THE TEXT CLASSIFICATION OF HERBAL DISEASES AND PICTURE EXTRACTION OF AYURVEDIC RECIPES

Project ID: 2020-122

Project Proposal Report

K.A.G.Y. Nadee Kumari – IT17014250

B.Sc. (Hons) Degree in Information Technology

Specializing Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

February 2020

THE TEXT CLASSIFICATION OF HERBAL DISEASES AND PICTURE EXTRACTION OF AYURVEDIC RECIPES

Project ID: 2020-122

Project Proposal Report

B.Sc. (Hons) Degree in Information Technology

Specializing Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

February 2020

DECLARATION

We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our 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.

Name	Student	Signature
K.A.G.Y. Nadee Kumari	IT17014250	

The supervisor/s should certify the proposal report with the following declaration. The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

………………………………… …………………………………

Signature of the supervisor: Date

(Mrs. Lokesha Weerasinghe)

………………………………… ……………………………….

Signature of the Co-supervisor Date

(Mrs. Ishara Weerathunga)

ABSTRACT

Ayurvedic means a science of Life well- being with its unique approaches of social and spiritual life is in practice science centuries in the Indian sub-content. Some treatments failed in the western medical approach, there are most of the people believe Ayurvedic approach is the best way of providing treatments because of having a less error prone of giving a feedback in any type of treatments and Although some of unknown diseases, bone fractured are curved in better condition way and it keeps in reducing much number of side effects. Usually, Ayurvedic medicine plants are used for treatments are undergone long period of time after taking the western medical approach. The proposed system will provide many features as the services of giving traditional treatments in Ayurvedic. As a special feature of this Arogya app is identifying the category type of disease in an unknown medicine plants giving its special characteristics/properties. After displaying the category, the system will display the most used medicine sample of each category and visualizing some of herbs that are included. Hence, there visualized images can be taken from the already used a database using in Image processing stage. In here, users can select dropdowns giving in the text area according to the herb's properties. If not mentioning relevant traits, typing field is given for including special herb characteristics within limited words. All the needed variables are stored in a relevant database. Although using some machine learning algorithm for analyzing the categories of each type of medicine plants. Feature extraction and filtering options should be done under the preprocessing stage around Natural Language processing (NLP). Nowadays, people mostly used day to day tasks depend on variety of technologies, therefore identifying the ingredients in the prescription medicine/ recipes is a very tedious task for people who are depend on technology. So that, this proposed system will support to predict most usable prescription for given category and visual appearance of its included ingredients.

Arogya App is the most important to find unknown medicine plants and identify the better solutions for health-issues neglected in western approaches and the long period of traditional ayurvedic treatments.

Keywords:
Machine Learning, Feature Extraction, Characteristics, Categories, Natural Language Processing

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

NLP: Natural Language Processing.

IAMP: Intelligent Ayurvedic Management Platform

NLTK: Natural Language Toolkit.

1. INTRODUCTION

Ayurvedic medicine is more affordable for a long time of health-issues in living beings rather than the approach of western medicine. Some countries like India, Sri Lanka believe those treatments can be reduced more side effects and avoiding less chance of having a threat to their lives. Although those are helping to provide stabilizing hormones and metabolism, natural healing and Strength in the immune system. Traditional Ayurvedic treatments are mostly going on generation to generation by a specific person of a family. Some type of Ayurvedic treatments can support avoiding either the dangerous effects of health-issues having in western approaches.

Arogya is a mobile application app for an Intelligent Ayurvedic Management Platform (IAMP). It provides more services in Ayurvedic treatments for undergone alongside health-issues by neglecting in a western approach. The process of medical plants identification by giving particular herbal leaves, creating a full-detailed information of prescription medicine samples of each category type of diseases such as Cancer, Diabetics, Arthritis etc. [2] in provided herbal plants, displaying locations where the identified medical plants are spread by using a geographical map and analyzing the category of diseases and picture extraction of ingredients in prescription medicine such are the main components of focusing on this Arogya app. Our proposed system concerns with these four main components for giving better solutions for the neglecting of health issues in western approach and traditional health issues in Ayurvedic treatments. There are Most of the people who don’t know which are what to use, details of curving process and what are the best prescription medicine (recipes, samples) etc. Although not able to spend time on finding different types of herbal remedies or medical plants within a short period of treatment, therefore this Arogya App will provide a good chance for these problems are going on busy daily routines.

1.1 Background

Arogya mobile app is going to overcome some of the problems mentioned in above. apart from there is a special feature of texting a description including characteristics of herbal remedies/ plants. Mainly, that option will be provided for identifying each type of herbal remedies / plants which represent in what category of diseases who may be seen or ever seen either knowing of their characteristics well. Proposed system will give some dropdowns for selecting needed attributes/traits/properties/characteristics consist in relevant text area. If not there having a space to mention it in briefly. According to the application, there are some multiple categories to separate data which are entered by user. Then analyzing what type of category after checking and rechecking using machine learning algorithms (Naïve Bayes, decision tree, etc.)

In this App going to implement an automation text classification on Machine Learning algorithms related applications following below steps.

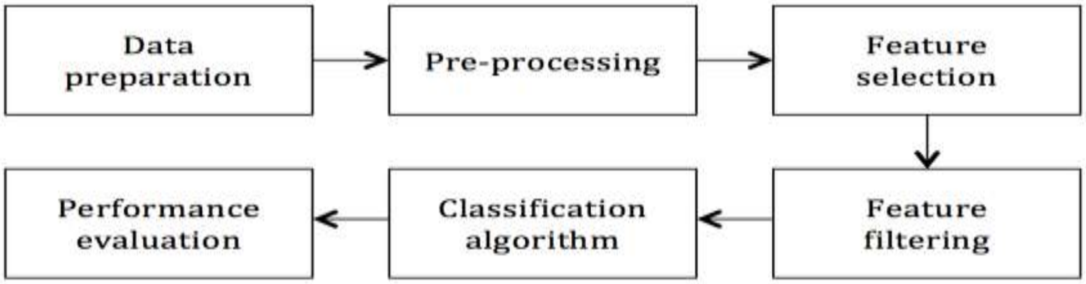


Figure 1.1- overall analysis process components

So that, this proposed system is more important to identify the herbal remedies when people failed in its identification and not understanding in better ayurvedic samples/prescriptions medicine even though in Offline status.

1.2 Literature Review

In ancient Indian medical system, also known as Ayurveda, is based on ancient writings that rely on a natural and holistic approach to physical and mental health. This is also world’s oldest medical systems and remains one of India’s traditional health care system. Every person engaged in various type of activities/ treatments internet either by using social media or online identification activities. Many numbers of activities go through on internet as well as treatments in Ayurvedic and western approach medicine for health-related issues.[1]. Nowadays, many online communities are focused on health-related issues in western medicine approach and Ayurveda treatments. In order to that it clarify different types of diseases such as diabetes, Arthritis, Cancers etc. [7] However, they introduced dynamic exercise plans, nutritious to treatment and diagnosis for curving everything else related diseases. [2]. Therefore, today Created many no of application for each of different tasks but our proposed system will given special options as clarifying a texts identifying its category type. According to this Text classification and picture extraction function is somewhat different from the extract process of text classification apply in this application. There interface displays with dropdown icons for selecting needed plants. Hence, there Analyzing a categorized process is on feature extraction stage using Naive Bayes classifiers. [4]

Comparison of the point of text classification and picture extraction in recipes between Arogya App and existing Mobile apps.

	Agrobase	Plantex	PlantSnap	MedLeaf	Arogya
Provide a space for typing a characteristic of desired herbal plants.	✗	✗	✗	✗	✓

Identifying a category type of diseases using a text classification	✗	✗	✗	✗	✓
Displaying recipes related to the classified category	✗	✗	✗	✗	✓
Picture Extraction of ingredients in recipes	✗	✗	✗	✗	✓
Getting results in Offline stage	✗	✗	✗	✗	✓

Table: 1.3.1 – Comparison between existing apps with Arogya app

1.3 Research Gap

Ayurvedic medicine systems are mostly produced to introduce some herbs details and prescriptions for either daily diseases or traditional diseases. The Existing Apps were introduced only for detailed information regarding methods of ayurvedic treatments.

1.3.1 Research Problem

But Our proposed system will provide many services for busy daily schedule of lives. According to this function “Analyzing the category of diseases and picture extraction of prescription medicine” is proceed on offline status. So that people can take this service whatever place they used. apart from that identify the ingredients in the prescription/ recipes is a tedious task for people who are depend on technologies. If we consider the medical leaves, most of the people are not familiar with their visual appearance. So, that this system will support to display most appropriate recipes in predicting categories and visual images of its appearance of included ingredients.

OBJECTIVES

The purpose of the Document /Text classification is to separate the contents of texts or documents for a one or multiple categories. It is mainly supported in providing information retrieval, document association and management. According to this function, “Text classification of herbal diseases and Picture extraction of Ayurvedic recipes” have much of goals including this proposed system.

1.4 Main Objectives

Users can identify the category type of diseases, just including its specific properties of needed medicine plants and ability to observe the visual appearance of ingredients in prescription which is most probably used.

2.2. Specific Objectives

There are some sub objectives related to this feature as per in below.

1. Users can identify better prescription medicine/recipes which are already not used in before
2. In any place can search because it can be activated either in off status
3. Identify more prescription samples according to the relevant category types.
4. Can be seen rare or unknown medicine plants which are included in as ingredients of medicine recipes.

2. METHODOLOGY

2.1 System Overview

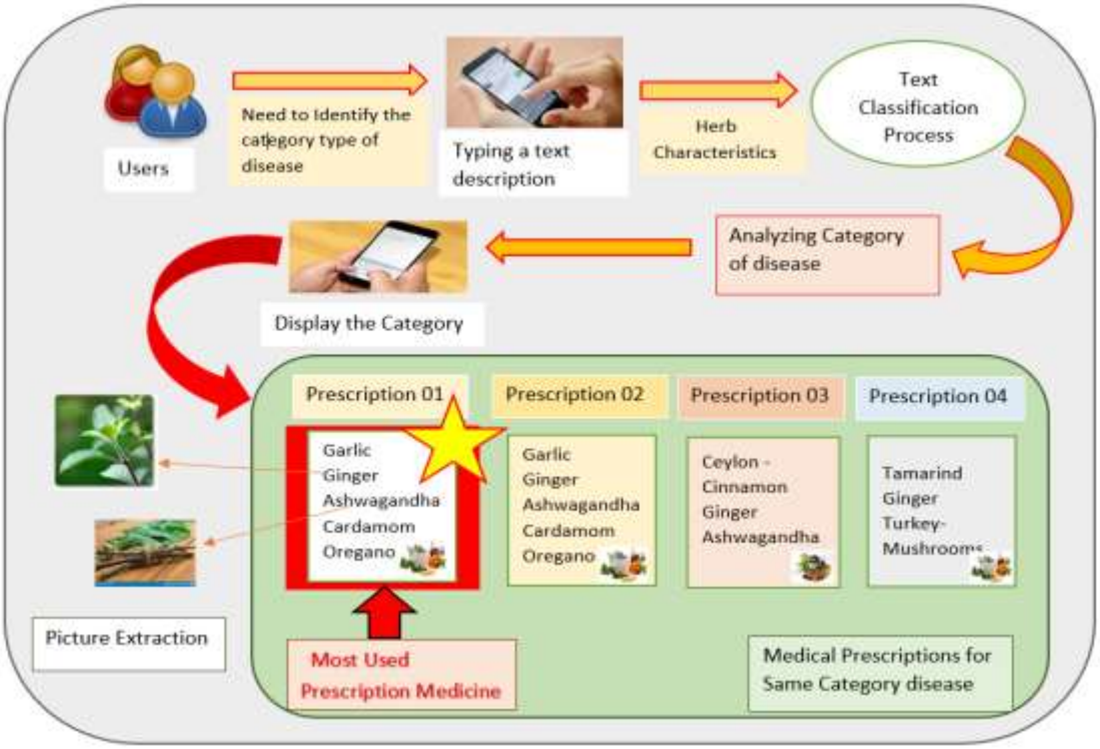


Figure 3.1

The outcome of this research project will produce a mobile application that including special features for Ayurvedic medicine. According to the "Analyzing the category of herb diseases and picture extraction of prescription medicine" function, there text analyzing is a separate process going on to identify unknown characteristics of herbal remedies via using machine learning algorithms. Then it will be predicting its own category type according to the giving details.

As considering the system overview of this function, it provides a text area to filling characteristics/properties of herbal plants. Hence, people can text whatever things they known about herbal plants. But this proposed system will be identified special points of herbal plants/remedies. Then System will display the name of the category disease according to the provided details after going on an analyzing process. In order to that Using some specific details should be gathered each of herb plants. Then displaying some of prescriptions which are relevant to the displayed category will be shown under the each of categories. Among those categories, system will display most usable prescription according to the user ranks and extract some pictures of herbals are included in.

2.2 Background Process of Text Classification using NLP

2.2.1 Process Model for supervised Text classification (High-Level Diagram)

Text Classification is an example of supervised machine learning task since a labelled dataset containing text documents and their labels is used for train a classifier.

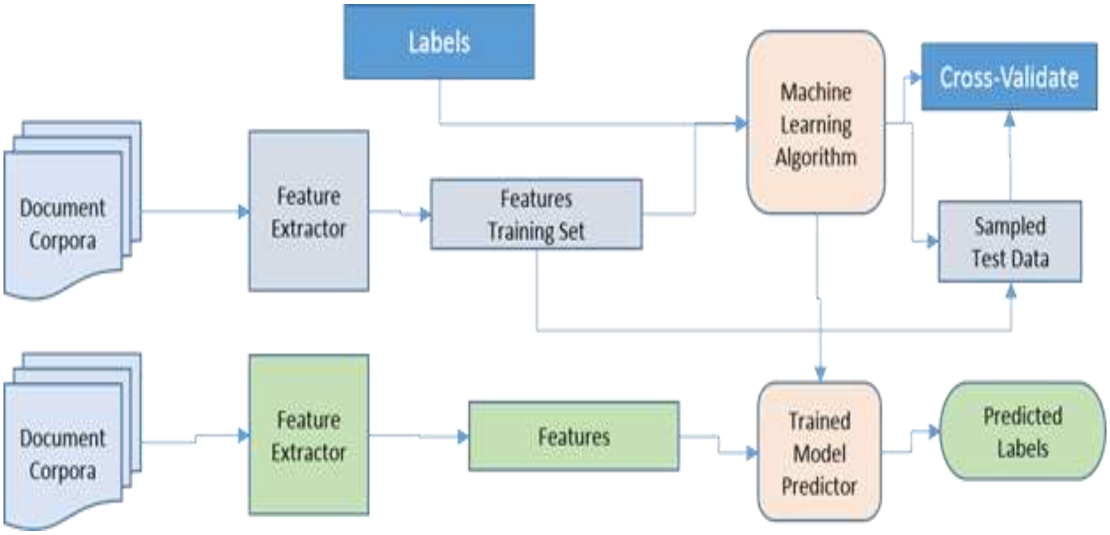


Figure 3.2.1- Process Model for supervised Text classification

This is using for a Text classification process while typing its sentiments of needed herbs. Generated data has a variety of tabular data columns have either numerical or categorial data. NLP (Natural Language Processing) is applicable in several problematic from speech recognition, language translation. Classifying documents to information extraction. This is helps identified sentiment of herbs, finding entities in the sentence, category of texts. NLP enables the computer to interact with human in a natural manner. It helps the computer to understand the human language and derive meaning from it.

2.3 System Architecture

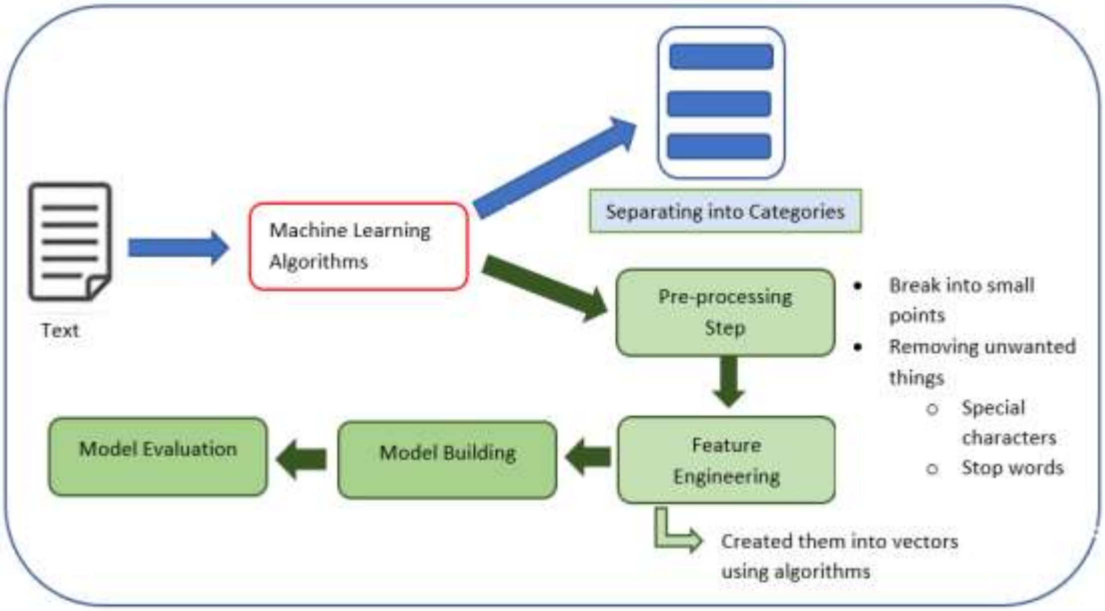


Figure 3.3 - System Architecture

Text classification process regarding the flow is implemented under the NLP. In order to that, NLP supports here to identify the category of texts and Information extraction. There end to end text classification which is a pipeline consists three main components. According to the Figure 3.2. It represents Dataset Preparation (Pre-processing Text), Feature Engineering, and Model Training are the major components of this classification process. Therefore, consisting some points regarding the each of levels describing briefly as per in below.

1. **Dataset Preparation**:

Loading a relevant dataset and performing basic pre-processing. If there including unwanted things which should be ejected on this stage. Hence, the dataset split into train and validation sets. The Traditional text classifiers usually break the documents into small word fragments(n-grams) and locate them as separate dimensions in the fragment hyperspace. These steps will be used for a typing field provide for special properties/features

According to the application, Giving an Interface including a separated dropdown buttons for selecting most appropriate words related to the medicine plants. After selecting all the dropdowns, there will be consisted a space for typing a special property if there have.

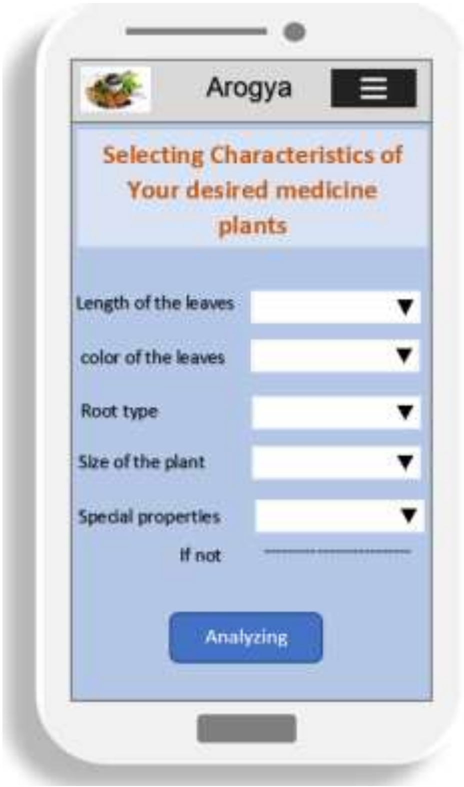
2. **Feature Engineering:**

Raw dataset transforms into flat features using Machine Learning model and This process is going on creating new features from the existing data. Raw text data will be transformed into feature vectors and new features will be created using an existing text data. According to the new features from dataset can be identified different ideas. Such as, Count vectors as features, TF-IDF Vectors as features, Word as features, Text NLP based features, Topic model as features. But this function not able to do such type of bored things to classify. Hence, Naïve Bayes algorithm is used for this all the NLP based TF-IDF.

According to the application, providing a same dropdown option to selecting the specific characteristics if the plants be consisting on. If this dropdown will not be included in needed characteristics, then giving a limited space for typing it with main points. So that, there only typing field must consider for the feature extraction part. There is a space for typing an any special feature of the medicine plant in provided area. In there should be considered about filtering options such as removing emoji, punctuation marks, spaces, special characteristics etc. Then it will be transformed into feature vector. This field is not a required filed and it’s not use in always because there are most of the special characteristics are included in dropdown field.

3. Model Training:

Machine learning model is trained on a labeled dataset. In there, data which are taken from dropdown fields should be moved to analyze options using Algorithms. Other data which are from typing area should have to label and categorize according to the relevant types of category diseases. Before that these types of data coming after doing previous steps. (Pre-processing and Feature Extraction). After going on both components, should have to Improve the performances of text classifiers.



According to the application,

This is the Planned Mobile UI of this Analyzing the category of diseases part. All the Fields should be required excepting a provided typing area. Finally, all the filed be filed then pressing a button "Analyzing" move to the next interface. Then displaying a Category after taking some time period of loading. Apart from that this interface will be displayed some recipes related to the displayed category. Then user can select most usable recipe and giving a chance of visualizing there some of rare ingredients which are included.

3.4 Text Classification Using Tools

Both NLTK and TextBlob performs well in Text Classification processing.

3.4.1 Text Classification - NLTK

NLTK is a very big library holding 1.5GB and has been trained on a huge data with proving different dataset in multiple languages which can deploy according to the functionality its be required. NLTK is a powerful Python package that provides a set of diverse natural language algorithms.

3.4.2 Text Classification - TextBlob

TextBlob library which is a python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as noun phrase extraction, sentiments analysis, classification, translation etc. this function will be applied for classification task. Under the Features there are some points to follow on.

1. Classification – Using Naïve Bayes and Decision Tree
2. Tokenization – Splitting text into words and sentences
3. Spelling correction
4. Emojis

3.4.3 Comparison of Text classification tools

There are lots of tools to work with NLP. NLTK, Spacy, Stanford Core NLP. These are the comparison of properties of tools

	NLTK	Spacy	Stanford Core NLP	TensorFlow	Allen NLP
Build an end-to-end production application	✓	✓	✓	✓	✗
Efficiency on CPU	✓	✓	✗	✗	✗
Train models from own data	✓	✓	✓	✓	✓
Different neural network architectures for NLP	✗	✗	✗	✓	✓

Table: 3.4.3 – Comparison of text classification tools

3.5 Text Classification Using Algorithms

Consisting of the most common algorithms such as *tokenizing*, part-of-speech tagging, topic segmentation, named entity recognition, etc. when considering this function, it helps the computer analysis, pre-process, and understand the written text. In This part, only use for the typing field and this field is consist with limited words as key words of herbal characteristics. Because if we give a large description there should have more stuffs further.

Text Classification Steps for typing field:

1. Loading data and Creating classifiers:

First create custom classifiers using TextBlob module. Before that creating some training and test data. Then creating a Naïve Bayes Classifier for passing the training data into Constructor.

2. Loading data from Files:

Loading data from common file formats including CSV, JSON, and TSV

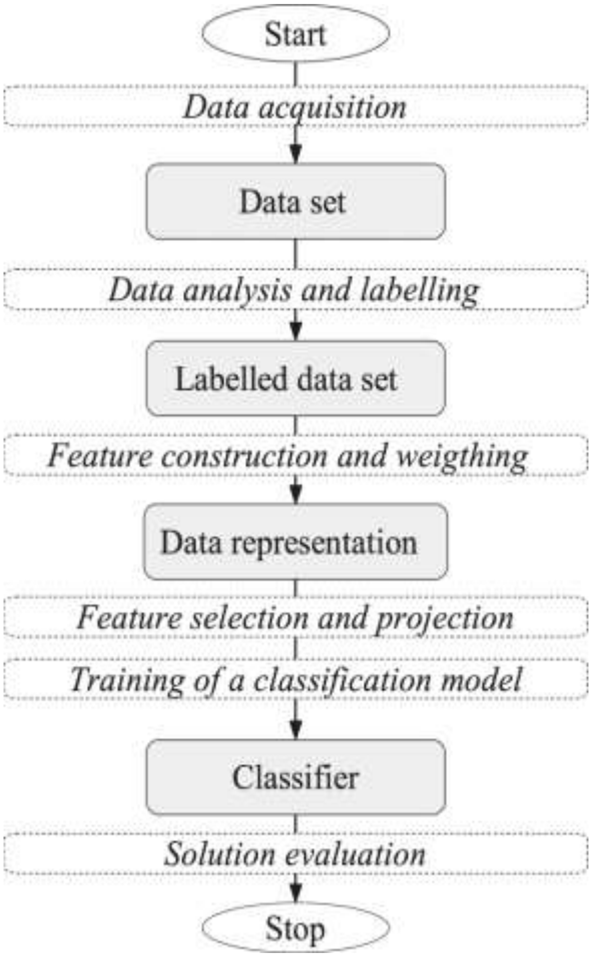
3. Classifying Text or Classifying TextBlobs
4. Evaluating Classifiers

   Compute the accuracy of the test data using a relevant method

5. Updating Classifiers with New Data
6. Feature Extractions.

We can use our own extractors to identify each specialty.

Text Classification – Flow Chart


Start
Data acquisition
Data set
Data analysis and labelling
Labelled data set
Feature construction and weigthing
Data representation
Feature selection and projection
Training of a classification model
Classifier
Solution evaluation
Stop

This application will classify only 3 or 4 category type of diseases such as Diabetics, Arthritis and Cancer or Sugar. According to the above categories giving by the users(dataset), has been labeled into some variables like color, length, shape etc.

Then using feature extraction functionalities like locate vectors etc. and for classifying them into model.

So that easily predicting a solution on model evaluation step.

Figure 3.5 – Text classification flow chart

After that selecting a Most usable medicine recipe and visualize ingredients as Feature extraction. Here take those images from the database which is used to store all the images in initial stage

3.6 Work Break Down Chart – Individual part

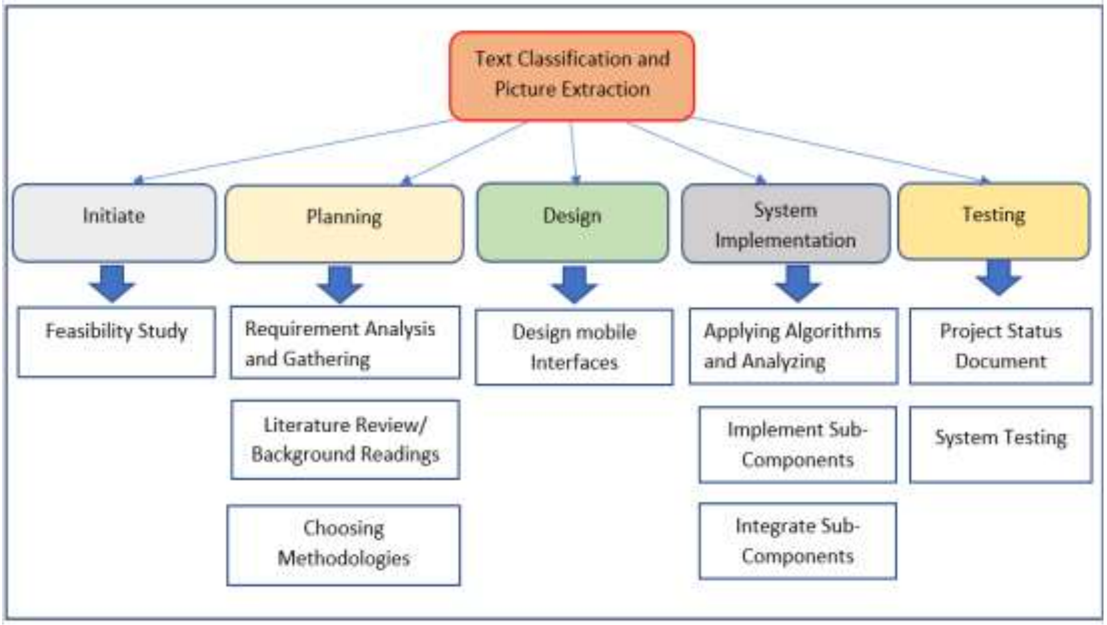


Figure 3.6 – Work break down chart

Stage	Time Estimation	Workload Distribution
Initiate	Nov- Dec (2019)	▪ Review Research Papers
Planning	Dec – Jan (2020)	▪ Analyzing Function part ▪ Background Readings ▪ Choosing methodologies
Design	March (2020)	▪ Creating wireframes ▪ Analyzing other existing UI
System Implementation	March – Sep (2020)	▪ Applying algorithms ▪ Implementing Mobile Interfaces
Testing	Sep – Nov (2020)	▪ Testing Subcomponents – unit testing ▪ Integration Testing ▪ System testing

Table: 3.6 – Work break down chart work allocation

3.7 Gantt chart – Individual Part

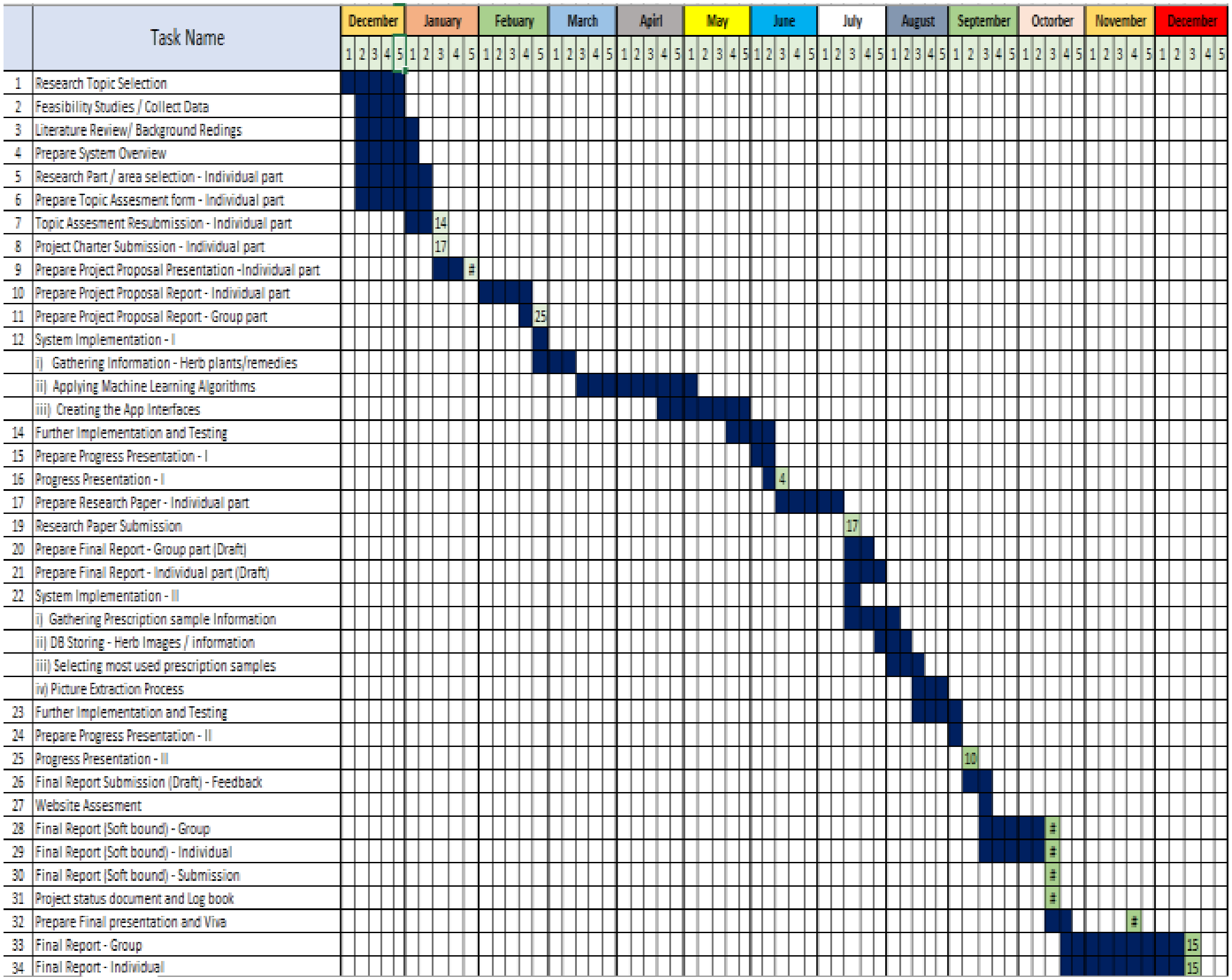


Figure 3.7 – Gantt chart

4. PROJECT REQUIREMENTS

4.1 Functional and Non-Functional Requirements

Functional Requirements	**Non-Functional Requirements**
Progress should be Efficient	Results should be efficient
Algorithms give prompt responses	Accuracy should be required
	Interfaces should be User-friendly
	Less Time-consuming for searching details of medicine plants

Table: 4.1 – Functional and non-functional requirements

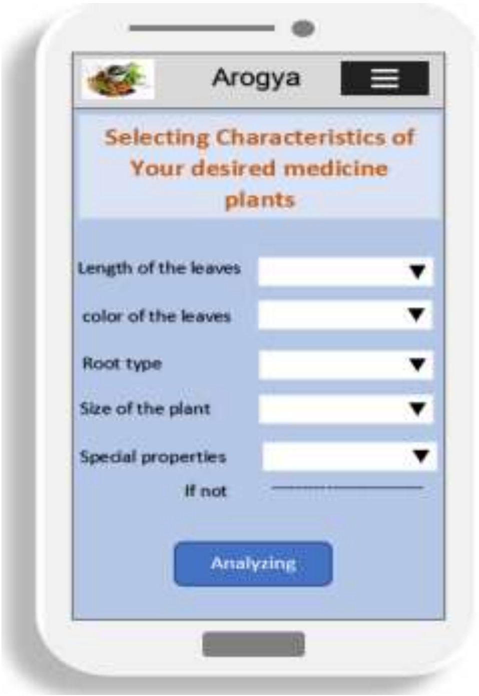
4.2 User Requirements

- Have a medicine plant which was known one or unknown one but familiar with the characteristics/properties

4.3 System Requirements

- If want to updated DB data, then use strong strength of Internet connection.

4.4 Wireframes



Arogya

Selecting Characteristics of Your desired medicine plants

Length of the leaves ▼

color of the leaves ▼

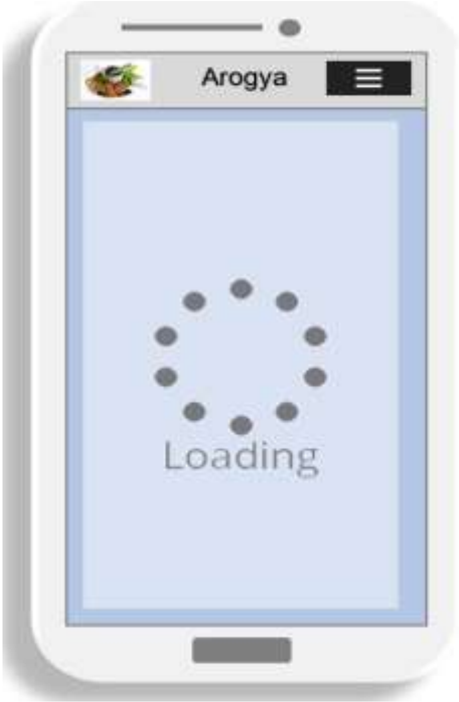
Root type ▼

Size of the plant ▼

Special properties ▼

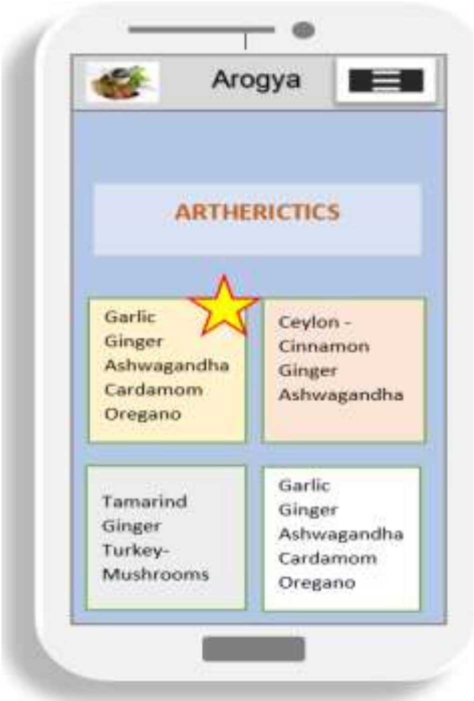
If not

Analyzing



Arogya

Loading



Arogya

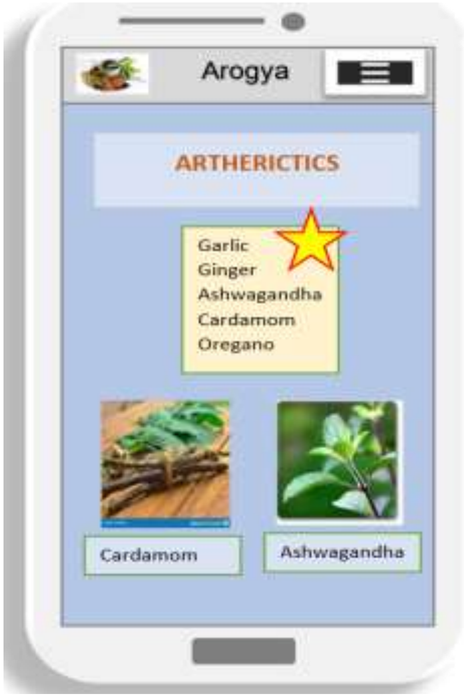
ARTHERICTICS

Garlic
Ginger
Ashwagandha
Cardamom
Oregano

Ceylon - Cinnamon
Ginger
Ashwagandha

Tamarind
Ginger
Turkey-Mushrooms

Garlic
Ginger
Ashwagandha
Cardamom
Oregano



Arogya

ARTHERICTICS

Garlic
Ginger
Ashwagandha
Cardamom
Oregano

Cardamom

Ashwagandha

5. BUDGET PALN AND BUSINESS PROTENCIAL VALUE

5.1 Budget plan and Justification

Arogya App is a mobile application for an Intelligent Ayurvedic Management Platform (IAMP) providing specific features to users who are willing to giving Ayurveda treatments.

- Finding specific medicine plants and their detailed- information
- Meet Ayurveda doctors and discuss with Ayurveda medicine treatments
- Finding rare plants in diseases categories and analyzing its characteristics

	Sri Lankan Rupees
Ayurvedic Doctor Charges (per appointment)	Rs. 3000/=
Data collection travelling chargers	Rs. 4000/=
Documentation printout cost	Rs. 3000/=
Total	**Rs 10000/=**

Table 5.1: Budget Plan and Justification

5.2 Commercialization

People who can access to Arogya in offline or online with friendly user interfaces and Giving better solutions for the critical health-related issues in Ayurveda. Considering the function related text classification and feature extraction that can be commercialized for saving time of people who are spending a tedious lifestyle.

5.2.1 Target Audience

- People who use ayurvedic treatment
- Ayurvedic plant sellers
- Doctors, Students, locals and foreigners

5.2.2 Market Space

- No age limitations for the users
- No need of advance computer literacy
- No need of advance knowledge in Ayurveda field

6. REFERENCES

[1]H. J. Gunathilaka, P. Vitharana, L. Udayanga, and N. Gunathilaka, “Assessment of Anxiety, Depression, Stress, and Associated Psychological Morbidities among Patients Receiving Ayurvedic Treatment for Different Health Issues: First Study from Sri Lanka,” BioMed Research International, vol. 2019, pp. 1–10, Nov. 2019, doi: 10.1155/2019/2940836.

[2]J.-R. Reichert, K. L. Kristensen, R. R. Mukkamala, and R. Vatrapu, “A supervised machine learning study of online discussion forums about type-2 diabetes,” in 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), 2017, doi: 10.1109/healthcom.2017.8210815.

[3] B. Ravishankar, “Journal of Ayurveda Medical Sciences – Peer Reviewed Journal for Rapid Publication of Ayurveda and Other Traditional Medicine Research,” Journal of Ayurveda Medical Sciences, vol. 1, no. 1, pp. 01–02, Oct. 2016.

[3] “Facts and Figures” [Online]
Available: https://spacy.io/usage/facts-figures

[4] “Natural Language Processing (NLP) Techniques for Extracting Information” [Online]
Available:
https://www.searchtechnologies.com/blog/natural-language-processing-techniques

[5] “Text Analytics for beginners Using NLTK” [Online]
Available:
https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

[6] “Machine Learning, NLP: Text Classification Using scikit-learn, python and NLTK” [Online] Available:
https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a

[7] “Ayurveda modern medicine interface: A critical appraisal of studies of Ayurveda medicines to treat osteoarthritis and rheumatoid arthritis” [Online]
Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3087360/

[8] Efficacy and Safety evaluation of Ayurvedic treatments (Ashwagandha power & sidh Makardhwaj) in rheumatoid arthritics patients: apilot prospective study [Online]
Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4405924/