



جامعة طيبة

كلية علوم وهندسة الحاسب الآلي

(قسم الطلاب)

KINGDOM OF SAUDI ARABIA

Ministry of Education

Taibah University

College of Computer Science and

Engineering

(Male Section)

**AI-Powered Requirements Analysis Modeling**

**Graduation Project 2**

**by**

Ali Talal Al-Ahmadi 4100379

Mohammed Hadi Al-Harbi 4101704

Rami Ramadan Al-Mohammadi 4101758

**A project submitted in partial fulfilment of the requirements for the degree of Bachelor of Science Computer Science**

**Supervised by**

Dr. Mohammad M. Alsuraihi

1st Semester - Academic Year 1445 (2023/2024)

# Abstract

The project may face issues with excessive resource consumption, such as time or costs, and frequent changes in requirements can lead to project delays and increased expenses. Additionally, sometimes the techniques or tools used in the analysis and design process are ineffective or outdated. Software maintenance constitutes a pivotal stage within the software development lifecycle, encompassing a substantial portion, varying from 40% to 80% of the total expenses associated with software development. It's worth highlighting that a notable 60% of the overall maintenance expenditure is dedicated to the improvement of existing software functionalities. Consequently, it becomes imperative to meticulously prepare appropriate software documentation at each developmental phase in order to alleviate the financial burdens of maintenance. Using artificial intelligence (AI) to analyze requirements and generate drawings or models can be an effective solution for saving time and costs in the design and analysis processes. The crux of reducing maintenance costs lies in enhancing one's grasp of the software system, as understanding a software system accounts for roughly 50% of the time spent in the maintenance phase.

**Keywords** NLP; ML; AI-Powered; Requirements; Analysis; UML

# Acknowledgement

Alhamdulillah. First and foremost, all praise is due to Allah SWT the worthy of all the praises and compliments for giving us the strength and ability to complete this project. Next, we express gratitude to our parents and families. We would like to extend our thanks to our project supervisor, Dr. Mohammad M. Alsuraihi, who guided us in executing this project, providing invaluable advice, assisting us in challenging times, and significantly contributing to the project's completion.

We also want to express our appreciation to all the professors at Taibah University, College of Engineering and Computer Science, who have been part of our academic journey at the university.

Finally, and not least, we would like to thank those who worked on this project: Ali Talal Al-Ahmadi, Mohammed Hadi Al-Harbi, and Rami Ramadan Al-Mahmoudi, for their contributions to this project. I appreciate their hard work and dedication.

**Contents**

[Abstract ii](#_Toc158653165)

[Acknowledgement iii](#_Toc158653166)

[List of Figures vi](#_Toc158653167)

[List of Tables vii](#_Toc158653168)

[List of Abbreviations viii](#_Toc158653169)

[1 Chapter 1: Introduction 1](#_Toc158653170)

[1.1 Introduction 1](#_Toc158653171)

[1.2 Problem Definition 1](#_Toc158653172)

[1.3 Project Aim and Objectives 2](#_Toc158653173)

[1.4 Project Methodology 2](#_Toc158653174)

[1.5 Project Timeline 4](#_Toc158653175)

[1.6 Document Organization 5](#_Toc158653176)

[1.7 Summary 6](#_Toc158653177)

[2 Chapter 2: Literature Review 7](#_Toc158653178)

[2.1 Introduction 7](#_Toc158653179)

[2.2 Research Methodology 7](#_Toc158653180)

[2.3 AI 9](#_Toc158653181)

[2.4 Machine Learning 10](#_Toc158653182)

[2.4.1 Neural Networks (NNS) 10](#_Toc158653183)

[2.4.2 Supervised Learning 11](#_Toc158653184)

[2.5 NLP 12](#_Toc158653185)

[2.6 System Development for AI 12](#_Toc158653186)

[2.6.1 NLP application Development 12](#_Toc158653187)

[2.6.1 ML application Development 14](#_Toc158653188)

[2.7 Similar Tools 15](#_Toc158653189)

[2.7.1 System 1 15](#_Toc158653190)

[2.7.2 System 2 15](#_Toc158653191)

[2.7.3 System 3 15](#_Toc158653192)

[2.8 Summary 24](#_Toc158653193)

[3 Chapter 3: System Analysis 26](#_Toc158653194)

[3.1 Introduction 26](#_Toc158653195)

[3.2 Methodology 26](#_Toc158653196)

[3.2.1 SDLC 26](#_Toc158653197)

[3.2.2 Analysis 27](#_Toc158653198)

[3.3 Analysis of Existing Systems 29](#_Toc158653199)

[3.4 Requirements Elicitation 30](#_Toc158653200)

[3.4.1 System Requirements 31](#_Toc158653201)

[3.4.2 Functional Requirements 31](#_Toc158653202)

[3.4.3 Non-Functional Requirements 32](#_Toc158653203)

[3.5 Requirements Specification 33](#_Toc158653204)

[3.6 Requirements Modeling 37](#_Toc158653205)

[3.6.1 Use case diagrams 37](#_Toc158653206)

[3.6.2 Class Diagram 39](#_Toc158653207)

[3.7 Summary 40](#_Toc158653208)

[4 Chapter 4: System Design 41](#_Toc158653209)

[4.1 Introduction 41](#_Toc158653210)

[4.2 Design Methodology 41](#_Toc158653211)

[4.3 Architectural Design 42](#_Toc158653212)

[4.4 Component Design 43](#_Toc158653213)

[4.5 Data Modeling Design 44](#_Toc158653214)

[4.6 User Interface Design 44](#_Toc158653215)

[4.7 Summary 46](#_Toc158653216)

[5 Chapter 5: Conclusion and Future Work 47](#_Toc158653217)

[5.1 Conclusion 47](#_Toc158653218)

[5.2 Goals Achieved 47](#_Toc158653219)

[5.3 Lessons Learnt 48](#_Toc158653220)

[5.4 Limitations and Future Work 48](#_Toc158653221)

[5.4.1 Limitations 48](#_Toc158653222)

[5.4.1 Future work 49](#_Toc158653223)

[6 References 50](#_Toc158653224)

# List of Figures

[Figure 1: Project Methodology 4](#_Toc165834733)

[Figure 2: Phases of requirements classification pipeline. 14](#_Toc165834734)

[Figure 3: Number of requirements per label. 15](#_Toc165834735)

[Figure 4: Top 10 most important features. 16](#_Toc165834736)

[Figure 5: Results of binary classification (FR and NFR). 17](#_Toc165834737)

[Figure 6: Results of NFR classification. 18](#_Toc165834738)

[Figure 7: Results of FR and NFR classification (12 granularities). 18](#_Toc165834739)

[Figure 8 DC-Builder System architecture. 20](#_Toc165834740)

[Figure 9 Proposed Solution Design 22](#_Toc165834741)

[Figure 10 Process flow with NLP And ML Model 24](#_Toc165834742)

[Figure 11 Sample dataset 25](#_Toc165834743)

[*Figure 12 confusion matrix for the case of binary classification* 29](#_Toc165834744)

[Figure 13: waterfall for project. 32](#_Toc165834745)

[Figure 14: Description of stages of analysis 33](#_Toc165834746)

[Figure 15: user use-case diagram 41](#_Toc165834747)

[Figure 16: class diagram. 42](#_Toc165834748)

[Figure 17:Design methodolgy 44](#_Toc165834749)

[Figure 19: Architectural Design 46](#_Toc165834750)

[Figure 20: Home Page 53](#_Toc165834751)

[Figure 20: Input Page 54](#_Toc165834752)

[Figure 22: Diagram Page 54](#_Toc165834753)

# List of Tables

[Table 1: Project plan v1 4](#_Toc153058531)

[Table 2: Project plan v2 5](#_Toc153058532)

[Table 3: Related Work Comparison 33](#_Toc153058533)

[Table 4:use-case (Upload Dataset) 37](#_Toc153058534)

[Table 5: use-case (Load Dataset) 37](#_Toc153058535)

[Table 6: use-case (train Model) 37](#_Toc153058536)

[Table 7:use-case (Analysis) 38](#_Toc153058537)

[Table 8:use-case (test Model) 38](#_Toc153058538)

[Table 9: use-case (Enter requirements Document) 39](#_Toc153058539)

[Table 10: use-case (analysis Document) 39](#_Toc153058540)

[Table 11: use-case (Create analysis model) 39](#_Toc153058541)

[Table 12: use-case (save diagram) 40](#_Toc153058542)

[Table 13: use-case (Draw use-case model) 40](#_Toc153058543)

[Table 14: use-case (Draw class model) 40](#_Toc153058544)

# List of Abbreviations

UML Unified Modeling Language

ML Machine Learning

NLP Natural Language Processing

AI Artificial Intelligence

UI User Interface

RNN Recurrent Neural Network

SDLC Software Development Life Cycle

NNS Neural Network Systems

NER Named Entity Recognition

# Chapter 1: Introduction

## Introduction

Nowadays, the field of software development is using the capabilities of artificial intelligence to improve the phases of this field, including prediction and decision-making assistance. Artificial intelligence has become instrumental in guaranteeing the high quality of software development phases. Additionally, it has the capacity to significantly reduce the time needed for software development, leading to increased overall productivity. [1]

Software development involves creating and maintaining computer programs using various tools and methodologies to produce high-quality software efficiently this field involves using various programming languages, implementing different development methodologies like Agile and Waterfall

This project focuses on using Machine Learning for software development, specifically for converting software requirements and specifications into analysis models. The project was split into two phases. The first phase aimed at completing the analysis and part of the design for our proposed system. The work on that phase was tested and a number of notes were taken. These are the most important notes:

* Give more focus on related work and systems that employ Machine Learning for software analysis and modeling.
* Writing scenarios on applying Machine Learning and Natural Language Processing algorithms on chosen data samples at the analysis phase.
* Revise the analysis models (function and object model).
* Enhance the architecture design.

It was also recommended to use the Python programming language and libraries (e.g., scikit-learn library for predictive data analysis

This chapter highlights the goal and objectives of this project, outlines our methodology to achieve these objectives, gives the plan for the project’s tasks, provides an overview of the subsequent chapters in this report, and finalizes with an overall summary of the chapter.

## Problem Definition

Software maintenance constitutes a pivotal stage within the software development lifecycle, encompassing a substantial portion, varying from 40% to 80% of the total expenses associated with software development. It's worth highlighting that a notable 60% of the overall maintenance expenditure is dedicated to the improvement of existing software functionalities. Consequently, it becomes imperative to meticulously prepare appropriate software documentation at each developmental phase in order to alleviate the financial burdens of maintenance. The crux of reducing maintenance costs lies in enhancing one's grasp of the software system, as understanding a software system accounts for roughly 50% of the time spent in the maintenance phase. To facilitate this understanding, various modeling languages have surfaced, enabling graphical representations that substantially contribute to an enhanced comprehension of software systems. [2]

The project may face issues with excessive resource consumption, such as time or costs, and frequent changes in requirements can lead to project delays and increased expenses.

Using artificial intelligence (AI) to analyze requirements and generate drawings or models can be an effective solution for saving time and costs in the design and analysis processes.

## Project Aim and Objectives

The objective of the project is to deepen our understanding of the related works in terms of design and code to utilize the algorithms and technologies used in the related works, improve the system analysis process, complete the design phase with a focus on the data model and components, and program and test the system. All of this is aimed at accelerating and facilitating the work of the system modeling analysts and saving their time.

To achieve this goal, we must achieve the following objectives:

1. To review and delve deeper into related work in terms of design and code.
2. Improve the system analysis process.
3. Complete the design phase with a focus on the data model and components.
4. Program and test the system.
5. Conclude with the lessons learned and the knowledge and experience gained from working on this project.

## Project Methodology

Reviewing and deepening past works in terms of design and code is a key step in the research phase of our project. By examining current research and projects in our project field, we can gain valuable insights into the latest developments, identify potential challenges, and build on the knowledge and results of others.

Improving the analysis phase is important for the development of our project. During this phase, we thoroughly examine project requirements, analyze current systems and processes. It provides the basis for all subsequent project activities, ensuring that we have a clear understanding of what to achieve and how to achieve it.

After completing the Analysis phase, we move on to the preliminary Design tasks. This phase involves translating the gathered requirements and insights into a comprehensive system design. We will define the system's architecture, data flows, and functionality.

After completing the design phase, we move to the implementation phase and validate the system.

Concluding the project with a comprehensive review of lessons learned, knowledge gained, and experiences accumulated is essential for continuous improvement. This reflection allows us to identify areas of success and areas for improvement in project management, technical implementation, and collaboration. By documenting our insights and best practices, we can enhance our future projects and contribute to the growth of our team's expertise.



Figure 1: Project Methodology

## Project Timeline

This section presents the timeline plan versions of our project as the following:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **May** | | **April** | | | | **March** | | | | **February** | | | | **January** | **Months**  **(2024)** |
| **2** | **1** | **4** | **3** | **2** | **1** | **4** | **3** | **2** | **1** | **4** | **3** | **2** | **1** | **4** | **Weeks** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Plan | **Tasks** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Analysis and design of the system |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Complete implementation and validate phase |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Documentation |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Presentation |

Table 1: Project plan

1. **Plan (3 weeks):**

"We will define the main objectives of the project and review all the notes from previous discussions. Additionally, we will review the existing systems to understand what languages were used in developing the previous systems and how they function.".

1. **Analysis and design of the system (2 weeks):**

We will work on improving the system analysis process to ensure we get the most accurate results. We will enhance the system design and draw data models and components.

1. **Complete implementation and validate phase (6 weeks):**

We will program the system and test it to ensure it works as expected.

1. **Documentation:**

This task spans throughout the project's duration as it involves continuous documentation of the work.

1. **Project Presentation:**

The project should be presented during this time to showcase the results.

## Document Organization

* **Chapter 1.** In this chapter we discussed the plan and objectives and identified the problem that the project solves.
* **Chapter 2.** In this chapter, we reviewed existing systems.
* **Chapter 3.** Our focus in this chapter was on identifying functional and non-functional requirements using UML diagrams, as well as identifying user requirements and search methodologies.
* **Chapter 4.** In this chapter, we designed the system, identified the components of the system, modeled the data, and designed the user interfaces.
* **Chapter 5.** In this chapter, we programmed the system and conducted a series of tests to ensure that the system is working well.
* **Chapter 6.** A presentation of the project's conclusion was given, along with a proposal for new work to improve the current work, and a statement of whether the original objectives of the project had been met.

## Summary

In this chapter, we documented the feedback from the project's first graduation thesis and outlined the project's goals. We also defined a plan to address the feedback and execute the project modifications.

In the second chapter, we will delve into similar previous works, gathering and discussing all relevant information.

# Chapter 2: Literature Review

## Introduction

In this chapter, we meticulously explore the current scholarly landscape within our field of study. Our primary objective is to conduct a precise examination of the existing literature and research pertaining to our subject matter. We will study and analyze the methodologies and techniques employed in this context.

Furthermore, we will undertake a comprehensive review and comparative analysis of systems and research akin to our project. Through these reviews and comparisons, we aim to extract valuable insights that contribute to understanding the historical and contemporary developments in our field, and elucidate the patterns, trends, and innovative approaches that have shaped the discourse and research.

By conducting a meticulous evaluation of prior works and engaging in a profound comparison with our research approach, we aim to direct our efforts towards delivering novel contributions and deepening our understanding of our study's domain.

## Research Methodology

Our data collection process involves using the Google Scholar search engine to find diverse scientific literature sources, such as research papers, conference papers, articles, and e-books that can assist us in understanding similar studies and identifying their strengths, weaknesses, similarities, and differences. For citation and referencing, we use digital libraries from IEEE, ACM, the university library, trusted websites (e.g., .org, .edu), and the Saudi Digital Library. Additionally, we use ChatGPT to assist us in translating and comprehending texts.

We will determine the Software Development Life Cycle (SDLC) model, whether we will use the Waterfall or Agile model, in Chapter 3 after researching topics related to our project and previous systems.

1. **IEEE** [3]**:** The IEEE style is a numerical style, where citations are numbered according to their appearance order. This number guides the reader to a complete reference in the reference list at the end of the work. The citation number should be placed within square brackets on the same line as the text, before any punctuation, with a space before the square brackets, such as: [x]. Once cited from a source, the same number is reused for all subsequent citations from the same source. [4]
2. **ACM Library Online** [5]**:** The ACM Digital Library is a research, discovery, and networking platform that offers full-text access to all ACM publications, including journals, conference proceedings, technical magazines, newsletters, and books. It also maintains a carefully curated collection of full-text publications from a limited number of publishers. The ACM Guide to Computer Literature is a searchable database dedicated solely to computing literature, encompassing a complex network of relationships among authors, works, institutions, and specialized communities.

## AI

Artificial intelligence (AI) is the branch of computer science concerned with making computers behave like humans (Computers with the ability to mimic or duplicate the functions of the human brain). [6]

Artificial intelligence systems encompass individuals, processes, hardware, software, data, and the essential knowledge required to develop computer and machine systems that exhibit intelligent characteristics. [6]

**Techniques used in AI:**

1. **Knowledge Representation**: It is the science of translating real-world knowledge into a form that can be used by computers.
2. **Search:** A technique for selecting the best solution from all possible solutions.
3. **Automated Reasoning:** The process of achieving a specific goal based on prior knowledge.
4. **Planning:** The ability to make a good sequence of actions to achieve our objectives.

## Machine Learning

Machine science is a branch of artificial intelligence that aims to facilitate human life by replicating human behaviors. It is used to train the machine to train itself and to teach the machine to deal with data accurately. One of its uses is to predict future results based on known or previous data. Some of its types include supervised learning, unsupervised learning, and neural networks. [7] [8]

Some machine learning models that we might use include: "There are several methods used in machine learning, including **K-NN (K-Nearest Neighbor)**, which classifies new data based on the distance between it and the items in the database, then identifies the closest k items and calculates their average for regression problems or obtains the mode for classification. Then there’s the **Support Vector Machine (SVM)**, which creates a linear boundary between two classes to minimize the likelihood of misclassifying new instances. This boundary is established with the maximum margin separating the two classes. **Logistic Regression (LR)** uses an independent variable to predict a dependent variable and can be multiclass if the dependent variable has more than two classes, and binary if it has two classes. Finally, **Multinomial Naive Bayes (MNB)** estimates the conditional probabilities for a specific class based on the input data. It assumes that input features are independent and is specifically used for classifying documents and texts.

## NLP

Natural Language Processing (NLP) is a branch of artificial intelligence that involves processing and analyzing text data and includes machine learning to understand and interact with human language. [9]

**How does a computer understand languages?**

Computers are emotionless machines, so we must convert natural languages into numbers. This numerical transformation allows the computer to perform mathematical operations on language data, enabling it to comprehend human language. [9]

### Normalization

Text normalization is the process of preparing and organizing textual data for use in Natural Language Processing (NLP) and analytics systems. It involves various steps, including tokenization, where the text is divided into individual tokens like sentences or words. Tokenization is just one aspect of text normalization. Other techniques in this process include case conversion, fixing spelling errors, removing irrelevant words like articles and pronouns, as well as stemming and lemmatization. The goal is to clean and standardize the text, making it suitable for input into NLP and analytics applications. [10]

### Vectorization of Text

Do machine learning algorithms expect input as a binary matrix, where rows represent instances and columns represent features. We need to convert our examples and documents into vector representations to implement machine learning.

Some examples of Text Vectorization, which involves analyzing text, are:

1. **Bag of Words (BoW):**

The Bag of Words (BoW) model is a technique for feature extraction from text, and it is one of the simplest and most powerful methods. This model transforms textual documents into vectors, where each document becomes a vector representing the frequency of each unique word in the document's vector space. Therefore, the document "J" is expressed by the vector. [10]

xi,j It refers to the weight of feature **i**, calculated from the frequency of term **i** in document **j**, where **n** indicates the number of terms in the requirements dictionary.

1. **TF-IDF:**

It is a model that combines two measures. The first is the initial frequency value of a specific term in a document, and the second is the inverse document frequency for each term. The inverse document frequency can be mathematically represented by the following formula:

When combining the two measures, TF-IDF can be mathematically represented by the following formula: [10]

Where **tf** represents the term frequency, **idf** is the inverse document frequency, **i** is for the term, and **j** is for the document.

1. **CHI2:**

It is a common statistical test that measures the deviation from the expected distribution if one assumes that the occurrence of the feature is actually independent of the class value. [10] It assesses the independence between the term t and the class c and is determined by the formula:

where **N** is the total number of documents, **A** is the number of times **t** and **c** co-occur, **B** is the number of times the **t** occurs without **c**, **C** is the number of times c occurs without t and **D** is the number of times neither **c** nor **t** occurs.

## System Development for AI

In this section, we will discuss the system development for Artificial Intelligence applications and the stages followed in the development of Natural Language Processing (NLP) and Machine Learning (ML) applications, along with the languages used in development.

### NLP application Development

Developing Natural Language Processing (NLP) applications is crucial in the field of artificial intelligence. This work involves several sequential steps to ensure the creation of an effective application that leverages human language understanding. We will take a look at each step to comprehend it. [11]

1. **Data Acquisition:**

In this stage, we aim to acquire a dataset that reflects the linguistic diversity and complexity relevant to the application's subject. This step forms the foundation for understanding language and its varied usage.

1. **Text Cleaning:**

After obtaining the data, it's time to clean it from linguistic errors and distortions. This step is crucial to ensure accuracy and quality of results in later stages.

1. **Pre-processing:**

This part involves an initial analysis of the data and the application of basic natural language processing techniques. This step is fundamental to comprehend the language structure and prepare the data for training phases.

1. **Feature Engineering:**

This step focuses on utilizing features derived from the data in a way that contributes to the understanding of models and enhances their performance.

1. **Modeling:**

Modeling encompasses the training of models using processed data, refining them to achieve greater efficiency in understanding and processing language.

1. **Evaluation:**

After model training, their performance is evaluated using various tests to ensure their effectiveness and robustness in dealing with a wide range of scenarios.

1. **Deployment:**

Following model evaluation, the application is deployed for widespread use and application in real-world contexts.

1. **Monitoring and Model Updating:**

In this step, the application's performance is monitored post-deployment, and models can be updated based on changes in data or usage requirements.

### ML application Development

Developing machine learning applications is an exciting challenge in the modern tech world. These steps represent the journey developers go through to build smart applications based on the learning capabilities of automated systems. We’ll overview the main steps in developing these applications and how to achieve effective integration between the process components.

1. **Data Acquisition:**

Developers begin by examining and gathering data that the system will learn from. This data forms the foundation for understanding the application's context and training the model.

1. **Data Analysis and Cleaning:**

Developing machine learning applications requires a thorough examination of the data, ensuring its cleanliness and freedom from errors to ensure quality training.

1. **Feature Selection:**

Identifying key features in the data that will contribute to the system's understanding of information and achieve outstanding performance.

1. **Data Splitting:**

Dividing the data into sets for training and testing ensures an effective evaluation of the model and avoids overfitting problems.

1. **Model Selection:**

Modeling encompasses the training of models using processed data, refining them to achieve greater efficiency in understanding and processing language.

1. **Model Training:**

The model is trained using data, learning and benefiting from patterns and details within the data.

1. **Performance Evaluation:**

The model's performance is assessed using test data to ensure its effectiveness and efficiency in dealing with new cases.

1. **Improvement and Adjustment:**

Based on the evaluation, developers adjust enhance the model's performance and efficiency in facing different challenges.

1. **Application Deployment:**

After ensuring the model's quality, the application is deployed for public use and interaction with users.

1. **Performance Monitoring and Updating:**

Developers continue to monitor the application's performance after deployment, making necessary updates to keep up with changes and ensure continuous improvement.

## Similar Tools

### Software Requirements Classification Using Machine Learning Algorithms

Text classification is an attempt to organize documents based on specific properties and features. This supervised learning task involves determining categories for new documents based on training with a set of previously categorized documents. Software requirements are classified into two types: functional requirements and non-functional requirements.

**Methodology:**

Five steps are employed to perform software requirements classification, as illustrated in the following figure.

A diagram of a process

Description automatically generated

Figure 2: Phases of requirements classification pipeline.

[10]

The language used is Python, and the NLTK library was used, and the database is PROMISE\_exp, consisting of 969 requirements with their configurations represented as follows:



Figure 3: Number of requirements per label.

[10]

1. **Normalization**

In the first phase, we conducted the normalization process by converting all words to lowercase. Subsequently, we removed words with little or no significance, such as in the transformation from "The system shall refresh the display every 60 s” to “System shall refresh display every second”. Additionally, words were transformed into their root form, for example, "users" changed to "user".

1. **Feature extraction**

Both BoW and TF-IDF were used for feature extraction, and a comparison was made during the classification stage to observe which one led to an improvement in algorithm performance. The table below displays the top ten words that received scores for each technique. It can be noted that the words are the same, but the difference in importance starts from the third position.



Figure 4: Top 10 most important features.

[10]

1. **Feature selection**

After the feature extraction stage, we move on to feature selection, involving a filtering process to remove less important features using statistical methods such as CHI squared. Two parameters were utilized: the first one, max\_df, ignores terms with a frequency higher than the specified threshold, while the second one, max\_df, disregards terms with a frequency lower than the specified threshold.

1. **Classification**

Four algorithms were used for training and testing performance: LR, MNB, SVM, and K-NN. In LR, MNB, and SVM algorithms, a parameter called class\_weight was employed, which utilizes class label values to automatically adjust weights inversely proportional to the class frequencies in the input data. All hyperparameters for the classification algorithms were selected using a function called GridSearchCV, which tests all possible parameter combinations and returns the set that achieved the best results.

The scikit-learn tool was utilized, as it integrates various machine learning algorithms that were selected because it contains the algorithms used.

1. **Performance measure**

The data was divided into 10 folds, where 9 folds were used for training the algorithm, and 1 fold was used for testing. [10] The following performance measures were employed:

True positives (TP), false positives (FP).

True Negatives (TN), False Negatives (FN)

The F1 or F-measure combines precision and recall into a single metric. The F1 score is the harmonic mean, giving much more weight to low values compared to the regular average, which treats all values equally. The result is not high unless both recall and precision are elevated. [10]

**Results:**

The metrics used to evaluate the mentioned machine learning algorithms are Precision Equation, Recall Equation, and F-measure Equation, and the results of the evaluation are presented in the following figure:



Figure 5: Results of binary classification (FR and NFR).

[10]



Figure 6: Results of NFR classification.

[10]



Figure 7: Results of FR and NFR classification (12 granularities).

[10]

The results showed that the combination of TF-IDF and LR gives the best performance in binary classification, non-functional requirements classification, and general requirements classification. The F-value was 91% in binary classification, 74% in 11-category classification, and 78% in 12-category classification. [10]

### From user requirements to UML class diagram

This scientific paper focuses on extracting separation schemes from text requirements using NLP technologies and domain ontology. To analyse a given text, the most Natural Language Processing (NLP) systems are based on the following levels: Morphological level, lexical level, syntactic level, semantic level, discourse level and pragmatic level. Ontology is a branch of computer science and information science that is interested in studying how knowledge and information and their relationships are represented in a given field. [12]

This study accepts, as an input, textual data expressed in natural language and representing the user needs then identify the classes’ names, their attributes, and associations between them in order to classify them in a structured XML file. [12]

**The framework used:**

GATE is an open source framework developed using Java programming language used in this scientific paper. It is used to develop software components that process natural language. It can provide a set of natural language analysis tools that can take text inputs in English and as a result give basic forms of words and their parts of speech, etc., determine the structure of sentences in terms of phrases and dependencies of words, and state which name phrases refer to the same entities. [12]

GATE has an information extraction system (IE) called ANNIE (The New Quasi Information Extraction System) which contains many language processing as stated in this paper [12]:

* **Sentence splitter:**

the sentence splitter separates each sentence from the input string and returns a list of strings. [12]

* **Tokenizer:**

the tokenizer takes each sentence as an input and splits them into tokens such as words and punctuation. [12]

* **Syntactic parser:**

transforms sequences of words into structures indicating how sentence units relate to each other. This step helps us identify the main parts of a particular sentence such as the object, subject, verb...etc. [12]

**System architecture:**

**A diagram of a language analysis block

Description automatically generated**

Figure 8 DC-Builder System architecture.

[12]

Figure 8 shows DC-Builder System architecture. The natural language analysis block processes the requirements descriptions submitted by the user using the framework GATE, and specially: Sentence splitter. [12]

The classes of candidates can then be extracted by considering the name phrases in the text of the requirements. Candidate relationships can be found in the same way by considering verb phrases. For example, by analysing the phrase "the doctor gives medicines to the patient" we can know three candidate classes (doctor, medication, patient) and one candidate relationship (give). In this context, heuristics can play a fundamental role to facilitate such task [12].

Given a parts-of-speech and their functions in sentences, Chen [13] proposed eleven rules in order to translate NL requirements description written in English from natural language (English) to an entity-relationship diagram. The proposal of Chen seems to be the first attempt using linguistic. [12]

**The rules described below will be used to facilitate the extraction of classes’ names for the diagram:**

* All nouns are converted to entity types. [12]

It can be concluded that all nouns can be mapped onto the names of the classes’; By nouns, we mean all types of nouns such as common nouns and collective nouns. [12]

* A gerund may indicate an entity type which is converted from a relationship type.

Firstly, A gerund can be defined as a noun which consists of a verb and an “ing”. It is often called an -ing word or a verbal noun. [12]

* a specialization’s relationship between entities: sentence’s structure “is a” can relate two nouns A and B to one another. [12]
* every proper noun (Person name, Location name …) is ignored to be a class. [12]

This rule can help to perform a partial filtering in order to obtain an accurate set of classes’ names for the class diagram. [12]

For attributes extraction, there some heuristics:

* A noun such as “vehicle\_number”, “group\_no”, “person\_id” and “room\_type” may refer to an attribute type. [12]
* A noun phrase succeeding the “has/have” verb phrase may indicate the presence of attribute types. [12]

For associations’ extraction, there Three heuristics:

* A transitive verb can be a candidate for relationship type. [12]

Transitive verb, in syntax, is a one that requires an object to complete its meaning. [12]

* if a verb is equal to one of the following list {“include”, “involve”, “consists of”, contain, “comprise”, “divided to”, “embrace”}, therefore, this relationship can be aggregation or composition. [12]

After that as an input, the concerned module produces an initial XML file that should be refined. [12]

**Refinement using domain ontology:**

The previous module produces an initial model, in XML form, including concepts related to classes, attributes, and associations. This model can contain erroneous elements which should be treated. In fact, the constructed ontology will help eliminate irrelevant elements. [12]

### AI Based UML Diagrams Generator

This study will focus on providing solution to generate Use-Case diagram and Class diagram against the particular business requirement. system will read and understand the business requirement using Natural Language Processing and Machine learning and identifying entities and relationships on that for generate use case diagram. And also, classes and relationship between classes will also be identified to generate class diagram. Then system will show generated diagram with including user interacting feature where user can add additional element and add changes to use case or class diagram or edit existing element and its attribute or relations. [13]



Figure 9 Proposed Solution Design

As showing in the Figure 6 proposed solution design, user can upload a text file containing a scenario or simply copy and paste the text inside. The system extracts word fragments and named entities using the NLP module. Subsequently, the ML module will identify and extract the usecases, actors, classes, associations according to the ML algorithms which have implemented. Plant UML modeling draws the use case and class diagram. The system allows editing of the generated diagrams and the user can modify and adjust the use cases, actors, classes, attributes and all types of relationships. Therefore, the output is a highly customizable output that the user will be able to achieve the desired output according to the business requirements provided. [13]

**Methodology**

The text assure the importance of efficiently extracting information from text data, noting that manual text searching can be difficult and time-consuming, which may lead to important information being skipping. Therefore, providing an automated way to extract information from texts and present it in an organized manner can provide many benefits and save time spent browsing text documents. Information mining technology aims to achieve this goal by systematically analyzing texts and identifying relevant information. The example provided assure the importance of understanding the parts of speech in a sentence to accurately interpret the meaning. [13]

**Algorithmic design**



Figure 10 Process flow with NLP And ML Model

Figure 6 shows the process flow with Natural language processing and Machine learning model where several preprocessing stages has involved before proceeding to the ML based information extraction. [13]

The important information in requirement texts usually follows certain patterns. When we look at the Parts of Speech tagging of a sentence, they follow specific patterns too. If we understand these patterns, we can easily get the information we need. This helps us create use case and class diagrams for the project. [13]

In algorithmic design, we can choose any Parts of Speech tagger modules for English literature. We mark words in a story based on their meanings and how they're used in specific parts of speech. [13]

Some POS tagging examples can be mention as: Coordinating Conjunction (CC), Cardinal Digit (CD), existential (EX), adjective (JJ), modal (MD) such as could, will etc., proper noun (NNP), predeterminer (PDT), possessive pronoun (PRP), etc. (Rachiele, G.2018) To assign grammatical information to each word in a sentence we used POS taggers. [13]

Most of the information in a requirement text is made up with verb, gerund/present (VBG), noun plural (NNS), TO, verb base (VB), noun (NN), preposition or subordinating conjunction with these pos taggers. But they come with different patterns. [13]



Dataset would be like as Figure 7 shown below according to the above requirement text.



Figure 11 Sample dataset

For now, according to the given sample, there are 3 kinds of specific patterns that can identify:

• {VBG>NNS>TO>VB>NN}

• {NN>IN>DT>NN}

• {VB>DT>NNS>TO>VB}

The process involves finding the key terms for class or use case diagrams using machine learning, which makes this task a classification one. Data is divided into two sets: training and testing. The trained model uses features from the training instances to predict terms for unseen data. Comparing predicted and actual terms helps evaluate the model's performance. Statistical techniques rely on adjustments between classification results using probability and costs for decision-making. [13]

Preprocess Functional requirement texts

The text discusses the process of text preprocessing, where ambiguity resulting from the use of multiple forms of a certain verb or the singular/plural form of a word is reduced. Additionally, common words like "a," "the," "of," and "is" do not contribute much information to our summarization goal. Below are several operations commonly used for document preprocessing. [13]

**Document segmentation:**

A text is divided into several paragraphs to find where each sentence is placed in its respective paragraph. [13]

**Stemming**:

We apply stemming to bring a word to its root or base form. The examples include use of a singular form rather than using a plural, or removal of ‘ing’ from a verb. To this end, Stanford NLP stemmer is employed in this paper. Paragraph segmentation. Paragraph segmentation divides a paragraph into sentences using sentence tags. [13]

**Word normalization:**

Each sentence consists of multiple normalized words. Through normalization and lemmatization individual words become one common form, stemming down to their roots. Ambiguities are removed by Porter’s algorithm. [13]

**Stop word filtering:**

Stop words can be filtered out after performing other preprocessing steps. There is no uniform rule for selecting a stop word because it depends on individual tasks. In this work, words such as a, is, in, the, of are selected as stop words and are filtered out from the document. In text mining applications stop word filtering is considered as a standard step. [13]

Machine Learning Techniques to Identify Diagram Elements

1. Classification Model
2. Recurrent neural network Unit
3. RNN based Sequence to Sequence Model
4. Strategies of training

**IMPLEMENTATION:**

The initial dataset is divided into subsets containing functional requirement sentences for each element in the diagram, specifically actors, use cases, and classes. Each subset is further divided into validation, training, and test pairs. Separate machine learning models are built for each diagram element using the training subsets. During implementation, data preprocessing is conducted using Python programming, which involves cleaning and preparing the data for analysis. [13]

check the words in the English dictionary and removing unused punctuation, removing stop words etc. are the preprocessing stages done in the implementation.

Thereafter extracting feature task was perform where it checks existence of whole keyword and also existence of all parts of known keywords were performed.

The recurrent words such as propositions don’t participate as key terms. Calculating the term frequency and inverse document term frequency (TF X IDF), these expressions can dominate other meaningful expressions. Consequently, to deal with some challenging terms we can eradicate from script by allocating probability assessment of zero and also, numerals and nonalphanumeric characters will be removed.

Thereafter naïve based classifier has been implemented and as previous stages clean and tokenization tasks also perform before generate predicted tags based on decision rule

In the decision rule method, it takes the functional text, and the keywords from the early stage perform and also stopwords. First it gets scores of each whole keywords from the functional text provided. Then get scores for each known keywords with all parts in the requirement sentence. After that, three steps performed as.

1. adding tag if posterior probability is higher than or equal to 0.5
2. add highest scoring to TF-IDF keywords if not already added.
3. Add ‘c#’ if there no tags (‘c#’ just a special keyword to identify it clearly from other keywords)

**Identifying Use case and Class Diagram Relationships with NLP and ML techniques**

In this implementation our focus mainly on identifying relationship in both use case and class diagram related components separately. This is a combination of two models where one model will be identifying relationship involved entities. As an example of use case diagram first model will identify what are the actors and use cases involve in the relationship. Second model will be identifying the relationship type of that particular relationship. These two models need to build for both use case and class diagram separately. Two models per each diagram type to identify relationship involved in particular functional requirement text. [13]

Similar to the previous implementation, in this iteration, we again divide the initial dataset into subsets containing functional requirement sentences for each diagram element. This time, we focus on use case relationships and class relationships. Each dataset is then separated into two subsets for the two models: one with entities involved in the relationship and the other with the relationship type along with the functional text. These subsets are divided into validation, training, and test pairs, with the same distribution as before. Two separate machine learning models are built for each diagram type using these training datasets. [13]

**Draw diagram with identified diagram elements**

With this implementation we want to generate the usecase and class diagram according to identified elements through previous two implementation. Usecase and class diagram elements that has been identified from functional requirement text has outputted as JSON file format. [13]

**Evaluation approach**

To evaluate the Model, we can use the term called Model Evaluation Metrics.

Model evaluation metrics are needed to measure model performance. The choice of evaluation parameters depends on the machine learning task provided. In this project that is classification. [13]

**Classification metrics**

When performing predictions through classification, four kinds of result that could occur as follow. [13]

* True positives: are when we assume an observation belongs to a particular class and actually it is belonging to that exact class.
* True negatives: are when we assume an observation is not belonging to a particular class and that actually not belonging to that particular class as expected.
* False positives: occur when we predict our observation belongs to a one particular class but in reality, it is not belonging to that class.
* False negatives: occur when we assume an observation is not belonging to a particular class but actually it is belonging to that same class



*Figure 12 confusion matrix for the case of binary classification*

Accuracy, precision, and recall are the three main metrics that can be used to evaluate a classification model in generally. [13]

1. Accuracy is the percentage of correctly predicted instances in the test data, calculated by dividing the number of correct predictions by the total number of predictions.
2. Precision is the fraction of true positives (relevant examples) among all examples predicted to belong to a particular class, indicating how many relevant items were predicted.
3. Recall is the fraction of test data predicted to belong to a particular class compared to all test data truly belonging to the class, indicating how many relevant items were predicted.

## Summary

In this chapter thoroughly examines the research context, starting with a detailed review of existing literature. The emphasis is on a meticulous analysis of Very precise. The research methodology, which includes Google Scholar, digital libraries for data collection, is outlined succinctly.

The chapter then delves into the domains of Artificial Intelligence (AI), covering its definition, conduct, and techniques such as knowledge representation, search, automated reasoning, and planning. Machine Learning (ML) is introduced as a branch of AI, focusing on supervised learning, neural networks, and their applications like NLP.

The discussion on NLP includes text tokenization, word embedding, and techniques like TF, TF-IDF, and Word2Vec. Named Entity Recognition (NER), parsing, pronoun resolution, and semantic analysis are explored in the context of NLP.

The chapter transitions to system development for AI, outlining the stages for developing NLP and ML applications. Python and various libraries are highlighted for development. Similar tools in the field, such as Lucidchart, Visual Paradigm, StarUML, Enterprise Architect, and MagicDraw, are introduced, emphasizing their AI-powered features.

After investigating these topics, we will be prepared in Chapter 3 to analyze the system and find both functional and non-functional requirements.

# Chapter 3: System Analysis

## Introduction

In this chapter, we will focus on defining the system requirements using simple diagrams. We will use UML tools to explain how the system operates and its requirements. We will specify both functional and non-functional requirements and discuss the development methodology. We will also analyze an existing system.

## Methodology

### SDLC

In our project, we used the waterfall methodology for project execution. In the first step, the project plan. In the second step Literature Review. in the third step, analyzed the system. in the fourth step, system design.

Stability of Requirements: The Waterfall model requires defining project requirements early. It fits well when requirements are stable and undergo minimal changes.

Path Determination: The Waterfall model follows a linear and organized approach, allowing the team to clearly outline the workflow and task sequence. This structure is beneficial for projects requiring precision and strict timelines.



Figure 13: waterfall for project.

### Analysis

In the analysis phase we have identified 4 basic activities to fully analyze the components of the system. Before starting with system design in general we must collect the requirements and analyze them correctly. During the analysis phase of the system, we used this existing strategy in Figure 16.



Figure 14: Description of stages of analysis

In the process of systems analysis existing, we will compare work related to our system and identify points of difference between them and mention the advantages of each work like our system.

In the process of requirements elicitation, we will describe system requirements and functional and non-functional requirements in the natural language, and we will Elicit them through brainstorming, use similar systems requirements, use the Lucid platform tool to generate ideas for the requirements.

In the process of requirements specification, we will construct tables to describe both functional and non-functional requirements.

In the process of requirements modeling, we will draw use cases that show all possible interactions with the system and write the description for them.

## Analysis of Existing Systems

At this stage, we compared our system with similar systems through features. Then we used some of the features of similar systems at the stage of the elicitation requirements.

The checkmark  represents that the advantage will be achieved in the system whereas the  shows that it will not be achieved in the system. “**?**” it shows that it is possible to try to make a feature in the system. The differences between our system and other comparable work are shown in Table 3.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| System  Feature | Lucid platform | Visual paradigm | StarUML | Enterprise Architect | MagicDraw | ChatUML | Our project |
| Create UML diagrams |  |  |  |  |  |  |  |
| Code Generation |  |  |  |  |  |  | ? |
| Smart Requirements Management |  |  |  |  |  |  |  |
| Automated generation of drawings. |  |  |  |  |  |  |  |
| Intelligent pattern recognition |  |  |  |  |  |  |  |
| Real-time Diagram validation |  |  |  |  |  |  |  |
| Suggesting Relevant Elements |  |  |  |  |  |  |  |
| Detecting Errors |  |  |  |  |  |  |  |
| Predictive Modeling |  |  |  |  |  |  |  |
| Conflict Analysis |  |  |  |  |  |  |  |
| Idea Generation |  |  |  |  |  |  |  |
| Idea Sorting |  |  |  |  |  |  |  |
| Idea Summarization |  |  |  |  |  |  |  |

Table 3: Related Work Comparison

We have identified some features that will not be achieved in the system for many reasons for each feature:

* Intelligent pattern recognition: because we have identified that the system analyzes the text or document through the context of the text through Named Entity Recognition. It classifies specific words such as names, locations and objects. It is divided into two sections: structural and semantic information.
* Real-time Diagram validation: because our system allows the user to enter text by typing and by upload file and at upload file, we do not need to process changes in real-time.
* Detecting Errors: because our system allows a certain percentage of requirements errors to be ignored by the user.
* Predictive Modeling: because our system does not predict future data and trends.
* Conflict Analysis: because our system does not identify inconsistencies between requirements and models but only generates them and displays them to the user.
* Idea Generation and Idea Sorting and Idea Summarization: because the system does not have the functions of generating ideas and sorting the ideas generated and not summarizing them.

## Requirements Elicitation

Requirements engineering is the process of discovering all system services, restrictions imposed on it, and services that meet user needs, analyzing, documenting, and verifying these services [14]. It includes high-level activities, including elicitation, analysis, and other activities. Requirements elicitation is the collection of all requirements, including system requirements and user requirements [14]. System requirements describe all the requirements that we want the system to perform [14]. After the elicitation process, we extracted the system requirements, and from them we identified the functional and non-functional requirements through brainstorming, using similar requirements tools, and using the Lucid platform tool.

### System Requirements

A structured document setting out detailed descriptions of the system’s services and operational constraints [15].

Defines what should be implemented so may be part of a contract between client and contractor [15].

* 1. To build an AI website system to receive software development requirements texts and documents and use NLP techniques to analyse them.
  2. To extract classes, functions (methods / procedure) and relations between classes and function.
  3. To generate UML diagrams: use-cases and class diagram.

### Functional Requirements

The requirements determine the system's behavior, what it should and shouldn't provide, and how it interacts with inputs [14].

1. To build an AI website system to receive software development requirements texts and documents and use NLP techniques to analyses them.
   1. The system shall allow the user to enter requirements either by typing directly or by uploading a document.
   2. The system shall allow the user to create analysis model.
   3. The system shall allow the user to choose type of UML diagram such as (class, use case).
   4. The system shall prepare to clean and organize the initial data to perform the text analysis process.
   5. The system should analyze the text by doing Tokenization, Parsing and Relation Extraction.
2. To extract classes, functions (methods / procedure) and relations between classes and function.
   1. The system shall analyses the requirements context using NLP techniques.
3. To generate UML diagrams: use-cases and class diagram.
   1. The system shall draw the diagram after generated by NLP techniques.

### Non-Functional Requirements

It describes the characteristics of the system and the constraints imposed on the services provided by the system such as time constraints and other constraints. It also describes specific standards for some of the system's services [14].

1. To build an AI website system to receive software development requirements texts and documents and use NLP techniques to analyses them.
2. The system must be fast, ensuring that the processing of requirements and drawing (use case or class diagram) does not exceed (write the time).
3. The resulting diagrams in the system should have an accuracy of no less than (write the percentage).
4. It should be able to handle (write the range of users) simultaneously.
5. The system should be user-friendly and adaptable to various devices to provide a seamless experience.
6. The system should perform its functions with minimal steps.
7. To extract classes, functions (methods / procedure) and relations between classes and function.
8. The system should be able to recognize and analyze a minimum of (write the percentage) of user requirements entered, with a permissible margin for error not exceeding (write the percentage).

## System Scenario

A black and white rectangular object with text

Description automatically generated

## Requirements Specification

## Requirements Specification

A Use case represents a class of functionality provided by the system

Use case can be described textually, with a focus on the event flow between actor and system [17]

The textual use case description consist of 6 parts:

* Unique name
* participating actors
* Entry condition
* Exit condition
* Flow of event

### Usecase 1:

1. **Name**:

analysis documents:

1. **Actor:**

User

1. **Entry condition:**

Data is available for analysis

1. **Exit condition:**

the analysis is completed

1. **Flow of event:**
2. user entered the requirements.
3. analysis documents:
4. Elementsfinder
5. Relationshipfinder :
6. After analysis is completed generate Diagram Specs get the analysis like number the class and the relation between them and then ready to draw.

### Usecase 2:

1. **Name**:

Enter requirements Document

1. **Actor:**

User

1. **Entry condition:**

The user enters the website.

1. **Exit condition:**

The requirements are entered successfully

1. **Flow of event:**
2. The user clicks on the text box.
3. The user enters requirements by typing.

Or The user clicks on the select doc.

### Usecase 3:

1. **Name:**

create analysis model

1. **Actor:**

User

1. **Entry condition:**

The system will determine whether it is the context or Document to Draw the Diagram.

1. **Exit condition:**

The selected model type is displayed on the screen.

1. **Flow of event:**
2. The user clicks on select type.
3. The system shows to type of model (class model, use-case model)

If the user chooses class model

1. The system shows class model on screen after the user start generate.

If the user chooses class model

1. The system shows use-case model on screen after the user start generate

### Usecase 5:

1. **Name:**

use-case diagram

1. **Actor:**

User

1. **Entry condition:**

The system is ready and prepared to start drawing the diagram

1. **Exit condition:**

The diagram has been successfully drawn by the system.

1. **Flow of event:**
2. After analysis completed the system draws a use-case model after the requirements.
3. The system displays use-case diagram screen.

### Usecase 6:

1. **Name:**

Class digram

1. **Actor:**

User

1. **Entry condition:**

The system is ready and prepared to start drawing the diagram

1. **Exit condition:**

The diagram has been successfully drawn by the system.

1. **Flow of event:**
2. The system draws a class model after the analysis requirements are completed.
3. The system displays class models on screen.

## Requirements Modeling

### usecase

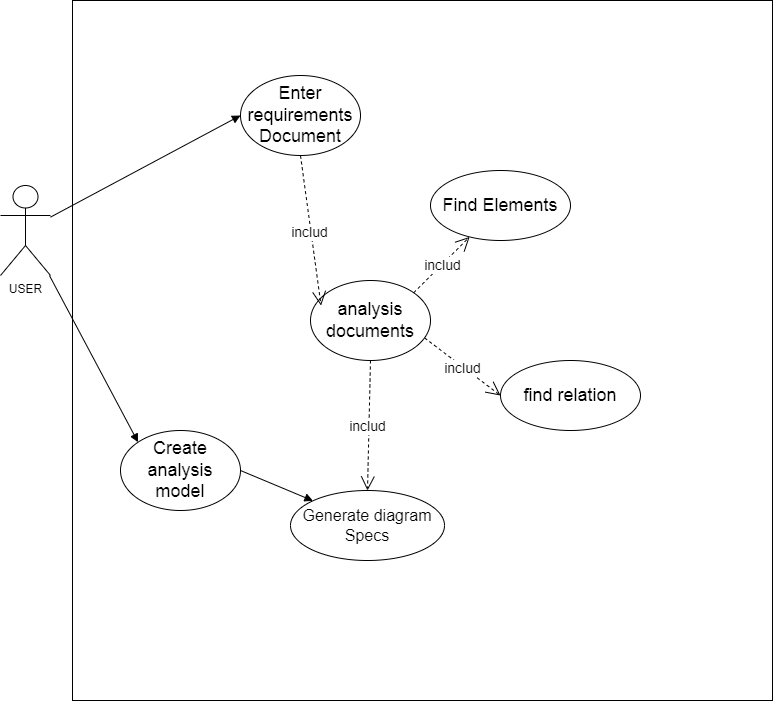


Figure 15: user use-case diagram

The user can enter requirements and then Upload data and then load data to analyze by normalization and Tokenization, feature (extraction, Selection) and then After analysis is completed generate Diagram Specs get all the Analysis specifications and draw analysis model uses these specifications to draw the diagram whither it is Usecases diagram, class diagram depends on the user what need.

### Class Diagram



Figure 16: class diagram.

The class DocumentAnalysis consists of two classes:

1. elementsFinder class:

This class finds diagram elements such as use cases, actors, and classes.

1. RelationshipFinder class:

This class identifies the relationships between actors, use cases, and classes among each other.

The GenerateDiagram class retrieves data from the DocumentAnalysis class after it has identified the elements and relationships, and then sends it to PlantUML to draw the use case and class diagrams. User class allows the user to entre requirement. DBMS class it can upload and load the data.

## Summary

In this chapter, we analyzed the system, identified the methodology of analysis and development, identified the functional and non-functional requirements, and displayed them in the format of tables and UML diagrams such as use case and class diagram.

After analyzing the system, we will be ready in chapter 4 to design the system and determine the architecture of the system and design the user interface.

# Chapter 4: System Design

## Introduction

This chapter covers system design, including structural design, component design, data modelling, and user interface design. Structural design focuses on organizing the system's structure to ensure efficient performance, while component design addresses the interaction of software components to achieve functional goals. Data modelling design deals with efficiently organizing and storing data, and user interface design highlights improving the user experience through an effective and attractive interface.

## Design Methodology



Figure 17:Design methodolgy

**Architectural Design:**

The process of determining the overall structure and organization of a software system, involving high-level decisions on system configuration, key component specification, and their interactions.

**Component Design:**

The process of dividing a system into manageable units, with each unit representing a specific module or function within the software.

**Data Modeling Design:**

The process of defining and organizing data requirements for a system, creating a conceptual representation of data, defining relationships between entities, and specifying data storage and access.

**User Interface Design:**

The process of creating an easily understandable and visually appealing interface for users to interact with software, including the design of structure, navigation, and visual elements. Goals include enhancing user experience through user-friendly and efficient interface design.

## Architectural Design

The architectural design is concerned with understanding how a system is organized and designing the overall structure of the system, it involves identifying major system components and their communications.



Figure 19: Architectural Design

We opted for the client-server architecture because it allows for the distribution of tasks between the client and the server. In our system, on the client side, the user enters a file containing textual requirements or requirements entered via the keyboard. Then, user requests are received through the user interface and sent to the server.

On the server side, the server sends the file or user request to the NLP class, which processes the requirements and analyzes the texts using the SpaCy library. A UML file is created based on the results extracted from the text analysis, and then the file is sent to the PlantUML tool to convert it into an image. Finally, the server sends the file generated by UML to the client to display the image to the user via the user interface.

## Component Design

Our system is divided into two main parts: document analysis and diagram drawing. Each part contains several categories. The document analysis part includes finding elements of use cases and class diagrams and identifying relationships. The diagram drawing part includes drawing an analysis model, a use case diagram, and a class diagram.

### First main component: document analysis

**class ElementsFinder:**

**attributes:**

1. Punct.
2. Keywords.
3. Stopwords.

**method:**

1. **get\_unused\_puncs:**

A list named puncs is created that contains all possible punctuation marks. Unused punctuation marks in each keyword are identified and removed from the puncs list. re.compile is used to create a regular expression that includes unused punctuation marks.

Libraries: String

1. **clean\_tags\_and\_titles:**

The titles and tags in the dataset are cleaned: The tags are converted to lowercase and split into a list of words. All tags are merged into one array and duplicates are removed. The function get\_unused\_puncs is called to get unused punctuation marks. A list of English stop words is loaded. The titles are converted to lowercase and unused punctuation marks are replaced with spaces. The titles are split into words and stop words are removed from them.

Local variables: train[‘Tages’], known\_kws.

Libraries: String, nltk, re, numpy.

1. **existence\_of\_whole\_keywords:**

An initial value is set for the variable index to 0. The data type in tag\_kws and title\_kws is checked, and if it’s float, it is ignored and the next iteration is proceeded. The difference between the two sets title\_kws and tag\_kws is printed. The keywords that exist in the title only are identified and stored in the dictionary keywords under the key kw with an initial value of 0. The value of the key ‘title’ within the dictionary keywords for the keyword kw is increased. “second” is printed. The keywords that exist in both the title and tag are identified and stored in the dictionary keywords under the key kw with an initial value of 0. The value of the key ‘both’ within the dictionary keywords for the keyword kw is increased. “Third” is printed. The keywords that exist in the tag only are identified and stored in the dictionary keywords under the key kw with an initial value of 0. The value of the key ‘tag’ within the dictionary keywords for the keyword kw is increased.

Local variables: index.

1. **existence\_of\_all\_part\_of\_known\_keywords:**

Each keyword in the tag\_kws list is analysed. If the keyword contains a ‘-’ badge, it is removed. The code checks if all parts of kw\_new are present in title\_kws. If that’s true, it adds 1 to the keywords dictionary under the ‘both’ key for this keyword.

1. **Calculate\_scores:**

The variable n is defined, which is calculated using the length of the DataFrame train index. The variable removeKeys is an empty list that will be used to add the words that will be removed. A for loop is used to browse all the keywords in the keywords dictionary. Setdefault is used to set default values for ‘title’, ‘both’, and ‘tag’ to 0 if they are not present. Total\_tag and total\_title are used to calculate the total number of tags and titles. If total\_tag or total\_title is zero, kw is added to the removeKeys list. In case both are non-zero, the conditional probability p is calculated using a special formula.

Local variables: n, removekeys, total\_tag, total\_title, p(posterior probability).

Libraries: numpy.

1. **nb\_classify:**

The variable unused\_puncs is determined using the get\_unused\_puncs function, Stop words are extracted using the nltk library. The title of each text in the DataFrame is converted to lowercase, unused symbols are removed, then it is split into a list of words and stop words are removed. After that, the decision\_rule function is used. The decision\_rule function, which is not defined in this part of the code, is used to perform the classification.

Local variables: unused\_puncs, test[‘Title’], pred, pred[‘tags’]

Libraries: pandas, nltk.

1. **decision\_rule:**

**the three rules:**

* + - 1. adding tag if posterior probability is higher than or equal to 0.5.
      2. add highest scoring to TF-IDF keywords if not already added.
      3. Add ‘c#’ if there no tags (‘c#’ just a special keyword to identify it clearly from other keywords)

It accepts four parameters: title\_hws, keywords, stopwords. title\_hws: This is the title that has been converted to lowercase and split into a list of words after removing unused symbols and stop words. keywords: These are the keywords that are used to identify the tags. stopwords: These are the empty words that are used to identify the tags. Two lists are used to save points for each keyword, one for testing using all the words in the title, and the other using part of it. Iteration is used to add points to the two scoring lists based on their match in the keys present in the “keywords” list. Filtering and anonymous works (lambda) are used to add points to the two scoring lists based on their match in the keys present in the “keywords” list. In the end, the tag that gets the highest number of points is returned. If there is a tie, the tag that gets the highest number of points using all the words in the title is returned. If there is a tie again, the tag that appears first in the “keywords” list is returned. If no tag is found, the value None is returned.

Local variables: nb\_scores, ti\_scorea, kws\_to\_tag.

**class RelationshipFinder:**

**attributes:**

1. temp.

**method:**

1. **transform\_input\_text:**

It accepts one parameter: texts. texts: These are the texts that are to be transformed. It is expected to be a list of texts. Each text in texts is converted into a sequence of numbers using a dictionary that links each word to a number. It appears in the code as self.input\_word2idx. Each text is split into words, and each word is converted to the number associated with it in the dictionary. If the word is not present in the dictionary, the number 1 is used. Texts that contain more than self.max\_input\_seq\_length words are truncated. Then, the pad\_sequences function is used to make all sequences the same length (self.max\_input\_seq\_length). Zeros are added at the beginning if the sequence is shorter than self.max\_input\_seq\_length. In the end, the shape of the transformed sequences is printed and returned.

Local variables: x, wid.

Libraries: keras.preprocessing.sequence.

1. **spilt\_target\_text:**

The function takes itself (self) and a variable called texts as parameters. The function creates an empty list called temp that is used to save the processed texts. The function analyzes each line from the texts variable. The function adds ‘START’ and ‘END’ tags to the beginning and end of each line. The function converts all characters to lowercase. The function splits the line into words and adds them to a list x. The size of x is checked, if it is larger than self.max\_target\_seq\_length, ‘END’ is added to x and it exits the iteration. In the end, x is added to temp. The function finally returns temp.

Local variables: x, line2.

Libraries: keras.preprocessing.sequence.

1. **generate\_batch:**

The function takes itself (self) and three other parameters: x\_samples, y\_samples, and batch\_size. Three empty lists are initialized: encoder\_input\_data\_batch, decoder\_input\_data\_batch, and decoder\_target\_data\_batch. A while True loop is used to create batches of data until all samples are processed. In each iteration, a for loop is used to extract each sample and the target words from x\_samples and y\_samples respectively. Then, another loop is used to add the encoding code to decoder\_input\_line using the dictionary self.target\_word2idx. After that, w2idx\_next is calculated using the same dictionary and the encoded code, and it is linked to decoder\_target\_label. Decoder\_input\_line and decoder\_target\_label are added to their respective updated batches. At the end of each rotation, if line\_idx equals the batch size, yield will return the optimized batches of encoder\_input\_data\_batch, decoder\_input\_data\_batch, and decoder\_target\_data\_batch.

Local variables: encoder\_input\_data\_batch, decoder\_input\_data\_batch, decoder\_target\_data\_batch, target\_words, x, decoder\_input\_line, w2idx, w, w2idx\_next, decoder\_target\_label, line\_idx.

Libraries: numpy.

1. **convolution\_and\_pooling:**

Filter size determination: The filter size is determined in the for loop where each filter size in the list self.filter\_sizes is iterated. Variable scope definition: tf.variable\_scope is used to define the variable scope, which is useful to avoid name conflicts in variables. Weights and biases definition: The weights W and biases b are defined using tf.get\_variable. A truncated normal distribution is used to initialize the weights. Optimization process: tf.nn.conv2d is used to perform the optimization process on the input using the defined weights and biases. Activation function application: The activation function ReLU is applied using tf.nn.relu. Pooling operation: The pooling operation is performed using tf.nn.max\_pool. Results aggregation: At the end of each iteration, the pooled output is added to the list pooled\_outputs.

Local variables: pooled\_outputs, filter\_shap, w, b, conv, ksize, h, pooled, strides, padding, name.

Libraries: The following TensorFlow functions are used: **tf.variable\_scope:** It is used to define the variable scope, which is useful to avoid name conflicts in variables. **tf.get\_variable:** It is used to define weights and biases. **tf.nn.conv2d:** It is used to perform the optimization process on the input. **tf.nn.relu:** It is used to apply the ReLU activation function. **tf.nn.max\_pool:** It is used to perform the pooling operation.

1. **compute\_logits\_loss\_accuracy:**

The function uses softmax, which is used in classification. softmax\_w is a variable used for softmax weights, and it is defined using a truncated normal distribution. softmax\_b is the bias and it is a constant variable. Adding L2 regularization: L2 regularization is added to the output layer. self.l2\_loss is the value of the L2 regularization loss. Calculating logits: logits are calculated. self.logits is a list of logits that are calculated using the matmul function for matrix multiplication. Calculating cross-entropy loss: cross-entropy loss is calculated. self.loss is a list of cross-entropy losses with the addition of L2 losses. Calculating classification accuracy: classification accuracy is calculated. correct\_predictions is calculated using the tf.equal function to verify the correctness of the classifications. self.accuracy is calculated using the tf.reduce\_mean function to calculate the average..

Local variables: softmax\_w, name, softmax\_b, self.predictions, self.cost, self.correct\_num, self.12\_loss, self.logits, prediction, losses, correct\_pred, self.accuracy

Libraries: TensorFlow.

### second main component: diagram drawing

**class ElementsFinder:**

**attributes:**

**method:**

1. **GeneratingUsecase:**
2. **GeneratingClassDiagram:**

## Data Modeling Design

The program converts the text to lowercase, then divides it into sentences and stores each sentence in a row called 'requirements'. After that, the program outputs various elements such as 'actor', 'use case', 'relation between actor and use case', 'class', 'method', 'attribute', and 'relation between classes'. Finally, it creates a file with the extension 'diagram.uml' to draw the digram.

And we used this dataset.

• Software requirement dataset can be find in keggle

(<https://www.kaggle.com/iamsouvik/software-requirements-dataset>) .

• Labeled requirement dataset can be find in zenodo

(<https://zenodo.org/record/268542#.X05hwMgzZhF>) .

• Freely available Natural Language Requirements Dataset

(<http://fmt.isti.cnr.it/nlreqdataset/>) .

## User Interface Design



Figure 20: Home Page



Figure 20: Input Page



Figure 22: Diagram Page

## Summary

In this chapter, we have created the Architectural design and utilized the client-server pattern. The component and data modeling design will be developed in the next phase. Finally, we have initiated the design of the user interface.

In the upcoming chapter, we will summarize everything we have accomplished since the beginning of the project. We will discuss how we achieved our goals, the lessons Learnt, and future work.

# Chapter 5: Conclusion and Future Work

## Conclusion

In this section, we will summarize all the work we have done from the beginning to the end of the project. We started by identifying the problem and how we would solve it, then defined the project's goals, methodology, and project plan.

In the second chapter, we researched topics related to the problem we are solving, including AI, ML, and NLP, and how to develop a system for artificial intelligence. We also explored similar systems.

In the third chapter, we analyzed existing systems, identified requirements from these systems, conducted brainstorming sessions, and then created the use-case diagram. We described the use cases and drew the class diagram.

In the fourth chapter, we developed the architectural design and user interface design.

## Goals Achieved

We have successfully achieved our project goals. Here's a overview of the objectives and how we accomplished them.

1. **To review previous works that can be related to our project domain:**

We conducted research on technologies related to our project and similar systems.

1. **To finish the Analysis phase of the proposed system to be built:**

We identified requirements after analyzing similar systems and used brainstorming techniques. We also created a use-case and described them.

1. **To do the preliminary Design tasks of the suggested system:**

We developed initial designs for the project, including class diagrams, architectural engineering design, and the user interface.

1. **To conclude with the lessons learnt and knowledge and experience gained from working on this project:**

We documented any new knowledge and experience gained from the beginning to the end of the project.

## Lessons Learnt

1. **Working as a team-** This project helped us to work together, learn and benefit from each other when we needed each other.
2. **Project planning**- involves setting the goals of the project, creating a task list for each goal, and developing a weekly schedule. Tasks are reviewed and assigned every weekend, and completed work is presented to the supervisor for feedback and approval.
3. **Learn NLP**- NLP is a new system for us, and we did not study anything about it in our university courses.
4. **Improve writing skills-** One of the most important things we learned in our project is to improve writing skill.
5. **Learning System Development for AI (NLP, ML)**- taught us the development steps in the fields of Natural Language Processing (NLP) and Machine Learning (ML).
6. **Researched and Read papers**- Acquired knowledge of methods for searching for reliable references and extracting information from them.
7. **Writing References**- Acquired knowledge of how to write references in the IEEE style.
8. **Every software project has different types of requirements**.

## Limitations and Future Work

### Limitations

The work we haven't done yet is component design and data modeling design due to time constraints. We will implement them in the second phase of the project.

### Future work

The tasks we will undertake in the second phase of the project include:

1. Improving analysis and design, which we worked on in this phase of the project.
2. Component design
3. Data modeling design
4. Project development
5. Project testing

# References

|  |  |
| --- | --- |
| [1] | H. Sofian, N. A. M. Yunus and R. Ahmad, "Systematic Mapping: Artificial Intelligence Techniques in Software Engineering," *IEEE Access,* vol. 10, pp. 51021-51040, 2022. |
| [2] | R. S. e. Bashir, "UML models consistency management: Guidelines for software quality manager.," *International Journal of Information Management,* vol. 36, no. 6, pp. 883-899, 2016. |
| [3] | T. w. l. t. p. o. f. t. a. o. technology, "The world's largest technical professional organization for the advancement of technology," IEEE. |
| [4] | university of BATH, 2017. [Online]. Available: https://www.bath.ac.uk/publications/library-guides-to-citing-referencing/attachments/ieee-style-guide.pdf. |
| [5] | ". R. M. A. D. Library. [Online]. Available: https://dl.acm.org/. |
| [6] | G. F. Luger., Artificial intelligence: structures and strategies for complex problem solving., Addison-Wesley Publishing CompanyUnited States, 2008. |
| [7] | 2. W. A. Hany H Ammar1, Software Engineering Using Artificial Intelligence Techniques: Current State and Open Problems, 2013. |
| [8] | S. Shafiq, A. Mashkoor, C. Mayr-Dorn and A. Egyed, "A Literature Review of Using Machine Learning in Software Development Life Cycle Stages," *IEEE Access,* vol. 9, pp. 140896-140920, 2021. |
| [9] | Y. Vasiliev, Natural language processing with Python and spaCy: A practical introduction., No Starch Press , 2020. |
| [10] | E. &. C. M. B. Dias Canedo, " Software requirements classification using machine learning algorithms.," *Entropy,* vol. 22, no. 9, p. 1057, 2020. |
| [11] | S. Vajjala, B. Majumder, A. Gupta and H. Surana, "Practical Natural Language Processing: A Comprehensive Guide to Building Real-world NLP Systems.," O'Reilly Media, 2020. |
| [12] | H. a. A. W. B. Herchi, "From user requirements to UML class diagram," *arXiv preprint arXiv:1211.0713,* 2012. |
| [13] | K. A. D. O. K. K. Arachchi, AI Based UML Diagrams Generator, 2021. |
| [14] | S. Ian, Software Engineering, 11th ed., Addison-Wesley,, 2015. |
| [15] | R. J. Leach, Introduction to Software Engineering, 2016. |
| [16] | G. L. S. C. &. H. O. S. Moon, "Automated construction specification review with named entity recognition using natural language processing," *Journal of Construction Engineering and Management,* vol. 147, no. 1, p. 04020147, 2021. |
| [17] | S. &. C. C. Jaf, "Deep learning for natural language parsing," *IEEE Access,* vol. 7, pp. 131363-131373, 2019. |
| [18] | R. P. S. C. E. &. T. R. Sukthanker, "Anaphora and coreference resolution: A review," *Information Fusion,,* vol. 59, pp. 139-162, 2020. |
| [19] | C. Manning, "Stanford lecture (CS224n) by Christopher Manning," 2019. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture10-coreference.pdf. |
| [20] | S. G. L. &. H. C. Loáiciga, "What is it? Disambiguating the different readings of the pronoun 'it.," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pp. 1325-1331, 2017. |
| [21] | H. a. Y. N. A. M. a. A. R. Sofian, "Systematic mapping: Artificial intelligence techniques in software engineering," *IEEE Access,* vol. 10, pp. 51021-51040, 2022. |
| [22] | Z. a. L. F. a. Y. W. a. P. S. a. Z. J. Li, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems,* 2021. |
| [23] | V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b,* vol. 4, pp. 51-62, 2017. |
| [24] | J. T. Catanio, "Requirements analysis: A review," *Advances in Systems, Computing Sciences and Software Engineering,* pp. 411-418, 2006. |
| [25] | B. Arendse, "A thorough comparison of NLP tools for requirements quality improvement," 2016. |
| [26] | M. O. I. Bashir, "ResearchGate," [Online]. Available: https://www.researchgate.net/figure/Different-branches-of-AI-13\_fig1\_357512563. |
| [27] | JayeshBapuAhire, "Data Science Central," [Online]. Available: https://www.datasciencecentral.com/the-artificial-neural-networks-handbook-part-1/. |
| [28] | P. M. Marta Maślankowska, "neurosys," [Online]. Available: https://neurosys.com/blog/intro-to-coreference-resolution-in-nlp. |
| [29] | "UML/Code Generation Tool," visual paradigm, [Online]. Available: https://www.visual-paradigm.com/features/code-engineering-tools/. |
| [30] | "Defining design pattern," visual paradigm, [Online]. Available: https://www.visual-paradigm.com/support/documents/vpuserguide/26/36/6246\_definingdesi.html. |
| [31] | "lucid collaborative ai," innovation training, [Online]. Available: https://www.innovationtraining.org/lucid-collaborative-ai/. |
| [32] | "lucidchart vs lucidspark," innovation training, [Online]. Available: https://www.innovationtraining.org/lucidchart-vs-lucidspark-when-why-and-how-to-use-them-both/. |
| [33] | "starUML in 2022," Pat research, [Online]. Available: https://www.predictiveanalyticstoday.com/staruml/#content-anchor. |
| [34] | "Introduction to Enterprise Architect," sparx systems, [Online]. Available: https://sparxsystems.com/enterprise\_architect\_user\_guide/13.0/. |
| [35] | "MagicDraw," Dassault Systemes, [Online]. Available: https://www.3ds.com/products-services/catia/products/no-magic/magicdraw/. |