Cairo University

Faculty of Computers and Artificial Intelligence

Department of Computer Sciences

PathFinder

Supervised by

Dr. Mohammed ElRamly

TA. Mohammed Atta

Implemented by

|  |  |
| --- | --- |
| 20200193 | Ziad Ibrahim Galal |
| 20200162 | Hamza Abdul-Hameed Ali |
| 20200348 | Omar Tarek |
| 20200636 | Yahia Ashraf |
| 20201222 | Yahia Mahmoud |

Graduation Project

Academic Year 2023-2024

Midyear Short Documentation

Table of Contents

[Abstract 4](#_Toc158645704)

[1 Background 4](#_Toc158645705)

[1.1 Project scope 4](#_Toc158645706)

[1.2 Motivation and Beneficiaries 5](#_Toc158645707)

[1.3 Main Techniques and Technologies 5](#_Toc158645708)

[2 Problem Definition 5](#_Toc158645709)

[3 Market Study 6](#_Toc158645710)

[4 Data Collection 6](#_Toc158645711)

[4.1 Web Scraping 7](#_Toc158645712)

[4.2 Crowd Sourcing 7](#_Toc158645713)

[4.2.1 PathFinder Tool 7](#_Toc158645714)

[5 Project Specifications 8](#_Toc158645715)

[5.1 Functional Requirements 8](#_Toc158645716)

[5.2 Non-Functional Requirements 9](#_Toc158645717)

[5.2.1 Performance Requirements 9](#_Toc158645718)

[5.2.2 Software Quality Attributes 9](#_Toc158645719)

[5.3 Stakeholders 9](#_Toc158645720)

[5.4 Use-Case Diagrams 10](#_Toc158645721)

[5.4.1 Student use case diagram 10](#_Toc158645722)

[5.4.2 Admin use case diagram 10](#_Toc158645723)

[5.4.3 University use case diagram 11](#_Toc158645724)

[5.4.4 Use-Case Tables 11](#_Toc158645725)

[Use chatbot use case table 11](#_Toc158645726)

[Add/Remove University use case table 12](#_Toc158645727)

[6 System Design 13](#_Toc158645728)

[6.1 Component Diagram 13](#_Toc158645729)

[6.2 Sequence Diagrams 15](#_Toc158645730)

[6.2.1 Student Sequence Diagrams 15](#_Toc158645731)

[6.2.2 Admin Sequence Diagrams 18](#_Toc158645732)

[6.2.3 University Sequence Diagrams 19](#_Toc158645733)

[6.3 ERD 20](#_Toc158645734)

[6.3.1 Main ERD 20](#_Toc158645735)

[6.3.2 PathFinder Tool ERD 20](#_Toc158645736)

[7 Machine Learning Models 21](#_Toc158645737)

[7.1 Conversational AI 21](#_Toc158645738)

[7.1.1 How it works 21](#_Toc158645739)

[7.1.2 Role and Benefits 21](#_Toc158645740)

[7.1.3 What is Rasa? 21](#_Toc158645741)

[7.1.4 Why Rasa? 22](#_Toc158645742)

[7.2 Chatbot Model 22](#_Toc158645743)

[7.2.1 NLU Training Data 22](#_Toc158645744)

[7.2.2 Rasa NLU Pipeline 23](#_Toc158645745)

[7.2.3 Rasa Core Policies 24](#_Toc158645746)

[7.3 Recommendation System 25](#_Toc158645747)

[7.3.1 Knowledge-Based Recommendation Systems (KB RS): 26](#_Toc158645748)

[7.3.2 Hybrid Recommendation Systems: 27](#_Toc158645749)

[8 Work Plan 28](#_Toc158645750)

[8.1 Task Table 28](#_Toc158645751)

[8.2 Gantt Chart 29](#_Toc158645752)

[9 References 29](#_Toc158645753)

List of Figures

[Figure 1 - PathFinder Tool 7](#_Toc158641061)

[Figure 2 - Student use case diagram. 10](#_Toc158641062)

[Figure 3 - University use case diagram. 11](#_Toc158641063)

[Figure 4 - Admin use case diagram. 11](#_Toc158641064)

[Figure 5 - Component Diagram. 14](#_Toc158641065)

[Figure 6 - Student Use Chatbot Sequence Diagram 16](#_Toc158641066)

[Figure 7 - Student Sign-Up Sequence Diagram 17](#_Toc158641067)

[Figure 8 - Student Log-In Sequence Diagram 18](#_Toc158641068)

[Figure 9 - Student Search College Sequence Diagram 18](#_Toc158641069)

[Figure 10 - Student Compare Colleges Sequence Diagram 19](#_Toc158641070)

[Figure 11 - Admin Review Feedback Sequence Diagram 19](#_Toc158641071)

[Figure 12 - Admin Add University Sequence Diagram 20](#_Toc158641072)

[Figure 13 - University Sign-Up Sequence Diagram 21](#_Toc158641073)

[Figure 14 - University Announcement Sequence Diagram 21](#_Toc158641074)

[Figure 15 - University Add College Admin Sequence Diagram 22](#_Toc158641075)

[Figure 16 - Main ERD 23](#_Toc158641076)

[Figure 17 - PathFinder Tool ERD 24](#_Toc158641077)

[Figure 18 – Rasa pipeline and policies. 25](#_Toc158641078)

[Figure 19 – Recommendation system types. 29](#_Toc158641079)

[Figure 20 - Gantt Chart 32](#_Toc158641080)

List of Tables

[Table 2 - Student Use Chatbot Use Case Table 13](#_Toc158583017)

[Table 3 - University Admin Add/Update Info Use Case Table 13](#_Toc158583018)

[Table 4 - Task Table 31](#_Toc158583019)

# Abstract

When students complete high school, especially Egyptian students, many often find themselves uncertain about which university or college they should enroll in to pursue a professional education. This lack of direction and data can lead to confusion and anxiety among students who are eager to embark on a fulfilling career path.

After high school, we are dedicated to assisting students in finding the perfect college that aligns seamlessly with their interests and aspirations. Our approach involves meticulous research and evaluation of numerous colleges and universities, ensuring we provide comprehensive and detailed information to empower students to make well-informed decisions about their higher education.

Our support extends to offering personalized recommendations and suggesting colleges that we believe are an excellent fit based on the student's preferences, academic ambitions, and future goals. We are committed to guiding students toward a path that meets their educational needs and inspires their success.

There are many tools needed to achieve our goals, such as Django REST framework, Rasa Open Source, MongoDB, React, and machine learning.

# Background

## Project scope

A place for high school students to get personalized recommendations and suggestions for colleges that we believe are an excellent fit based on the student's preferences, academic ambitions, and future goals. We are committed to guiding students toward a path that meets their educational needs and inspires their success.

The main area of the application is our chatbot which interacts with the user answering their questions and gathering information about the student, and a recommendation system uses the gathered information to recommend a college for that student.

## Motivation and Beneficiaries

The primary motivation behind our project is to reduce the confusion and anxiety of high school students in university selection, especially in Egypt. By providing a comprehensive College inquiries and Recommendation System, we aim to help students with personalized recommendations and detailed insights into various colleges and universities.

The primary beneficiaries of our project are high school students looking to enter higher education. Through our project, we aim to help them make an informed decision about their future careers.

## Main Techniques and Technologies

Our system is composed of a web application that includes a chatbot. The main techniques and technologies used in our system are:

* **Rasa Open source**: Rasa is an open-source conversational AI framework designed for building contextually aware conversational AI.
* **Application**: Chatbot Development
* **Django REST framework**: Django REST framework is a Python framework used to build restful web API.
* **Application**: Web Application’s Backend
* **React.JS**: React.JS is a popular JavaScript library for building dynamic, and interactive user interfaces.
* **Application**: Web Application’s Frontend
* **MongoDB**: MongoDB is a flexible and scalable NoSQL database that uses a document-oriented data model.
* **Application**: Data Storage
* **MySQL**: MySQL is a popular open-source relational database management system (RDBMS) known for its reliability, performance, and ease of use.
* **Application**: Data Storage
* **Machine Learning libraries**: Machine learning libraries encompass a wide range of tools and frameworks used for building and training machine learning models and providing access to various other algorithms and techniques.
* **Application**: Building the Recommendation System

# Problem Definition

High school students, particularly in Egypt, face significant uncertainty and anxiety when it comes to selecting a suitable university or college for their higher education. The lack of accessible and comprehensive information about various institutions, coupled with the absence of personalized guidance, often leaves students feeling overwhelmed and directionless.

As a result, there is a need for a platform that offers tailored recommendations and detailed insights into colleges and universities, aligning with students' preferences, academic aspirations, and future career goals.

# Market Study

When facing the decision of choosing a college, people try several solutions:

1. **Asking other people:**

People tend to ask other people about their opinions regarding colleges they consider candidates. This is a decent solution. However, people around may just be a non-representative sample. They may mislead the student and guide him towards an unsuitable decision.

1. **Online research:**

A lot of websites provide data about colleges such as <https://www.dbse.co/universities>, <https://egecmena.com/ar>, <https://bigfuture.collegeboard.org/>, <https://www.appily.com/> and many more. However, those websites have a lot of problems such as outdated and insufficient data. It’s usually the case that you an extensive search is needed to know little information that help in taking a decision.

1. **Visiting multiple colleges:**

Visiting colleges is a very effective solution. However, it requires a lot of time and effort.

1. **Chatbots**:

Options like ChatGPT are available for students to use in order to help them decide. However, ChatGPT has a more general use and is not very good when addressing the domain at hand. Other specialized options are emerging such as <https://mainstay.com/>. However, they only support universities and colleges in foreign countries.

Our solution facilitates the process of choosing a suitable college through a single web app. This app is equipped with data about all colleges, a chatbot and a recommendation system. Hence, it saves a lot of time and effort in comparison with the various solutions stated above. It basically works as a hub that includes data of all colleges, ratings of other people, on-demand answers to inquiries and personalized recommendations to high school students.

# Data Collection

During our project, we needed to collect data such as university information, questions that the students might ask, and answers to these questions. While collecting these data we faced several challenges related to data, including:

* Lack of Centralization: The required data was scattered across multiple locations rather than being centralized in one place.
* Outdated Data: A big portion of the data we needed was not maintained and updated regularly.
* Accessibility Issues: Some data needed for our project was not readily accessible, requiring a big effort to acquire it ourselves.

To address these challenges, we used two primary approaches:

1. Web Scraping: by using web scraping techniques, we extracted relevant data from various online sources, gathering scattered info into a more centralized location for use in our project.
2. Crowdsourcing: due to the data unavailability, we needed to gather the information from diverse sources and contributors by ourselves.

In the following sections, we talk about the strategies used in more detail and the outcomes achieved through these two approaches.

## Web Scraping

The technique we used for web scaping is HTML parsing in which we send HTTP requests to websites, retrieving the HTML content of the page then parse the HTML to extract the desired data. This is done by using libraries such as BeautifulSoup for Python.

This approach helped us solve a big portion of the challenges associated with the lack of centralization and outdated data. Scrapping data from multiple websites and then cleaning and storing them helped us make the data more centralized. Also, during the process of scrapping these websites, we encountered multiple websites that contained the same data, which allowed us to identify and select the most recent and relevant information.

## Crowd Sourcing

To overcome the challenge of data accessibility, we spent a considerable effort on crowdsourcing. The process consisted of two steps.

Firstly, we sent forms to our colleagues and institutions, including other universities. These forms consisted of the questions which may be asked to our bot such as general information about colleges, etc.

Secondly, we developed a specialized tool to help us gather answers to these questions more easily. To be able to view and clean these answers more easily. We will talk about the tool in more detail in the following section.

Of course, crowd-sourcing has multiple drawbacks such as the quality of data due to different individuals contributing to the data. However, these problems can be mitigated by an effort to clean the data and double-check it before using it.

### PathFinder Tool

PathFinder tool is a simple website designed to facilitate the gathering of responses to the collected questions that our chatbot may encounter from diverse sources, including colleagues and institutions. it consists of multiple simple components:

Figure 1 - PathFinder Tool

* Frontend component: a simple web interface that displays a randomized question from a predetermined pool of questions to the user. The user is then asked to either answer the question or skip it. If the user answers the question and submits then the answer is sent to the backend.
* Backend component: the backend receives the question's ID and the question's answer from the frontend via an HTTP POST request. It then validates that the answer is not empty. Finally, it stores the answer in the database using the question's id.
* Database: the database is a MySQL database hosted on pythonanywhere. It follows the schema previously defined by the ERD in figure 16.

PathFinder Tool serves as a complementary tool to web scraping and other data acquisition methods. By combining insights gathered through the tool with other sources of data. We can ensure that the data is well-rounded for our use.

# Project Specifications

## Functional Requirements

For Students

* Supported by the Chatbot
  + Input preferences and queries in natural language to receive personalized recommendations.
  + Receive detailed information about recommended colleges.
  + Access chat history.
* Supported by the webapp
  + Sign up for the system by providing personal details and email.
  + Log in to the system using email and password.
  + View comprehensive information about colleges and universities.
  + Search for colleges based on name or its university and display it in an organized manner.
  + Select and compare two colleges easily, in an organized manner, side by side.
  + Access news updates related to universities' events, scholarship opportunities, and more.
  + Filter displayed news content based on college/university.
  + Provide feedback on college choices and the recommendation process.

For Universities/Colleges

* Sign up by providing the necessary details.
* Sign in to the system using the provided credentials.
* Edit or delete uploaded information about colleges and courses.
* View all uploaded information and products.

For Admin

* View the University's details and upload information.
* Add/remove/edit universities’ details.
* Review feedback received from students

## Non-Functional Requirements

### Performance Requirements

* **Response Time** The system shall respond to user interactions within an average response time of 3 seconds, measured from the initiation of the request to the display of the corresponding response.To enhance user satisfaction and engagement, improving overall user experience.
* **Concurrent User Handling** The system shall support a minimum of 100 concurrent users without degradation of response time beyond 2 seconds. The platform must handle a significant number of users simultaneously to accommodate peak usage times without compromising performance.
* **Database Query Performance** Database queries should execute within an average time of 200 milliseconds, considering a reasonable load and dataset size. Efficient database performance is crucial for data retrieval and application responsiveness.
* **Chatbot Response Time** the Chatbot shall generate responses to user inputs within 1 second, providing a quick and natural conversation flow. Swift responses from the Chatbot maintain engagement and encourage continued interaction.
* **Error Handling** The system shall display clear error messages within 3 seconds of an error occurrence, aiding users in understanding and resolving issues. Quick and informative error handling is crucial for seamless user experience and effective troubleshooting.

### Software Quality Attributes

* **Usability** The system shall provide intuitive user interfaces, clear navigation paths, and contextual help features to ensure that users can easily understand and utilize the system's functionalities.
* **Reliability** The system shall achieve a system uptime of 99.9% availability, calculated monthly.
* **Maintainability** The system shall be designed and implemented in a modular fashion, allowing for easy maintenance and future enhancements by facilitating the isolation and replacement of individual components without impacting the overall system functionality.
* **Portability** The system shall be compatible with the latest versions of major web browsers (e.g., Chrome, Firefox, Safari) on Windows, macOS, and iOS, Android operating systems.

## Stakeholders

1. **End-user**
   * 1. **Student** high school graduates who need college and career advice based on their preferences, desires, and career goals.
     2. **University/Colleges** who want students to know more details about their programs, scholarships, events, and tuition fees.
     3. **Admins** who verify and suspend users, verify universities and handle feedback.
2. **Developers**.

Collect the data needed for the system, design, build, deploy, and test the system to be ready for users.

## Use-Case Diagrams

### Student use case diagram

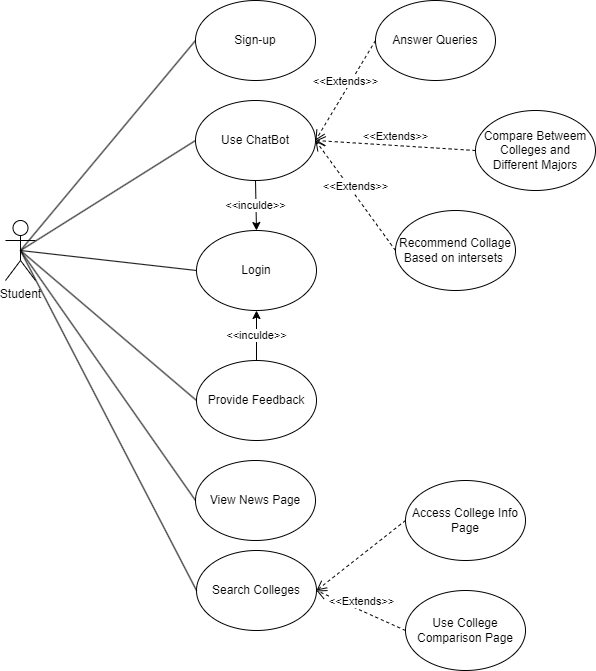


Figure 2 - Student use case diagram.

### Admin use case diagram

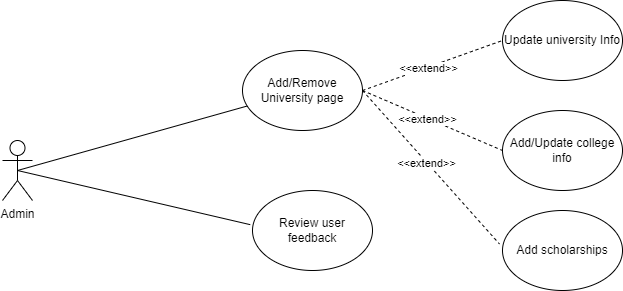


Figure 4 - Admin use case diagram.

### University use case diagram

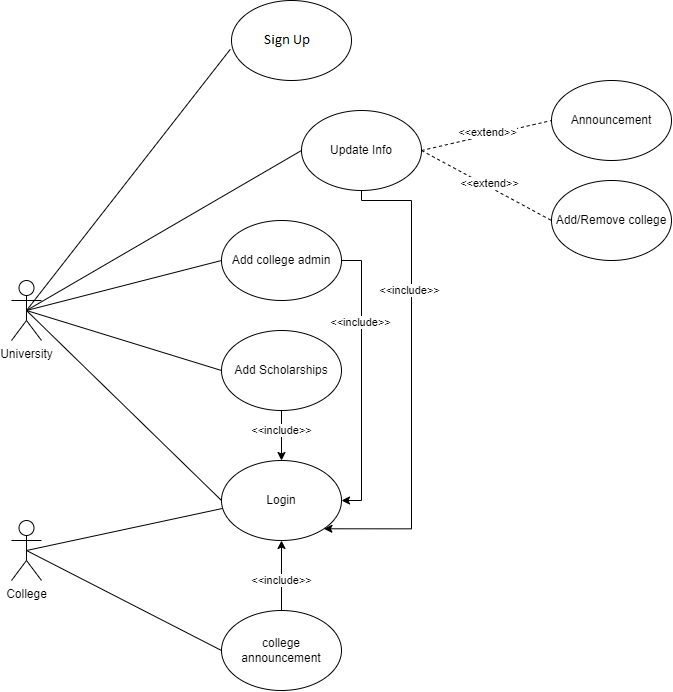


Figure 3 - University use case diagram.

### Use-Case Tables

Use chatbot use case table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Description | Student wants to query the chatbot for recommendations or guidance | | | | | |
| Actor | Student | | | | | |
| Preconditions | Student has already signed up for an account on the webapp and the webapp is already opened | | | | | |
| flow | # | | Actor Action | System Response | | |
| 1 | | Student chooses the chatbot option from the webapp | | | The webapp asks the student to login using his account |
| 2 | | Student enters his login credentials | | The webapp redirects the user to the chatbot interface | |
| 3 | | Student asks the chatbot for college recommendation based on his info and interest | | The chatbot recommends a college to the student based on the provided info | |
| Alternative Flow #1  Starts at step 3) | # | Actor Action | | System Response | | |
| 3 | Student asks the chatbot a general question about computer science | | The chatbot answers the question with the correct info | | |
| Alternative Flow #2  (Starts at step 3) | # | Actor Action | | System Response | | |
| 3 | Student asks the chatbot for comparison between two colleges | | The chatbot shows the student a comparison of the two colleges | | |
| Exception | The student’s login credentials are wrong so they aren’t able to use the chatbot | | | | | |

Table 2 - Student Use Chatbot Use Case Table

Add/Remove University use case table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Description | Admin adding or removing university page | | | | | |
| Actor | Admin | | | | | |
| Preconditions | Admin is logged in | | | | | |
| flow | # | | Actor Action | System Response | | |
| 1 | | Admin chooses to add a new university page | | | The webapp shows a page with sections to add the university info |
| 2 | | Admin adds info about university | | Webapp stores the data in the database | |
| 3 | | Admin clicks save | | University page is saved on the webapp | |
| Alternative Flow #1  Starts at step 2) | # | Actor Action | | System Response | | |
| 3 | Admin chooses to update info about a college in the university | | Webapp shows the college page for the admin to edit its info | | |
| Alternative Flow #2  (Starts at step 3) | # | Actor Action | | System Response | | |
| 3 | Admin chooses to add scholarships | | Webapp shows a page for scholarships so that the admin can edit | | |

Table 3 - University Admin Add/Update Info Use Case Table

# System Design

## Component Diagram

Figure 5 - Component Diagram.

There are four main subsystems in our component diagram:

1. Frontend subsystem

This subsystem represents an intermediate layer between the end user and the rest of the system. It includes two components:

* Chatbot interface

The main page of the chatbot that the user interacts with. This component sends the user message directly to the chatbot backend component.

* Webapp frontend

The user interface for the web application. It allows the user to interact with application features.

1. Backend subsystem

A layer that responsible for the server-side logic and handling requests from the frontend subsystem.

* Chatbot backend  
  It handles validation and preprocessing to the user message before passing it to the chatbot.
* Webapp backend: It handles business logic such as user authentication, data retrieval and receiving feedback.

1. AI models subsystem

It represents the AI components of the system such as the chatbot and recommendation system.

* Chatbot

A conversational agent that handles user inquiries. It is responsible for understanding the user messages and providing helpful responses.

* Recommendation system

It uses user requests and preferences to provide personalized recommendations.

1. Data storage subsystem  
   This subsystem represents data storage and management. It has only one component which is the database. The database stores the universities’ data, students’ data, and students’ feedback. The database follows the schema defined in the ERD of figure 16.

## Sequence Diagrams

### Student Sequence Diagrams

1. A white sheet of paper with black lines

   Description automatically generatedUse chatbot.  
   The student will need to login up first to use the chatbot, click on the chatbot, then the chatbot interface will be displayed, then the student can ask the bot for an college recommendation based on his preferences.

Figure 6 - Student Use Chatbot Sequence Diagram

1. Sign-up  
   student signs up by using his info. An OTP will be sent to their mail. Upon correctly entering it they will be successfully registered

A diagram of a program

Description automatically generated

Figure 7 - Student Sign-Up Sequence Diagram

1. Log-in  
   student logs in by entering his mail and password correctly.

A diagram of a login process

Description automatically generated

Figure 8 - Student Log-In Sequence Diagram

1. Student search for college  
   student clicks on search to see all universities/colleges, Then filter the universities/colleges accordingly.

A diagram of a diagram

Description automatically generated

Figure 9 - Student Search College Sequence Diagram

1. Student compare colleges  
   student selects two colleges to show a shallow comparison between the two of them.

A diagram of a diagram

Description automatically generated

Figure 10 - Student Compare Colleges Sequence Diagram

### Admin Sequence Diagrams

1. Admin review feedback  
   Admin click on review feedback, then all the student feedbacks will be listed to the admin.

A diagram of a webpage

Description automatically generated

Figure 11 - Admin Review Feedback Sequence Diagram

1. Admin add university  
   Admin select on add University, then enter University details, then the University is added to system

A diagram of a webpage

Description automatically generated

Figure 12 - Admin Add University Sequence Diagram

### University Sequence Diagrams

1. University announcement  
   University clicks on colleges, to display all colleges, click on a college then add admin to this college.

A diagram of a diagram

Description automatically generated

Figure 14 - University Announcement Sequence Diagram

1. University adds college admin  
   university chooses college from list of all colleges then adds an admin to this college

A diagram of a company

Description automatically generated

Figure 15 - University Add College Admin Sequence Diagram

## ERD

### Main ERD

1. Important Tables:

University Admin Table: responsible for managing University data.

College Admin Table: responsible for managing College data.

Student Table: utilized for collecting student data for chatbot purposes and authentication.

A diagram of a diagram

Description automatically generated

Figure 16 - Main ERD

### PathFinder Tool ERD

A black and white diamond with black text

Description automatically generated

Figure 17 - PathFinder Tool ERD

# Machine Learning Models

## Conversational AI

Conversational AI is a type of artificial intelligence (AI) that can simulate human conversation. It is made possible by natural language processing (NLP), a field of AI that allows computers to understand and process human language.

### How it works

Conversational AI works by using a combination of natural language processing (NLP), foundation models, and machine learning (ML).

Conversational AI systems are trained on large amounts of data, such as text and speech. This data is used to teach the system how to understand and process human language. The system then uses this knowledge to interact with humans in a natural way. It’s constantly learning from its interactions and improving its response quality over time.

### Role and Benefits

When people think of conversational artificial intelligence, online chatbots and voice assistants frequently come to mind for their customer support services and omni-channel deployment. Most conversational AI apps have extensive analytics built into the backend program, helping ensure human-like conversational experiences.

Conversational AI helped a lot of companies grow as it has many benefits such as:

* Reduce costs and increase productivity
* Increase operational efficiency through automation
* Conversational AI can automate tasks that are currently performed by humans and thereby reduce human errors and cut costs.
* Deliver better customer experience, achieve higher customer engagement and satisfaction.
* Conversational AI can provide a more personalized and engaging experience by remembering customer preferences and helping customers 24/7 when no human agents are around.

### What is Rasa?

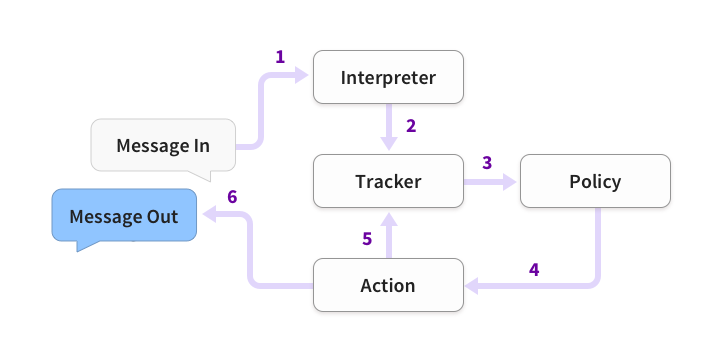
**Rasa** is an open-source framework that helps businesses improve their conversation and interaction with their customers or audience. It is based on natural language understanding, dialogue management and interactions.

Using Rasa, it becomes easier to build conversational AI and improve it over time. The best part? It follows the interactive learning method for improving which makes it easier for you and saves you time, money and resources. It can be used to build great text- as well as voice-based virtual assistants efficiently. In fact, Rasa claims that they provide standard infrastructure for AI chatbots so that all developers and companies can easily understand and develop one of their own virtual assistants.

### Why Rasa?

* Easy to integrate and customize
* Integrate into existing systems
* Supports various intents
* Multiple deployment environment
* Analytics and reports

## Chatbot Model

Our Arabic chatbot is built using Rasa. Rasa is an open-source conversational AI framework designed for building contextually aware conversational AI. Rasa chatbots are split into two main parts, Rasa NLU and Rasa Core.

Rasa NLU interprets the user input and extracts entities and intent with the help of various pipelines. It then converts the input into a dictionary that includes the original text, intent, and entities identified, which is then sent to Rasa CORE. Rasa CORE is responsible for the chatbot’s action. It selects the appropriate action as per the user input and the specified policy then sends it back as a chatbot response.

Figure 18 – Rasa pipeline and policies.

### NLU Training Data

NLU training data consists of example user utterances categorized by intent. Training examples can also include entities. Entities are structured pieces of information that can be extracted from a user's message. You can also add extra information such as regular expressions and lookup tables to your training data to help the model identify intents and entities correctly. Training data is split into two parts, Training examples and Conversation data.

#### Training Examples

* Intent Training Examples: in which each intent has examples that help identify the correct intent.
* Entities: Entities are structured pieces of information that can be extracted from a user's message. They are annotated in training examples with the entity's name.

#### Conversation Data

Stories and rules are both representations of conversations between a user and a chatbot. They are used to train the dialogue management model. Stories are used to train a machine learning model to identify patterns in conversations and generalize to unseen conversation paths. Rules describe small pieces of conversations that should always follow the same path and are used to train the RulePolicy.

### Rasa NLU Pipeline

In Rasa, a pipeline is a set of data processing units arranged in series, where the output of one element serves as the input for the subsequent element. Incoming messages undergo processing through a sequence of components defined in the config.yml file. These components are executed sequentially, comprising the natural language understanding (NLU) pipeline. Choosing an NLU pipeline allows customization and fine-tuning of the model on the dataset. The pipeline consists of five types of components, each serving a specific purpose:

* **Tokenizer**
* **Featurizer**
* **Intent Classifier**
* **Entity Extractor**
* **Response Selector**

#### Tokenizer

Tokenization refers to the process of converting a sequence of text into smaller parts, known as tokens. These tokens can be as small as characters or as long as words. The primary reason this process matters is that it helps machines understand human language by breaking it down into bite-sized pieces, which are easier to analyze.

The tokenizer we use in our Arabic chatbot is the “WhitespaceTokenizer”, which efficiently breaks down Arabic text into individual tokens based on whitespace characters such as spaces, tabs, and line breaks. This tokenizer is particularly useful in Arabic text processing pipelines as it allows for straightforward segmentation of text into meaningful units, which can then be further processed for other tasks.

#### Featurizer

Featurizers in Rasa are components responsible for converting tokens from the tokenizer into numerical features that machine learning models can understand and learn from. These features capture various aspects of the input text.

In our Arabic chatbot, we use multiple featurizers to capture different aspects of the input and increase our chatbot robustness since if one featurizer performs poorly the others may compensate for its performance. Finally, the output features are fused to create a combined feature representation of our tokens. The featurizers are as follows:

* RegexFeaturizer: During training the RegexFeaturizer creates a list of regular expressions defined in the training data format. For each regex, a feature will be set marking whether this expression was found in the user message or not.
* LexicalSyntacticFeaturizer: The LexicalSyntacticFeaturizer is a component meant to create features that are useful when detecting entities. It moves with a sliding window over every token in the user message and creates feature configurations such as, whether the current token is a title, or is the token before/after is a title, and if it is at the end or beginning of a sentence, in general, these features are useful to know to be able to detect entities in the user message such as names, cities, etc
* CountVectorsFeaturizer: The CountVectorsFeaturizer builds a vocabulary based on tokens observed in the training data and each token becomes a feature. The CountVectorsFeaturizer then generates a feature vector based on the presence or absence of each token. In our pipeline, we use two of this featurizer
  + Word level CountVectorsFeaturizer in which word token counts are used as a feature.
  + Char level CountVectorsFeaturizer in which n-chars of token counts are used as a feature.

#### Intent Classifier

Intent classifiers are components responsible for determining the intent behind user messages, based on feature vectors, according to the intents defined in the training data, and then return rankings of intents. In our Arabic Chatbot, we use two intent classifiers DIETClassifier and FallbackClassifier

* DIETClassifier: DIET (Dual Intent and Entity Transformer) is a multi-task architecture for intent classification and entity recognition. It utilizes transformer-based architectures which is a type of neural network commonly used in NLP tasks.
* FallbackClassifier: The FallbackClassifier classifies a user message with the intent nlu\_fallback in case the previous intent classifier wasn't able to classify an intent with confidence greater than or equal to the threshold of the FallbackClassifier (usually 0.5).

#### Entity Extractor

Entity extractors extract entities, such as person names or locations, or any other information from the user message. We use EntitySynonymMapper along with the DIETClassifier from the previous component to extract the entities.

* DIETClassifier: which is also used in entity extraction.
* EntitySynonymMapper: this component maps extracted entities into specified values. For example, it maps both NYC and New York City to NY

#### Response Selector

Response selectors are components responsible for handling retrieval actions, which involve selecting a response from a predefined set of responses based on the predicted intent, its subintents, and the context of the conversation. Retrieval actions are used for tasks like frequently asked questions (FAQs, or any other cases where intent can be split into sub-intends

The response selector used in Rasa is called ResponseSelector. It works by ranking all subintents within the specified intent, which is defined using the retrieval\_intent parameter in the pipeline configuration. The retrieval\_intent parameter allows you to designate certain intents as retrieval intents, indicating that they are associated with a set of predefined responses.

### Rasa Core Policies

In Rasa, policies are responsible for making decisions about what actions the chatbot should take based on the current conversation state and user input. Policies determine the bot's behavior by mapping predicted intents and entities, and taking into consideration the chat history, to appropriate dialogue actions, such as responding with a message, asking for clarification, or executing an external function.

In our chatbot, we use multiple different policies to choose the action and the one with the highest confidence level is chosen, and in case there is a tie it is decided using policy priorities where a higher number means higher priority. The used policies are:

* RulePolicy (priority: 6): The RulePolicy is a policy that handles conversation parts that follow a fixed behavior (e.g. business logic). It makes predictions based on any rules you have in your training data. It also handles the fallback intent in which the user asks for something outside the scope of the chatbot.
* MemoizationPolicy (priority: 3): The MemoizationPolicy remembers the stories from your training data. It checks if the current conversation matches the stories in your stories.yml file. If so, it will predict the next action from the matching stories of your training data with a confidence of 1.0. If no matching conversation is found, the policy predicts None with confidence 0.0.
* UnexpecTEDIntentPolicy (priority: 2): UnexpecTEDIntentPolicy allows your bot to react to unlikely user turns. It has the same model architecture as TEDPolicy. The difference is at a task level. Instead of learning the best action to be triggered next, UnexpecTEDIntentPolicy learns the set of intents that are most likely to be expressed by the user given the conversation context from training stories. It predicts the action action\_unlikely\_intent only if the intent is unlikely to happen in the conversation flow.
* TEDPolicy (priority: 1): TEDPolicy is a machine learning policy that uses several transformer encoders to learn the best action that should be done given the dialog’s history.

## Recommendation System

Figure 19 – Recommendation system types.

A recommendation system is a data filtering tool that recommends the most relevant items to a particular user or a customer using machine learning algorithms. In our case, items are colleges and users are high school students.

There are four main types of recommendation systems or engines: Content-Based filtering, Collaborative filtering, Knowledge-Based filtering and Hybrid recommendation systems. We will care the most about Knowledge-Based recommendation systems.

### Knowledge-Based Recommendation Systems (KB RS):

A knowledge-based recommendation system is a system that make recommendations based on specific queries made by the user, not on a user’s rating history. It prompts the user to give some rules or guidelines on what the results should look like. The user may also give an example of an item. The system searches through its item database and tries to find a similar match. There are two main types of knowledge-based recommendation systems:

#### Constraint-Based:

It uses existing knowledge bases that contain explicit rules about how to relate user requirements with item features. It is very similar to an expert system in knowledge base systems. It is considered a constraint satisfaction problem.

#### Case-Based:

It depends on letting the user choose a target or an anchor item the algorithm finds a similar item to recommend. Results are usually treated as new target cases with some interactive modifications. This is similar to Content-Based systems where items similar to ones the user previously liked are suggested. The main difference is that most knowledge-based systems depend on the description of the items in the form of relational attributes in knowledge bases rather than as text keywords like in Content-Based systems.

#### Similarity Metrics:

Similarity metrics are needed to retrieve examples similar to the specified item. For continuous variables, similarity can be the difference between two numbers. For example, two different colleges have similar fees if the difference in fees is close to zero. It can also be more complex and require the use of statistical measures such as the standard deviation. When it comes to categorial variables, similarity is much more challenging. Domain hierarchies are often used in this case. Domain hierarchies are tree graphs where each category can have a parent category and subcategories. The similarity of two items can then be measured by the length of the path between the items.

#### Critiquing:

Once some result is found using similarity metrics and recommended to the user, the user is able to provide feedback and customizes the results to match what he is looking for. The user specifies a change request on some attributes of the item they like. The change request can be a directional critique or a replacement critique. A directional critique provides feedback on the direction in which the recommendation should be adjusted. For example, if the system recommended a college that is 100 km away, the user may request a shorter distance. On the other hand, a replacement critique suggests alternative items to replace the initially recommended item. For example, if the system recommended a computer engineering college, the user may suggest recommending a computer science college instead.

#### KB RS Strengths:

* Complex Item Domain: KB RS is very proficient when the items have many complex aspects to consider. Those aspects often need expert knowledge in the domain. KB RS effectively captures those complex aspects by utilizing knowledge bases crafted by domain experts. This strength is especially significant when the item is of great importance and requires a lot of thought, such as a university or a college.
* Avoids Cold Start Problem: User data is not needed because what the user wants is explicitly defined. The recommendation process is accurate and can start and work well without requiring existing rating data.

#### KB RS Weaknesses:

* Knowledge Acquisition Bottleneck: A big problem lies in the creation of the knowledge base. It requires the conversion of the knowledge possessed by domain experts into consumable representations.

### Hybrid Recommendation Systems:

A special type of system that combines multiple recommendation techniques such as collaborative, content-based filtering, and other approaches and uses them simultaneously. This provides more accurate and diverse recommendations to users as it utilizes the strengths and overcomes the weaknesses of different approaches. It can be used to overcome major problems such lack of data. Some hybridization techniques include:

* Cascade
* Feature Combination
* Feature Augmentation
* Meta-Level
* Weighted
* Switching
* Mixed

# Work Plan

## Task Table

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Task Title | Description | Task Statues |
| 1 | Determine project idea | Decide what the main idea for the project is and see if there are any problems with it. | Completed |
| 2 | Searching and studying related work | Search for similar projects and see how they were completed. | Completed |
| 3 | Determine project features | Decide what features our project will have. | Completed |
| 4 | Study needed technologies | Learn needed technologies to be able to implement our project. | In Progress |
| 5 | Web scraping | Collect data using web scraping. | In Progress |
| 6 | Pathfinder tool development and deployment | Develop and deploy the Pathfinder Tool to help with data collection. | Completed |
| 7 | Crowdsource data | Send the Pathfinder tool to other people to crowdsource data. | In Progress |
| 8 | Clean collected data | Clean the collected data to ensure its quality. | In Progress |
| 9 | Develop Prototype | Develop a chatbot prototype along with a simple interface | Completed |
| 10 | Write Documentation | Design UML Diagrams, and finish mid-year documentation to show progress | Completed |
| 11 | Develop Backend | Develop web application’s backend | Planned |
| 12 | Develop Frontend | Develop web application’s frontend. | Planned |
| 13 | Chatbot development | Develop the Arabic chatbot and ensure its accuracy in predicting user intents. | Planned |
| 14 | Build recommendation system | Build the recommendation machine learning model. | Planned |
| 15 | Testing and Evaluation | Test the chatbot, recommendation system, and the web application as a whole | Planned |

Table 4 - Task Table

## Gantt Chart

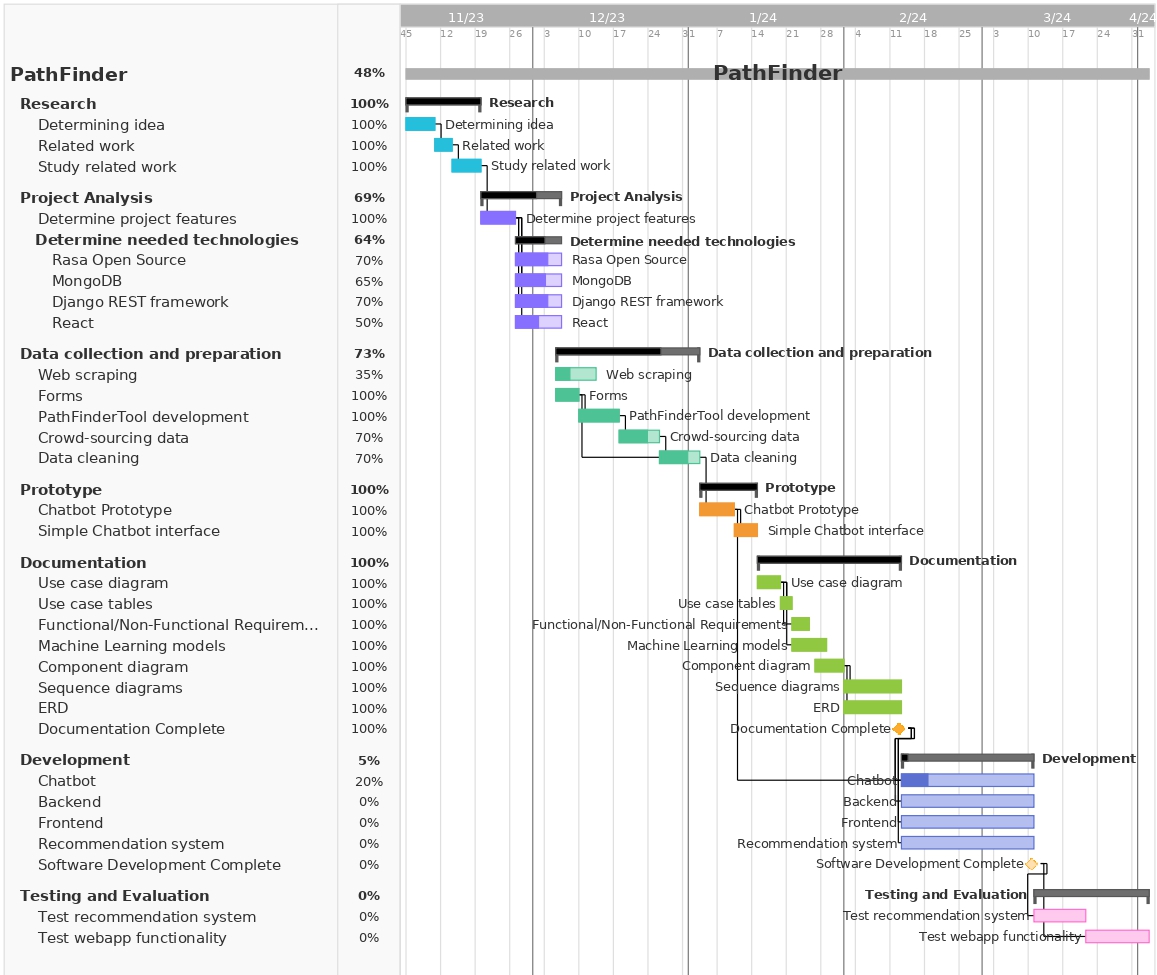


Figure 20 - Gantt Chart

# References

Paul0. (n.d.). Knowledge-Based RS. Medium. <https://medium.com/@paul0/knowledge-based-rs-b4547b84108a>

jwu2. (n.d.). Knowledge-Based Recommender Systems: An Overview. Medium. <https://medium.com/@jwu2/knowledge-based-recommender-systems-an-overview-536b63721dba>

Rasa. (2024). Rasa Open Source Documentation. <https://rasa.com/docs/rasa/>

Rasa. (n.d.). Conversational AI with Rasa, Pipelines and Policies. <https://learning.rasa.com/conversational-ai-with-rasa/pipeline/>

Knoldus. (n.d.). Understanding the Rasa NLU Pipeline. Knoldus Blog. <https://blog.knoldus.com/understanding-the-rasa-nlu-pipeline/>