Measurement (TBSM). They further mention Corpus Based Similarity Measurement (CSBSM) and Knowledge Based Similarity Measurement (KBSM) and articulate the various similarity algorithms that are used to realize text object similarity measurement. Analysis of these approaches is based on Natural Language Processing concepts which underpin the theory of comparing text similarity using syntactic, semantic, lexical and conceptual analysis.

2.1.1. Similarity Measurement Based on Strings:

Similarity measurement based CBSM consider character to character comparative analysis. For instance the Longest Common Substring (LCS) technique performs similarity comparison by comparing two strings and determining the adjacent chain of characters that exists in both the strings[30]. The Dama-Levenshtein algorithm computes the distance between two words considering again the commonality of characters amongst and defines the insertions, deletions, or substitution of a single character or transposition of two adjacent characters as a measure of difference. The N-gram similarity measures two strings by noting a succession of N characters from a given order of the two compared text objects. The N denotes the number of character(s) to differentiate with. Uni-gram would note a difference of one character between two compared strings and a bi-gram would notice a measure of two sequential characters.

Term Based Similarity Measurement computes a word into a numerical weight based on its frequency and rarity in a document, query or corpus [30]. The parameters used to numerical deduce the term weight is called the term frequency (*tf*) which implies the number of times the term appears in a document or question or sentence. Another term used is the inverse term document frequency (*idf*) which is used to describe the rarity or value of word to discriminate a sentence or document in a given corpus. The two parametric terms are computed as illustrated in equations 1, where $f_{j,k}$ represents the number of times a term j appears in a document k which is found in a corpus. Equation 2 then computes the term frequency as normalized value which divides frequency term $f_{j,k}$ in equation 1 by the most common term in the document to yield term weight for the computed string or word is numerical computed as in equation 2. Liu, et al. [31] defines term frequency of a term in a document as the as the number of occurrences of the term between the document and the posed query.

$$f_{j,k} = \text{Frequency of term } j \text{ in document } k \dots \dots \dots (1)$$

$$tf_{j,k} = \frac{f_{jk}}{\sum_k f_{jk}} \dots \dots \dots (2)$$

Terms that emerge in numerous different documents in a corpus are less pinpointing of overall documents needed. Therefore there is need to compute for rare terms which are effective in picking the relevant documents in a corpus. The *df* is computed by the documents frequency which calculates the number of documents with the same term as described in equation 3

$$df_{j,k} = \text{document frequency of term } j \dots \dots \dots (3)$$

$$idf_k = \text{Log} \left(\frac{N}{df_k} \right) \dots \dots \dots (4)$$

$df_{j,k}$ Document frequency of term j indicates the number of documents containing term j appears in the corpus. idf_k the inverse document frequency computes the weight of N indexed by the search engine where $df_{j,k}$ is the document frequency and is calculated as in equation 4.

$$W_{jk} = tf * idf_k \dots \dots \dots (5)$$

The two text feature properties idf_k and idf_k when combined together as product, complement their measuring capabilities to form a term weight composed of the term frequency weight as computed in equation 5.

The approach to create a numerical value facilitated the similarity measurement approach TBSM which has ushered a number of similarity measuring algorithms like the Dice Coefficient, accard Coefficient, Cosine Similarity, Euclidean Distance, Matching Coefficient and overlap Coefficient[30]. These approaches form the bulky of similarity measures adopted in many conventional informational retrieval systems of course with variations here and there. Similarity measurement using Term Based Similarity Measure is based on a representation of text information in a corpus as term weight to document matrix.

In the context of this research the document would stand for an FAQ question and hence representation of such information in the FAQ corpus is represented as FAQ questions in column and the constituent words of the FAQ question being computed term weights cutting across the entire FAQ question as illustrated in figure 2. The information represented is Vector Space Model (VSM) where term weight in the corpus forms a scalar unit.

	FAQ1	FAQ2	...	FAQk
Term1	w_{11}	w_{12}	...	w_{1k}
Term2	w_{21}	w_{22}	...	w_{2k}
Term3	w_{31}	w_{32}	...	w_{3k}
:	w_{i1}	w_{i2}	...	w_{ik}
Termj	w_{j1}	w_{j2}	...	w_{jk}

Figure 2: Representation of an FAQ Corpus as Term-Weights in Term Document Matrix

The FAQ question or document, if all terms weights that constitute a question or document are added, and forms a vector which has a specific direction in a given corpus spatial representation. The user queries are also computed by considering the terms they have, which coincide with terms in the corpus thus a query vector is represented in the corpus spatial. Equation 6 and 7 relate the computation of both the question/document and user vectors as illustrated in figure 5.

$$FAQ_1 = W_{11} + W_{12} + \dots + W_{1k} \dots \dots \dots (6)$$

$$Q_1 = W_{11} + W_{12} + \dots + W_{1k} \dots \dots \dots (7)$$

Based on the concept of spatial representation of corpus terms and computation of both question and document vectors, similarity measurement can be done between these two textual objects as illustrated in figure 3. The corpus terms term1, term2 and term3 are all or partly comprising the question/documents FAQ1, FAQ2 and the user query Q. An angle of relatedness between the query and the FAQ documents is defined as θ_1 and θ_2 .

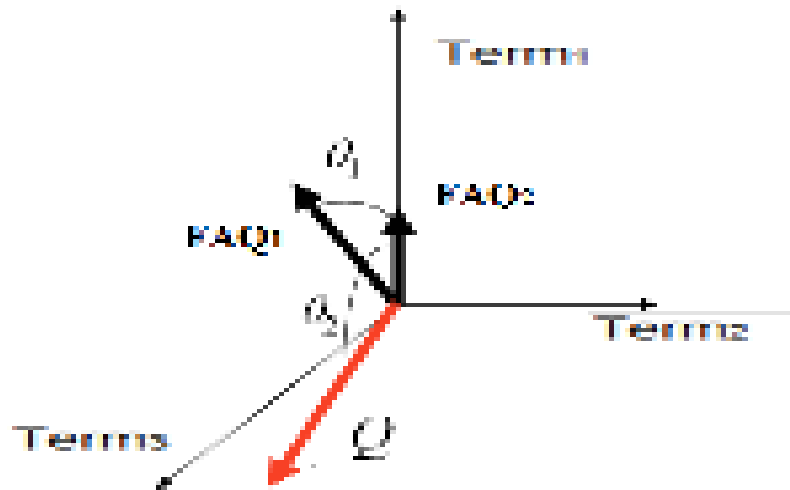


Figure 3: Spatial representation of question/documents and user query vectors

Apparently, θ_2 is greater than θ_1 implying that the query Q is closely related to FAQ1 than FAQ2. This illustration already describes Cosine similarity measure which is implied by angle of difference as illustrated in figure 3 and computed in equation 8. Cosine $\theta = 0$ means the documents or question do not bear any relation at all and Cosine $\theta = 1$ would mean the documents do bear a 100% similarity.

$$Cos(\theta) = Sim(FAQ_n, Q_m) = \frac{\overline{FAQ_n} \cdot \overline{Q_m}}{|FAQ_n| \cdot |Q_m|} = \frac{\sum_{x=1}^y W_{kn} \cdot W_{km}}{(\sum_{x=1}^y W_{kn}^2 \cdot \sum_{x=1}^y W_{km}^2)^{\frac{1}{2}}} \dots \dots \dots (8)$$

Similarity measure can also be based on the Euclidean Distance which measures the distance between a query and a question/document $Disatnce(FAQ_n, Q_m)$ as represented by their term vectors FAQ_n

and Q_m respectively. The Euclidean distance is computed as in equation 9: Distance represents a measurement of distance between the vectors where $W_{x,m}$ and $W_{x,n}$ are the respective term weights computed based on the VSM of the query and question/documents.

$$Distance (FAQ_n, Q_m) = \left(\sum_{x=1}^y W_{x,m} - W_{x,n} \right)^2 \dots \dots \dots (9)$$

The accard coefficient is used to measure similarity between two sets textual objects by calculating the number of words common to the two compared objects and dividing by the sum of the words found in the objects. Term weights can used to calculate this similarity measurement as computed in equation 10:

$$Sim (FAQ_n, Q_m) = \frac{\sum_{x=1}^y W_{kn} \cdot W_{km}}{\sum_{x=1}^y W_{kn}^2 \cdot \sum_{x=1}^y W_{km}^2 - \sum_{x=1}^y W_{kn} \cdot W_{km}} \dots \dots \dots (10)$$

TBSM similarity computations base the selection of similarity measurement on the rate of occurrence of a term in a document corpus and also its rarity. The method implements a crude syntactical analysis which implies that the term is present in the other textual object or not. This approach could be appropriate if it can be implemented and complemented by another similarity measure which can factor the dimension of semantic and conceptual analysis of relations between text since it caters for the statistical relationship of terms. This approach is supported by Gomaa et al “...without resolving word-level redundancies i.e. synonymy and ambiguities i.e. polysemy a similarity computation cannot accurately reflect the implicit semantic connections between words” [30]

2.1.2. Knowledge Based Similarity Measurement

Knowledge based similarity measurement is guided by the principle of semantic similarity measures were similarity algorithms like res, lin, lch, vector and many more are used [30]. This approach divides similarity measurement into key concepts like measure of similarity and measure of relatedness. Tools used to support knowledge based measurement are WordNet for English words, HowNet for Chinese words and many lexicon dictionaries: Similarity and Natural Language Toolkit (NLTK). WordNet is lexical database of English Nouns, Verbs, Adjectives, Adverbs and they are grouped into key hierarchical and cognitive synets which define their level of similarity and relatedness. Synet is defined by [32] as an inter-connected set of words which bear similar conceptual, semantic and lexical relationships.

Ideally knowledge based similarity measure tend to support information retrieval techniques like query expansion and automatic relevance feedback [33]. Key words in a query initiate the process as they are identified and presented to the Knowledge Base so that terms with a similar in meaning form a conceptual, semantic and lexical relatedness are retrieved. For instance, if an English information retrieval system is used, WordNet is implemented so that initial query words are used to retrieve conceptual, semantic and lexical related words from the Knowledge Base. These words are added to the existing and initial query key words. This process logical intends is to increases the number of query key words so that many documents bearing the same words can be retrieved from document repository. This maximizes the chances of attaining a better yield with similar in meaning or correct documents from the information retrieval repository.

Use of KBSM for HIV/AIDS FAQs could have been an ideal approach as users express or ask in different words but implying to a single concept or object or matter. However it is unfortunate that in the domain of HIV/AIDS there is no an authentic and published or researched lexicon dictionary for HIV/AIDS therefore its application in this experimental research is not feasible or it is not an experimental option. WordNet can be used but its disadvantage is that is an English lexicon knowledge Base bound to give general knowledge about terms in HIV /AIDS, rather than alternative and pertinent words. For example one would not say a condom is a latex or rubber though they have a room of relationship in the sense of material used to manufacture a condom. Building a specific knowledge base of HIV/AIDS like Ontology is not possible in the scope of research.

2.1.3. Corpus Based Similarity Measurement (CBSM)

Similarity measurement based on a CBSM is also based on the principle of semantic measurement where the query keywords are compared to words gained according to information from the corpora[30]. Similarity measuring algorithms used are the Hyperspace Analogue Language (HAL), Latent Semantic Analysis (LSA), Generalized Latent Semantic Analysis (GLSA) and the Explicit Semantic Analysis (ESA) and many others. It is out of context for the research to discuss all the algorithms but an exception is made to the LSA model which has features appropriate for technical implementations that could be used for applying similarity measuring and matching technicalities using conceptual, semantic and pragmatic concepts.

Latent semantic analysis (LSA) is a practice in natural language processing, implemented for vector based semantic analysis of relations between a set of documents and the terms in corpus. The intent is to determine a data structure with similar in meaning and related documents to terms in of a corpus. The assumption implied by the LSA analysis is that the terms or strings of words that bear a close relationship to each other will often occur in similar piece or body of knowledge. From this perception a document to term matrix that computes these words is created using mathematical model singular value decomposition (SVD). Comparison of similarity measurement using LSA is computed by using angle differentiation since the vectors have been used as units of document and query representation. Alternatively the inner product technique is also used to compute the similarity between a query and documents.

Latent Semantic Analysis is mostly used in document reduction techniques in particular where analysis has been done of the corpus and a Sparse Vector Space Model is the outcome. Turney et al [2] describes document reduction techniques as an approach that is ideally used on sparse matrix or huge matrix as a way to achieve latent meaning, noise reduction, high-order co-occurrence and sparsity reduction. However the most vital aspect of this approach is its ability to derive and discover latent information in terms of a copra has the same meaning semantically and conceptual. The term latent is as defined by the Oxford online dictionary as hidden or concealed [34]. Thus the LSA uses the SVD mathematical model to compute and relate words into key thematic words that brings about the latency of hidden words in the corpus into key thematic words which defines the corpus similarity semantically and conceptually.

The Principal Component Analysis (PCA) is an advanced state of the LSA and it also uses the SVD model to manifest the latent words into thematic words of the corpus. The advantage of PCA is its ability to arrange the derived thematic words in a determined, orderly and sequential manner through use of eigenvalues. This orderly approach also defines the importance of each thematic word in a numerical and also ranking in descending and ascending thus enabling selection of thematic words key in the corpus. Heuristic rules, knowledge and simulations tools like MATLAB are used to select the appropriate and needed thematic terms depending on the need.

The approach of using PCA incorporates semantic and conceptually similarity measurement through a mathematical approach of SVD using document reduction and therefore corpus similarity matching and measurement is achieved. The approach of using Vector Space Model uses statistical

representation thus adopting feature selection property through use of features like term frequency and its variants. In summary, TBSM is done at surface level thus learning of the presence of the key words more to say crude syntactic similarity matching based on a binary model. Huang et al [27] remarked some researchers have incorporated lexical, semantic and conceptual analysis through using ad hoc approaches however the best and most principled way is to adopt learned similarity measures which is done through machine learning. The authors' further remark that the approach of learned similarity matching involves document representation, feature extraction and similarity calculation based semantic, conceptual, syntactical and pragmatic measurement of textual objects.

In our approach we intend to adopt TBSM since it covers the aspect of document representation and feature extraction and also include a crude syntactic similarity measures by virtue of document representation. As for the similarity measurement based on semantic, conceptual, contextual and pragmatic the research adopts CBSM by implementing a SDV to compute a PCA matrix. The TBSM VSM shall be an input for the PCA matrix thus syntactical, semantic, conceptual and pragmatic learning is done. The PCA model shall be used to conduct supervised learning thus training the artificial neural networks with sample thematic words of the HIV/AIDS FAQs questions.

2.2. Approaches to FAQ Information Retrieval

In general [35], comments that a key parameters that characterize an information retrieval are the document and queries representation, similarity measurement or matching of the relevant documents to the user query, methods of ranking the query output and finally the mechanism for acquiring user relevance feedback. These characteristics form the basis for explaining and defining the main information retrieval models mainly used for the FAQ retrieval

FAQ retrieval systems can be implemented in various approaches that encompasses different information retrieval models i.e.(i) Statistical analysis of the user query and document collection using statistical and similarity techniques [36-38]. (ii) Template-based retrieval using computer annotated user input query to match with predefined question templates stored in the knowledge base(iii) Relevance feedback to enhance precise retrieval of correct question using manual or automated technique[39, 40].(iv) Probabilistic model techniques implying usage of the probability of a question existing in a given corpus and then retrieve the relevant documents. (v) Boolean model representing user and existing documents using binary values and perform binary similarity computation to retrieve relevant document(vi) Fuzzy Set Model does not use crisp, statistical or

probabilistic values instead uses fuzzy terms which measure the uncertainty quantification to match and retrieve documents based on user questions. (vii) Language model information retrieval which has its origination from the probabilistic information retrieval model. The difference is language model creates the probability model of each given document and computes the probability of a query by randomly sampling from this model. (viii) Latent Semantic Indexing Model (LSI) FAQ Information Retrieval is another approach which uses thematic words in corpus to retrieve similar in meaning words. The thematic words represent similar in meaning words found in the documents in the corpus. The thematic words are determined through training process which uses the Latent Semantic Indexing algorithm (ix) Machine learning retrieval model an approach that learns from examples or huge databases of given domain knowledge to discover knowledge and provide a solution to a query posed alternatively it groups items of same similarity in concepts in well defined categories.

2.2.1. Statistical Based FAQ Information Retrieval

Statistical based FAQ retrieval approach is based on the concept of a bag of words explained as a collection of unstructured words which have defined frequency of appearance. The words or key words might be weighted document terms or weighted query terms. If the key words are weighted terms then numerical computation is done by tf representing the number of times words appears and then idf being the word rarity in the document and these text features are used to compute the term weighted W_{jk} . Weighted terms are processed using NLP techniques that include tokenization and stemming. The key words are represented as numerical values by using text word representation techniques like Term Frequency (TF), Inverse Document Frequency (DF_T), Term Frequency Inverse Document Frequency (TF-IDF), [41, 42]. By virtue of representing key words in these numerical values word or sentence similarity is possible. For un-weighted terms they are simply stemmed and their representation would be done using a binary value 1 or true to indicate the presence or false being 0 to indicate none presence .

A formal and structured representation weighted terms is through use of a Vector Space Model (VSM) which would relate each weighted term to its documents through the term document matrix structure. The VSM is an information similarity model used to compare user query and documents in a corpus[43] and is commonly used for similarity measurement approach in many statistical FAQ retrieval. The FAQ questions and the user query are added to represent vectors that would be used to realize similarity comparison based on an angle between the two vectors representing as entities in

FAQ corpus space [44]. Two versions of VSM exist: the standard approach and classical. The classical incorporates weighted terms. The standard approach is used to represent the un-weighted terms using Binary values and is termed Binary VSM.

Cosine correlation and $INNER(d, q)$, inner product, accard Coefficient, Dice Coefficient are algorithms commonly used to compute for similarity measure between an FAQ document and user query for classical VSM where the document and query vector are represented as in:

$$FAQ_j = W_{1,j} + W_{2,j} \dots \dots W_{k,j} \dots \dots \dots (11)$$

$$Q_i = W_{1,i} + W_{2,i} \dots \dots W_{m,i} \dots \dots \dots (12)$$

If the FAQ document and user query contain the same keywords then the similarity computation value would be 1 else it would be zero meaning it does not contain similar keywords. Any value in between is a relative value where there is need to define a cutoff point to define a similarity or none similarity.

VSM is used to represent document to term relations and is also used to facilitate query term expansion by computing words of similarity from an ontology knowledge base[45, 46]. VSM is used to represent both user query and indexed FAQ documents for search by using TF-IDF as the weight term determiner and sentence similarity is incorporated to cater for semantic meaning. This address the short coming of the VSM based on TF-IDF method[25].

Yongming [47, 48] et al uses Extended VSM (EVSM) to enable similarity computation of documents in an XML by defining anodal structure. Liu, et al. [49] uses a Pattern VSM to enable semantic analysis of words where the left contexts and the right contexts of two words are respectively similar, and the degree of synonymy between them is high. VSM is used to represent both user query and indexed FAQ documents for search by using ITDF as the weight term determiner and sentence similarity is incorporated to cater for semantic meaning hence address the short coming of the VSM based on TF-IDF method [25].

This method considers the statistical properties of words in the context, ignores the conceptual and semantic information of sentences. Therefore it's important to add methods that caters for semantic meaning use i.e. use of Latent Semantic Indexing (LSMI) or Principal Component Area, usage of knowledge bases like ontology and Neural networks methods which are capable to resolve semantic, lexical, word disambiguation and incomplete questions.

2.2.2. Template Based FAQ Information Retrieval

Template based FAQ Retrieval system makes use of knowledge annotations which are machine readable sentences that query a knowledge base. Annotations mimics a defined structure depending on the knowledge base used[46]. Pattern matching retrieval technique is used with NLP, in particular syntactic and lexical analysis. Xiaola et al [50] precisely describes the concept of template based retrieval as a pattern is used to customize the user question, the pattern has an influence on the question parser for finding similar questions and answers and finally refinement of the answer through clustering and fusion.

A user question is defined into a pattern of strings called an annotation of the format for instance `subject relation object` by using a lexicon and is used to match with a similar FAQ pattern stored in the knowledge base an answer is retrieved[17, 38, 50]. A user query is expanded using ontology and transformed into a structured language query which is then used to select the appropriate FAQ pattern for querying the ontology base [46]. Sung et al [51] adopts an advanced approach in template based retrieval by implementing algorithms that i) generate templates by sequence alignment; ii) select templates based on a filtering process; iii) apply a matching algorithm for determining question's category. The problem with such systems is that the knowledge base is hand crafted through knowledge rules and many patterns needs to be generated. The knowledge base is costly in terms of time and maintenance as it would require constant human intervention and expertise for new knowledge and queries.

2.2.3. Relevance Feedback Based FAQ Information Retrieval

Information seekers have a tendency to pose short questions that do not have sufficient data to provide accurate answers. In view of this behavior, combinations of irrelevant and relevant documents are retrieved as answers. If some of the retrieved documents are relevant to the query, terms from those documents can be added to the query in order to be able to retrieve more relevant documents; this defines the principle of relevance feedback information retrieval.

Famous methods for implementing this approach include (i) Rocchio feedback which uses an algorithm to implements relevance feedback by modifying the vector space model, also known as the query vector modification (QVM) technique. (ii) Query expansion, where improvement of retrieval results is accomplished by adding synonyms related terms to the query and (iii) Using sources for related terms: manual thesauri, automatic thesauri and query logs.

Yin et al [52], combines the QVM, probabilistic Bayesian inference-based and Frequency feature Selection (FRE) to enhance the process of relevance feedback by developing a hybrid system, and their conclusion was the average precision rates obtained using the proposed model were significantly higher than those obtained using the traditional methods. By extracting and using keywords from the user question relevance feedback is implemented using the QVM technique, however this approach requires an expert to perform the relevance feedback as use of keywords is implemented[53]. Pseudo feedback technique imitates a user from the first retrieval and implements the QVM by computing expansion terms based on probability and reformulates the new query using query expansion to expand capacity of generating more relevant documents[54].

2.2.4. Probabilistic Based FAQ Information Retrieval

FAQ information retrieval using probabilistic model is based on the Probabilistic Theory where information retrieval is based on user questions containing uncertain information and terms used to index the documents. The theory uses mathematical models to quantify uncertainty. The following models are used to compute the similarity and ranking over user questions which have uncertainty in comparisons to documents in given corpus, (i) Probability Ranking Principle, (ii) Binary Independence Model, (iii) Bayesian Networks.

Probability Ranking Principle, implies that for a given user query q , there is a set of documents d_j within the document collection which contain exactly the relevant documents and they constitute the ideal answer set. The mathematical model: $\text{sim}(q, d_j) = P(d_j \text{ relevant-to } q) / P(d_j \text{ non-relevant-to } q)$ define the maximal existence of relevant d_j documents and minimal irrelevant documents based on the query q . A more detailed and implementable model is availed in other literature.

Binary Independence Model (BIM) represent documents and queries as binary incidence vectors of terms as $d_j = (W_{1,j}, W_{2,j}, \dots, W_{n,j})$ and $q_i = (W_{1,q}, W_{2,q}, \dots, W_{m,q})$ $W_{i,j}$ is the weight term being 1 if the term is present and 0 if term is not present, d_j is the document and q_i is the query. The model operates on the same principle of relevance of documents to a posed query, however the term weights used are binary not continuous. One of the assumptions made by the model is terms in documents and queries are totally independent of each which is not practical at all. The BIM was originally designed for small corpus of fairly consistent size and works reasonably.

A Bayesian Network is a graphical representation of nodes (variables) with causal links (relationships) between random nodes (variables), which allow inference on the nodes. Associated with each node and link is a distribution of values for the random variable and a conditional probability table describing the probability distribution of the random variable and dependent on the probability distribution of the parent node. The conditional probabilities are used to compute a priori probability of any instance.

Bayesian Model information retrieval model adopts this approach by defining a document network and also for each query posed; a query network is compiled and attached to the document network where layers of document nodes, document term nodes and concept nodes are defined and also the same for the query. A subset of d_j 's which maximizes the probability value of query node is computed and documents related are retrieved as the answer to query. Bayesian-based method is based on probabilistic principles and is used to determine a class for FAQs in order to reduce searching space [55, 56].

Some probabilistic methods use an initial estimate value to enable retrieval of a set of documents which are then refined through user feedback or an autonomous process i.e. poor query expansion mechanism[57]. The model needs to carry a lot of information along to support reasoning because of independence assumption. Limited representation of documents and queries since terms are regarded as single words what about for advanced text analysis which considers text phrases.

2.2.5. Boolean Model Based FAQ Information Retrieval

Set theory and Boolean algebra define another information retrieval model where key words or terms are used and with logical operators to formulate a question. Logical operators AND, OR and NOT are used to formulate a query for searching documents in a collection. For instance, a question might consist of a Boolean expression, such as“(Disease OR Symptom) AND HIV/AIDS”. The search retrieves all the documents that match the query expression which contain the term Disease or Symptom and associated with the term HIV/AIDS.

Documents are represented by the index terms assigned to the document and there is no indication on which terms are more important than others. Term weights $W_{i,j}$ are discrete, i.e., $W_{i,j} = 1$ if the term is there or $W_{i,j} = 0$ if the term is not there. A query q_i and document $-j$ are viewed as a Boolean vector expression i.e. $q_i = (1, 1, 0, 1)$ and $-j = (0,1,1,0,1)$ and this forms the basis for comparison. Euclidean distance technique is used to determine content similarity and ranking.

Song [58] uses Boolean vectors to represent a query and FAQs document to perform statistical similarity between the two. Gao et al [59] transforms a traditional Boolean Model and Vector Space Model using a matrix, hence combining the advantages of both these models to improve the retrieval performance of the FAQ system.

Main problems noted with Boolean system as described by Larson [37] are that one needs to be an expert in the domain to conduct a search. The functions AND and OR can restrict information provided or tend to overload information provided respectively. The model is not sensitive to semantic analysis at all and documents satisfy the query to the same degree making ranking a difficult [57].

2.2.6. Fuzzy Model Based FAQ Information Retrieval

Conventional information retrieval systems evaluate user queries and retrieve/rank documents based on matching keywords in user queries with words in documents[60]. However this is an ideal approach as most of day to day questions are full of uncertainty. Fuzzy Model Information Retrieval is based on implementing algorithms that use uncertainty quantification to match and retrieve documents based on these questions. It employs the theory of sets whose boundaries are not well defined i.e. they are vague, but the notion is the degree of membership associated with the elements of a set. This degree of membership varies from 0 to 1 and allows modeling the notion of marginal

membership. Membership is a gradual notion, contrary to the crispy notion enforced by Boolean Information Retrieval Model.

Queries and documents are represented by sets of index terms and matching is from the onset by virtue of the representation of the documents and queries i.e. degree of membership for elements in a set. This vagueness can be modeled using a fuzzy framework where each term is associated to a fuzzy set and each document bears a degree of membership in this fuzzy set. Lynn et al [60] retrieve documents using a fuzzy set IR model and rank retrieved documents for any vague query using the “vagueness score” of the documents based on the word senses as defined in WordNet.

2.2.7. Language Model Based FAQ Information Retrieval

Another approach to information retrieval is to think of words that would likely appear in a relevant document, and to use those words as the query. Language modeling embraces this theory as it defines that a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query words often.

A language model is a function that puts a probability measure over strings drawn from some model. The concept behind is to generate a probability language model for each document and computing the probability that the query was generated by randomly sampling from this model. heng et al [61] sums the language model concept by stating that language modeling is to estimate the likelihood or probability of a word string like words, sentences and documents

Various forms of language models are used to realize the concept of language modeling and they are summated as n-gram. The unigram language model computes the probability of each term occurring in a given query. Bigram language model describes the probability of occurrence for two sequential terms in query based from a given query. The essence of language model is the ability to capture the probability of occurrence of words in their sequence and retain syntactical representation or lexical meaning hence provides more concept and semantic meaning. For example when using the bi-gram approach the phrase white house’s probability could be computed easily and the meaning be retained intact in contrast to other information retrieval models which retain the word white and then house then tries to make sense out of it.

heng et al[61] uses a combination of Ontology with unigram language model to retrieve an initial document set and improve the precision of top N ranking documents by re-ranking document set and it resulted in a good performance than all runs in NTCIR-3. Wen and Li [62]uses a modified language model to retrieve documents based on phrases and co-occurrence terms through mapping ad agency relation. The distant relation in a document and the results were tested on five TREC test collections, it was observed that language model improves the performance of information retrieval.

As much as the Language Model is close to the normal usage of user queries it has limitations, it assumes that the document and the query are almost the same but this is rare in many cases. The Language Model does not regard the notion of relevance to query content as exhibited by the probabilistic model where it has borrowed its core foundation.

2.2.8. Latent Semantic Indexing Model (LSI) FAQ Information Retrieval

The LSI principle uses a different approach in effecting document similarity compared with all document management Information Retrieval models. Its bases of document retrieval are simply based on the context of documents sharing the same concept and that having similar terms and therefore they relate to the same idea. Such documents are retrieved as relevant to the submitted user query. The approach to LSI is to map every question or document and also user queries into key thematic concepts of the whole corpus. Thus the corpus is analyzed using a mathematical model and key thematic concepts are indentified and user queries related to the concepts. Store documents or questions belonging to those related concepts are retrieved.

2.2.9. Machine Learning Based FAQ Information Retrieval

Machine learning is when a computer system is trained from examples and is able to pick knowledge patterns, trends, relations from the training examples and predicate, classify, cluster responses in concepts of similarity or perform regression analysis. The learning rules are stochastic or symbolic and use mathematical models which exhibit cognitive simulation, parallel processing, probabilistic computations, neuroscience techniques to extract knowledge or discover data patterns in huge databases or user queries[39].

Three common machine learning methods are supervised, unsupervised and recurrent methods. The supervised learning method uses learning algorithms which rely on training examples and to find out from these examples how to generalize the answering of user queries. Unsupervised learning method learns on its own to classify and cluster data into groups of similarity or of same concepts. Reinforcing learning method uses set of defined procedures through examples which facilitate the system to learn and produce answers and reward for good set of procedures to a correct answer.

Various models of machine learning exists and are classified according to the nature of the training rule, and the most common models include Bayesian networks, Hidden Markov, Decision tree, Nearest-neighbor classifier, Artificial Neural Network (ANN), Genetic Algorithms and Support Vector Machines.

2.3. Approaches to FAQ Information Retrieval

The research shall adopt ANN machine learning approach because it mimics the way a human being learns new knowledge and relate to questions in reference to the new knowledge providing generalized responses and specific responses. Michie D [63] state that “Neural network approaches combine the complexity of some of the statistical techniques with the machine learning objective of imitating human intelligence: however, this is done at a more “unconscious” level ...” The Research adopts the supervised learning approach as it shall use existing HIV/AIDS FAQ questions to train the neural network to answer other unforeseen HIV/AIDS FAQ questions. Furthermore it is anticipated some of the user HIV/AIDS questions would be arbitrary or colloquial or full of grammatical errors, the e neural network approach is best renowned for its ability to resolve incomplete and noisy information to its original state through classification and recognition capabilities which emulate a human being thinking and solving mental capacity.

CHAPTER 3

3. Artificial Neural Network Approach

Artificial Neural Network (ANN) is a technology inspired by the human brain functioning mechanism, where biological neurons are interconnected and process stimuli in parallelism to generate knowledge and have an appropriate response[64]. ANN is composed of neurons grouped and interconnected through synapses weight layer which store knowledge. The groups are defined into layers namely the input layer, hidden layer and output layers which form a neural network system architecture. Two types of ANN are defined as feed-forward networks where the neurons and layers are connected without feedback and the recurrent/feedback architecture which features a loop to connect the derived output to the neurons and the preceding layers.

The building block of neural network architecture is an Artificial Neuron also referred to as the node. Practically the ANN is not tangible neither the basic building block the neural node, but however their functionality are best expressed and felt by the mathematical models which relate how they function. Abram comments that “The human brain provides proof of the existence of massive neural networks that succeed at those cognitive, perceptual, and control tasks in which humans are successful”[65] and this suffices as evidence and concrete proof of existence, prowess of neural network in human kind.

The next section shall articulate the mathematical models that explain the functionality of artificial neuron when processing information and extended to artificial neural network architectures when they learn or are trained with domain specific information so that they perform complex nonlinear functions like cluster, classify, predicate, and recognize patterns, associate incomplete data with wholesome data[66]

3.1. Artificial Neuron

An artificial neuron is a basic building block for ANN and is analogous to the biological neuron. Artificial neural networks are made up of layers of interconnected neural nodes. Input to a single node is derived from another node or is a direct input from source data. Each node generates a non-linear function of its input as an output. Output from a preceding node is input to pending node and through such a mechanism, processed information is propagated through tiers of interconnecting nodes. The entire neural network therefore constitutes a set of inter dependent nodes and layers which

features a measurable degree of nonlinearity computing, allowing modeling of complex and nonlinear behavioral functions.

Therefore, mathematical modeling of complex and nonlinear functions using artificial neural is done through complex mathematical calculations that modify neural nodes behavior in a reaction defined per input and output environment. The neural nodes mimic the real biological neurons by exhibiting a learning behavior. The ability for neurons to learn was articulated by Hebb’s learning law in 1949, which explains that when a neural cell A is repeatedly and persistently sub ected to a firing neural cell B, then A’s efficiency in firing B is increased thus implying changes of behavior and depicting some learning[67].

This behavioral process of learning in a neural node is mathematical modeled as illustrated in figure 4 and is computed as computed in equation 13

$$Y(i) = F(\sum_{k=1}^n W_k(i) \cdot X_k(i) + b) \dots \dots \dots (13)$$

where $X_k(i)$ is a data source a given discrete duration i where k goes from 0 to n , $W_k(i)$ is weighted value of the training or input data source discrete duration i where k goes from 0 to n , b is bias factor , F is the transfer function and finally $Y(i)$ is a computed output at duration moment i , The $F(X)$, is an activation function or at times referred to as transfer functions, chosen based on the functionality to be performed by the ANN that is linear function or nonlinear function.

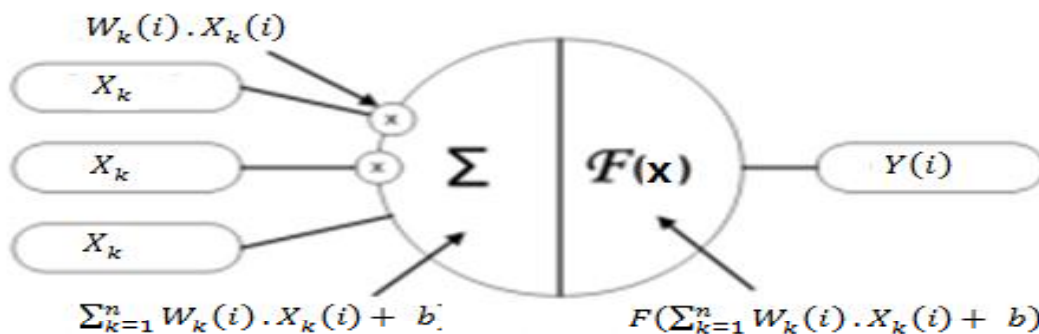


Figure 4: Artificial Neural Neuron Mathematical Model[67]

3.2. Artificial Neuron Activation Functions

An activation function is utilized to transform the activation level of a neuron into an output signal [68]. A variety of common activation functions are used in artificial neural networks (ANN). In any neural network architecture each neuron has an activation function which detects the output of a neuron based on a particular data input per that period.

The research is attempting to determine the mapping of a given question to one or many similar questions which share the same meaning and are stored in a neural network memory thus exhibiting a hetroassociation property. This functionality own its own cannot be done by non linear transformation mechanism. The research has to review techniques and approaches that are able to deal with nonlinear transformation even under the neural activation functions. Karlik and Olgac [69] mentions that most commonly used activation functions for solving non-linear problems include: Uni-polar sigmoid, Bi-polar sigmoid, Tanh, and Radial Bases Function (RBF). The research study reviews these functions to understand which particular function can be adopted for implementing the IHAFR system.

Mathematically the activation function or neural node in neural network can be defined as $Y_i = F(x, w_j)$ where the output Y_i is a product of the synaptic weight layer w_j and the input data x of a neuron i for activating the state or triggering of the neuron i . Karlik and Olgac [69] defines the neural node activation functions qualitatively and comments “The most important unit in neural network structure is their net inputs by using a scalar-to-scalar function called “the activation function or threshold function or transfer function”, output a result value called the “units activation”.”

3.2.1. Uni-Polar Sigmoid Function

The Uni-Polar Sigmoid Function switches outputs either in the positive or negative range hence the name single polarity or uni for example the large inputs that can be squashed be 0 and 1 despite whatever input. Their inputs are an infinite range of numbers. The functionality of the sigmoid is explained by formula 14. This function is normally used for activation output functions in neural networks trained by back-propagation algorithms. Figure 5 illustrates how the bipolar functions as it activates the hidden neuron in between the ranges of 1 to 0 and formula 11 models the functionality.

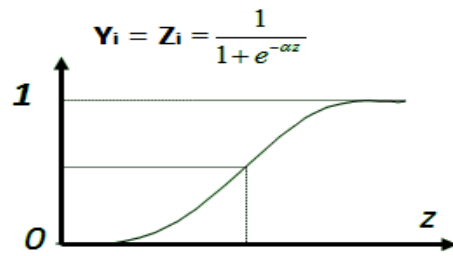


Figure 5: Uni-Polar Sigmoid Function Model

$$Y_i = F(R_i) = \frac{1}{1+e^{-\alpha z}} \dots\dots\dots (14)$$

3.2.2. Bipolar Sigmoid Function

The Bipolar Sigmoid Function is similar to the Uni-Sigmoid function, the difference comes from the output range of the activation function which produce output values in the range of [-1, 1]. This property defines the Bipolar attribute of the activation as it is able to give both positive and negative values squashed in the range of [-1, 1]. Mathematically the Bipolar Sigmoid Function operates as illustrated in equation 15 and figure 6 illustrates the switching ranges for the activation function.

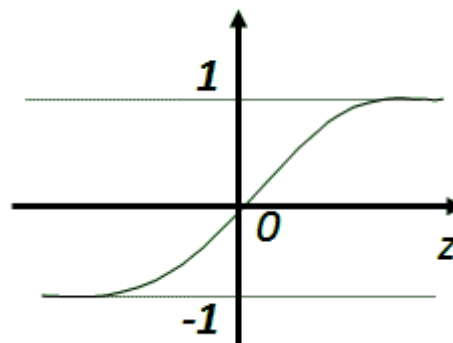


Figure 6: Bipolar Sigmoid Function Model

$$Y_i = F(R_i) = \frac{1V^{-q \alpha z}}{1+e^{-\alpha z}} \dots\dots\dots (15)$$

3.2.3. Hyperbolic Tangent Function

The Hyperbolic Tangent function is a ratio of two trigonometric or hyperbolic sine and the cosine functions which enable to increase the magnitude of the output to be expanded from the range that exceed [1, -1] to another range [N,-N] where the N is an integer number. Alternatively it can be expressed as the ration of half-difference and half-sum of two exponential functions in the points X and X. Mathematically the Hyperbolic Tangent Function operates as illustrated in equation 16 and figure 7 illustrates the switching rages for the activation function

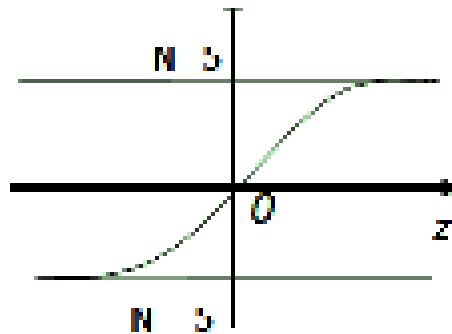


Figure 7: Hyperbolic Tangent Function Model

$$Y_i = F(R_i) = \frac{-N - q \cdot \infty Z}{-N + q \cdot \infty Z} \dots \dots \dots (16)$$

3.2.4. Radial Bases Function Neural Network (RBFNN)

This RBFNN are based on the Gaussian activation function. The Gaussian activation function for RBF networks is primarily based on the mathematical model specified in equation 17 where $k = 1, \dots, N$ and X is the input vector data and the N is the number of hidden neurons, μ_k and Σ_k are the mean and the covariance matrix of the k_{th} Gaussian function.

$$F_k(X) = \exp[-(X - \mu_k)^S \Sigma_k^{-1} (X - \mu_k)] \dots \dots \dots (17)$$

The RBF's are determined by their center and an activation which are considered to be μ_k and Σ_k respectively. The hypersurface represents the covariance matrix which has the diagonal values. The Mahalanobis or Euclidean distance from the center shall be used to determine the activation of the neuron basis function hence[70].

In summary, the behavior of a neuron is changed when it acquires knowledge through a training rule or algorithm that changes the synaptic weight $W_k(i)$ of a neuron node and that the synaptic weight layer constitute as the memory storage device of an artificial neural network. The transfer function serves per neuron in parallel computation serves as the cut-off point to determine what is relevant or not so as to trigger the next neuron hence storing relevant knowledge or discarding.

The research task is mapping a given HIV/AIDS FAQ query or arbitrary question and retrieves a similar in meaning cognitively and then measure relevance and precision of retrieved HIV/AIDS FAQ to the posed query. The recall and precision values shall assist in determining a correct answer for the user since the HIV/AIDS FAQ question is stored together with its answer as illustrated in figure 1. This task is a typical nonlinear mapping functionality and more close to classification therefore the research intends to use nonlinear activation functions to engage proper and accurate processing of similarity matching between posed query and FAQ questions stored

3.3. Artificial Neural Network Architectures

An Artificial Neural Network (ANN) is an information processing system that is composed of neural nodes specified in defined layers. These neural nodes are highly interconnected in layers and from layer to layer are responsible for processing data per required functionality and results are displayed at the output layer. Two main architectural layouts define the interconnection of neurons in an ANN system the feedforward and Feedback architectures.

The importance of understanding and critiquing the architecture of an ANN is primarily based on the influence an architecture bears on the functionality intended as commented by Michie who says "...the complete network therefore represents a very complex set of interdependencies which may incorporate any degree of nonlinearity, allowing very general functions to be modeled"[71]. This prowess of an artificial neural network is also complemented by the ability to store knowledge as weighted links or synapses weighted layers in the multilayered neurons which span from the input layer to the output layers[72]. Such an understanding would enable the researcher to select an appropriate architecture hence a learning process which is used to facilitate learning of the ANN system on domain information to perform predication, classification, association, clustering pattern recognition and many more[73]

3.3.1. Artificial Neural Network Recurrent / Feedback Architecture

Recurrent neural networks were primary designed to work like the way a human brain’s memory work that is the ability to remember or recall an occasion through an association. For instance, one can recognize a familiar pattern or activity through associating it with an experience that once occurred, e.g. remembering all known azz music experienced through a single azz track playing. The human brain habitually associates one event with another as long as they share a commonality. Recurrent neural network because of their feedback property i.e. its outputs linked back to its inputs, they develop a profound capability of learning and recalling with the same demeanor like the human mind.

The Recurrent or Feedback is neural network architecture with a feedback loop and connects each neural network output to its respective input layer at neural node level. There are instance where neuron provides a feedback link to its self. Two key Recurrent Feedback architectural categories exist as the single layer recurrent network and the multi layer recurrent network and these are illustrated in figure 8 and 9. Typical examples of recurrent or feedback neural networks are the competitive networks, Kohonen’s Self Organizing Map (SOM), the Hopfield network and ART networks.

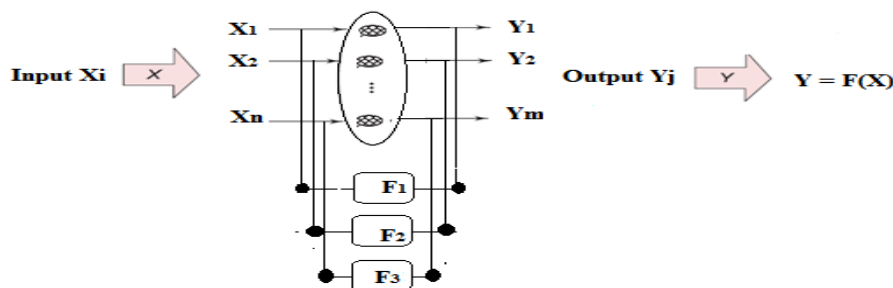


Figure 8: Single Layer Recurrent Neural Network Model

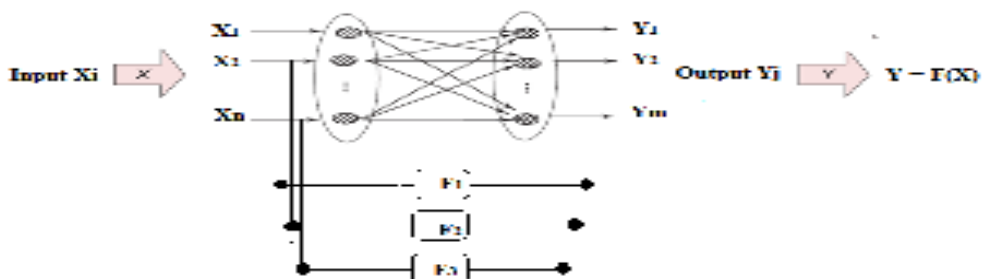


Figure 9: Multi Layer Recurrent Neural Network Model

Recurrent neural networks tend to have a setback of memory instability and limited memory capacity beside their auto-associative memory. In 1982 John Hopfield solved the problem by devising “physical principle of storing information in a dynamically stable network” still there was a limit to what the Hopfield Recurrent neural network could store as defined by equation 18 which states that the maximum memory M_{max} is limited to the number of neurons in the neural network represented by n .

$$M_{max} = 0.15n \dots \dots \dots (18)$$

An auto-associative memory based Recurrent neural network is capable of retrieving same data pattern or image based on an incomplete or very noisy input of the same data pattern but it lacks the inability to associate with lookalike or similar in meaning data pattern. To complement this shortfall inherent in Hopfield networks Bart Kosko proposed a Bidirectional associative memory (BAM) which is a hetroassociative neural network capable of essentially associating inputs to similar or lookalikes stored the in the neural network.

The BAM is a Recurrent neural network with a capability to associate an input pattern on one set of neurons and producing a related, but different, output pattern on another set of neurons. It associates patterns from one set, set A, to patterns from another set, set B, and vice versa. Like a Hopfield network, the BAM has a maximum number of associations to be stored in the BAM and they not exceed the number of neurons in the smaller layer. Another error also inherent with the BAM is the inability to correctly converge to a stable output.

The Kohonen’s Self Organizing Map (SOM) is another type of the Recurrent Neural Networks as illustrated in figure 10. The horizontal interconnections of output neurons for the SOM are used to create a contesting environment amongst and between neurons and the fittest survive as the output for a given input. The horizontal feedback amongst and between neurons connections initiates an excitatory or inhibitory effects, based on the distance from the winning neuron. The model for judging the winner neuron is simply based on the Mexican hat function which explains that the synaptic weights between neurons in the Kohonen layer. In the Kohonen network, a neuron learns by shifting its weights from inactive connections to active ones where the winning neuron and its neighborhood are allowed to learn, If a neuron does not respond to a given input pattern, then learning cannot occur in that particular neuron.

Training of neural nodes in a self organizing network begins with the fittest neighbourhoods of a fairly large size in weights $W_k(i)$ compared with the input data or query X_k . The neuron nodes learn by changing weights $W_k(i)$ from an inactive connection to proactive ones. The fittest neural node and its neighbourhood are allowed to learn and a neural node which does not react to a given data or input pattern, then learning does not occur in it becomes an inactive neuron node.

The Kohonen neural network architecture has a defined number of input nodes which formulate the input layer $X_1, X_2 \dots X_n$ and these translate or map the data into an upper-dimensional level of the Kohonen layer which is defined as the output $Y_1, Y_2 \dots Y_m$.

The competitive learning rule in the Kohonen relates a change in the neural node weight as ∇W_{ki} as applied to synaptic weight W_{ki} by applying the computation as in equation 19, where x_k is the training data and α is the learning rate of the system

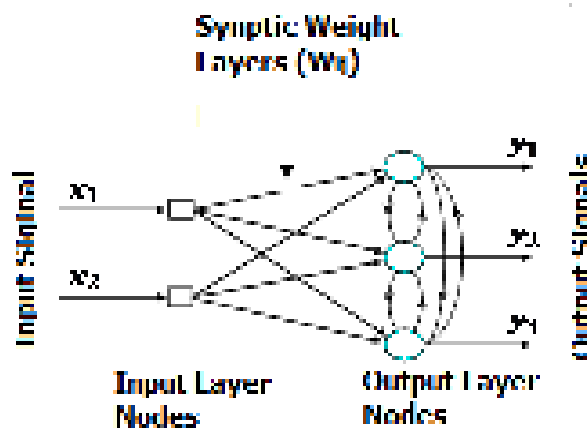


Figure 10: Structural Layout of Self Organizing Neural Network Architecture

$$\nabla W_{ki} = \begin{cases} \alpha (x_k - w_{kj}) & \text{if the neuron node } j \text{ win the competion} \\ 0, & \text{if the neuron node } j \text{ does not win the competion} \end{cases} \dots \dots \dots (19)$$

The principle of operation in computing the synaptic weights is based on the measuring of distance between the input weight x_k and the target fittest neuron node weight j in the output layer using the Euclidean Distance similarity measure d between a given pair of input data and the fittest winning neural node neuron node j as in equation 20

$$d = \|X - W_j\| = \left[\sum_{i=1}^n (X_i - W_{ij})^2 \right]^{\frac{1}{2}} \dots \dots \dots (20)$$

As the training is conducted the neighbourhood size or population of fittest weights gradually decreases as the competitive learning rule computes synaptic weights based on pre-synaptic. The post-synaptic activities and the numerical value of the synaptic weight W_{ki} between two neurons augment if the neural nodes are in the same state and diminish in numerical quantity if they are of different states. To determine the fittest and winning neural nodes in the Kohonen output layer, an input vector X is applied and n stands for the number of neural nodes in the Kohonen output layers. The updating of the synaptic weights is computed as illustrated in equation 21.

$$i_x = \min \|X - W_i\|, i = 1, 2, 3 \dots n \dots \dots \dots (21)$$

$$W_{ij}(s + 1) = W_{ij}(s) + \nabla W_{ij}(s) \dots \dots \dots (22)$$

$\nabla W_{ij}(s)$ is the computed difference between the input data and output neural node that has responded at iteration s and this is determined by the applied competitive learning rule as defined in equation 22.

The commonly used competitive learning rule as stated by [67] is the ‘‘Hebbian learning procedure ...refers to unsupervised learning in which the synaptic strength (weight) is increased if both the source and destination neurons are activated ... the synaptic strength is chosen as:.. where N is the number of neurons of the network accommodating a storage of N patterns. Hebb’s rule always leads to symmetric synaptic coupling.’’ Competitive (Kohonen) learning rule is based on the principle that the output neurons compete amongst themselves to be activated and the one activated at a given time is the winner also known as the winner-takes all neuron or simply the winning neuron. Stochastic rule increments link weights using the probabilistic approach[74]. However the best operational form of the SOM is to cluster and choose the winning neuron without considering an element of generalization or look alike and in this research, this forms a critical dimension of consideration.

3.3.2. Artificial Neural Network Feedforward Architecture

The feed forward networks are categorized into three unique generic classes Single-layer, Multilayer and Radial Basis Function neural networks architectures. Feedforward neural network FF networks are structural organized into one or more layers composed of neural nodes and are all connected and processes information in a forward direction.

The starting point of operation and basis of feedforward neural network is a single layer ANN which has one or more neurons in the layer. The single layer FF has a single weight which between the input and the neuron layer acting as the output with function $y = F(X)$ as illustrated in figure 11.

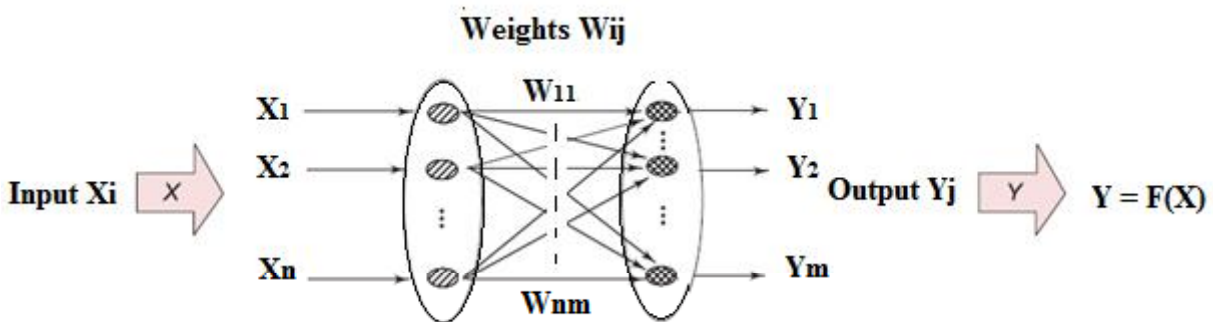


Figure 11: Architecture of Single Layered FeedForward.

Mathematical the single layer FF network accepts inputs $X_1, X_2, \dots, \dots, X_n$ as inputs. Based on the mathematical model of the operation of a neuron each input is multiplied with the weights $W_{11}, W_{12}, \dots, \dots, W_{ij}$ of each respective neuron. The weighted values $X_n \cdot W_{nm}$ are summed $\sum X_n W_{nm}$ and subjected to an activation function that computes similarity between the input and the expected output. This process constitute a learning process as implied by Hebb's learning law in 1949, which explains that when a neural cell A is repeatedly and persistently subjected to a firing neural cell B, then A's efficiency in firing B is increased thus implying changes of behavior and depicting some learning[67].

The mathematical model for a Single FF network is given by equation 23 where B is used a bias function. Equation 24 illustrates the summarized equation for modeling the behavior and functionality of a Single Layer FF neural network

$$Y = F(X) = F(X_1.W_{11} + X_2.W_{12} + X_3.W_{13} \dots \dots \dots X_n.W_{nm} + B) \dots \dots \dots (23)$$

$$Y = F(X) = F(\sum_{n=1}^m(X_n.W_{nm} + B)) \dots \dots \dots (24)$$

Multilayer Feedforward neural network also referred to as a **Multi-Layer Perceptron MLP** is composed of an input layer of neurons, one or more layers of hidden neurons and one layer of output neurons. This hidden layer results in two or more weights layers which function as the knowledge storage devices for the MLP. Each neural node computes a weighted summation of the inputs as stated in the Single layer FF network and this summation is subject to the neuron activation. Mathematical model for the MLP simply encodes the principle of the Single layer FF network by computing each MLP layer output and it becomes the feeder for the next layer until the processed information traverses the entire network and an output is derived at the output layer node. An example of one hidden layer MLP architecture is illustrated in figure 12.

It is clearly noted that two weights layers V_{ij} and W_{jk} being the **Input Hidden Weights Layer** and the **Output Hidden Weights Layer** respectively are created by virtue of the one **Hidden Layer** of neurons V_j . The mathematical model to explain the functionality of the one hidden layer MLP is derived and illustrated below. Equation 28 models the functionality of one Hidden Layer by factoring the inputs of predecessor layers to finally derive the learning or output of the MLP

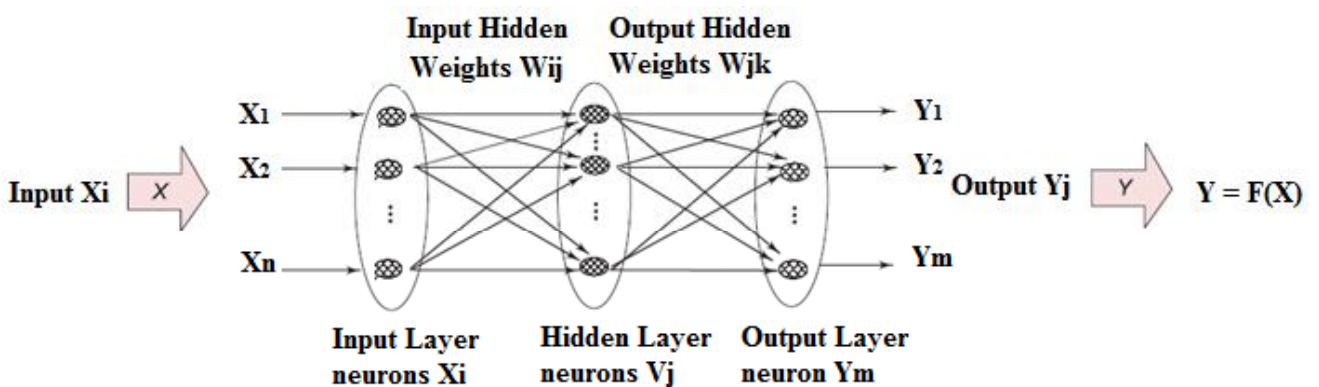


Figure 12: Illustrate the Architectural Concept of Multi-Layered FeedForward Neural Network

$$\text{Input Layer} = Y_1 = F_1(X_n^1.W_{nm}^{1,1} + B_1) \dots \dots \dots (25)$$

$$\text{Hidden Layer} = Y_2 = F_2(Y_1.W_{nm}^{2,1} + B_2) \dots \dots \dots (26)$$

$$\text{Output Layer} = Y_3 = F_3(Y_2 \cdot W_{nm}^{3,2} + B_3) \dots \dots \dots (27)$$

$$F(Y_3) = F_3(F_2(F_1(X_n^1 \cdot W_{nm}^{1,1} + B^1) \cdot W_{nm}^{2,1} + B^2) \cdot W_{nm}^{3,2} + B_3) \dots \dots \dots (28)$$

Radial Basis Function neural networks architecture is another multilayered neural network. It has a similar architecture with MLP architecture; the only difference is the RBF neural network has one hidden layer only which is comprised of neurons that are always apply the Gaussian Radial Basis Function and an output layer with linear activated function neurons as illustrated in figure 13.

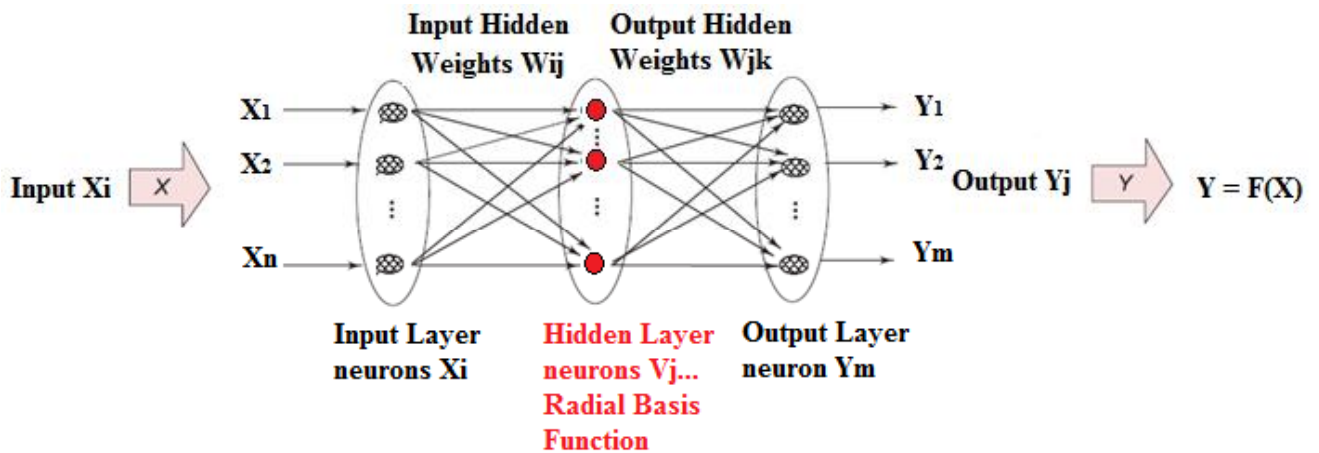


Figure 13: Illustrate the Architectural Concept of Radial Basis Neural Network

The concept of data processing in the hidden layer is based on clustering, implying if the center to the data being processed is known then distance measurement can be done to judge the relevance and a value 1 would indicate a related data. However the main assumption is that the considered area is radial symmetrical around the cluster centre, so that the non-linear function becomes known as the radial-basis function. With reference to figure 14, the mathematical model implies that a radial basis neuron processes an input X_n being a data unit from center T and computed as $\varphi_\sigma(\|X_n - T\|)$. The hidden neuron is very sensitive to data points close to the center and using the Gaussian RBF this sensitivity is ad usted by a variable *sigma* σ . The further away a data unit is, the less sensitivity. When training the neural networks with domain data, it is very critical that special attention is paid to the following matters, the centers T of the RBF activation functions, the spread *sigma* σ of the Gaussian RBF activation functions and manipulation of the output hidden weights layer W_{jk}

Recurrent neural network are very good pattern recognition system in various applications because of the feedback link to the neuron and therefore recalls directly. This feature is directly inferred from their architecture and also from the mathematical model as illustrated in equation 26 and figures 5 and 6. The nature of memory supported by this behavior is called an auto associate memory.

The auto associative memory links directly to what it knows, however it lacks the ability to have attached associations that is generalization. For instance it would be able to associate a question like “Is HV contagious disease” to a question like “Is HIV a contagious disease”. Thus it is able to complete the whole from a given noisy or incomplete input.

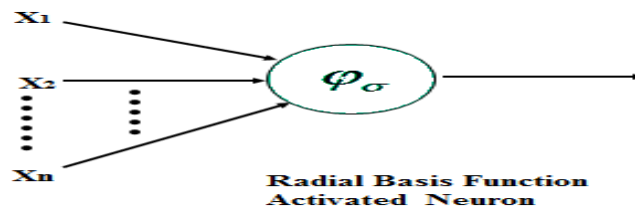


Figure 14: Operational Concept of Gaussian RBF activated neuron in the Hidden Layer

One grievous problem about the Recurrent neural network memory is instability. There is need to establish a proper working mechanism that leads to a performance that can be understood in a rational manner.

$$F: R^+ \rightarrow R^+ \dots \dots \dots (29)$$

Considering the mathematical model analysis of the FF neural network architectures it can be deduced that the FF network perceives an M dimensional space and divides it into respective sections called classes. These classes constitute the domain data categories. An M dimensional space can be properly represented by selecting a correct number of neurons and a proper neuron activation function for the neurons. Mapping in FF neural network architectures could be modeled by mapping a vector R^P to be associated with another vector R^+ in another space thereby exhibiting a property called hetroassociation memory as illustrated in equation 29. This neural network architecture ability relates the property to recall from the memory space of other associated vectors found in another set and sharing the same commonality semantically, contextually or pragmatically as illustrated in equation 30.

$$F: R^P \rightarrow R^+ \dots \dots \dots (30)$$

The set of input patterns are related to a particular class of a given corpus data. Suppose a set of documents D defined as: $D = (d_1, d_2, d_3, d_4, \dots \dots \dots d_N)$. If these documents are processed and into terms per each documents in the corpus and all are defined in a set t defined as: $t = (t_1, t_2, t_3, t_4, \dots \dots \dots t_M)$. The terms are further decoded into a set of term weights: $(w_{11}, w_{12}, w_{13}, w_{14}, \dots \dots \dots w_{JK})$ then a term weight to document matrix sub ected to statistical analysis, latent semantic analysis and finally principal component analysis to induce semantic and conceptually can be created. Each document and term is regarded as a vector in the spatial representation of the term document matrix.

Term-documents matrix are used to train the neural network so that it would create a hetroassociative memory in the weight layers of the feedforward neural network. A specific vector x with some terms $x = (x_1, x_2, x_3 \dots \dots \dots x_K)$ extracted from the term-documents matrix is presented as shown to the neural network in put as shown in figure 15. The Vector which represents an input of query terms is mapped to a document d_1 .

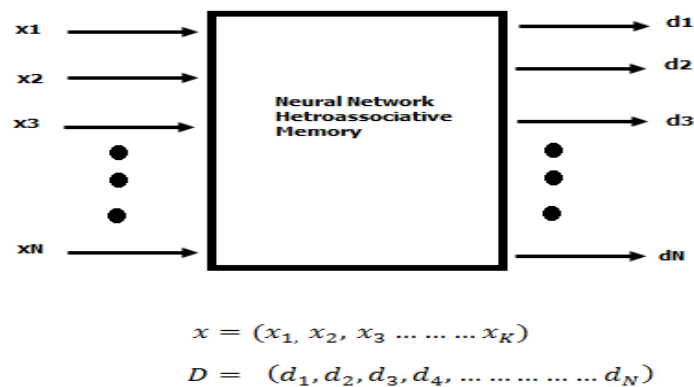


Figure 15: Mapping of Input Vector to a Documents Using MLP Neural Network.

Associative memory permit parallel explorations of stored objects that are similar in meaning or alike based on a given query and retrieve the responses entirely or partially. The process of retrieval is also called recalling and equation 31 defines the mapping function. In hetroassociative memory the input vector of x_K provides the document stored vector d_N , the mapping is distributed in the network. The mapping functions through a distributed arrangement of intense connections, feedback with or without nonlinear activation functions and retrieval algorithm process the content of the input vector to derive an output vector d_N

$$d_N = M(x_K) \dots \dots \dots (31)$$

In this research a single FAQ bears an element of uniqueness though it retains a degree of relatedness through conceptual, lexical and semantic properties to other FAQs in the corpus. In this instant we can say the HIV/AIDS FAQ corpus has 0 classes and the classifier should map a given input FAQ to 0 classes by defining the degree of similarity between the input and output. The output class 0 with the highest value denotes the class in which the input HIV/AIDS FAQ query is more similar or related to.

The research finds the Feedback Forward neural network architecture as an appropriate architecture for implementing mapping of a given HIV/AIDS FAQ query to a stored HIV/AIDS FAQs stored in a trained neural network based on the mathematical analysis. General literature review show that FF neural networks are very good approximators and have been widely used in all generic functions of pattern classification or mapping.

Feedback Forward neural network architecture system is able to bring an association between questions like “Is HIV a contagious disease” which is semantically and conceptually similar to the question “Can HIV be infectious”. For such a mapping the Feedback Forward neural network architecture is able to compute a very good degree of relevance and measure of similarity. The Single layered FF neural network architecture is not capable of resolving complex nonlinear functions compared to advanced architectures like the MLP and the RBF. However the RBF has the disadvantage of using many hidden neurons compared with the MLP, thus making the making the architecture complex and induce an over computation payload

3.4. Artificial Neural Network Training Algorithms

Learning is an essential ability of neural networks to learn through learning rules or algorithms which are used to determine suitable weights for neural network weight layers. ain et al [73] comments that learning is an integral aspect of intelligence, hence an artificial neural network needs to be trained so that it gets the relevant intelligence in a specific area. The training algorithm plays a critical role in determining how knowledge in an ANN is stored. Selection of a training algorithm depends on the intended task of the ANN.

Training algorithms or learning rules compute the weights in links connecting the neurons as they acquire new knowledge or store the knowledge[67]. Abraham et al[65] define a training algorithm or learning rule as a procedure for modifying the weights and biases of neurons in neural network so that they can store knowledge. Learning of any neural network is perceived as a nonlinear optimization problem for finding a set of network parameters that minimize the cost function for given examples belonging to a problem to be solved[75]. If a neural network is trained with examples of that domain knowledge then the neural network embodies a complex relationship that is characterized with the ability for generalization of responses when queried with any other knowledge from that specific knowledge. This property depicts the ability of neural network to exhibit the semantic, conceptual and pragmatic similarity measure that should be possessed by an intelligent retrieval system.

Training a neural network is determined best by first denoting the cost function which is then used to treat the training or learning of a neural network task as a standard parameter estimation process. The process of parameter estimation is carried out by training or learning algorithms and which are largely classified as supervised training, unsupervised training and reinforcement learning.

ain, et al. [73] further articulate reinforcement learning as a variant of supervised learning “...the network is provided with only the critique on the correctness of the network output, not the correct answer themselves” and unsupervised the network “... does not require a correct answer associated with each input pattern in the training data set” as illustrated in figures 16 and 17.

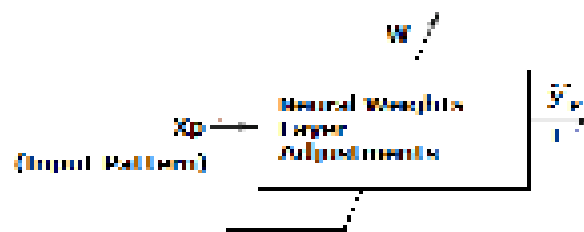


Figure 16 Unsupervised Learning Model

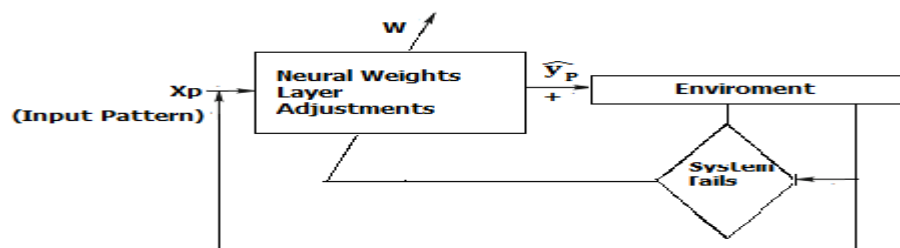


Figure 17: Reinforcement Learning Model

Supervised learning is widely used in classification, approximation, control, modeling and identification, signal processing, and optimization[75]. Unsupervised learning schemes are mainly used for clustering, vector quantization, feature extraction, signal coding, and data analysis[75]. Reinforcement learning is usually used in control and artificial intelligence[75]

ain, et al. [73]describes supervised learning as a learning process where “...the network is provided with a correct answer (output) for every input pattern ... weights are determined to allow the network produce answers as close possible to the known and correct answer” as illustrated in figure 18. Input pattern X_N constitute the input and the expected output is Y_P here explained as the correct answer or similar in meaning to input pattern. A general description of the supervised learning model notes the difference between input pattern X_N and output pattern Y_P to generate an error output E_N which shall be used to ad ust the neural network synaptic weights layers W until a minimal error E_N is attained based on the cost function implemented by the supervised training algorithm

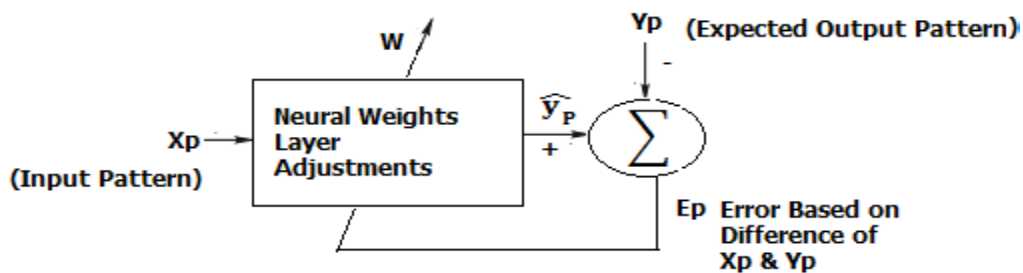


Figure 18: Supervised Learning Model

3.4.1. Perceptron Learning Rule

ain et al [73] mentions two renowned supervised training learning rules also termed as the error-correction learning rules, the Perceptron learning rule and the Backpropagation learning rule. The Perceptron Learning rule uses the Least Mean Square (LMS) cost function to compute the error between the input pattern and the out pattern. The LMS cost function is utilized as an adaptive filter which iteratively computes and models a relationship between an input pattern and output pattern to generate defining coefficient based on an algorithm and self-ad usts the filter coefficients to reflect the least error between an input signal and the output signal. A small marginal error would imply a very high similarity between the input pattern and the output pattern. However the key disadvantage with LMS is slow convergence due to eigenvalue spread and more so its suitable for liner regression analysis or linear task operations which do not involve complex and nonlinear functions

3.4.2. Backpropagation Training Rule

Backpropagation was derived by generalizing the Perceptron learning rule to apply on MLP networks which can resolve nonlinear functions through use of nonlinear activation functions like the sigmoid and other variants. There are a number of variations on the basic algorithm that are based on other standard optimization techniques, such as conjugate gradient and Newton methods.

The Backpropagation training rule uses the Mean Square Error (MSE) as a cost function to compute the error between the input pattern and the output pattern. Like the perceptron learning rule, the Backpropagation is a typical example of supervised training where the learning algorithm is provided with a set of examples of desired network behavior defined as follows $[(X_1, Y_1), (X_2, Y) \dots \dots (X_p, Y_p)]$ where X_p is the input pattern and Y_p is the target pattern or the expected output pattern as depicted in figure 12. As each input is applied the differential error E_p of input pattern X_p and Y_p minimize the average sum these factors as defined in equation below.

$$\text{MSE} = \frac{1}{P} \sum_{s=1}^P E(s)^2 = \frac{1}{P} \sum_{s=1}^P E(Y(s) - X(s))^2 \dots \dots \dots (32)$$

Statistically, the mean squared error (MSE) measures iteratively the average of the squares of error between the intended target (output pattern) and the given target (input pattern) until the lowest value is attained and this measure depicts how close the similarity between the target and the intended target is. Thus the MSE is a cost function, which computes matching of the target, output pattern or vector and input pattern or vector target by deriving the squared error loss or quadratic loss.

Correctly trained neural network using the Backpropagation training rule provides a practical output patterns which are almost similar or have a generalized similarity to the presented with input patterns which have never been “seen” by the neural network before. This generalization property exhibited by training is attained through a training process where a neural network is trained with a few representative input patterns or vectors/ output patterns or vectors pairs for the domain knowledge and thereafter obtain reasonable outcome patterns without training the network on all possible input/output pair and good training rule should empower the neural network with such a property.

Competitive (Kohonen) learning rule is based on the principle that the output neurons compete amongst themselves to be activated and the one activated at a given time is the winner also known as the winner-takes all neuron or simply the winning neuron. Stochastic rule increments link weights using the probabilistic approach[74]

3.5. FAQ Retrieval Using Neural Network

The neural network documents retrieval developed by [72] uses MLP ANN architectures. The first neural network is an MLP architecture which is trained with Backpropagation learning rule to process arbitrary user query words into domain keywords. The second auto-associative Hopfield net is a one layer processing ANN with auto-associative memory and query expansion capability using spread activation mechanism. The architecture is a Hopfield net with a capacity to reconstruct an incomplete input pattern to a complete pattern by using generated key words derived from the first neural network and output relevant documents. The Vector Space Model was used to represent the representative of arbitrary keywords to domain keywords for the first neural network and used to train the neural network as well. The binary VSM was used to represent the domain keywords to document for the Hopfield network and also training.

Pandey et al [26] implemented a neural network system that has an ability of identifying sentences of similarity which have lexicon grammatical diversity using a single layered binary recurrent Hopfield ANN. The first Hopfield network maps the distorted sentence words into their proper patterns, and the second binary Hopfield ANN uses the assembled keywords to identify a similar sentence. The ANN question representation is in binary and therefore system uses the one nearest neighbor classification rule. The system resolved word sense ambiguity, usage of phrases, dates, numbers, incomplete words and other special characters. Systems managed to convert input sentences to corresponding sentences with a precision of 92.2 %

Mital and Gedeon [76] Developed a neural network retrieval system based on a multilayer feed forward architecture and neurons representing every word and document in the legal Hypertext Assembly FAQ questions. The neurons in the input layer representing input keywords have a spread activation link to neurons in the output layer representing the documents. An initial computed weight value called Text-Associative link represents the domain knowledge and is computed from the word document frequency in the legal Hypertext Assembly. The supervised learning method is used and Backpropagation rule is used to train the ANN.

Des ardins et al [77] designed a self-organizing map with unsupervised neural network where neurons compete to classify the input query to various concepts or clusters in a visual grid or map. The connection weight maps documents in groups or clusters of the same similarity based on the input query. The input vectors or query is represented using term weights. The system used unsupervised learning and the Competitive (Kohonen) learning rule.

Chen et al [78] makes a distinction between Knowledge Based Similarity Measurement systems (KBMS) [37, 55] and Corpus Based Similarity Measurement learning based systems (CBMS) [72, 79] . They imply that the former acquire knowledge from human experts and perform as dictated and the later learn from examples or data sources using learning algorithms to discover knowledge and provide answers. Knowledge based systems use resources like ontology knowledge base or FAQs knowledge base containing indexed questions or Natural Language Database (NLDB) which stores answers in tables which are related or use graph-based techniques to store and represent the unstructured knowledge.

Sahay et al [80] implemented a precise search over medium-sized knowledge bases using constantly asked questions on medical domain knowledge by using the Case Based Reasoning approach in conjunction with the Hopfield network over a web platform. The medical knowledge base used graph based techniques to represent the information. The Hopfield neural net was used as an auto-associative tool to search for the main documents by using few basic keywords in the user query (an incomplete document) to retrieve wholesome and complete documents from the medical knowledge base. Anderson et al[81] noted a discrepancy for KBMS systems and remarked that it is impossible to provide all the answers needed to all questions provided especially in a dynamic environment where new knowledge is acquired continuously. This phenomenon renders the knowledge base incomplete hence no answers can be provided to user satisfaction.

Corpus Based Similarity Measurement implement artificial neural networks to learn pattern and trends in a data with a representative sample of input/output pairs is capable providing and increasing levels of automation in the knowledge engineering process, replacing much time consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data[82]. This approach which is also termed Learning Similarity Measurement [27] technique learn from training examples of the domain specific corpus induce

semantic, conceptual and embraces static knowledge. Neural network model is able to retrieve information by generalizing i.e. predicting new answers from the provided examples because the training data is preprocessed using Latent Semantic Analysis through the SVD model which tends to create thematic knowledge groups based on the content corpus or the use of Principal Component Analysis (PCA).

The primary advantage of using neural network approach as donated by [82] is that they are adaptable and nonparametric; predictive models that can be tailored to the data at a particular site. Artificial Neural Network has a high tolerance to noise and incomplete data, i.e. robust and fault tolerant to incomplete queries, noisy queries etc, and they have a high processing speed because of parallelism computing ability of neural nodes and also the architectures that can host properties like auto-associative memory and hetroassociative memory.

The suitability of machine learning also suites our problem research because we want to adopt an “intelligent behavior” emerging from the local interactions that occur concurrently between the numerous network nodes through their synaptic connections, i.e. emulating how the human brain process queries to resolve a given problem and further generalizing the resolution[79].

The research proposes to adopt the MLP layout because of their ability to handle and resolve nonlinear tasks through imparting the relevant intelligence by using a learning rule like Backpropagation. Another ideal property for the MLP is the use of the hetroassociative memory property which is the ability to resolve an input and associate it with similar or look likes, a very critical task in our research because the system is supposed to find an equivalent or similar in meaning HIV/AIDS FAQ pair based on a single posed query. This property is a key requisite for any classification or mapping system the ability to generalize responses.

Using the MLP choice implies using a nonlinear activation function in particular the uni-polar sigmoid function because a similarity measure is done between of document or FAQ question d_n and user query V_m is defined $im(d_n, V_m) = [0 \rightarrow 1]$ thus a positive range of values from 0 to 1. The query and document are not similar if the similarity measurement is 0 and is 1 if the query and document are 1 and any value in between represents a degree of similarity reflecting generalization.

Supervised training approach shall be used and general literature review notes that the training algorithm is normally used together with the Backpropagation training rule. The Backpropagation shall be used to train the IHAFR system with a sample of HIV/AIDS FAQs from a compiled corpus FAQs on HIV and AIDS. The Intelligent HIV/AIDS FAQ Retrieval System architectural layout is therefore proposed as in the next section.

3.6. Intelligent HIV and AID FAQ Retrieval Using Neural Networks (IHAFR) Architecture

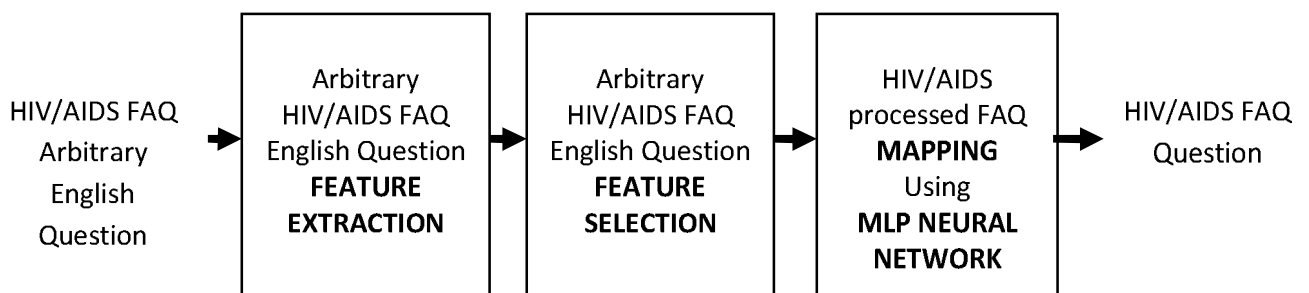


Figure 19: Intelligent HIV/AIDS FAQ Layout

Figure 19 illustrate the structural layout proposed for the IHAFR. The system shall accept arbitrary HIV/AIDS queries and subject to feature extraction process that entails stop word removal, stemming, term-weighting and creating the VSM for representing the term to document or FAQ questions for HIV/AIDS FAQs. The term frequency and inverse document frequency shall be adopted as term weight technique for both the query and HIV/AIDS FAQ question.

The created VSM is further processed for latent semantic analysis using the mathematical model SVD using Principal Component Analysis (PCA) to discover thematic keys words which bring about the semantic and conceptual analysis in the HIV/AIDS FAQ questions. The end result is a PCA matrix T_{PCA_MATRIX} which has thematic words against the total number of documents or HIV/AIDS FAQ questions in the corpus expressed as in matrix of equation 33. The P_n represents the thematic word ranked up to the least significant thematic word based on some heuristic or experimental approaches which has the PCA value W_{nm} to represent relevance of the document or FAQ question FAQ_m .

$$T_{PCA_MATRIX} = \left\{ \begin{array}{cccccc} & FAQ_1 & FAQ_2 & FAQ_3 & FAQ_4 & FAQ_m \\ P_1 & W_{11} & W_{12} & W_{13} & W_{14} & W_{1m} \\ P_n & W_{n2} & W_{n3} & W_{n4} & W_{n5} & W_{nm} \end{array} \right\} \dots\dots\dots (33)$$

Thematic words P_n representing the FAQ question would form a vector of which some shall be used as training samples for the neural network as inputs to adjust the weight of the MLP to map or classify a query to the appropriate FAQ question. The next chapter explains the implementation of the process.

CHAPTER 4

4. The Artificial Neural Network Development

The chapter 4 describes particular experimental tasks taken to implement the Intelligent HIV/AIDS FAQ Retrieval System (IHAFR) using a neural network. The implementation should map or classify an arbitrary HIV/AIDS query to a correct HIV/AIDS FAQ question stored in the IHAFR knowledge base. The chapter also inform on sourcing and compilation of HIV/AIDS FAQ questions which have been prescribed, approved, published by medical and associated experts. By the time of writing this research HIV/AIDS FAQ corpus was not many but however the research adopted FAQ questions from MASA, IPOLETSE and United Nations World Health Organization, and created an electronic HIV/AIDS corpus.

The chapter also details on the procedures taken to resolve and deduce an electronic HIV/AIDS electronic knowledge base of term weighted HIV/AIDS key words to their respective HIV/AIDS FAQ questions using a term - document matrix. The outcome was a Vector Space Model and Principal Component Analysis Matrices for HIV/AIDS FAQ using term frequency and principal component factors or thematic points for HIV/AIDS FAQ terms to their respective FAQ questions.

The chapter also reflects on the key procedures and outcomes taken to experimental determine the operational parameters of the IHAFR neural network. Key factors included in the determination of the operational parameters are number of input neural nodes, hidden neural nodes, output neural nodes. Other factors reviewed were the activation function, maximum training cycles, the appropriate MSE cost function value to use and finally the appropriate training Backpropagation learning rule. It is noted from general literature review that so many variants for the Backpropagation rule exists and an appropriate Backpropagation needs to be identified. The research study also performed a validation of the IHAFR neural network using MATLAB functions like linear regression analysis test and generalization test to determine whether the neural network performs per intended specifications.

Appropriate evaluations of the IHAFR are done to authenticate the relevance of all HIV/AIDS FAQ questions mapped per user HIV/AIDS user queries. Judgment of relevance and accuracy was done in conjunction with the answer per each classified or mapped HIV/AIDS FAQ question. The approaches taken are the golden standard rule or the grounded truth was used to evaluate the output of the standard and traditional information retrieval system used as benchmark against the IHAFR.

4.1. HIV/AIDS FAQ Questions Source and Selection

The MASS, IPOLETSE and United Nations World Health Organization HIV/AIDS FAQs were used to compile build a corpus of correct HIV/AIDS FAQ questions and their answers. All the questions were allocated an identity number. A typical extract of an HIV/AIDS FAQ question from the IPOLETSE HIV/AIDS FAQ booklet and its answer arrangement is illustrated below.

190 Can I breastfeed my baby if I have HIV?
 No. HIV can be transmitted through breast milk, so it is not safe to breastfeed your baby if you are HIV-positive. You could infect, or reinfect, your baby with HIV.

Techniques used to build an electronic knowledge base of HIV/AIDS FAQs for the IHAFR were document reduction an approach demonstrated and explained section 4.2. Techniques like text feature extraction, feature selection and dimension reduction are used to create the Vector Space Matrix of Key Words for every FAQ question in the corpus and corresponding HIV/AIDS FAQ question using a AVA algorithm. Figure 20 gives a summary of the generated HIV/AIDS FAQ questions and the keyword list VMS.

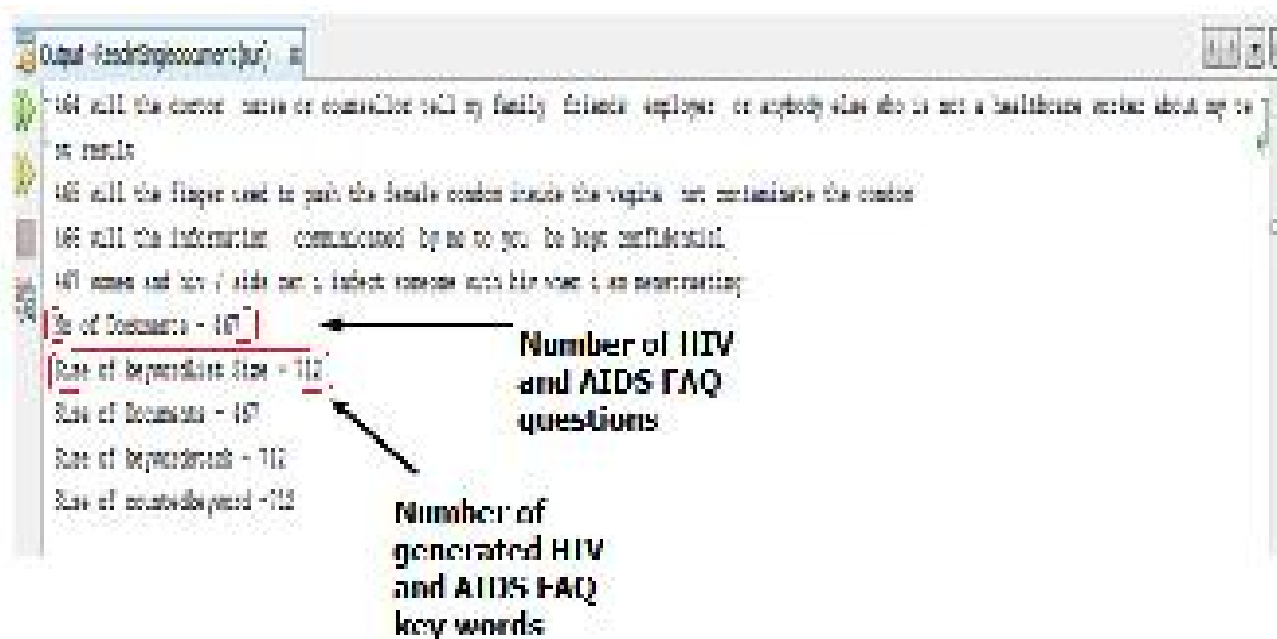


Figure 20: Total number of HIV/AIDS FAQ questions and the Key words [1]

4.2. HIV/AIDS Questions Processing (Document Reduction)

Document reduction is the processes whereby a document is removed irrelevant and redundant features. Khan et al [83] mentions two key steps meant to attain document reduction as feature extraction and feature selection.

4.2.1. Feature Extraction of HIV/AIDS FAQs: Experiment I

Feature extraction involves techniques like tokenization, stop word removal and stemming. The research has used feature extraction techniques to “clean” HIV/AIDS FAQ questions used to train and create knowledge in the IHAFR.

4.2.1.1. Tokenization

Tokenization, a Java String Tokenizer has been used to split the words in a query or question and create a string array for each HIV/AIDS FAQ question. For example, an HIV/AIDS FAQ question “what are my responsibilities as an HIV infected person?” is demarcated to tokens as illustrated in figure 21. The right side shows the question and its ID as FAQ number 296 and the left side shows the tokenized question where each tokenized word still retains the FAQ ID.

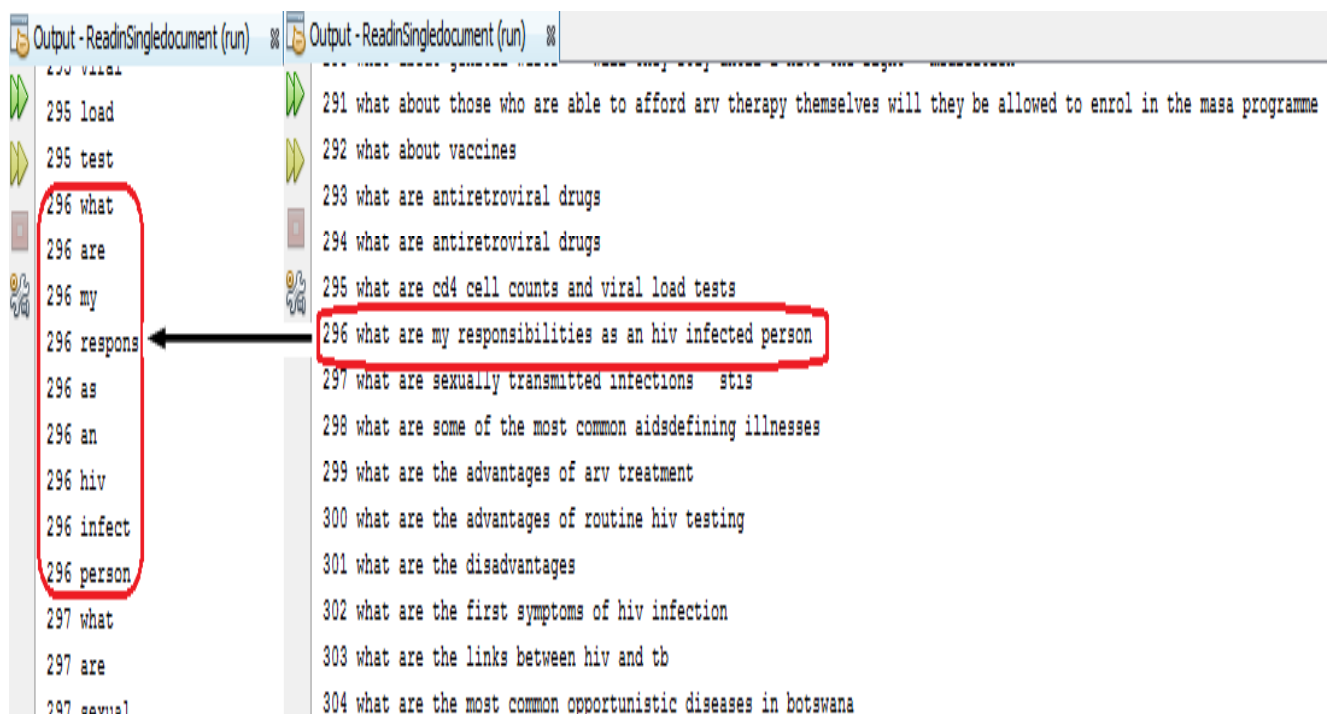


Figure 21: Tokenized HIV/AIDS FAQ Questions and its Tokens.[1]

4.2.1.2. Stop Word Removal

Stop word removal is a technique used to remove noisy words or common words in a document. Common words do not assist in identifying and discriminating a document from other documents in a system repository. HIV/AIDS FAQ questions were processed as shown table 1 and 2 below. The HIV/AIDS FAQ questions are tokenized and stop words removed. Highlighted rectangles show tokens per each processed HIV/AIDS FAQ or query. Key words from processed HIV/AIDS question or query and are retained as shown in table 2. Stop words removed for instance words like I, will, Shall etc, they form what is termed as noisy words. However words like What, How, Where, What, Who, Why, When, Do and Is although they are defined as key words in this research we retained them as key words.

Table 1: Tokenized and Stop Word Removed AIDS FAQ Questions[1]

HIV/AIDS FAQ Question	Tokenized & Stop Word Removed
1 what is hiv	1 hiv
2 what is aids	2 aids
3 how does the hiv test work	3 hiv test work
4 what are my responsibilities as an hiv infected person	4 responsibilities hiv infected person

4.2.1.3. Stemming:

Stemming reduces a word to its generic or root word. AVA Lucene Stemmer was used to reduce the tokens to their stems. The ava class snowball analyzer with English language as a selected language was used for stemming. Figure 20 shows all stemmed tokens of the HIV/AIDS FAQ questions collection and a total of 712 stemmed words from 467 HIV/AIDS FAQ questions in the collection were stemmed. The HIV/AIDS FAQ questions stemmed questions are given as example in table 2 below.

Table 2 Examples of Stemmed Words for the HIV/AIDS FAQ questions[1]

HIV/AIDS FAQ Question	Tokenized & Stop Word Removed
1 what is hiv	1 hiv
2 what is aids	2 aid
3 how does the hiv test work	3 test hiv work
4 what are my responsibilities as an hiv infected person	4 respon hiv infect person

4.2.2. Feature Selection of HIV and HIV FAQs: Experiment II

General documents in a given corpus exhibit properties like term frequency per document $tf_{t,d}$ being a term weight measuring number of times a token appears in a document, df_k relates number of documents bearing a term, idf_k explains term weight related to rarity of a term in an entire collection of documents, (TF – IDF) describes normalized appearance and rarity of a term weight in a document collection. Feature selection involves using methods term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TF-IDF), many others to represent text into numerical values. Vector Space Matrix (VSM) depicts a space visualization and representation of documents in their spatial form. Documents key words are aligned to corresponding documents in spatial text or numerically. Principal Component Analysis (PCA), Latent Semantic Indexing (LSI), information gain (IG) and many others are used to “compress” key words with similar meaning into a key thematic term and these are document reduction techniques.

The research adopted mathematical models for implementing feature selection techniques (TF – IDF), document length normalization using cosine normalization, VSM and PCA to represent HIV/AIDS FAQ questions and key words in numerical and statistical forms as explained below.

4.2.2.1. HIV/AIDS FAQ Key Word using Term Frequency and Inverse Document Frequency

Term frequency gives a count of how many times does a term appear in a given document within a given document collection. This measure tends to give an indication of the relevance of that particular term to the entire collection of documents therefore measuring a property of relevance which is annotated as $tf_{j,k}$ and calculation is based on equation 34. Disadvantage of term frequency is it does not consider the importance of term in a collection which is a good measure to select relevancy of documents.

$$tf_{j,k} = \frac{N_{jk}}{\sum_k N_{jk}} \dots \dots \dots (34)$$

The property inverse document frequency idf_k measures the rarity of a term in the entire document collection. idf_k points to the importance of a term to discriminate all other documents and select relevant documents. Computationally it is calculated as given in equation 35 .

$$idf_k = \text{Log} \left(\frac{N}{df_k} \right) \dots \dots \dots (35)$$

What is clearly noticeable is that as the term occurrence value gets small the idf_k gets high i.e. pointing to rarity of a term and equally the same as it rises towards the total number of documents in the collection N gets to zero, indicating none commonness of the term. To mitigate the shortfall of $tf_{j,k}$ and idf_k , we have adopted (TF and IDF). This term weight is used to measure both properties of term frequency and term rarity. Calculation of $TF - IDF$ is computed as in equation 36.

$$W_{jk} = TF - IDF = tf_{j,k} \times idf_k \dots \dots \dots (36)$$

The overall document FAQ_j weight is the summation of term weights W_{jk} in the documents and is computed as given in equation 37.

$$FAQ_j = W_{11} + W_{12} + W_{13} + \dots W_{jk} \dots \dots \dots (37)$$

4.2.2.2. HIV/AIDS FAQ Question Length Normalization

Some term weights W_{jk} for the HIV/AIDS FAQ questions (document) in the collection have been observed to have values that are greater than 1. This phenomenon could be attributed to variation of question lengths. Questions which have long lengths i.e. have many unique and repeated terms compared to question with few terms are bound to generate a cumulative and immense term weight. Therefore the research adopted method to normalize all documents in the corpus using a length normalization technique.

Several document length normalization techniques like the cosine normalization, maximum to normalization and byte length normalization are used to rationalize and standardize the document term weights to fall within a required range of value. The cosine normalization was adopted because the technique is already used to compute all discussed term weight parameters so far i.e. the $tf_{j,k}$ and idf_k . A cosine normalized term weight is computed as in equation 38.

$$W_{jk(nCAmP_iTRI)} = \frac{Xf_{jp} \times \text{Log} \frac{N}{tf_k}}{\sqrt{\sum_{k=1}^t (tf_k)^2 \left[\text{Log} \left(\frac{N}{df_k} \right) \right]^2}} \dots \dots \dots (38)$$

4.2.2.3. Vector Space Model

The principle to visualize representation of documents as mapped vectors in a space is attributed to the work of Salton's work[84]. Vector Space Model (VSM) relates term weight W_{jk} as a measure of importance for term j of document k . A Vector Space Model can be computed for each term weight W_{jk} calculated as in equation 38 were values like tf_{jk} , df_k and N are used. In our research study each tokenized term for each HIV/AIDS FAQ question is computed and represented as VMS matrix as shown in equation 39.

$$A_{(txd)} = \begin{pmatrix} W_{11} & W_{12} & W_{13} & W_{14} & W_{15} & W_{16} & \dots & W_{1d} \\ W_{21} & W_{22} & W_{23} & W_{24} & W_{25} & W_{26} & \dots & W_{2d} \\ W_{t1} & W_{t2} & W_{t3} & W_{t4} & W_{t5} & W_{t6} & \dots & W_{td} \end{pmatrix} \dots \dots \dots (39)$$

Vector Space Model is used to compute clustering, classifications and similarity measures such as cosine in other linguistic processing applications. Turney et al [85] comments that machine learning algorithms like neural networks can work with real valued vectors from VSM in performing linguistic processing so that they are able to do their own internal similarity measures. A Classical Vector Space Model for the HIV/AIDS FAQ corpus was created using a AVA code and the key words were weighted using the TF-IDF term frequency mathematical model as in equation 38. Each FAQ question could be represented as an HIV/AIDS FAQ vector as demonstrated in equation 34.

From the experiment conducted 712 key term weights for HIV/AIDS FAQ questions were computed and aligned to corresponding and constituting columns of the 467 HIV/AIDS FAQ questions. A classical Vector Space Model for HIV/AIDS FAQ questions and corresponding Key weighted terms for every FAQ question in the corpus was created and the VSM of HIV/AIDS 712 X 467 was generated and numerical represented as normalized.

The created VSM HIV/AIDS FAQ question had values ranging from 0 to 1. A value 0 indicates zero weight values relevance to the question and 1 means it has a strong value in the question selection. Values found in between expresses a level of weight relevance between a term and the

FAQ question. A summation of the weight terms in column reflects the question overall weight relevance to the FAQ question corpus. It has been observed that the created VSM HIV/AIDS FAQ is a highly sparsed matrix i.e. has a lot of zeros in the cells and it also imperative for the research to establish and determine the thematic words that bring about the semantic and conceptual relationship between the HIV/AIDS FAQ questions and their key words. This can only be done by using the SVD mathematical model which brings this latency to clarity by defining the thematic words.

4.3. Dimensional Reduction Techniques with Principal Component Analysis: Experiment III

The VSM HIV/AIDS FAQ generated is a sparse matrix of 712 rows (key terms of HIV/AIDS FAQ questions) by 467 columns (HIV/AIDS FAQ questions). Wikipedia [86]cite Stoer & Bulirsch 2002 by defining sparse matrix as a matrix primarily populated with zeros as elements of the matrix. Rosario [87] mentions of 0.00 to 0.002% as non zero entries of a sparse matrix that had 90,000 terms and 70,000 documents in TREC evaluation. This percentage magnitude gives a picture of what sparse matrixes are in terms of zero entries. To solve the issue of sparsity, the research considered to implement rank lowering method. The approach is a reduction technique and is more focused on dealing with sparse matrix.

The research adopted a Principal Component Analysis (PCA) to implement dimension reduction of the HIV/AIDS FAQ VSM sparse matrix with a view to remove ‘noisy’ zeros and also discover latency of concepts embedded in terms and documents in the HIV/AIDS corpus. Principal Component Analysis (PCA) is a method used to analyze variables of given corpus and deduce key thematic factors which portray data variance between questions and terms for purposes of retrieval. The thematic factor is considered to be a principle component, since it has captured all variants of the data under a key point. Can [88] describes PCA as a statistical tool that interprets variations of data and then reduces to meaningful key points. Basically, PCA is formed from SVD on the covariance matrix [89]. Principal Component Analysis is used to reduce multidimensional datasets to lower dimensions for analysis[90].

Turney et al [85]describes rank lowering as an approach that is ideally used on sparse matrix or huge matrix as a way to achieve latent meaning, noise reduction, high-order co-occurrence and sparsity reduction. Wikipedia [86] cites typical instances were rank lowering methods, as when matrix is.

Too large for the computing resources, therefore an approximated low rank matrix should be considered

Noisy thus anecdotal instances of terms must be eliminated and the result would be de-noisified matrix.

Overly sparse thus latent meaning could be used to use concepts between documents and terms

4.3.1.Principal Component Analysis via Eigenvector and Eigenvalue

There are many approaches to the PCA computations for a given VSM matrix representing a term to document of words in a corpus then using eigenvectors and eigenvalue the PCA components can be computed as demonstrated. This approach requires the normalization of the VSM Matrix $A_{(txd)}$ by deriving a mean Matrix X so that the mean point is in the origin.

$$A_{(txd)} = \begin{pmatrix} W_{11} & W_{12} & W_{13} & W_{14} & W_{15} & W_{16} & \dots & W_{1d} \\ W_{21} & W_{22} & W_{23} & W_{24} & W_{25} & W_{26} & \dots & W_{2d} \\ W_{t1} & W_{t2} & W_{t3} & W_{t4} & W_{t5} & W_{t6} & \dots & W_{td} \end{pmatrix} \dots \dots \dots (40)$$

The mean point is calculated as in equation 41 where the column term weights are aggregated and a mean value μ_n derived. This μ_n is subtracted from each column value as in equation 42. The covariance between columns in matrix X is derived by equation 43.

$$\mu = (\mu_1 \mu_2 \mu_3 \mu_4 \dots \dots \dots \mu_n) \text{ where } \mu_n = \frac{1}{m} \sum_{k=1}^m W_{td} \dots \dots \dots (41)$$

$$X = \begin{pmatrix} W_{11} - \mu_1 & W_{12} - \mu_2 & W_{13} - \mu_3 & W_{14} - \mu_4 & W_{15} - \mu_5 & W_{16} - \mu_6 & \dots & W_{1d} - \mu_n \\ W_{21} - \mu_1 & W_{22} - \mu_2 & W_{23} - \mu_3 & W_{24} - \mu_4 & W_{25} - \mu_5 & W_{26} - \mu_6 & \dots & W_{2d} - \mu_n \\ W_{t1} - \mu_1 & W_{t2} - \mu_2 & W_{t3} - \mu_3 & W_{t4} - \mu_4 & W_{t5} - \mu_5 & W_{t6} - \mu_6 & \dots & W_{td} - \mu_n \end{pmatrix} \dots \dots (42)$$

$$S = cov(X) = \frac{1}{m} \sum_{k=1}^m (W_{11} - \mu_1)(W_{ti} - n) = \frac{X^T X}{m-1} \dots \dots \dots (43)$$

$$S = cov(X) = \frac{1}{m} \sum_{k=1}^m (W_{11} - \mu_1)(W_{ti} - n) = \frac{X^T X}{m-1} \dots \dots \dots (44)$$

$$S\alpha = \gamma\alpha \dots\dots\dots (44)$$

To determine eigenvalues and eigenvectors of the matrix $A_{(txd)}$ where S is a real symmetric matrix (covariance matrix) so that a positive real number γ and a nonzero vector α can be found, equation 42 and 43 defines the model. γ is called an Eigenvalue and α is an eigenvector of matrix S . Suppose matrix S is an $n \times n$ matrix of full rank, n eigenvalues can be found such that $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 < \gamma_5 \dots \dots \dots < \gamma_n$

By using $(S\alpha = \gamma\alpha) \times \alpha = 0$, all corresponding eigenvectors can be found. The eigenvalues and corresponding eigenvectors will be sorted so that $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 < \gamma_5 \dots \dots \dots < \gamma_n$. Then the first desired PCA factors or thematic words of $d = n$ eigenvectors are selected using techniques like the cumulative percentage of variance, Scree plot and broken stick.

4.3.2. Principal Component Analysis via Singular Value Decomposition

Another computational approach to PCA is based on the SVD mathematical model. The SVD based PCA is more numerically stable. If a number of variables is greater than the number of observations then SVD based PCA will give efficient result[91]. Suppose, for a given VSM matrix $A_{(txd)}$ we derive a matrix X as the SVD outlining three matrices the $T_{(txn)}$ for term concepts, $S_{(nxn)}$ matrix representing the principal components and finally $D_{(nxd)}$ indicating the document concepts as in equation 47.

$$A_{(txd)} = T_{(txn)} \cdot S_{(nxn)} \cdot (A_{(nxd)})^T \dots\dots\dots (45)$$

$$X = \frac{1}{m-1} A_{(txd)} \cdot (A_{(nxd)})^T \dots\dots\dots (46)$$

$$\begin{aligned} \text{But } A_{(txd)} \cdot (A_{(nxd)})^T &= D_{(nxd)} \cdot S_{(nxn)} T_{(txn)}^T T_{(txn)} S_{(nxn)} (D_{(nxd)})^T \dots\dots\dots (47) \\ &= D_{(nxd)} \cdot S_{(nxn)} S_{(nxn)} (D_{(nxd)})^T \\ &= D_{(nxd)} \cdot S_{(nxn)}^2 (D_{(nxd)})^T \\ &= D_{(nxd)} \cdot \frac{S_{(nxn)}^2}{M-1} (D_{(nxd)})^T \end{aligned}$$

Matrix \mathbf{X} is a covariance hence it is symmetric therefore the eigenvectors of the covariance are the same as the SVD matrix's D_{dxm} called the right singular document in our case it is the document concept matrix. Therefore the eigenvalues of matrix \mathbf{X} can be computed from the singular values as in equation 48

$$\gamma_1 = \frac{\alpha_i^2}{m-1} \dots \dots \dots (48)$$

Calculate the covariance of the matrix $A_{(txd)}$ as matrix \mathbf{X} outlined in the equation 47. The eigenvectors of the covariance matrix \mathbf{X} are the same as the right singular vectors of the SVD of matrix $A_{(txd)}$, thus the eigenvalues of the matrix \mathbf{X} can be computed from the singular values. Suppose the eigenvalues are arranged in an ascending order

$$\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 < \gamma_5 \dots \dots < \gamma_n \dots \dots \dots (49)$$

If the corresponding eigenvectors are arranged accordingly into an $[n \times n]$ matrix \mathbf{T} so that the left most column corresponds to \mathbf{x}_1 , then matrix \mathbf{T} is defined as

$$T = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} & X_{15} & X_{16} & \dots & X_{1d} \\ X_{21} & X_{22} & X_{23} & X_{24} & X_{25} & X_{26} & \dots & X_{2d} \\ X_{t1} & X_{t2} & X_{t3} & X_{t4} & X_{t5} & X_{t6} & \dots & X_{td} \end{pmatrix} \dots \dots \dots (50)$$

A new dimensional reduced matrix $A_{(txd)}$ new using PCA is computed using the equation 50.

$$A_{(new1)} = TA_{(txd)} \dots \dots \dots (51)$$

The research used MATLAB as a tool to perform attained HIV/AIDS FAQ question classical VSM Matrix $A_{(txd)}$ to derive a the HIV/AIDS FAQ PCA matrix. . A MATLAB in built function Princomp() was used to derive PCA components. Some PCA components or thematic key weighted terms derived are not worth considering because of their insignificant values. However the problem that manifests is what would be the cut off point for this insignificance value because there might be the risk of discarding valuable thematic terms. The research experimented to determine the

appropriate maximum PCA factors or thematic weighted terms that could accurately represent the semantic and conceptual relationship of all HIV/AIDS FAQ questions in the corpus and their keywords.

4.4. Selection of the Principal Analysis Factors: Experiment IV

The research has implemented visual and computational approaches as back to back methods in determining and selecting relevant PCA component factors or the thematic words which are representative of all HIV/AIDS FAQ questions in the corpus.

The techniques used are Cumulative Percentage of Variance (CPV), Latent Root Criterion (LRC) and Scree Plot using MATLAB functions and these approaches enable determination of relevant and appropriate PCA components as they compute the range of weighted highest PCA values against the total number of key words in the input matrix in our case the HIV/AIDS FAQ question which is 712 weighted semantic words.

4.4.1. Scree Plot Approach

Scree Plot test considers PCA components values at the ‘elbow’ as the curve becomes flat or the value at the largest drop on the curve as pertinent values to represent the data set variation of the corpus on a plot of the all PCA components computed against the number of eigenvalues when ranked. Figure 19 illustrate the research computed highest PCA component at 1.27 against the ranked eigenvalues. The value at the largest drop indicates PCA factor value of 1.0194 and being 5 they account for 5% variation of the data set. The top most PCA components cumulative percentage is not sufficient to represent the data variance because of the low cumulate percentage value. However, elbow values close to the end of the curve gives a range of 250 to 360 PCA components as potentially representative values and have a cumulative percentage of 95% as illustrated in figure 22.

4.4.2. Cumulative Percentage Variance technique

Cumulative Percentage Variance technique uses a graph to determine the values of PCA to adopt as illustrated in figure 23. Heuristic knowledge on use of CPV points that, PCAs components are selected from a minimum value of 80% to 95%. This value was considered and according to figure 23, PCA component deduced were 280 to 360. These values almost tallies with graphical observation and deduction from the Scree plot

4.4.3. The Latent Root Criterion considers

This approach considers PCA components that contribute a value greater than 1. The individual PCA Component Value are represented in figure 24 which indicates a value of 5 PCA component values. However as argued previously the CPV for the first 5 PCA is not enough to warranty any consideration for determining the PCA components for the network.

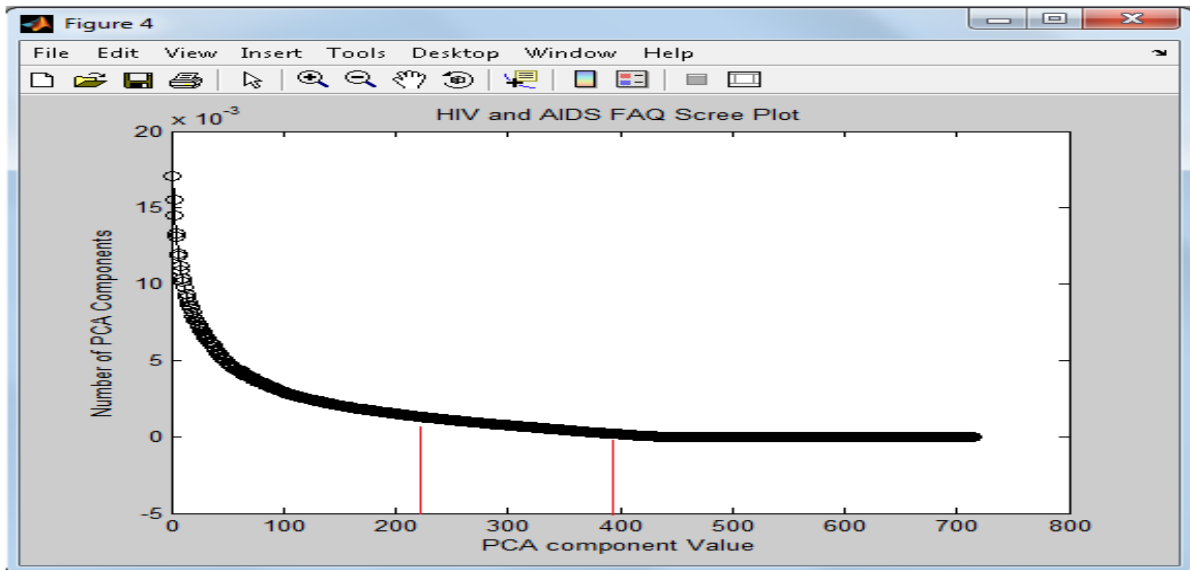


Figure 22 Scree Plot of PCA Components Against Eigenvalues

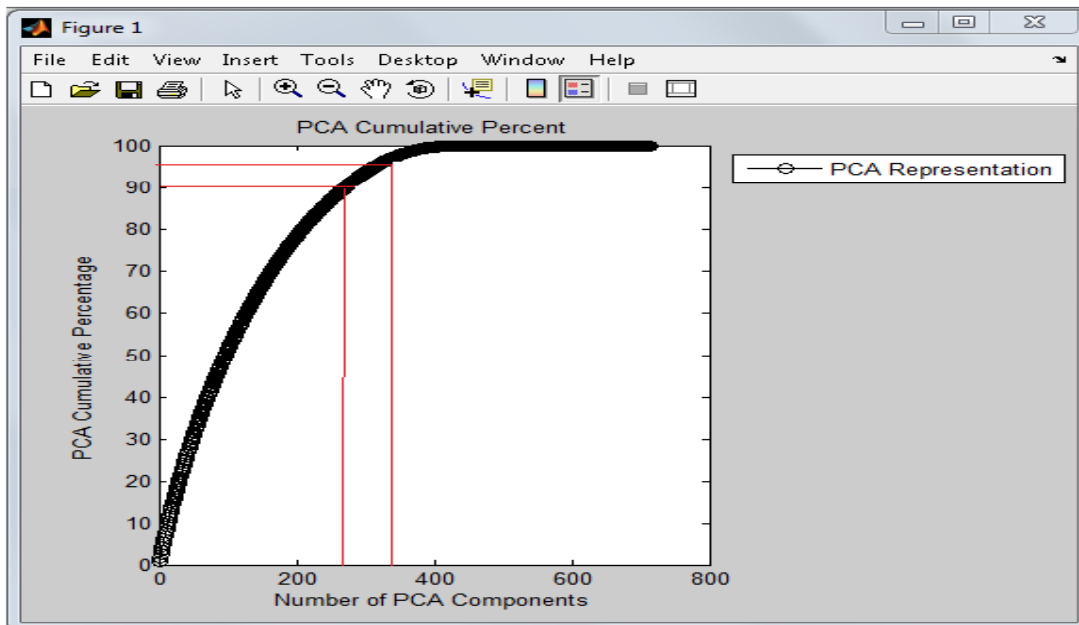


Figure 23: Cumulative Percentage Plot Variance against Eigenvalues

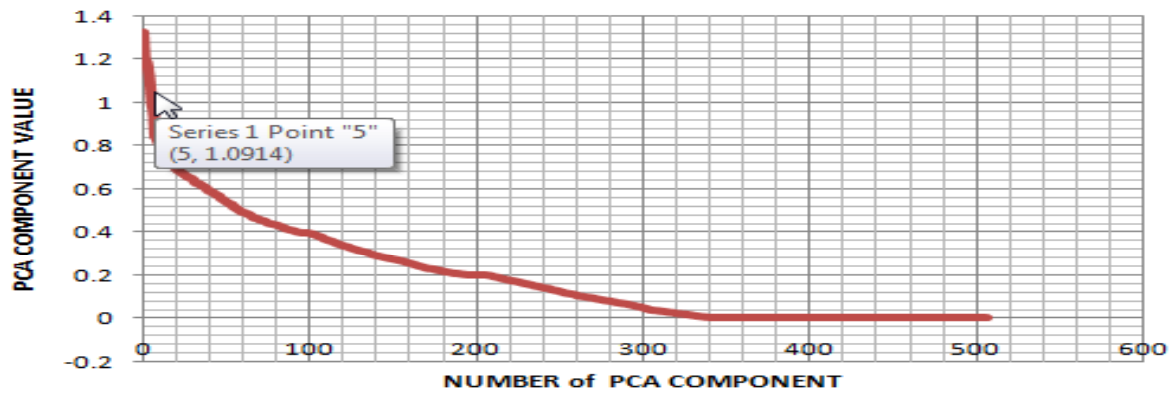


Figure 24: The Latent Root Criterion Lot against Eigenvalues

Table 3 lists summaries of the findings for the three experiments conducted to determine and select the optimum PCAs factors. The research shall use values from the screen plot approach and Cumulative Percentage Variance technique values because they cover a huge range of thematic components for the HIV/AIDS corpus with a total cumulative data variation of 95%. Thematic weighted terms or PCA components for the HIV/AIDS corpus account for 95% of the HIV/AIDS key words meaning the neural network shall have input nodes in the range of 250 to 360 compared to a scenario where the VSM HIV/AIDS FAQ were to be used then it would have neural input nodes of 712. This reflects a dimensional reduction of almost 50% to 65%

Table 3: PCA Component Determination and Specified Component Values

Technique	Criterion Selection	Range Specified	Number of PCAs to Considered
Scree Plot Test	“Large Drop” between consecutive PCAs	6%	5
	“Elbows” in the plotted curve...Defines a range	90 - 95%	280 - 370
Cumulative Percentage Variance Technique	Heuristics of researches done to select PCA components used an cumulative value with a minimum of 60%	80 to 95%	280 to 360

4.5. Determining IHAFR Neural Network Parameters

Structural the neural network architecture is implemented as a Multilayered FeedForward Neural Network (MLP) which is shown in figure 22. The MLP has three layers defined as input layer X_i , hidden layer Y_i and output layer Z_k . The MLP consists of two connection weight layers being the input hidden layer V_{ij} and the output hidden layer W_{jk} .

The first connection weight layer V_{ij} stores knowledge on key words for HIV and AIDS. This layer determines similarity amongst input key words of a query to respective and equivalent stored key words for HIV and AIDS. The stored HIV/AIDS keywords were stored during a neural network training process.

The second connection weight layer W_{jk} represents stored HIV/AIDS FAQ_N questions, where N represents the last stored FAQ_N question in this connection weight layer. The FAQ_N triggered key words will activate an HIV/AIDS question which comprises of all the triggered key words and it is activated as an output of the neural network.

Functionally the Multilayered Feedforward Neural Network based on figure 25 accepts inputs into the neural network depending on a presented user query which has been subjected to feature reduction and transformed with PCA function to dimensional reduced input to nodes X_1 to X_i . Thus the research needs to determine the optimal number of input neurons which constitute the input section experimental.

The hidden layers nodes formulate the system intuitive processor section and perform the matching and activation of stored key HIV/AIDS FAQ key words to HIV/AIDS FAQ questions represented by neural nodes Y_1 and Y_i . each node acts as an independent 'processor' but in tandem with others nodes when processing user query keywords and determining the equivalent and matching key words which are stored within this layer hence would activate relevant and corresponding HIV/AIDS FAQ questions. There is need to determine the exact and optimal number of neural hidden nodes else the network shall lose the ability to generalize and answer user queries.

The output of the neural network nodes simply and directly relate to the total number of words in the HIV/AIDS corpus. Thus it is a direct parameter to determine as the total number of the HIV/AIDS FAQ questions used is 467 FAQ questions so the system has the same number of output nodes and is represented symbolically as 1 to K

This entire process is described as mapping or classification through a function $FAQ_M = f(Q_N)$ which maps an input query Q_N to a required FAQ_M question. The output is noticed through output layer (Z_K) of the neural network.

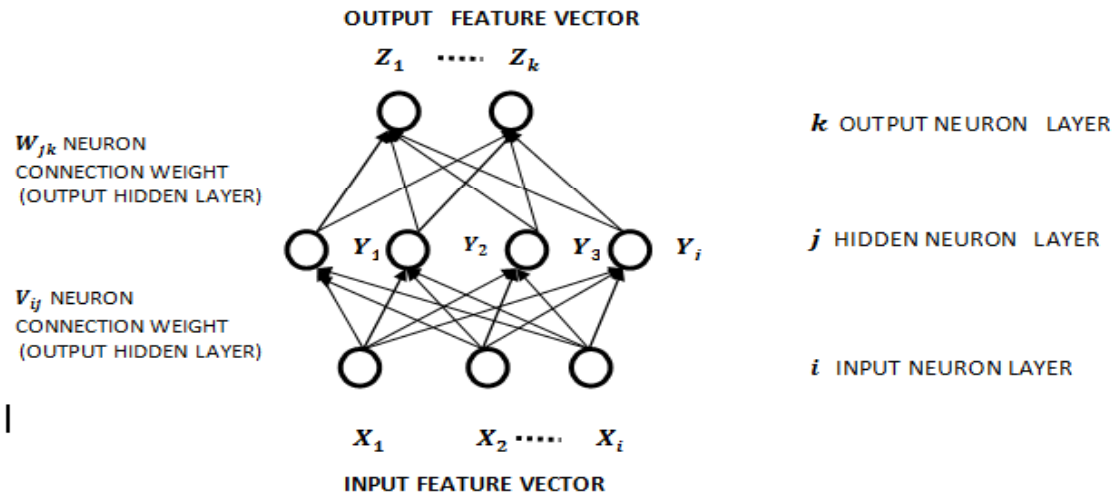


Figure 25: Multilayered Feedforward Neural Network Structure.

4.5.1. Input and Output Neural Nodes: Experiment V

The neural network inputs have been derived from the range of PCA components investigated and determined in experiment IV. These values span the range of a minimum of PCA250 to a maximum of PCA360 components as determined by experiments IV. The research shall experiment to determine the optimum input neural nodes within this range. The HIV/AIDS FAQ corpus has 467 FAQ questions to map too for every issued HIV/AIDS query, hence this translate to 467 neural network output nodes.

For each neural input size, the research altered the size of the neural network hidden layer node from magnitude 1 up to a maximum value specified by a heuristic thumb rule findings in table 5 which specifies a maximum of 1008 hidden neurons [92]. The variation of the hidden neural nodes is in accordance to the constructive technique, an experimental practice used to determine an optimum number of hidden neurons by incrementing hidden neurons until the least MSE is observed over a specified experimental range. Additional neural network inputs 360 to 400 have been are used to validate whether the MSE value would start to rise after the optimum PCA specified values as in experiment IV are surpassed. MSE value derived from the experiment are recorded for each set of input and hidden neuron as specified in table 4

Table 4: Number of Input, Output and Hidden Neurons versus MSE

No Number	Experimental Number of Inputs Neurons X_1 from X_2 (PCA Factors)	Number of output neurons	Mean Square Error (MSE) Goal = 0.1
1	250	467	0.098320951
2	260	467	0.098265774
3	270	467	0.098046142
4	280	467	0.097585801
5	290	467	0.097689624
6	300	467	0.097416083
7	310	467	0.096116329
8	320	467	0.095846446
9	330	467	0.096028206
10	340	467	0.097284434
11	350	467	0.095748815
12	360	467	0.095177702
13	370	467	0.096600541
14	380	467	0.108148126
14	390	467	0.097198808
15	400	467	0.097646273

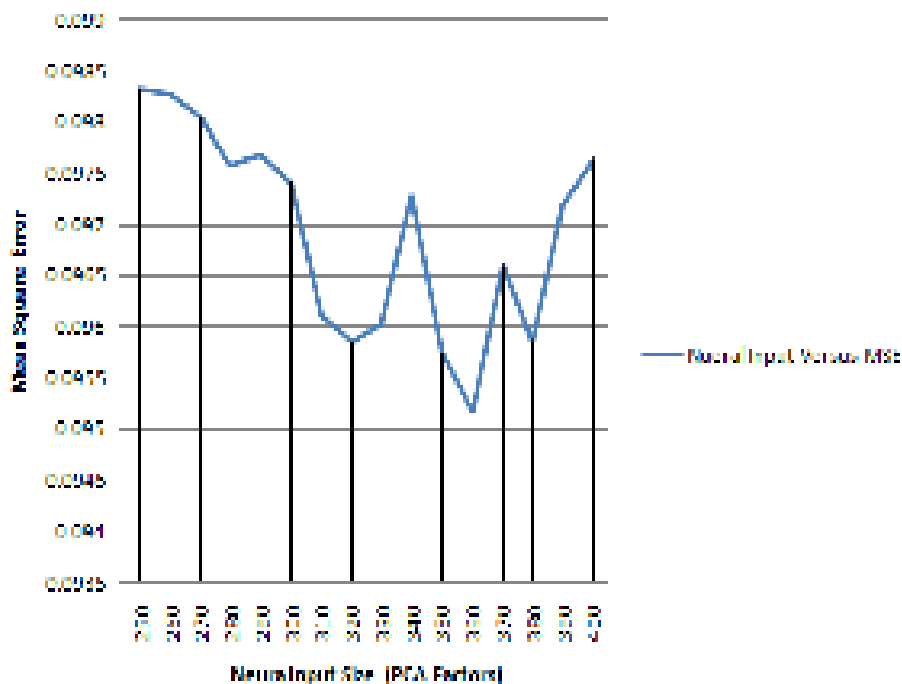


Figure 26: Excel Plot for the Table 5a: Number of Input Neurons versus MSE

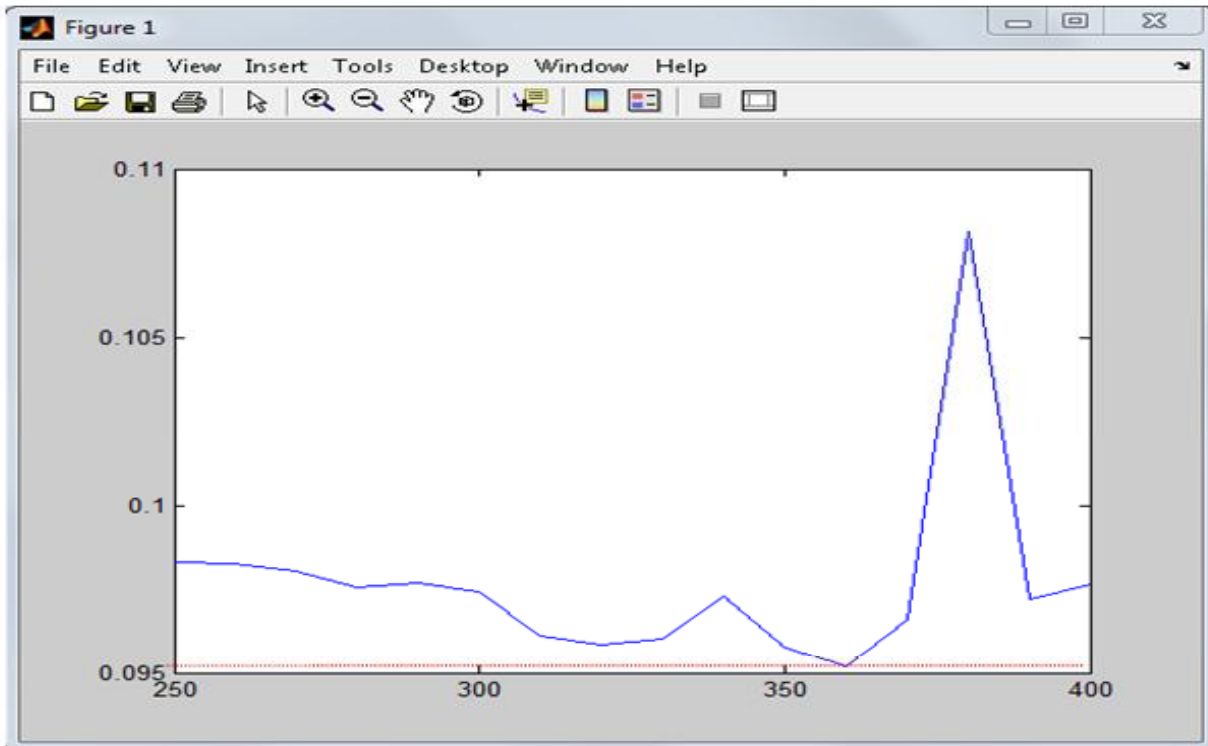


Figure 27: MATLAB Plot for the Table 5: Number of Input versus MSE

Tabulated results of table 4 and illustrations in figures 26 and 27 clearly indicate that input 360 PCA factor produce the least MSE results of the magnitude 0.09517702 for a defined goal of 0.1 MSE. The MSE starts to rise beyond the value of 400 PCA inputs thus confirming the optimum of 360 PCA factor.

4.5.2. Hidden Neural Nodes: Experiment VI

Heuristic thumb rules have been used to guide and establish the range of hidden neural network node to consider. It is noted that from these rules that the number of hidden neurons are calculated based on key parameters like neural network input and output neural nodes [64, 93-95] which experiment IV of this research has determined. Constructive technique is used to filter the best hidden neurons as deduced from the bounds set by heuristic knowledge in table 5

Constructive technique entails varying the hidden neurons of a neural network from 1 to a value that reflects best neural network performance. Overtraining and under training is a phenomena that afflicts neural network during the training process. As the HIV/AIDS neural network is being trained to establish appropriate hidden neural nodes, the neural network can also be affected by these problems

Table 5: Heuristic Thumb Rules:

Heuristic Thumb Rule	Formula: [Hidden Neurons (H), Input Neurons (I), Output Neurons (O)]	Number of Hidden Nodes Minimum Maximum
"A rule of thumb is for the size of this [hidden] layer to be somewhere between the input layer size ... and the output layer size ..." (Blum, 1992, p. 60)[92].	$H \leq O$ and $H < I$	1 410
"To calculate the number of hidden nodes we use a general rule of: (Number of inputs + outputs) * (2/3)" (from the FAQ for a commercial neural network software company)[92].	$H = (I+O) * (2/3)$	277 494
"you will never require more than twice the number of hidden units as you have inputs" in an MLP with one hidden layer (Swingler, 1996, p. 53)[92].	$H = < 2 * I$	10 820
"How large should the hidden layer be? One rule of thumb is that it should never be more than twice as large as the input layer." (Berry and Linoff, 1997, p. 323). [92]	$H = < 2 * I$	10 820
"Typically, we specify as many hidden nodes as dimensions [Principal Components] needed to capture 70-90% of the variance of the input data set." (Boger and Guterman, 1997)[92]	70-90% of Variance Input Data Set	434 554
Geometric pyramid rule was proposed by Masters and specifically address hidden neurons for 3 layered networks which have n inputs and m outputs.	$H = \sqrt{n \times m} \times 1.5$ OR $H = \sqrt{n \times m} \times 2$	765 1008
Bailey et al mentions that the range of hidden neural nodes should be 75% of the neural network input size (N). The range of neural network input is from 200 to 330 as provided in experiment III	$H = N \times 0.75$	150 248
Katz remarks that the range of hidden neural nodes should be 3 times or 1.5 e neural network input size (N). The range of neural network input is from 200 to 330 as provided in experiment III	$H = N \times 1.5$ OR $H = N \times 3$	300 990
Ersay remarks that the range of hidden neural nodes should be 3 times or 1.5 e neural network input size (N). The range of neural network input is from 200 to 330 as provided in experiment III	$H = N \times 2$	400 600

Early stopping is a technique used to prevent a neural network being over trained by co-opting a Validation Date Set. This data set is used to monitor and prevent network overtraining by stopping the neural network training if the validation error starts to rise. The effect of overtraining in neural network training leads to memorization of training set and hence the neural network is not able to answer any general question except those contained in the training data set. Early stopping was used to prevent overtraining of the neural network by creating a Validation Data Set as illustrated in figure 28 where during training as the neural network settles at the most optimum learning point the validation data set will immediately rise reflecting the most optimum number the neural network should have.

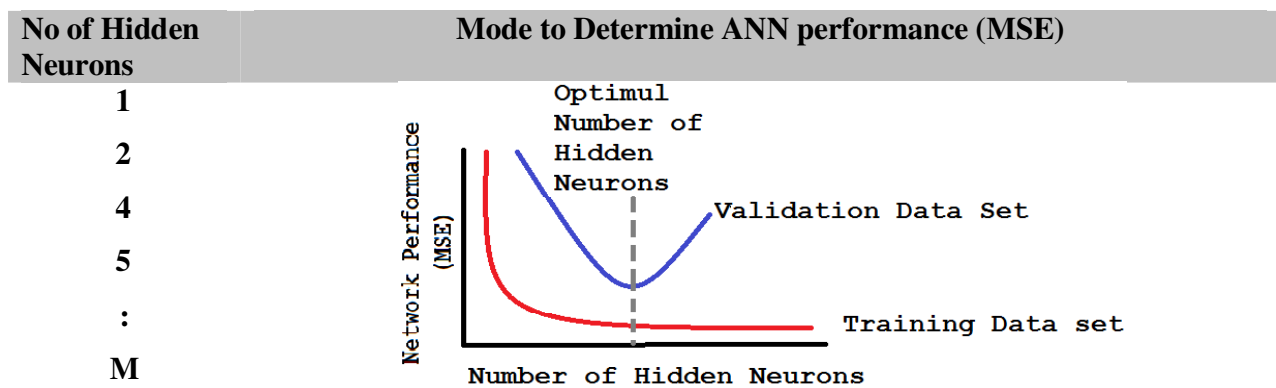


Figure 28: Determining the Hidden Neurons for the ANN Architecture using the Validation Data Set

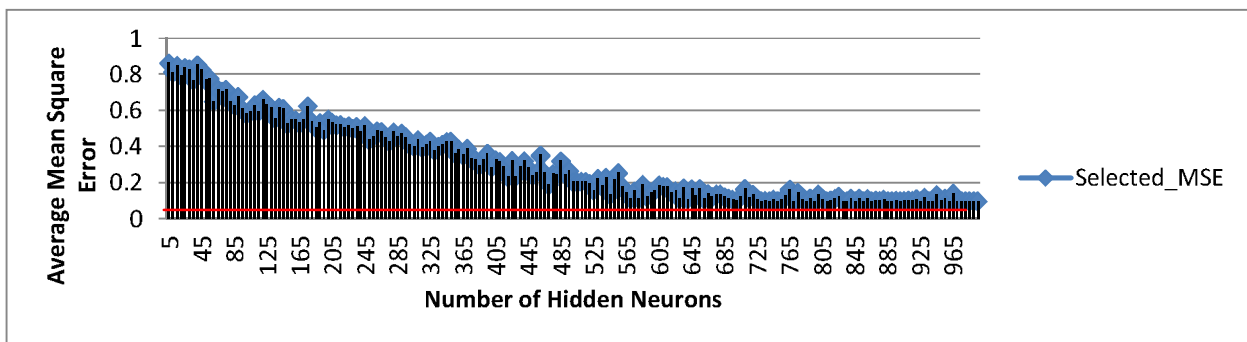


Figure 29: Plot of Hidden Neurons versus Least Mean Square Error

Based on the training done Figure 29 illustrate the results of the experiment using the constructive technique where input neural nodes were 360 and the output was 467. The number of hidden neurons in the middle layer was incremented from 1 up to 1008. It was established that as the number of neurons increased the network performance settles for constant performance and this starts at 720 hidden neurons where an MSE 0.13734 of is attained. The research adopted 720 as it has the same impact with other subsequent number of hidden neurons and also that it has the least MSE value and this phenomenon was also acknowledged Masters [96] when he developed NETALK neural network and noted that from 60 up to the maximum set 120 hidden neurons the MSE changed very slightly therefore 60 neural nodes were adopted for the hidden layer.

4.5.3. Determining Neural Network Training Epochs:

Heuristic knowledge has been used to determine the initial number of epochs during the initial experiment proceeding. Kaastra [95] mentions of 85 to 5000 runs should be conducted when training the neural network and in between the MSE function should converge to its minimum value. Walezka[94] states that a fully trained neural network should converge with epochs in between 5000 to 191400. The research adopted a maximum of 5000 iterations or epochs for training a neural network and observes whether it converges within this range. A plot of the average means square error and the recorded number of iterations is illustrated figure 30. The result indicates the network can easily converge to a MSE value of 0.1 and less within 127 epochs. This value shall be taken as parameter for defining the maximum epoch value for the IHAFR system

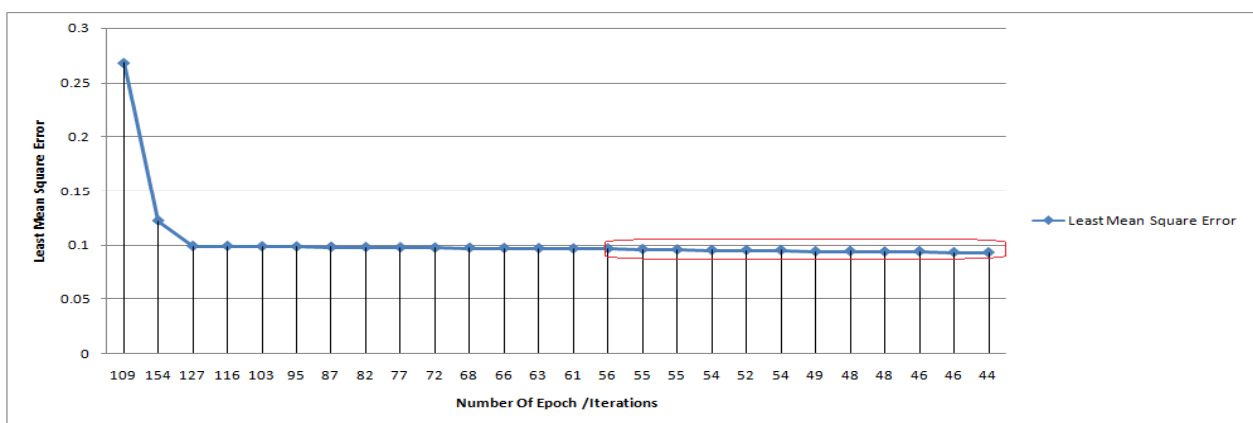


Figure 30: Illustrations of range of Epoch for IHAFR

4.5.4. Artificial Neural Network Activation Function and Rule: Experiment VII

The research experimented in conjunction the determination of the neural network activation function and the learning rule because these two parameters have a great influence on the training or learning phase of a neural network system. The learning rule computes and adjust the weights of the neural network layers based on the learning or acquired knowledge and then adjust the weight is determined by the neural node activation function whether it sufficient to derive an output or not.

4.5.4.1. Artificial Neural Node Activation Function

The literature review adopted the Uni-polar sigmoid neural network activation for activating the hidden neuron. The reason for this choice is that similarity measurement for textual documents is going to be guided by the criterion that if a stored document is similar to the queried document then a similarity measurement of 1 is attained else if the no similarity then the score should be 0. The last option is a degree of relevance which is measured between 0 and 1. The obtained neural network parameters from previous experiments were used to determine the best and effective activation function from the two uni-polar activation functions the Log Sigmoid and the Tan Sigmoid activation functions. Table 7 illustrates the attained results based on the experimented neural input nodes, optimum hidden neurons the epoch cycles.

4.5.4.2. Artificial Neural Network Learning Rule

Neural network training involves learning a given set of historical data which should be mapped into an intended or desired target. In our case the IHARF was trained with HIV/AIDS FAQ questions so that it would be able to map presented HIV/AIDS queries to similar and look alike HIV/AIDS FAQ questions, which in essence is a classification task spread over the number of questions in the HIV/AIDS corpus. IHAFR Neural network is trained to classify a given set of historical FAQ questions which are HIV/AIDS FAQ question Q_m and classifying into a corresponding stored HIV/AIDS FAQ question FAQ_n over the 467 classes representing the 467 HIV/AIDS FAQ questions. Therefore a data training set is defined as in equation 52 where T_i is training pair which constitute an input Q_m and a targeted output FAQ_n per given instance.

$$T_i = [(Q_i, FAQ_i)]_{i=1}^M \dots \dots \dots (52)$$

The learning actually takes place as the neuron connection weight layer changes in response to a given input of a sample of HIV/AIDS FAQs. Absolute learning is said to have occurred when the neurons connection weights changes have a small and insignificant mean square error (MSE) and this happens when an input approximately HIV/AIDS FAQ query stored in the neural network equals to the desired target.

Learning algorithms are responsible for changing of the neuron connection weights in connections with a learning function. The learning function provides a defined procedure on how the neuron connection weights changes depending on the neuron input. Some learning algorithms use the error derived between the output and static input as a parameter in changing the connections weights.

The research used the Backpropagation neural network training algorithm which works best with the supervised training method and Multi-layer Feedforward neural network. The learning rule autonomously extract the functional relationship between input data and expected output data embedded in a set of historical data set i.e. HIV/AIDS training data set and encodes it into connection weights i.e. (W_{ij}) and (W_{jk}) . Most neural network based information retrieval systems use Hopfield networks or Self Organizing Maps, these networks do not include hidden layers of neurons which processes non linearity of text data by mapping text terms to appropriate text document and Mandal [97] describes this hidden layer as the intuitive processor. However Backpropagation neural network training algorithm has so many variants, the research shall investigate an appropriate variant which would be suitable for training the IHAFR.

Training, validation and testing of HIV/AIDS FAQ questions was compiled using a MATLAB function where the question HIV/AIDS FAQ question X_p and corresponding to HIV/AIDS FAQ question in the corpus FAQ_N . A training pair set for each instance is reflected as per the PCA model input vector $PCA_{n \times m}$ of question X_p . For instance a pair T_i could be defined as input $X_p = W_{11} + W_{12} + W_{13} \dots W_{nm}$ and should yield an output $Y_p = FAQ_N = FAQ_{doc1}$ as the most relevant document and this only happens if the training cost function error E_p as the least and minimum error as illustrated in the figure 31.

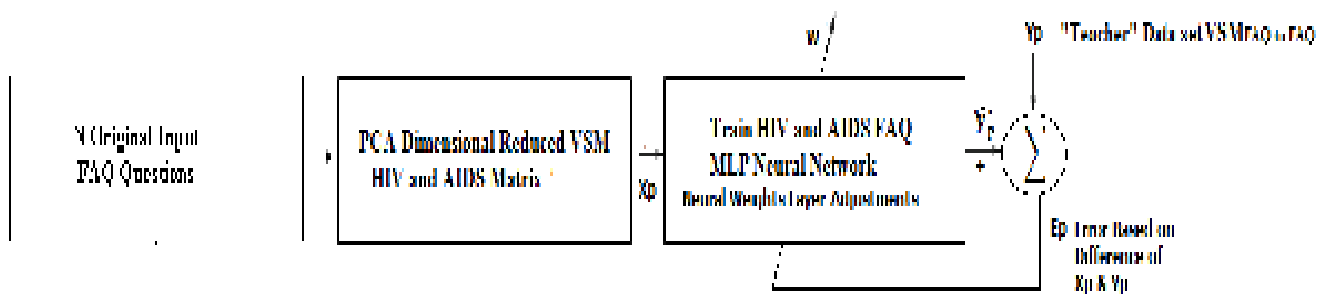


Figure 31: Training the Neural Network PCA dimensional reduced VMS for HIV/AIDS Matrix

MATLAB has so many functions to divide corpus data into training, validation and testing data and these are divider, divideblock, divideint and divideind. Dividerand randomly divides the data into the three respective categories. Divideblock creates contiguous data blocks in the respective classes. Divideint divides into categories by interleaving with specified integer value and divideind categorizes into the groups using a specific index value. The research shall implement the MATLAB Dividerand() function to split the data in the HIV/AIDS FAQ corpus into training data, testing data and validation data randomly so that we induce a dimension of equally representation of all sorts of HIV/AIDS FAQ questions in the corpus.

The “teacher” data set for HIV/AIDS FAQ questions was derived by implementing VSM algorithm for the whole HIV/AIDS FAQ questions and compared to each question and all the FAQ questions to derive a document to document VSM. An algorithm computed in MATLAB in appendix 3 was used to create the VSM data set with ‘teaching’ FAQ questions.

The “teacher” data set for HIV/AIDS FAQ questions computation was implemented by adding the sum of each FAQ term weight to derive the total question term weight. Similarity comparison is done per and against each FAQ question to determine their order and strength of similarity. Cosine Similarity Metric, was used to measure similarity between any two FAQ question in the corpus illustrated in equation 53

$$\text{Sim}(Q_M \text{FAQ}_N) = \frac{\sum_i^t w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \sqrt{\sum_{i=1}^t w_{i,q}^2}} \dots \dots \dots (53)$$

This mathematical model was implemented using a MATLAB code with a “teacher” data set for HIV/AIDS FAQ questions Vector Space Mole of 467 by 467 was created and stored as an excel CSV document. The VSM is also subjected to the `diverand()` MATLAB functions so that corresponding FAQ question could be matched to selected FAQ questions for training, testing and validation purposes.

The “teacher” data set has a measure of each and every FAQ question in the HIV/AIDS corpus expressed as a degree of relevance between and each and every HIV/AIDS FAQ question in the corpus as $W_{FAQ+..FAQk}$ and as 1 unto the its self 1 as indicated in equation 54. The input will be the PCA matrix input as per the number of PCA components defined as thematic values and the output as per the $VSM_{FAQ\ to\ FAQ}$ matrix.

$$VSM_{FAQ\ to\ FAQ} = \begin{pmatrix} 1 & W_{FAQ1..FAQ2} & W_{FAQ1..FAQM} \\ W_{FAQ2..FAQ1} & 1 & W_{FAQ2..FAQM} \\ W_{FAQN..FAQ1} & W_{FAQN..FAQ2} & 1 \end{pmatrix} \dots\dots\dots (54)$$

4.5.4.3. Activations Functions and Backpropagation Learning Algorithm Variants: Experiment VIII

The research experiment adopted neural node activation functions in conjunction with the Backpropagation learning algorithm’s variants so that best selection is made. Backpropagation function is endowed with many learning algorithm variants. These learning algorithms are categorized as general gradient descent, heuristic based gradient descent and standard numerical optimized learning algorithms. A training algorithm that could provide an output at the shortest possible time and with the highest accuracy could be a practical solution in a real world. The variants of the Backpropagation variants were used and the best variants were selected for further testing. The output and selection of the best variants is based on the results obtained and noted as in appendix 3 It was noted that the `Trainsgf`, `Traincgp` and `Trainscg` provided the best results. These were further tested with derived network parameters and the two neural activation functions and the results are illustrated in table 6.

Table 6: Performance Comparison with Different Transfer Activation Functions and Learning Rule

Training Algorithm	Number of Input Neurons	Number of Hidden Neurons	Transfer Activation Function	Network Performance Measure (MSE)	Epoch	Time
Traincgf	720	360	Log Sigmoid	0.00492599	50	40.685000
		360	Tan Sigmoid	0.00101584	130	94.832000
Traincgp	720	360	Log Sigmoid	0.00239480	67	62.104000
		360	Tan Sigmoid	0.00061608	130	112.945000
Trainseg	720	360	Log Sigmoid	0.00453794	53	53.072000
		360	Tan Sigmoid	0.000422944	130	131.056000

The Log sigmoid activation function produces the lowest epoch and a fast execution time though a high MSE in comparison to the Tan sigmoid function. In contrast, the Tan sigmoid has an execution time which stretches beyond 130 epochs which is beyond the determined and expected epoch from experiment V. Network training with Tan sigmoid does not reach a validation stop and also performance goal is not met which might indicate probability of an overstrained neural network. The training activation function TRAINCGP has the highest accuracy compared to all training active functions is therefore adopted as the training activation function for IHAFR neural network system. Besides the Tan Sigmoid activation use Bi-polar and is not suitable to represent the learning outcomes properly as it would switch ranges from [-1, 1].

The Backpropagation gradient algorithms that produced better results exhibited the lowest MSE, faster convergence time with the least epochs. The research shall adopt the conjugate gradient training algorithm as the learning rule for the IHAFR in particular the traincgp which upon further testing yielded an MSE result of 0.0020982 from the intended 0.1 goal. The final designed neural network had the following parameters as its key design specifications, as listed in table 7 below

Table 7: IHAFR Neural Network Parameters

Network Parameter	Network Value
Number of inputs	360
Number of Hidden Neurons	720
Number of Hidden Layers	1
Number of Output Neurons	467
Training Rule	Supervised/Backpropagation
Training Algorithm	TRAINCGP
Training Activation Function Hidden Layers	Log Sigmoid
Training Activation Function Output Layer	Purelin
Ranking Cut off point	M 0.5
Network Goal	0.0020982
Network Epoch	Max 130

4.5.5. Similarity Matching of HIV/AIDS FAQ Questions: Experiment IX

The question to question mapping technique is used to identify an FAQ question by virtue of inputting key query words which would trigger stored HIV/AIDS key terms and in turn the intended FAQ question. The PCA HIV/AIDS FAQ has specifications of the term weights of all the FAQ questions thus the thematic input for training the neural network are expressed as in equation 55 where Q_M^T is the vector and X_1W_{IN} are the numerical weight terms of the query (refer to theory of PCA computation section 4.3.2).

$$Q_M^T = X_I^T = (X_1W_{11}, X_1W_{11}, \dots \dots \dots X_1W_{11}) \dots \dots \dots (55)$$

The research needs to compute FAQ questions in a numerical format to represent the target data used as training instances. Equation 56 specifies T_i as a training data set for an instance input of a query Q_M and target FAQ_N .

$$T_i = (Q_M, FAQ_N) \dots \dots \dots (56)$$

Question to question similarity comparison in the neural network is generally realized through the Backpropagation training which is done in two phases. In the first phase (forward-propagation phase), suppose we have input X_l^T as feature vectors consisting of term weights for each query key word Q_M transformed to thematic words through a PCA function and is fed into the input layer as described in equation 57.

A corresponding output vector representing activated neurons for each document question in FAQ_N is derived based on estimated general weight update in the two weight layers of the network. The objective is to minimize error function E , Mean Square Error (MSE) for mapping each sample $T_i = (Q_M, FAQ_N)$ in the training set by continuously changing the weights so that all input vectors are correctly mapped to their corresponding output vectors [98]. The error is computed based on the derived output of the network Z_{il} and the expected outcome of the network per given input FAQ_{il} as given in equation 58 as shown in figure 32

$$FAQ_N^T = Z_{ij}^T = (FAQ_{1N}, FAQ_{2N} \dots \dots \dots FAQ_{iN}) \dots \dots \dots (57)$$

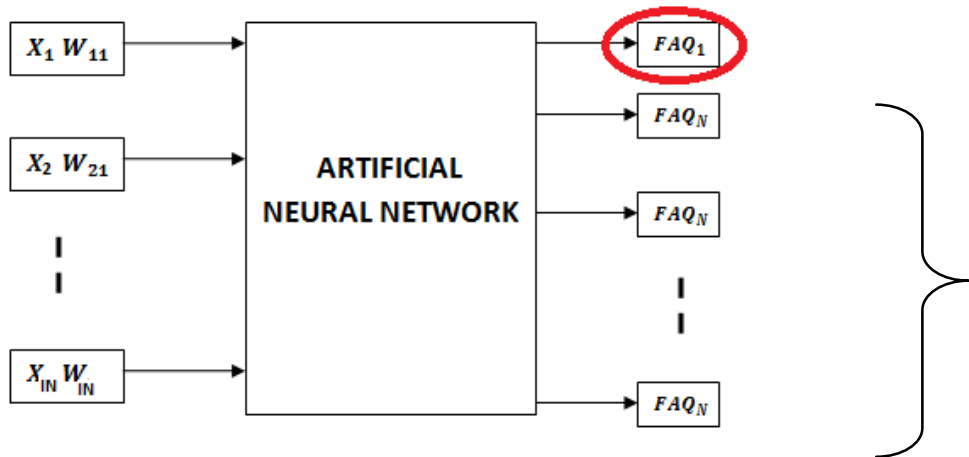


Fig 32: Question Q_M to Question FAQ_N () Mapping using Neural Network.

$$E = \sum_{i=1}^n \sum_{l=1}^m \frac{(FAQ_{il} - Z_{il})^2}{2} \dots \dots \dots (58)$$

In the second phase (back-propagation phase), a gradient descent in the weight space, ΔW_{ln} , is performed to locate the optimal solution by changing and direction of the ΔW_{ln} as indicated in formula 59, where the learning rate should in the range $0 < \varepsilon < 1$ and it controls algorithm's convergence rate.

$$\Delta W_{ln} = \frac{-dE}{dW_{ln}} \times \varepsilon \dots \dots \dots (59)$$

The value E (MSE) is propagated back layer by layer from output units to the input units in the second phase and also weight adjustments are determined on the way of propagation at each level. These stages are done for each iteratively until the value E converges. Fig 33 shows the basic illustration of the Backpropagation algorithm

```

While
  E (MSE) is unsatisfactory AND computational bounds are not exceeded
  Do
    For each input pattern  $X_l, 1 \leq l \leq L$ ,
      Compute hidden node inputs
      Compute hidden node outputs
      Compute inputs to the output nodes
      Compute the network outputs
      Modify outer layer weights
      Modify weights between input and hidden nodes
    End-for
  End-while

```

Figure 33: Backpropagation training rule for the multilayered feed forward ANN

The stored knowledge of a neural network is encoded in the set of connection weight neurons of input and output hidden layers V_{ij} and W_{jk} after the training phase has been completed. In a sense, these set constitutes a single representation in which representations of all learned patterns are superposed or mushed together. [99]. In our case knowledge of arbitrary English FAQ question terms to logical and syntactical correct FAQ_N terms and finally to semantic, conceptual and pragmatic English HIV/AIDS questions.

4.6. IHAFR Validation Performance

The research pre-evaluated the IHAFR system in order to determine its effectiveness prior to testing with unforeseen data and targeted users. Regression analysis measures were used to give an indication of the how the network shall classify any given inputs into respective outputs.

MATLAB was used to determine the classification rate using the Regression analysis. The test determines neural network ability to respond to a given input and map it into its proper target. Averagely from the training data used for training the system it yielded a correlation rate of $R= 0.658$ which translate to 66% as shown by the correlation coefficient value R -value indicated in figure 34 below. This value measures the degree of variation generated by the general network's performance of mapping an input and to a corresponding output through a specification value ranging from 0 to 1. A value of zero indicates that there is no relation between the given value and its output, whilst a value 1 indicates a true and correct copy of the input.

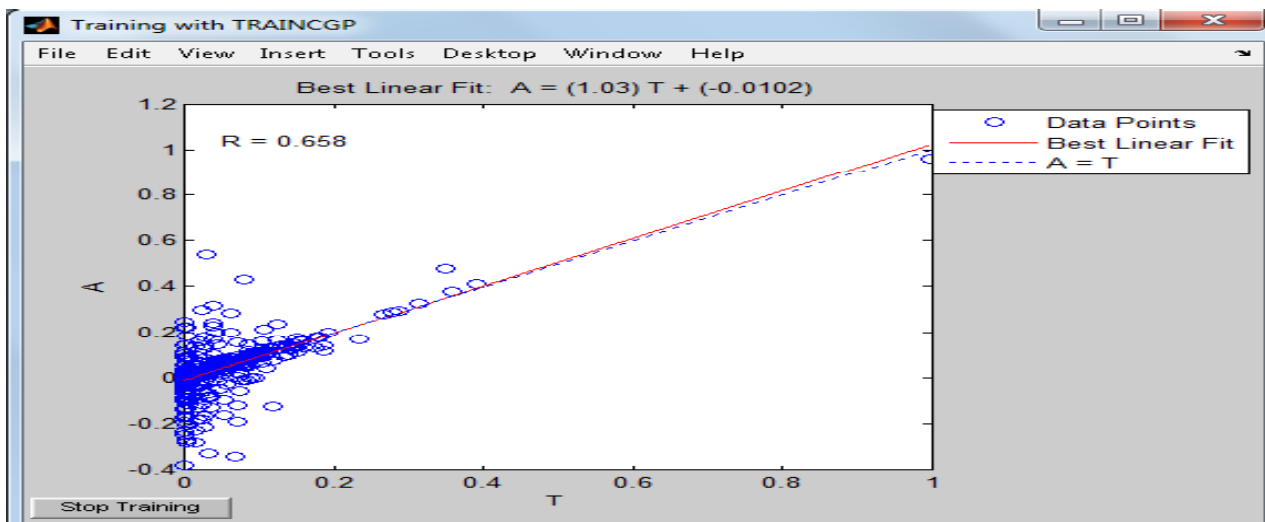


Figure: 34 Linear Regression Analysis for the Classification

The system is trained with neural some training examples selected from the HIV/AIDS FAQ knowledge base. Training data, Test data and Validation data were created from the 467 FAQ questions and were used to train as explained in section 4.5.4.2 paragraph 6. TRAIN function in MATLAB was used. Early stopping technique was used to detect overtraining of the neural network by using the created validation and testing data. Figure 35 shows the neural network training performance and clearly indicates no

overtraining occurred as the validation and testing curves are almost the same. Overall training occurred in 39 epochs with a time span of 14.5006 seconds.

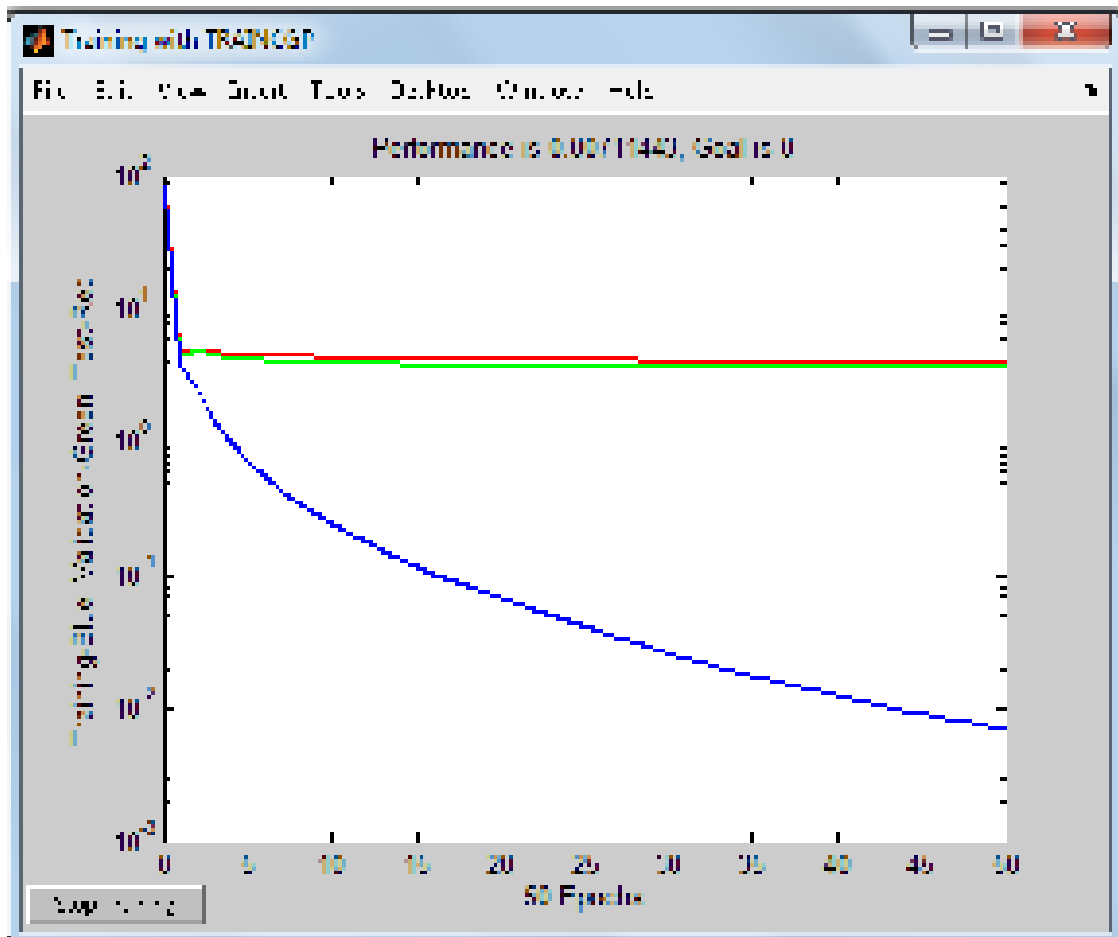
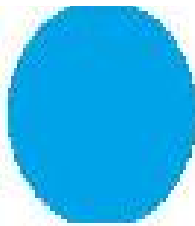


Figure 35: The Training Diagram for the IHAFR Based on the Experimental Determined Parameters.

4.7. IHAFR Generation of Responses to HIV/AIDS Queries

IHAFR system was deployed on a web based interface and was subjected to 120 twenty HIV/AIDS questions which were not used during the training, testing and validation stage from the developed HIV/AIDS FAQ corpus. The derived training outputs in conjunction with the appropriate generated weights were captured in a MYSQL database and simple PHP interface and script were used to capture and process an HIV/AIDS query and search was done to derive appropriate responses. Figure 36 shows an example of the query obtained.



HIV & AIDS FAQ RETRIEVAL SYSTEM

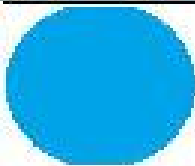
SEARCH HIV and AIDS FAQ QUESTION:

what is aids

[Search](#)

[Go Back](#)

Figure 36a: The query interface for entering the HIV/AIDS query.



HIV & AIDS FAQ RETRIEVAL SYSTEM

FAQ Query	FAQ Question	FAQ Question Weight
What causes AIDS?	294 what causes aids	0.9754000
What causes AIDS?	80 can an aids vaccine cause aids	0.7299000
What causes AIDS?	152 how does hiv cause aids	0.7220000
What causes AIDS?	272 what are hiv and aids	0.6900000
What causes AIDS?	312 what is aids	0.6527000

[Go Back](#)

Figure 36b: Response Interface to Show all Possible Answers[1].

4.8. Testing and Collection of Results for IHAFR System

In order to determine the efficiency and effectiveness of the IHAFRS neural network system to classify FAQ questions the research implemented another question to question matching system using classical VSM method which is a traditional and standard information retrieval system technique. The purpose of the system is to facilitate a comparison mechanism that would enable measurement of efficiency and effectiveness of the IHAFRS.

The same HIV/AIDS FAQ corpus was used to create the classical VSM IR FAQ based system. The cosine similarity matching technique was used as the similarity matching technique. A set of HIV/AIDS question not used during the neural network training for testing, training and validation were used to query the systems. Their responses were noted and documented for presentation to human experts to judge the relevance of the FAQ responded with. The human expert judgment was used to determine the relevance and similarity matching of the question to the presented HIV/AIDS FAQ question per questionnaire attached as in appendix 2

The people used to judge the relevance and generalization of the two system responses to presented HIV/AIDS questions were HIV/AIDS coordinators deployed in tertiary institutions by Human Resource Development Council (HRDC) to educate youth/students about the HIV and AIDS. The students who participated in the workshops and activities were organized by HIV/AIDS HRDC coordinators. Another target population used to answer the responses were lecturers as well. Two institutions participated in the evaluation exercise namely Limkokwing University and Botho University located in Gaborone.

120 twenty questions were used to test the two systems and these were split into batches of ten questions per questionnaire so that they do not become cumbersome and overloading to the participants. A total of 100 students expressed their willingness to participate in the survey, 20 lectures and 2 HRDC HIV/AIDS coordinators from the two tertiary institutions were involved. Not all sent questionnaires were replied but however a total response rate of 80% was attained from the sent questionnaires.

The IHAFR was sensitized into five cutoff points by adjusting the cut off point for training the neural network with values from 0.1 up to 0.5 as was experimented. The label for the responses was indicated as in table 8. The IHAFRS cut off point have been established as 0.5 and above. The system will retrieve FAQ questions from cut off points ranging from 0.1 to 0.5 as practical deduced from experiment IV. The same values shall be used for the VSM based retrieval system so that consistency is maintained in all the systems. By adjusting the cutoff point the system for the neural network would respond with optimized answers

Table 8 Categorization of IHAFR Systems based on training cut-off points

VSM IR	IHAFR = 0.1	IHAFR = 0.2	IHAFR = 0.3	IHAFR = 0.4	IHAFR = 0.5
S1	S2	S3	S4	S5	S6

4.9. Evaluation of the IHAFR System

General information system retrieval works on the principle that for each user question posed to the system they could be some documents which bear a relevance to the user’s question. FAQ information retrieval system adopts a slight twist over this approach by implying that for every user’s FAQ question posed they should be a ‘right FAQ question-answer’ that best addresses the user query[100]. heng et al [101] also subscribes to this notion by stating that users FAQ query can be retrieved in three kinds of situations from an FAQ information retrieval systems, that is 1: Having the same question in FAQ database; 2 Having similar questions in FAQ database and finally 3: No same or similar question in FAQ knowledge base. If an FAQ question is regarded to fulfill the conditions 1 or 2 then that user FAQ query has an equivalent FAQ question answer in the FAQ knowledge base . If the user query falls in the category 3 then no corresponding FAQ is present and there is no response.

Anderson et al [102] developed and evaluated a Healthcare FAQ information retrieval system using the same approach as [101]. They categorized their evaluation in two distinct scenarios as Top1 and Top5. For Top1 the system presents the highest ranked FAQ question answer and this response should have a ‘useful’ response to the user FAQ query i.e. similar in meaning or exact and the response is regarded as correct. For the Top 5 scenario, if they exists a ‘useful’ FAQ question answer amongst the sequential ranked five responses then an FAQ query has answer. Recall was used to measure the ability of the system to find the right answer and was defined as the percentage of questions for which the FAQ

systems retained a correct answer. Re action rate was also used to measure the FAQ system’s inability to provide a correct answer and was computed as percentage of unanswered questions.

Burke modified the traditional evaluating metrics for FAQ information retrieval systems and defined recall rate as a measure which specifies a ratio of all user FAQ questions with right answers N_C to all user FAQ questions with correct answers N_R as computed in equation 60. Re action rate is introduced to replace precision rate and is defined by equation 61 where N_{C_N} represents the number of questions which the system has completely failed to get an answers for and N_{R_N} indicates the total number of questions which the systems does not have the right answers or do not exists

$$\text{Recall Rate} = \frac{N_C}{N_R} \dots\dots\dots (60)$$

$$\text{Re action Rate} = \frac{N_{C_N}}{N_{R_N}} \dots\dots\dots (61)$$

heng et al [101] also modified the precision metric and computed the metric with a slight modification as shown in equation 62, where N is the total amount of test questions (namely users questions), a_i is the value that whether the answer to question i is true or not, if true , then $a_i = 1$, else $a_i = 0$.

$$P_{rec1} = \frac{1}{N} \sum_{i=1}^N a_i \dots\dots\dots (62)$$

The IHISM-FAQ information-retrieval system was evaluated by using the re action rate and the recall rate. Recall was computed as a percentage of questions for which the FAQ system finds the right answer when one exists and re action as the based on a defined cut off point [102]. FAQ FINDER [103] uses a question to question mapping methodology to determine FAQ questions, used the success or re action rate metrics to determine its effectiveness. Sneiders [38] uses the re action rate to evaluate an Automated FAQ answering system and he describes the evaluation metric as the system’s ability to report garbage if there is no answer to a paused question. Sneiders [38] further describes the traditional recall and precision metrics as the ability of the system to show a share of relevant FAQs among all retrieved FAQs and precision as a share of correctly reported none answers when there is no answers in the FAQ knowledge base. The two evaluation metrics are used to depict what the FAQ knowledge has at that moment and is capable of delivering since FAQs are dynamic.

The research identifies a number of common elements in the evaluation of all mentioned FAQ systems in comparison with the IHARS in the sense that it is a FAQ system. Therefore in evaluating the IHARS the research shall adopt the reaction rate metric which already has been defined as a performance measure which depicts the system's ability to show whether an answer exist or not in the system i.e. In addition the research shall also adopt the traditional performance measuring metrics as being recall rate. These evaluation metrics show how the system is able to relate its performance in terms of retrieving relevant FAQ question answers and coined to the system's ability in also failing to pick relevant FAQ question answers as well.

The IHAFRS neural network was evaluated using the reaction and recall rate measurements. The same evaluation metrics were used by Anderson et al [81] who implemented a similar FAQ retrieval system. Recall is calculated as the percentage of FAQ questions which are right, correct and are found in the retrieval system. Reaction is defined as percentage of questions the IHAFRS system without an equivalent matching question in the system. Precision is also defined as the questions returned by the systems and are found to be relevant or have similar meaning to the posed question. The next chapter gives an overview of the results discussion and evaluation.

CHAPTER 5

5. Experimental Evaluation

This chapter provides an overview discussion of the results achieved and constraints for the design and implementation of the IHAFR neural network system. The discussion is related to findings on the theoretical matters of HIV/AIDS FAQ retrieval system. The discussion is primarily based on the results achieved with regard to conformance and discrepancies to the general literature review on FAQ retrieval using neural network.

The IHAFR neural network system was sensitized into five ranking cutoff points which are shown in table 9 below. The ranking cutoff point is the total weight of HIV/AIDS FAQ when computed by the neural network system during similarity comparison. This value is used to retrieve HIV/AIDS FAQs based on a condition which is given to the system during the retrieval process. HIV/AIDS FAQs that have a ranking cut off point lower than the stated values are retrieved by the system. For instance, if we consider IHAFR = 0.1 it means retrieve all documents that have a HIV/AIDS FAQs weight of 0.1 and below shall be retrieved.

Table 9 Categorization of IHAFR Systems based on training cut-off points

VSM IR	IHAFR = 0.1	IHAFR = 0.2	IHAFR = 0.3	IHAFR = 0.4	IHAFR = 0.5
S1	S2	S3	S4	S5	S6

The advantage of sensitizing the system over a broad range of 0.1 to 0.5 values was meant to observe and detect where the system performance is best and optimum. The research study expects the system performance to be optimized at 0.1 ranking cut off points. At this ranking cut off point the IHAFR system should retrieve HIV/AIDS FAQs that have a similarity of 90% and above per posed user HIV/AIDS query. The importance of sensitizing the system and testing it at various cut off points shall enable the research study to observe the system performance and therefore propose other techniques to enhance the system performance suppose it does not show better performance.

5.1. IHAFRS Results

Based on the results obtained as tabulated in table 10 for the recall rate and the rejection rate metrics, graphical plots are done in figure 37 and 38 respectively. The graphical plots facilitate research to establish a qualitative explanation of the IHAFR behavior per various sensitized points per each system.

5.1.1. Recall Rate Performance

The recall rate expresses the percentage of HIV/AIDS queries for which the IHAFR system finds the 'right answer' when one exits[1] and the IHAFR research results for the sensitized systems are tabulated in table 10 and figure 37. The IHAFR pitched at a cutoff point 0.5 provides the least recall rate 73.33% and the highest is 79.1% at 0.2. All the IHAFR pitched at 0.4 to 0.1 cutoff points indicate similar performance. The system averagely performs the same in the Top1, Top5, Top10, and Top15 categories. They show an average performance recall rate of 38.33% in Top1, 61.67% in Top 5, 75.33% in Top 10 and finally 77.17% in Top 15. The Top 5 to Top 15 show better recall rate.

The HIV/AIDS FAQ keyword based system used as a benchmarking system, reflects a poor performance in all the ranking cut off points' categories of Top1, Top5, Top10 and Top15 as tabulated in table 10 and illustrated in figure 37. The system shows the least recall rate of 18.33% for Top1 to 55.83% for T5 compared to the IHAFR systems.

5.1.2. Rejection Rate Performance

The IHAFR systems reflect a very low rejection frequency rate than the key word based HIV/AIDS FAQ information retrieval system. Burke et al [103] defines rejection rate as a measure which represent the percent of questions that an FAQ system cannot correctly report as being unanswered in the file. The rejection rate for the key word based HIV/AIDS FAQ information retrieval system is very high compared to the rejection rate of the all the IHAFRS neural network based systems as shown in table 10 and figure 38. The key word based HIV/AIDS FAQ information retrieval system shows a good ability to reject if there is no answer in the FAQ system repository. The research need to test the key word based HIV/AIDS FAQ information retrieval system with none HIV/AIDS questions and see how it performs and secondly there is need to increase the FAQ collection so that the depth and coverage of FAQ questions becomes sufficient[1].

Table 10: Recall and Re action Rate Values

FAQ Information	Recall Rate				Re action Rate			
	Top 1	Top 5	Top 10	Top 15	Top 1	Top 5	Top 10	Top 15
VSM Based	18.33%	38.33%	48.33%	55.83%	82.50%	62.50%	52.50%	45.50%
Neural Network Cut Off Point 0.1	38.33%	62.50%	76.67%	77.50%	62.50%	38.33%	24.17%	23.33%
Neural Network Cut Off Point 0.2	38.33%	64.17%	76.67%	79.17%	62.50%	36.67%	24.17%	21.67%
Neural Network Cut Off Point 0.3	37.50%	61.67%	75.00%	77.50%	63.33%	39.17%	25.83%	23.33%
Neural Network Cut Off Point 0.4	38.33%	63.33%	75.83%	78.33%	62.50%	37.50%	25.00%	22.50%
Neural Network Cut Off Point 0.5	39.17%	56.67%	72.50%	73.33%	61.67%	44.17%	28.33%	27.50%

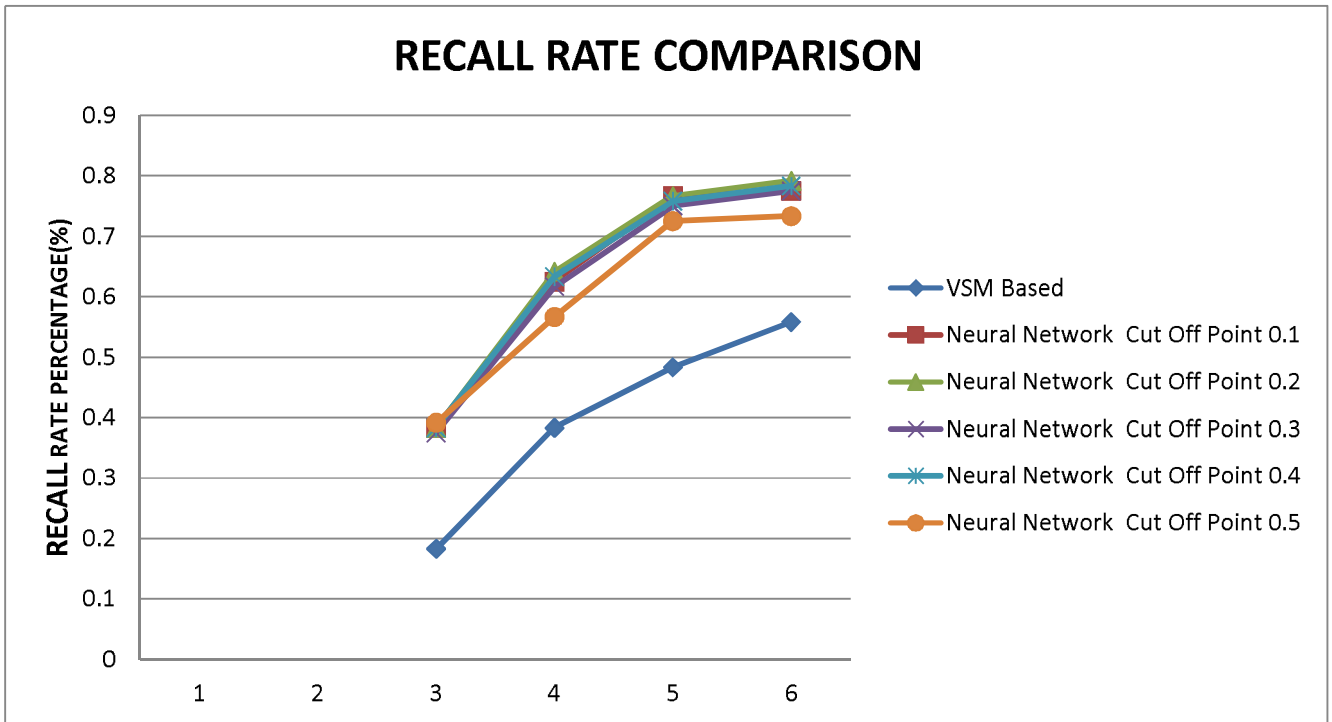


Figure 37: Plot of the Recall Rate Compared to the Responses of the Experimented Systems.

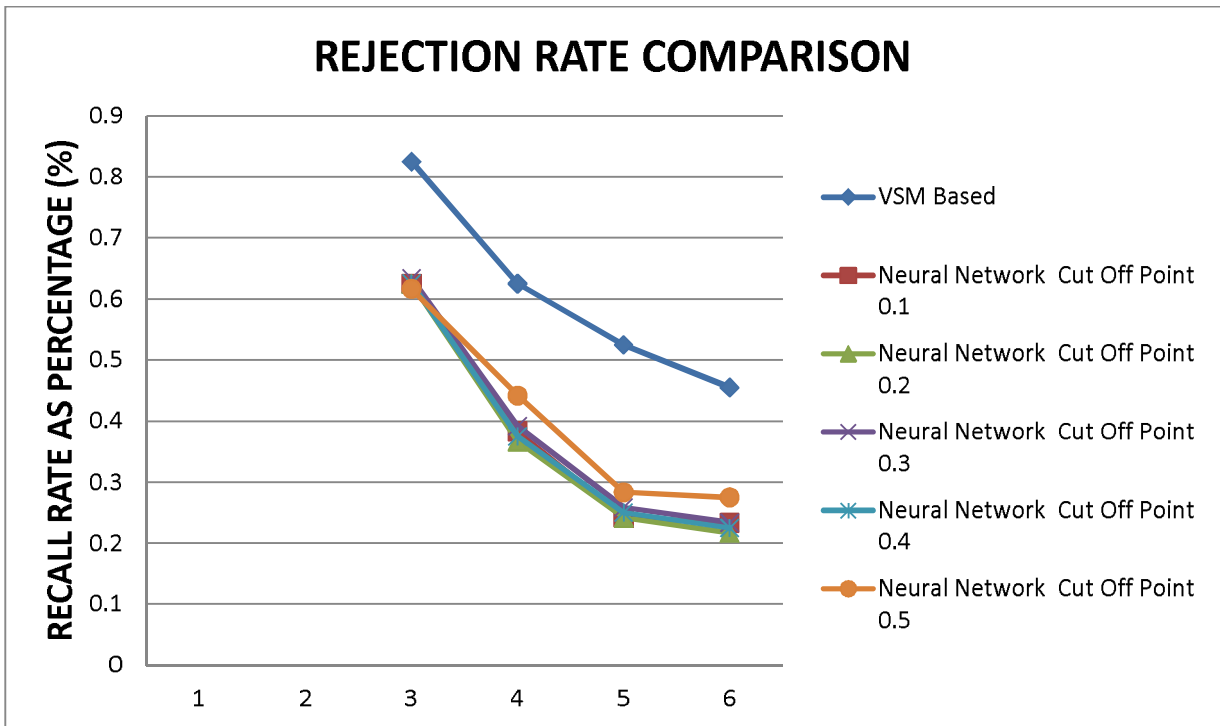


Figure 38: Plot of the Rejection Rate Compared to the Responses of the Experimented Systems[1].

5.2. Main Issues

The research study believes the amount of HIV/AIDS FAQs used were few and therefore might have an impact on the IHAFR neural network learning ability and generalization as remarked by Fei et al [104], the effectiveness of neural network retrieval can be enhanced by using larger training and test set to improve the generalization and recall rate of the system. As far as the research study is concerned there is no standardized and published HIV/AIDS FAQ corpus. A standardized and published HIV/AIDS FAQ corpus from the medical community might have increased the performance of the neural network training, testing and validation data samples as they would be many. The research compiled a paltry HIV/AIDS FAQ questions corpus largely from Botswana medical community in the form of two booklets the IPOLETSE and MASA HIV/AIDS FAQ questions and another source with scanty questions from WHO and New York City Health council. A corpus with 467 HIV/AIDS FAQ questions was compiled.

Obasa and Salim resolved that [105] “FAQ question answering system can conveniently answer about 80% of the questions asked by user if properly constructed and effectively managed” and in this research though the IHAFR neural network system constructed surpassed the performance of a key word based FAQ retrieval with good and remarkable recall rate [1]. However there is need to investigate and adopt complementing techniques that would improve the system to attain a better performance in yielding a better mapping for category of Top 1 for HIV/AIDS FAQ retrieval. It was observed that the recall rate for both the neural network and the key word based FAQ HIV/AIDS FAQ systems performance was below 50%. For the Top 5 and up to Top 15 the recall rate was good and contained a lot of similar and also exact HIV/AIDS FAQs to the provided HIV/AIDS query.

One approach that can enhance performance of neural network information retrieval besides the inherent heteroassociative memory capability of MLP is use of conceptually systems to compliment the functionality of the heteroassociative memory. This approach could be adopted to improve the neural network performance hence better recall and re cation rate could be attained. Huang et al [27] deliberate on a technique of creating concepts as units of knowledge, each bearing a unique knowledge and key advantages of this technique are they embed semantic knowledge, they disambiguate terms which have multiple meanings and the semantic property bound in the concepts can be numerically evaluated with related documents thus measuring similarity.

Huang et al [27] further elaborate the matter of concepts as unit of knowledge by stating that the concepts can be organized and structured according to relations among them and concept system can be created. Huang et al [27] mentions of successful conceptual systems such as the online electronic online encyclopedia Britannica , the Wikipedia and lexical ontology system like WordNet for English common words and HowNet for Chinese common words. An HIV/AIDS lexical ontology could be ideal complimentary conceptual systems that can used to improve the recall and re cation rate for neural network FAQ retrieval systems.

Another approach the research study could have implemented is to optimize the input HIV/AIDS query rather than use of the heavily manual preprocessing which involved HIV/AIDS query feature extraction and selection. Optimum query can be implemented used by using neural network systems like the Kohonen’s Self Organizing Map (SOM) using Recurrent Neural Network [77]with horizontal

interconnections of output neurons. The system will be train trained with samples of typical HIV/AIDS queries. The Kohonen's Self Organizing Map (SOM) neural network in turn will ad ust its self to select an optimized query based on the input HIV/AIDS query. This optimized query then shall be used as an input to the question to question mapping systems. Kohonen's Self Organizing Map (SOM) neural network creates a contesting environment amongst and between neurons and the fittest survive as the output for a given input. The best and fittest activated neuron would represent the optimum query to be submitted to the MLP for generalization of similar in meaning conceptually and pragmatic.

5.3. Experiment Limitations

The limitation that the research can account about is the unavailability of a published HIV/AIDS FAQs corpus. The research study believes such corpus could have sufficient HIV/AIDS FAQs that could have sufficient FAQs and that can be used to train the neural network system. As once mentioned earlier, this had a bearing on the performance of the IHAFR system.

Coupled to this limitation is also the challenge of adequate and precise knowledge designing neural network parameters. Many heuristic thumb rules for design and implementation of neural networks exists. The research took quite a lot of academic time to settle and understands the implications of designing and parameterizing a neural network so that it would accomplish classification task of HIV/AIDS FAQ questions from a presented query. Few research exists on knowledge of how to effectively design and derive appropriate neural network system parameters and the research took a lot of time on this area to derive the IHAFR system parameters.

CHAPTER 6

6.1. Conclusion

The aim of the research study was to design and implement an intelligent HIV/AIDS FAQ retrieval system using neural networks for retrieval of relevant and similar HIV/AIDS FAQs. The information retrieval system will use a question to question similarity matching technique to find related and similar FAQs

The research study compiled an electronic HIV/AIDS FAQ question corpus from reliable sources like MASA, IPOLETSE and United Nations World Health Organization HIV/AIDS FAQs booklets. An HIV/AIDS FAQ knowledge base was created using VSM and PCA techniques. An artificial neural network was designed and experimented for input and hidden neural nodes, activation function and training rule. Sampling based on the MATLAB divideint function was done to create the training data, testing and validation data. Backpropagation training rule using TRAINCGP was used to train the neural network. Early stop technique was used to prevent over training of the neural network and achieve better generalization.

Golden standard approach was used to evaluate the system performance by using HIV/AIDS experts who included HRDC HIV/AIDS training coordinators, lecturers participating in HIV/AIDS awareness campaigns and the students. HIV/AIDS keyword based retrieval system was used to benchmark HIV/AIDS FAQ retrieval system based neural network. General it was noted the HIV/AIDS FAQ retrieval system using neural network had better recall rate of 79.17% and traditional keyword based FAQ system 55.83% for equivalent or similar HIV/AIDS FAQs. Traditional keyword based FAQ retrieval system attained a re ection rate of 82.50%, compared to 61.67 % for neural network system

6.2. Research Study Conclusion

The study concludes the research based on analysis of attained results from the conducted experiment on the IHAFR system using question to question similarity matching compared to the traditional information retrieval using key words. The research study determined that IHAFR research findings general conform to other research studies [72, 76] who have also used the neural network technique for information retrieval.

The neural network FAQ retrieval system superiority is acknowledged on its ability to relate conceptually and semantic HIV/AIDS FAQ questions when using question to question similarity matching. This ability, to mimic how human being brains work [65] as they match and relate questions which have the same meaning but different words is a powerful trait found in artificial neural networks. The neural network retrieval system adopts this mode of working because of the training done and also the structural set up which has hidden neural nodes that act as parallel processors for non-linear data[71].

The ability by an information retrieval system which is neural network based, to generalize FAQs through question to question mapping reflects a very important characteristic. For instance in this research study, figure 33a shows an HIV/AIDS FAQ query presented to the system as “What causes AIDS”. In figure 33b it shows HIV/AIDS FAQs responses which are generalized and related but with different wording and this demonstrates the system ability to generalize. To HIV/AIDS FAQ retrieval this is an important characteristic as it is able find many related questions and thus provide more valuable information to the end user.

The most advantageous aspect of the neural network retrieval technique system is a mechanism to test whether the generated mathematical model is capable to generate or map the FAQ as expected by cross checking and validating with simulation [106]. The research used MATLAB functions like the linear regression analysis and correct classification error rate to check the feasibility of the system functionality in advance. This mechanism is not possible in all other traditional information retrieval techniques. This functionality is very important for HIV/AIDS FAQ retrieval questions because of the nature of the question we are dealing with i.e. health matters. More so to have a general estimation of how the neural network system is going to work without having to conduct an independent survey. In this research study the IHAFR managed to record a validation performance test of 66% percent in the laboratory. Such confidence tests are very important as they give a proof of better performance besides using other independent surveys. There is need to determine laboratorial that whatever is being deployed and whether it is meeting a defined and anticipated criterion.

For better performance in FAQ retrieval system there is need to optimize the presented FAQ query so that it is possible to attain the most prioritized and number ranked one FAQ question generated as a response. The starting point should be initiated by researching for standardized HIV/AIDS FAQ corpus like used by TREC for testing and evaluating QA systems. Another attention point is the development of an HIV/AIDS lexicon ontology [1] which can be used improve the functionality of neural network based information retrieval by implementing an ontology knowledge base as experimented in the following researches[107-109], they developed FAQ retrievals that had a high recall rate and good re ection rates.

6.3. Contribution of this Research

The research though works in the shadow of a partly researched area; neural network based information retrieval for HIV/AIDS FAQ retrieval. As far as the research is concerned no significant research work has been done in sharing knowledge on HIV/AIDS through FAQ questions using an FAQ information retrieval platform. It is envisaged that this information retrieval platform which uses an ‘intelligent’ approach can conveniently be used as the stepping stone to do more research work on sharing of HIV/AIDS information using information retrieval systems. It goes an untold that information sharing has become one of the effective ways used to mitigate the impact and spreading of HIV/AIDS and to date there is no cure for this disease so far. The research has gone to an extent of publishing a research paper [1] on the potential of utilizing neural network as an important tool for sharing information on HIV/AIDS in Health Informatics platform.

The research has created awareness for establishments of well researched and published electronic HIV/AIDS FAQ corpus [1]so that extended research could be done. There is need to have a research which should compile and publish an HIV/AIDS FAQ corpus like what is done in the TREC evaluation were a standardized corpus is used to test various QA systems. This development could lead to accelerated development of improved HIV/AIDS FAQ information retrieval systems. The HIV/AIDS corpus[1] could be extended to building an ontology for HIV/AIDS FAQ terms like the WORDNET and this could also accelerate the development of better information retrieval on FAQ systems for HIV and AIDS.

New knowledge and practice can be attributed to this research because it is one amongst the few researches which acts as the starting point of creating an electronic information retrieval system for HIV/AIDS FAQs using artificial neural networks. Currently other various means of information sharing HIV/AIDS have their deficiencies noted and therefore this approach introduces a complementary approach of sharing information in way similar to what an HIV/AIDS call centre does. The most important trait adopted by the IHAFR is the ability to mimic human beings who manner a call centre and providing related or similar HIV/AIDS FAQs bearing the answer to the user question using artificial neural networks.

6.4. Future Work

Improvements to neural network FAQ information retrieval could be done by pre-processing the HIV/AIDS user query using methods like competitive learning rules with apply Kohonen neural network method instead of using the tedious feature selection and extraction process used in this research. A comparison can be done with what this research has done and may better performance can be attained. Creation and implementation of an officially published and researched HIV/AIDS FAQ corpus since the epidemic has no cure and the best mitigating strategy is prevention through information sharing. Creation and implementation of an HIV/AIDS Lexicon Ontology that could help in expanding related words through getting their synonyms, hyponym and hypernym which could compliment the hetroassociative memory for the neural network by increasing the number of key words to train the neural network with.

References

- [1] G. Mlambo and Y. Ayalew, "Intelligent HIV/AIDS FAQ Retrieval System Using Neural Networks," in *Proceedings of the IASTED International Conference Health Informatics (AfricaHI 2014)*, 2014, pp. 304 - 308.
- [2] National AIDS Coordinating Agency, "HIV/AIDS in Botswana: Estimated Trends and Implications Based on Surveillance and Modeling," National AIDS Coordinating Agency, Gaborone, Botswana, 2008.
- [3] UNAIDS, "Global report: UNAIDS report on the global AIDS epidemic 2013," UNAIDS: Joint United Nations Programme on HIV/AIDS (UNAIDS), New York, United States, 2014.
- [4] USIAD Botswana. (2010, 2013 January 21). *USAID HIV/AIDS Health Profile for Botswana - September 2010* [Online]. Available: http://www.usaid.gov/our_work/global_health/aids/Countries/africa/botswana.html
- [5] National AIDS Coordinating Agency, "Botswana 2013 Global AIDS Response Report: Progress Report of the National Response to the 2011 Declaration of Commitments on HIV and AIDS," National AIDS Coordinating Agency, Gaborone, Botswana, Online, 2014.
- [6] AVERT. (2011, 2013 January 26). *HIV & AIDS In Botswana* [Online]. Available: <http://www.avert.org/aids-botswana.htm>
- [7] Ministry Of Health and Department Of HIV and AIDS Prevention & Care, "IPOLETSE Call Centre/ Hotline (Telephonic HIV/AIDS/STI Information and Counseling Services)," Gaborone, 2009.
- [8] Harvard School Of Public Health AIDS, "Knowledge, Innovation & Training Shall Overcome AIDS (KITSO)," Research & Programs, 2011.
- [9] A. Gordon. (2010, 2013 March 24). *Makgabaneng Radio Serial Drama* [Online]. Available: <http://www.comminit.com/en/node/123510/304>
- [10] I. Williams, L. Chigona, and J. Schatz. (2009, 30 April 2014). *An Assessment of HIV and AIDS Radio Campaign Messages in Botswana* [Online]. Available: <http://66.199.148.216/hiv-aids-southern-africa/content/assessment-hiv-and-aids-radio-campaign-messages-botswana>
- [11] Botswana Telecommunications Authority, "Annual Report 2010," Botswana Telecommunications Authority, Gaborone, 2010.
- [12] A. Masizana-Katongo and R. Morakanyane, "Representing Information for Semi-Literate Users: Digital Inclusion Using Mobile Phone Technology," *Community Informatics Research Network (CIRN) Conference, Prato, Italy, November, 2009.*, pp. 1 - 17, 2009.
- [13] M. Majewski and J.M. Zurada, "Sentence recognition using artificial neural networks," *Knowledge Based Systems*, vol. 21, pp. 629 - 635, 2008.
- [14] S. Quarteron and S. Manandar, "Designing an Interactive Open-Domain Question Answering System," *Natural Language Engineering* vol. 1 pp. 1-23, 2008.
- [15] Faqs.Org. (2011, 5 February 2013). *Usenet FAQ Archive Search* [Online]. Available: <http://www.faqs.org/faqs/>
- [16] A.J. Agrawal, "Using Domain Specific Question Answering Technique for Automatic Railways Inquiry on Mobile Phone " *5th International Conference on Information Technology: New Generations (ITNG 2008)*, Las Vegas, Nevada, USA, April 7-9, 2008,, vol. 205, pp. 1111 - 1116, 2008.
- [17] M.S.H Khyial, A. Khan, and S. Khalidb, "Mobile System Using Natural Language Annotations for Question Answering " *Computer Technology and Development, 2009. ICCTD '09. International Conference on* vol. 1 pp. 367 - 371, 2009.
- [18] M.R. Kangavari, S. Ghandchi, and M. Golpour, "A New Model for Question Answering Systems," *World Academy of Science, Engineering and Technology*, vol. 42 pp. 506 - 513, 2008.

- [19] Z. Yu, Y. Qiu, J. Deng, L. Han, C. Mao, and X. Meng, "Research On Chinese Faq Question Answering System In Restricted Domain," *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, 2007 International Conference , Hong Kong*, vol. 7, pp. 3927 - 3932, 2007.
- [20] A.I Isiaka and N. Salim, "Mining FAQ from Forum Threads: Theoretical Framework," *Journal of Theoretical and Applied Information Technology*, vol. 63, pp. 39 - 50, 2014.
- [21] K. Hammond, R. Burke, C. Martin, and S. Lytinen, "FAQ Finder: A Case-Based Approach to Knowledge Navigation," *Artificial Intelligence for Applications, 1995. Proceedings., 11th Conference on* pp. 80 - 86, 1995
- [22] J. Wen, J. Nie, and H. Zhang, "Clustering User Queries of a Search Engine," *ACM*, pp. 162 - 168, 2001.
- [23] Ministry of Health, *MASA Antiretroviral Therapy*,. Gaborone, Botswana: Government of Botswana, 2006.
- [24] A. Al_Molijy, I.Hmeidib, and I.Alsmad, "Indexing of Arabic documents automatically based on lexical analysis " *International Journal on Natural Language Computing (IJNLC)* vol. 1, pp. 1-8, 2012
- [25] L. Liu, M. Zhong, and R. Lu, "Measuring Word Similarity Based on Pattern Vector Space Model," in *Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on*, 2009, pp. 72-76.
- [26] E. Atlam, M. Fuketa, K. Morita, and J. Aoe, "Documents similarity measurement using field association terms," *Journal of Information Processing and Management* vol. 39 pp. 809–824, 2003.
- [27] L. Huang, D. Milne, E. Frank, and I.H. Witten, "Learning a concept-based document similarity measure," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, pp. 1593-1608, 2012.
- [28] V. A. Praksher, "Critical Challenges in Natural Language Processing," presented at the 3rd International CALIBER - 2005, Cochin, 2-4 February, 2005, © INFLIBNET Centre, Ahmedabad, 2005.
- [29] X. Zhang and H. Wang, "A Fast and Effective Method for Clustering Large-Scale Chinese Question Dataset," *The 3rd CCF Conference on Natural Language Processing and Chinese Computing*, vol. 334 - 345, 2014.
- [30] W.H. Gomaa and A.A. Fahmay, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications* vol. 68, pp. 975 - 8887, 2013.
- [31] K. Liu, W. Meng, and Y. Yu, "Discovery of similarity computations of search engines," presented at the Proceedings of the ninth international conference on Information and knowledge management, McLean, Virginia, USA, 2000.
- [32] S.Y. Yang, "An Ontological Multi-Agent System for Web FAQ Query," in *Machine Learning and Cybernetics, 2007 International Conference on*, 2007, pp. 2964-2969.
- [33] H. Imran and A. Sharan, "Thesaurus and Query Expansion," *International Journal of Computer science & Information Technology (IJCSIT)*, vol. 1, pp. 89 -97, 2009.
- [34] Oxford. (2012, 14 March 2015). *Oxford Dictionary*. Available: <http://www.oxforddictionaries.com/definition/english/online>
- [35] V.N. Gudivada, V.V. Ragahavan, W.I. Grosky, and R.Kasanagottu, "Information Retrieval On The World Wide Web," *IEEE Internet Computing*, pp. 58 -68, 1997.
- [36] H. Chen and J. Kim, "GANNET: A Machine Learning Approach to Document Retrieval," *Journal of Management Information Systems*, vol. 11, pp. 7-41, 1995.
- [37] R.R. Larson "Evaluation of advanced retrieval techniques in an experimental online catalog," *Journal of the American Society for Information Science* vol. 1, pp. 34 - 53, 1992.
- [38] E. Sneiders, "Automated FAQ Answering with Question-Specific Knowledge Representation for Web Self-Service," *IEEE Proceedings of the 2nd International Conference on Human System Interaction (HSI'09)*, pp. 298 - 305, 2009.
- [39] Y. Che-Yu Y, H. Cheng-Wei, C. Yu-Wei, and L. Yi-Chun, "Using Online Automated FAQ System to Promote Community Learning," *IEEE*, pp. 535 -540, 2008.

- [40] H. Kim and J. Seo, "Cluster-Based FAQ Retrieval Using Latent Term Weights," *IEEE Computer Society: Intelligent Systems* pp. 58 - 65, 2008.
- [41] L. Fangfang and L. Liu, "The Construction and Maintenance of the Frequently Asked Question," *IEEE*, pp. 296 -300, 2010.
- [42] Z. M. Juan, "An Effective Similarity Measurement for FAQ Question Answering System," in *Electrical and Control Engineering (ICECE), 2010 International Conference on*, 2010, pp. 4638-4641.
- [43] X. Liang and D.Wang, "Improved Sentence Similarity Algorithm Based on VSM and Its Application in Question Answering System," *IEEE*, pp. 368 - 371, 2010.
- [44] I. R. Silva, J. N. Souza, and K. S Santos, "Dependence among terms in vector space model," in *Database Engineering and Applications Symposium, 2004. IDEAS '04. Proceedings. International*, 2004, pp. 97-102.
- [45] C. Bratsas, V. Koutkias, E. Kaimakamis, P. Bamidis, and N. Maglaveras, "Ontology-based Vector Space Model and Fuzzy Query Expansion to Retrieve Knowledge on Medical Computational Problem Solutions," in *Proceedings of the 29th Annual International Conference of the IEEE EMBS Cité Internationale, Lyon, France, 2007*, pp. 3794 - 3797.
- [46] L. Hanxing, L. Xudong, and L. Caixing, "Research and Implementation of Ontological QA System based on FAQ," *Journal of Convergence Information Technology*, vol. Volume: 5 pp. 79 - 85, 2010.
- [47] G. Yongming, C. Dehua, and L. Jiajin, "An Extended Vector Space Model for XML Information Retrieval," in *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on*, 2009, pp. 797-800.
- [48] G. Wei, M. Bao, and S. Wu, "Research on Ontology-Based Text Representation of Vector Space Model," in *Database Technology and Applications (DBTA), 2010 2nd International Workshop on*, 2010, pp. 1-4.
- [49] L. Liu, M. Zhong, and R. Lu, "MeasuringWord Similarity Based on Pattern Vector Space Model," *IEEE Computer Soceity International Conference on Artificial Intelligence and Computational Intelligence*, pp. 72 - 76, 2009.
- [50] L. Xiaoli, W. Guoqing, J. Min, Y. Min, and W. Weiming, "Software architecture for a pattern based Question Answering system " *Software Engineering Research, Management & Applications*,, pp. 331 - 336 2007.
- [51] C.L. Sung , M. Day, H. Yen, and W. Hsu, "A Template Alignment Algorithm for Question Classification," *IEEE*, pp. 197 - 198, 2008
- [52] P. Yin, B. Bhanu., K. Chang, and A. Dong, "Reinforcement learning for combining relevance feedback techniques," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 510-515 vol.1.
- [53] W. Xiao-gang and L. Yue, "Web Personalization Method Based on Relevance Feedback on Keyword Space," in *Services Science, Management and Engineering, 2009. SSME '09. IITA International Conference on*, 2009, pp. 34-37.
- [54] P. Zhao-hui, Z. Jun, W. Shan, W. Chang-liang, and C. Li-zhen, "VSM-RF: A method of relevance feedback in Keyword Search over Relational Databases," in *IT in Medicine & Education, 2009. ITIME '09. IEEE International Symposium on*, 2009, pp. 738-744.
- [55] J. Fu, J. Xu, and K. Jia, "Domain Ontology Based Automatic Question Answering " *Computer Engineering and Technology*, vol. Volume: 2 pp. 346 - 349 2009.
- [56] Z. Yu, H. Zong, Y. Xu, J. Guo, Y. Mao, and X. Meng, "FAQ Extracting and Domain Filtering Based on Improved Bayes," in *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, 2009, pp. 108-112.
- [57] M. Farid and R. Yacine, "A document management methodology based on similarity contents," *Information Sciences*, vol. 158, pp. 15-36, 2004.
- [58] W. Song, M. Feng, N. Gu, and L. Wenyin, "Question Similarity Calculation for FAQ Answering," in *Third International Conference on Semantics, Knowledge and Grid*, 2007, pp. 298 - 301.

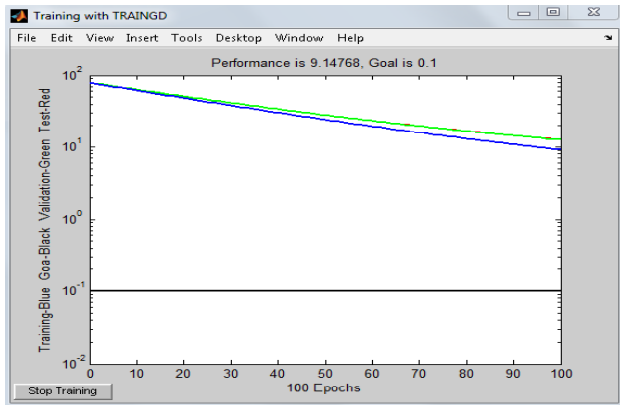
- [59] K. Gao, Y. Wang, and Z. Wang, "An efficient relevant evaluation model in information retrieval and its application," in *Computer and Information Technology, 2004. CIT '04. The Fourth International Conference on*, 2004, pp. 845-850.
- [60] S.Lynn and N. Yiu-Kai, "Using Vagueness Measures to Re-rank Documents Retrieved by a Fuzzy Set Information Retrieval Model," *IEEE Computer Society Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 39 - 43, 2008.
- [61] D. Zheng, T. Zha, F. Yu, S. Li, and H. Yu, "Research on Chinese Information Retrieval Based on a Hybrid Language Modelling," in *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006*, 2006, pp. 2586 - 2591.
- [62] J. Wen and Z. Li, "Improving Information Retrieval within Language Model Framework by Integrating Adjacent and Distant Relation," presented at the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.
- [63] Michie D, Spiegelhalter D.J., . (1994). *Machine Learning, Neural and Statistical Classification*
- [64] D. Svozil D, V. Kvasnicka, and J. Pospichal, "Introduction to Multi-layer Feed-forward Neural Networks," *Chemometrics and Intelligent Laboratory*, vol. 39, pp. 43 - 62, 1997.
- [65] A. Abraham, *Handbook of Measuring System Design*. Oklaham: John Wiley & Sons, Ltd, 2005.
- [66] B.M. Wilamowski, "Neural Network Architectures and Learning Algorithms," *IEEE Indurtsrial Electornics Magazine*, pp. 56 - 62, 2009.
- [67] P. S. Neelakanta and D. DeGross, *Neural Network Modeling: Statistical Mechanics and Cybernetic Perspectives*. Florida, United States of America: CRC Press, Inc, 1994.
- [68] P.Sibi, S.A.Jones, and P.Siddarth, "Analysis of Different Activation Functions Using Back Propagation Neural Networks," *Journal of Theoretical and Applied Information Technology*, vol. 47, pp. 1265 -1268, 2013.
- [69] B. Karlik and A. V. Olgac, "Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks," *International Journal of Artificial Intelligence And Expert Systems (IJAE)* vol. 1, pp. 111 - 122, 2010.
- [70] Z. Zainuddin and O. Pauline, "Function Approximation Using Artificial Neural Networks " *International Journal of Systems Applications, Engineering & Development*, vol. 1, pp. 173 - 178, 2007.
- [71] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification*: Ellis Horwood, 1994.
- [72] I. Mokris and L.Skovajsova, "Neural Network Model Of System For Information Retrival From Text Documents In Slovack Language " *ACTA Electrotechnica et Informatics*, vol. 5, pp. 1 - 6, 2005.
- [73] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial Neural Networks: A Tutorial," *IEEE*, pp. 31 - 44, 2006.
- [74] R. Chakraborty. (2010 2013 March 24). *Soft Computing* [Online]. Available: http://www.myreaders.info/html/soft_computing.html
- [75] K.L.Du and M.N.S.Swamy, *Neural Networks and Statistical Learning*. London: Springer-Verlag, 2014.
- [76] V. Mital and T. D. Gedeon, "A Neural Network Integrated with Hypertext for Legal Document Assembly," *IEEE*, pp. 533 - 539, 1992.
- [77] D. Guy, G. Robert, and P. Robert, "A Self - Organizing Map For Concept Classification in Infiramation Retrieval " in *Proceedings of International Joint Conference On Networks, Montreel, Canada, 2005*, pp. 1570 - 1574.
- [78] H. Chen and J. Kim, "GANNET: A Machine Learning Approach to Document Retrieval," *Journal of Management Information Systems*, vol. 11, pp. 7-41, 1995.
- [79] H. Chen, "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," *Journal of the American Society for Information Science*, vol. 46, pp. 194-216, 1995.

- [80] S. Sahay, B. Ravisekar, S. Venkatasubramanian, A. Venkatesh, P. Prabhu, and A. Ram, "iReMedI - Intelligent Retrieval from Medical Information," in *Advances in Case-Based Reasoning*. vol. 5239, K.-D. Althoff, R. Bergmann, M. Minor, and A. Hanft, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 487-502.
- [81] G.Anderson, S.D Asare, Y. Ayalew, D. Garg, B. Gopolang, A. Masizana-Katongo, O. Mogotlhwane, D. Mpoeleng, and H.O. Nyongesa, "Towards a Bilingual SMS Parser for HIV and AIDS Information Retrieval in Botswana," *IEEE*, pp. 1 - 5, 2009.
- [82] Y. Singh, P. K. Bhatia, and O. Sangwan, "A Review of Studies on Machine Learning Techniques," *International Journal of Computer Science and Security*, vol. 1, pp. 70 - 84, 2009.
- [83] A.Khan, B.Baharudin, L.H Lee, and K.Khan, "A Review of the Machine Learning Algorithms for Text Documents Classification," *Journal of advances in Information Technology*, vol. 1, pp. 2 - 20, 2009.
- [84] Z. Peng, J. Zhang, S. Wang, C. Wang, and L. Cui, "VSM-RF: A Method of Relevance Feedback in Keyword Search over Relational Databases," pp. 738 - 744, 2009.
- [85] D.P. Turney and P.Pantel, "Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* vol. 37, pp. 141 - 188, 2010.
- [86] Wikipedia. (2013, 2014 March 25). *Sparse Matrix: An Overview* [Online]. Available: http://en.wikipedia.org/wiki/Sparse_matrix
- [87] B.Rosario, "Latent Semantic Indexing: An Overview," *Infosys 240 Spring 2000*, pp. 1 - 16, 2000.
- [88] M. Can, "Principal Computer Analysis of Soft Computing and Neural Networks for Authorship Attribution," *Southeast Europe Journal of Soft Computing*, vol. 1, pp. 99 -113, 2012.
- [89] L. Muflikhah and B. Baharudin, "Document Clustering Using Concept Space and Cosine Similarity Measurement," in *Computer Technology and Development, 2009. ICCTD '09. International Conference on*, 2009, pp. 58-62.
- [90] O.Vikas, K.A. Meshram, G.Meena, and A.Gupta, "Multiple Document summarization Using Principle Component Analysis Incorporating Semantic Space Model," *The Association for computational Linguistics and Chinese Language Processing*, vol. 13, pp. 141 - 156, 2008.
- [91] A. Niemistö, "Statistical Analysis of Gene Expression Microarray Data," 2005.
- [92] FAQ ORG. (2012, 21 February 2014). *comp.ai.neural-nets FAQ, Part 3 of 7: Generalization* [Online]. Available: <http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-10.html>
- [93] K. G. Sheela and S.N. Deepa, "Review on Methods to Fix Number of Hidden Neurons in Neural Networks," *Mathematical Problems in Engineering* vol. 2013, pp. 1 - 11, . 2013.
- [94] S. Walezak and N. Cerpa, "Heuristic Principles for the Design of Artificial Neural Networks," *Information and Software Technology*, vol. 41, pp. 109 - 119, 1999.
- [95] I.Kaastra and M.Boyd, "Designing a Neural Network for Forecasting Financial and Economic Time Series," *Neuro-computing* vol. 10, pp. 215 - 236, 1996.
- [96] T. Masters, *Practical neural network recipes in C++*: Academic Press Professional Inc., 1993.
- [97] T. Mandal, "Vague Transformations in Information Retrieval," *Social Science Information Centre, Knowledge Management and Communication Systems*, vol. 6, pp. 312 -325, 1998.
- [98] L. Yu, S. Wang, and K. L. Keng, "Neural Network Metalearning for Parrallel Textual Information Retrieval," *International Journal of Artificial Intelligence*, vol. 1, pp. 57 - 73, 2008.
- [99] E.R Smith, "What Do Connectionism and Social Psychology Offer Each Other?," *Journal of Personality and Social Psychology*, vol. 70, pp. 893-912, 1996.
- [100] M. Zhang, T. He, and F. Yang, "The Model Research of FAQ Answering System Based on Concept," in *Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on*, 2009, pp. 7-10.
- [101] L. Zheng, Z. Chen, and H. Huang, "Design and Implementation of FAQ Automatic Return System Based on Similarity Computation," *Journal of Natural Sciences*, vol. 11, 2006.

- [102] G. Anderson, Y. Ayalew, P.A. Mokotedi, N.P. Motlogelwa, D. Mpoeleng, and E. Thuma, "Health Care FAQ Information Retrieval Using a Commercial Database in Management System," in *Proceedings of the 2nd IASTED Africa Conference on Modelling and Simulation (AfricaMS 2008)*, Gaborone, Botswana, 2010, pp. 307-313.
- [103] R. D. Burke, K. J. Hammond, V. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, "Question Answering from Frequently Asked Question Files," *American Association For Artificial Intelligence Magazine*, vol. 18, pp. 57 - 66, 1997.
- [104] T. Fei, W.J. Heng, K.C. Toh, and T. Qi, "Question classification for e-learning by artificial neural network," in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, 2003, pp. 1757-1761 vol.3.
- [105] A.I. Obasa and N. Salim, "Mining FAQ From Forum Threads Theoretical Framework," *Journal of Theoretical and Applied Information*, vol. 63, pp. 39 - 50, 2014.
- [106] M. H. Beale, T.M. Hagan, and B.H. Demuth, "Neural Network Toolbox Users Guide: R2013b," MathWorks, MA, USA, 2013.
- [107] F. Wang, G. Teng, L. Ren, and J.B. Ma, "Research on Mechanism of Agricultural FAQ Retrieval Based on Ontology," presented at the Computer Society: Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008.
- [108] C. Hsu C, S. Guo, R. Chen, and D. S, "Using Domain Ontology to Implement a Frequently Asked Questions system," *IEEE Computer Society, Proceedings of the World Congress on Computer Science and Information Engineering*, pp. 714 - 718, 2009
- [109] J. Yeh, M. Chen, and C. Wu, "Semantic Inference Based on Ontology for Medical FAQ Mining " *IEEE.*, pp. 710 - 715, 2003.

APPENDIX 1: Backpropagation Training Algorithm Variants and Results

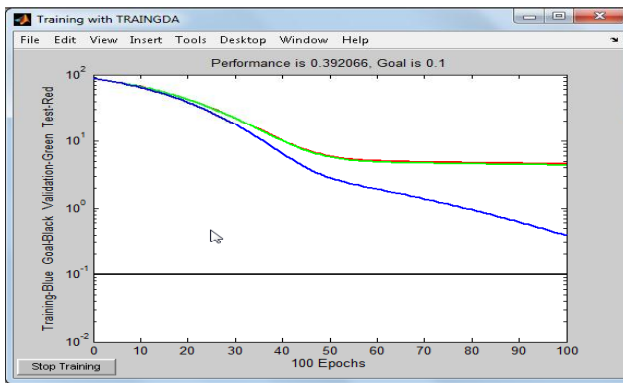
MATLAB Diagram



Training Parameters

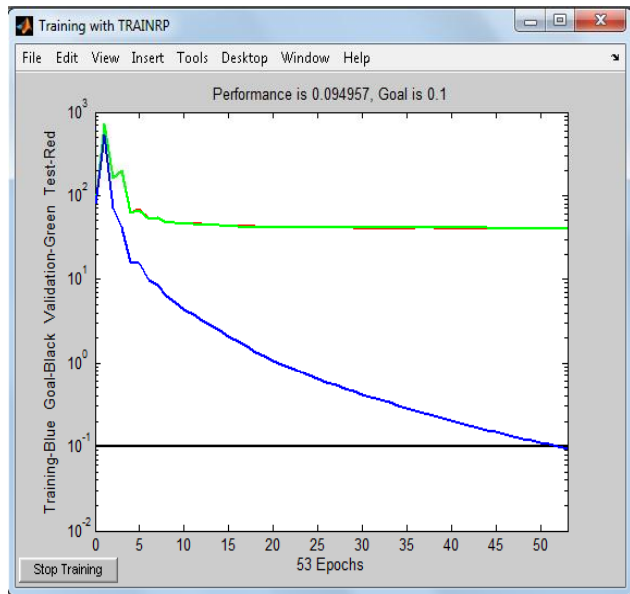
TRAINGD, Epoch 0/100, MSE 79.6573/0.1, Gradient 14.1431/1e-010
TRAINGD, Epoch 100/100, MSE 9.14768/0.1, Gradient 3.92637/1e-010
TRAINGD, Maximum epoch reached, performance goal was not met.

Elapsed time is 56.456000 seconds.



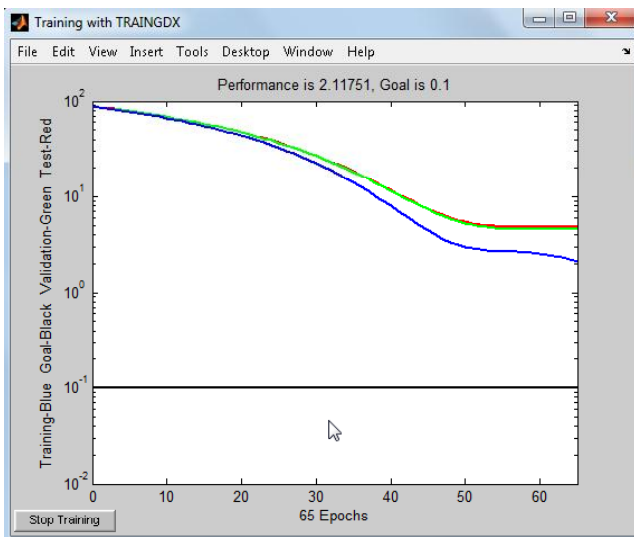
TRAINGDA, Epoch 0/100, MSE 89.085/0.1, Gradient 15.2139/1e-006
TRAINGDA, Epoch 100/100, MSE 0.392066/0.1, Gradient 0.120435/1e-006
TRAINGDA, Maximum epoch reached, performance goal was not met.

Elapsed time is 55.286000 seconds.



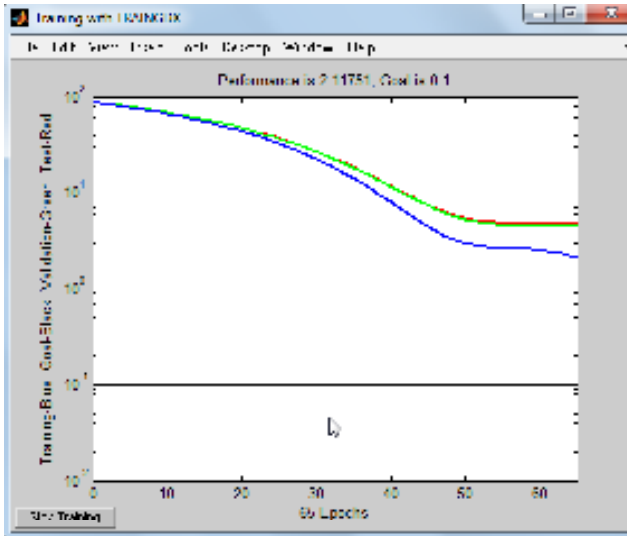
**TRAINRP, Epoch 0/100, MSE
79.6573/0.1, Gradient 14.1431/1e-006**
**TRAINRP, Epoch 53/100, MSE
0.094957/0.1, Gradient 0.0323344/1e-006**
TRAINRP, Performance goal met.

Elapsed time is 34.929000 seconds.



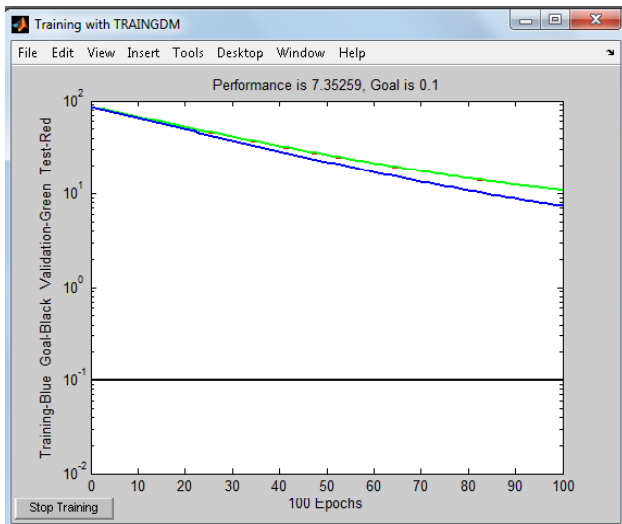
**TRAINGDX, Epoch 0/100, MSE
89.085/0.1, Gradient 15.2139/1e-006**
**TRAINGDX, Epoch 65/100, MSE
2.11751/0.1, Gradient 0.842048/1e-006**
TRAINGDX, Validation stop.

Elapsed time is 36.692000 seconds



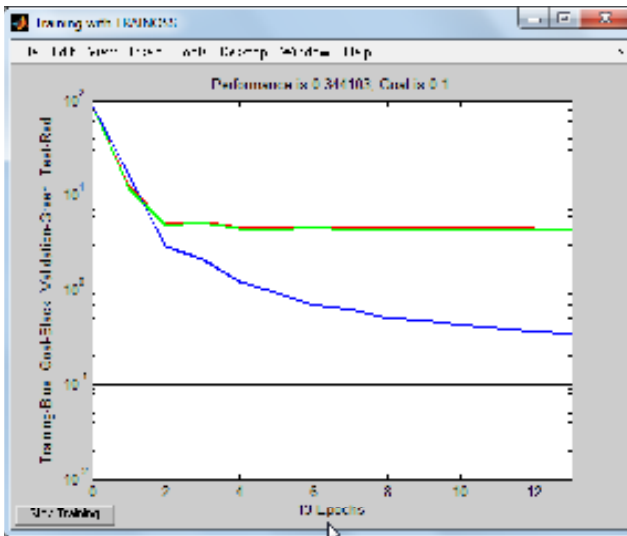
TRAINGDX, Epoch 0/100, MSE 89.085/0.1, Gradient 15.2139/1e-006
TRAINGDX, Epoch 65/100, MSE 2.11751/0.1, Gradient 0.842048/1e-006
TRAINGDX, Validation stop.

Elapsed time is 36.692000 seconds



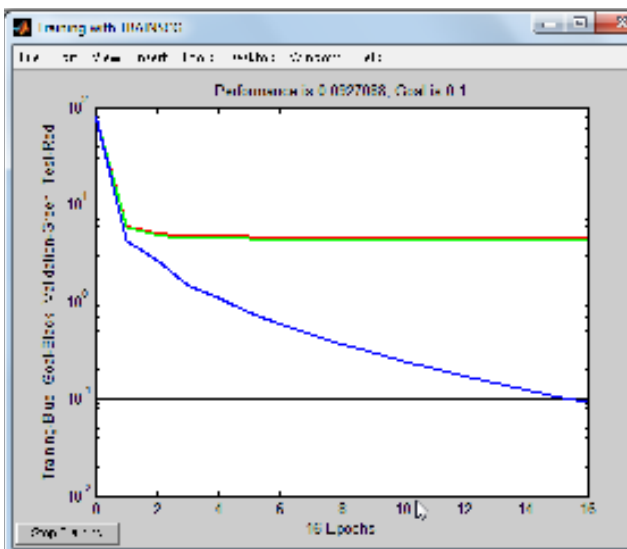
TRAINGDM, Epoch 0/100, MSE 86.2903/0.1, Gradient 14.9601/1e-010
TRAINGDM, Epoch 100/100, MSE 7.35259/0.1, Gradient 3.35944/1e-010
TRAINGDM, Maximum epoch reached, performance goal was not met.

Elapsed time is 55.287000 seconds..



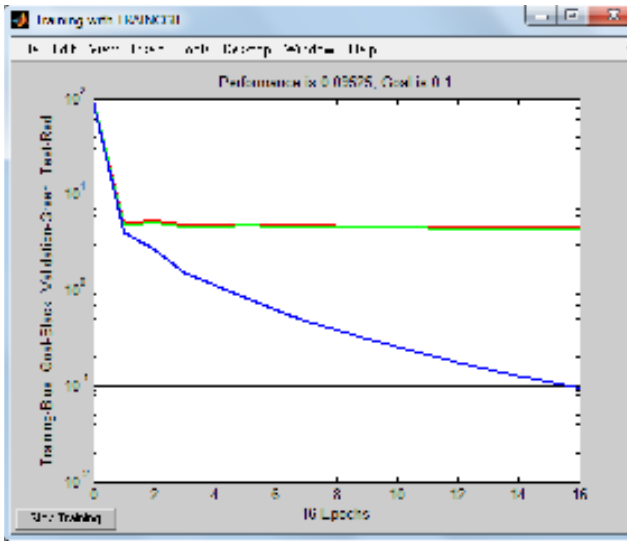
TRAINOSS-srchbac, Epoch 0/100, MSE 89.1071/0.1, Gradient 15.5719/1e-006
TRAINOSS-srchbac, Epoch 13/100, MSE 0.344103/0.1, Gradient 0.188978/1e-006
TRAINOSS, Validation stop.

Elapsed time is 15.085000 seconds.



TRAINSCG, Epoch 0/100, MSE 79.6573/0.1, Gradient 14.1431/1e-006
TRAINSCG, Epoch 16/100, MSE 0.0927088/0.1, Gradient 0.104696/1e-006
TRAINSCG, Performance goal met.

Elapsed time is 19.687000 seconds.

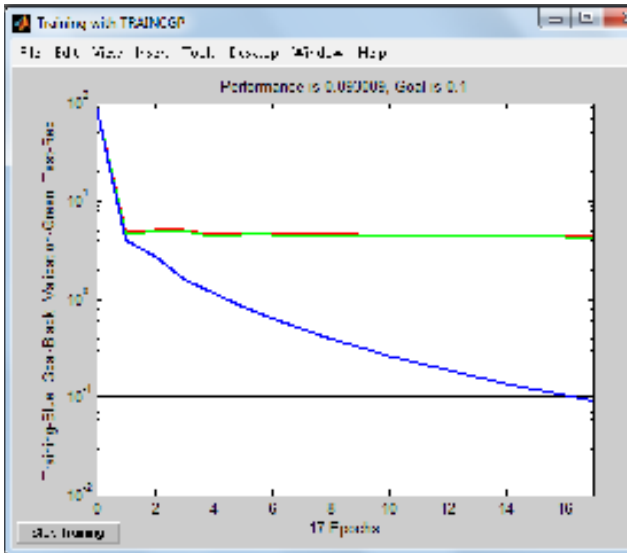


TRAINCGB-srchcha, Epoch 0/100, MSE 89.085/0.1, Gradient 15.2139/1e-006

TRAINCGB-srchcha, Epoch 16/100, MSE 0.09525/0.1, Gradient 0.106247/1e-006

TRAINCGB, Performance goal met.

Elapsed time is 16.536000 seconds

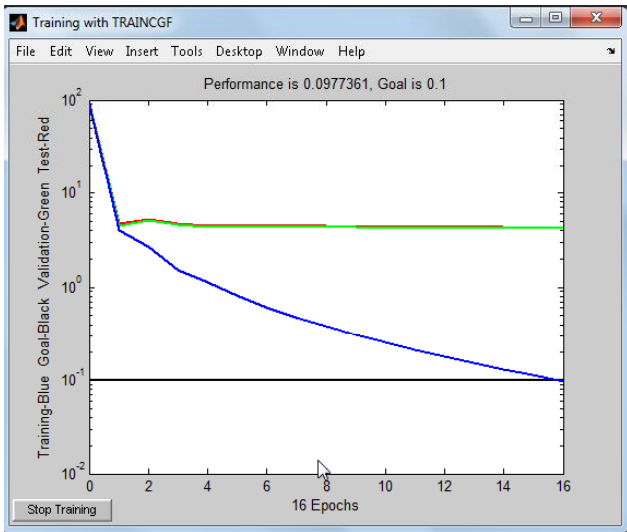


TRAINCGP-srchcha, Epoch 0/100, MSE 86.2903/0.1, Gradient 14.9601/1e-006

TRAINCGP-srchcha, Epoch 17/100, MSE 0.093009/0.1, Gradient 0.106837/1e-006

TRAINCGP, Performance goal met.

Elapsed time is 17.987000 seconds.



TRAINCGF-srchcha, Epoch 0/100, MSE 89.1071/0.1, Gradient 15.5719/1e-006
TRAINCGF-srchcha, Epoch 16/100, MSE 0.0977361/0.1, Gradient 0.108087/1e-006
TRAINCGF, Performance goal met.

Elapsed time is 16.317000 seconds.

APPENDIX 2: Research Questionnaire for HIV/AIDS FAQ Questions.

Godfrey Mlambo is conducting a research to answer Frequently Asked Questions (FAQs) on HIV/AIDS using Artificial Neural Network technique. A prototype system has been developed and has yielded results which must be judged by people to determine the ability of this system to generalize similarity matching of a posed FAQ to a stored FAQ question in the system. The systems accepts a posed FAQ question and finds equivalent FAQ question(s) in the system then determine the answers.

You are kindly asked to assist in evaluating the ability of the system to match a posed FAQ question tagged as unseen question by evaluating the system generated FAQ questions, tagged as target question. Your evaluations shall be kept confidential and shall only be used for improving the quality and effectiveness of this research.

Instructions on how to evaluate the FAQ questions:

From the questionnaire the column Unseen Question there is one question provided. In the column Target Question there are 15 or less questions provided which have been generated by the system in response to the question in the Unseen Question column. Determine exact or equivalent or similar in meaning question(s) in the Target Question column. Question under the title target question and write in order of correctness by inserting a numerical value.

In the first column which reads Best 1 select the one question in the Target Question column that best answers the unseen question indicating by writing 1.

In the second column which reads Best 5 select the best five questions in the Target Question column that best answers the unseen question indicating by writing 1 to the first and 2 to the second question until up to the fifth question. In the third column which reads Best 10 select the ten questions in the Target Question column that best answers the unseen question indicating by writing 1 to the first and 2 to the second question until up to the tenth question. In the third column which reads Best 15 select the fifteen questions in the Target Question column that best answers the unseen question indicating by writing 1 to the first and 2 to the second question until up to the fifteenth question.

No	Unseen Question	Target Question(s)	Best 1	Best 5	Best 10	Best 15
			Write number in preference of ranking order [S1]			
1	Apart from HIV, which other STDs are deadly	154 how does HIV infection differ from other viruses which infect human beings				
		5 are health care workers or people in other occupations at risk for HIV				
		6 are healthcare workers at risk from HIV through contact with infected patients				
		7 are lesbians or other women who have sex with women at risk of HIV				
		57 can I get HIV from a toilet seat by being bitten by an infected mosquito or from a swimming pool				
		58 can I get HIV from a toilet seat or by being bitten by an infected mosquito or from a swimming pool				
		60 can I get HIV from casual contact shaking hands hugging using a toilet drinking from the same glass or the sneezing and coughing of an infected person				
		61 can I get HIV from contact with my doctor dentist or other health care professional				
		66 can I get HIV from living in the same house as a person with HIV or AIDS				
		116 does the presence of other sexually transmitted diseases STD facilitate HIV transmission				

⋮

120	Are condoms a hundred percent effective to prevent the spread of HIV/AIDS?	25 can a mother keep taking DV A T after the baby is born for her own health				
		29 can a woman who has HIV pass the virus to her baby				
		149 how does a mother transmit HIV to her unborn child				
		200 how would I one know if a baby born to an HIV positive woman has the HIV infection				
		396 why do some HIV positive mothers transmit the virus to their babies while others do not				
		134 how can I know the HIV status of the person i am going to marry				
		177 how long can the virus live outside the human body				
		20 are not all babies born to positive mothers infected with HIV				
		28 can a woman give HIV to a man during vaginal intercourse				
		33 can doctors notify the partners of a patient with HIV without the patient's permission				