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 >

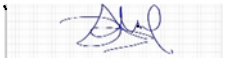
Approved for MSc:

……...

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

List of Figures

List of Tables

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