



Information extraction for business process enhancement using Natural Language Processing and Machine Learning

W.A.D.A. Wicrama Arachchi

Reg. No.: MS21901058

M.Sc. in IT

Supervisor: Ms. Anjalie Gamage

October 2022

**Department of Information Technology
Sri Lanka Institute of Information Technology**

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Ms.Anjalie Gamage

Approved for MSc. Research Project:

.....

Head/<Department >

Approved for MSc:

.....

Head – Graduate Studies-Dr.Anuradha Jayakody

DECLARATION

I declare that this is my own research project thesis, and this document does not incorporate without acknowledgement any material previously published submitted for a Degree or Diploma in any other university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Sign: 

Name: W.A.D.A .Wicrama Arachchi

Date: 25/11/2022

ABSTRACT

As part of the digital era, data became more important than ever. Especially the activities at work align with more text-based information. Over the past years, research on data has become a rapidly growing area with continuously innovative techniques. As a result, nowadays Business Intelligence, Big Data Analysis, No SQL Analysis, and other data science tools are processing huge amounts of data to provide business patterns and trends related to business fields. Studies on unstructured data such as text-based data, PDF documents, videos, and images were not captured properly to provide more insightful information.

Text-based data within a company can be an extremely rich source of information. Therefore, it is very important to extract insights from this unstructured data. Extracting information from unstructured documents like product catalogs can be a difficult task due to their unorganized nature. Currently, in the real world, there is no such system to gain insight into manuals or product catalogs easily. As part of the job activities of electrical engineers, referring to product manuals and catalogs is a recurrent task. They have to spend considerable time on this task. This directly impacts the efficiency and productivity of an employee. Especially in the electrical industry, engineers have to go through a lot of product catalogs to find more information on a single item.

Over the past ten years, Natural Language Processing and Machine learning have had a major impact on business processes. It is a known fact that NLP and ML are becoming the top enterprise-level technologies that enable to perform business tasks that were impossible to reach. There were many technologies introduced to fill the gaps and meet the requirements. However, NLP and ML are becoming more popular than the other technologies in the industry.

In this research, I'm providing a concept that gains more insightful information from unstructured data such as product catalogs. The research reading is to develop a digital assistant with the use of NLP and ML where electrical engineers can submit their queries and get the information about their products easily. This will increase the efficiency and productivity of the electrical engineer as it will provide a method to avoid time consuming activities such as reading product catalogs.

ACKNOWLEDGEMENT

I would like to dedicate this research project to my family members and my research supervisor, Ms. Anjalie Gamage. Also, I want to thank my supervisor for accepting this project. Also, I would like to thank my supervisor for working with me throughout the project and providing a successful solution to businesses using Natural Language Processing and Machine Learning. Also, I would like to extend my deepest gratitude to everyone who gave me support to make this project successful. Apart from that, employees in the IPD group have given me great support in identifying the requirements of this project. Therefore, I would like to extend my gratitude to all the employees.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT.....	iv
TABLE OF CONTENTS.....	v
List of Figures	viii
List of Tables	ix
List of Abbreviations	x
Chapter 1 Project Introduction	1
Content	1
1.1 Chapter Overview	1
1.2 Introduction	1
1.3 Research Contribution	3
1.4 Problem definition	4
1.5 Importance of the Digital Assistance	6
1.6 Features and outcome Digital Intelligent Assistant	7
1.7 Hardware & Software Requirements.....	8
1.8 Structure of the Thesis.....	9
1.9 Chapter Summery	9
Chapter 2 Project Formulation	10
2.1 Chapter Overview	10
2.2 Project Aim.....	10
2.3 Research Question.....	11
2.3.1 Academic research question.....	11
2.3.2 Cooperate research question.....	11
2.4 Main Objective	11
2.5 Sub Objectives.....	11
2.6 Research Approach	11
2.7 Chapter Summery	12
Chapter 3 Literature Review	13
3.1 Chapter Overview	13
3.2 Literature Review.....	13
3.3 Summary of Findings and Research Gap.....	18
3.4 Chapter Summary	20
Chapter 4 Methodology.....	21

4.1 Chapter Overview	21
4.2 Data Gathering.....	21
4.2.1 Empathize	22
4.2.2 Data Description	23
4.2.3 Data collection and Integrations.....	24
4.2.4 Data Records.....	24
4.2.5 Define.....	24
4.2.6 Ideate	26
4.3 Methodology Stages Description	28
4.4 Research Approach and Design	30
4.4.1 Prototype	30
4.5 System Architecture- Detailed Overview Diagram	32
4.6 Lemmatization	35
4.7 Word Tokenization.....	36
4.8 Stop Words Removal.....	36
4.9 Lowercasing	37
4.10 Digital Assistance Interface.....	39
4.11 NLP Powered Digital Assistants.....	40
4.12 Detail Architecture of Actual system	42
4.13 Data Flow Diagram.....	43
4.14 System Design and Experimental Results	44
4.15 Limitation	48
4.16 Research Outcome.....	48
4.17 Risk Management	48
4.18 Chapter Overview	50
Chapter 5 Result and Discussions	51
5.1 Chapter Overview	51
5.2 Experiment Results	51
5.3 Testing and Refinements	53
5.4 Testing Methods	53
5.4.1 Unit Testing.....	53
5.4.2 Developer Testing	54
5.4.3 Functional Testing.....	55
5.5 System Validation	56
Chapter 6 Working Plan and Time Schedule.....	59
Chapter 7 Feasibility.....	60

7.1 Chapter Overview	60
7.2 Financial Feasibility	60
7.3 Technical Feasibility	60
7.4 Resource and Time Feasibility.....	61
7.5 Purpose And Objective of The Development Impact assessments	61
7.6 Key Outcome of the Development Research.....	62
7.7 Achieved Objectives.....	62
7.8 Future work.....	63
7.9 Chapter Summary	63
Chapter 8 Conclusion	64
References	65
Appendix	66

List of Figures

Figure 1 -Industrial Intelligent Digital assistant	5
Figure 2 - Digital Assistants Techniques.....	19
Figure 3 Azure Cognitive Knowledgebase.....	31
Figure 4 Digital assistance basic interface	32
Figure 5-System Architecture	33
Figure 6 Lemmatization Technique.....	36
Figure 7 Stopwording Removal.....	36
Figure 8 Lower Casing Sample Code	37
Figure 9 Data Pre Processing Steps.....	37
Figure 10 -Product data	38
Figure 11 -Proposed Digital Assistant	41
Figure 12 Detail Architectural Diagram.....	42
Figure 13 Data Flow Diagram.....	44
Figure 14 Azure Resource Group	44
Figure 15 App Service Plan.....	45
Figure 16 Digital Assistant Resource Utilization	45
Figure 17 Service Utilization	46
Figure 18 Knowledgebase	46
Figure 19 Knowledgebase QnA pairs	47
Figure 20 Digital Assistant Interface	47
Figure 21 Azure Bot Framework JSON View	51
Figure 22 Web Caht bot	52
Figure 23 Sample Knowledgebase	52
Figure 24 Digital Assistant Interface	53
Figure 25 -Testing and Validation	55
Figure 26 System Understandability	56
Figure 27 System Interactivity	56
Figure 28 Date Retrieval.....	57
Figure 29 System Usability.....	57
Figure 30 System Overall Impression.....	58
Figure 31 Proposed Timeline	59

List of Tables

Table 1 -Summary of data collection methods 22
Table 2 - Risk Management 49

List of Abbreviations

NLP -Natural Language Processing

ML -Machine Learning

BI -Business Intelligence

BIAS – Business Intelligence Analytical Systems

NLPTK – Natural Language Processing Tool Kit

IE – Information Extraction

TC -Text Clustering

NER -Named Entity Recognition

AI -Artificial intelligence

LUIS - Language understand Intelligence Service

Azure ML -Azure Machine Learning

PIM- Product Information Management