Informatics Institute of Technology
in Collaboration with

University of Westminster, UK

Semantic Bot: A Semantic Question Answering System for Raw Documents

Key Words: natural language understanding, text similarity, question answering system

Project Proposal by

Mr. Adrian Anchelo

W1790020 | 20191176

Supervised by

Mr. Deshan Sumanathilaka

November 2022

This Project Proposal is submitted in partial fulfilment of the
requirements for the BSc (Hons) Computer Science degree
at the University of Westminster

Table of Contents

List of Figures

List of Tables

1. Introduction

In this research project, the author tries to identify an architecture to search within a non-pre-encoded document for user queries. The proposed novel architecture will try to make the user search a raw document semantically. This research considers a non-pre-encoded or non-pre-processed document as a raw document.

This document states the problem, research gap, challenge, and strategy the author wishes to pursue over the next few months. Furthermore, the justification of the research gap, problem, and existing work, are also discussed. Finally, in the Work Plan, the project's deliverables are presented together with their anticipated schedule.

2. Problem Domain

2. 1. Text Similarity

Natural language processing (NLP) has drawn much attention as social networks and news stories produce abundant data. Even within NLP, it is crucial to gauge how similar sentences are to one another (Yoo et al., 2021). Specifically, text similarity is considered in information retrieval, automatic question answering, machine translation, dialogue systems, and document matching (Wang and Dong, 2020).

The primary purpose of text similarity is to investigate and calculate how two text entities are close or outlying. Text entities can be simple tokens or terms, like words, or whole documents, which may include sentences or paragraphs of text (Sarkar, 2016). There are numerous methods of analyzing text similarity. However, all the methods come under two areas, lexical similarity and semantic similarity.

2. 1. 1. Lexical similarity

A measure of the similarity between two texts is provided by lexical similarity based on the intersection of word sets from the same or different languages. A score of 0 indicates no shared words between the two texts, while a score of 1 indicates that the vocabularies overlap entirely (Majumdar, 2022).

2. 1. 2. Semantic similarity

The primary goal of semantic similarity is to measure the semantic meanings between two words, phrases, sentences, or documents. Knowledge-based, statistical-based and string-based distributional methods are the basic ways to measure semantic similarity (Semantic Similarity of Two Phrases | Baeldung on Computer Science, 2021).

2. 2. Question Answering System

People seek answers to questions. Information-seeking questions from people have led researchers to study Question answering systems frequently. Therefore, this need for information has led the research community to look further for answers beyond paragraphs, articles and documents (Asai et al., 2021).

Question answering (QA) systems provide answers in response to questions in natural language. QA falls within information extraction, natural language processing, and information retrieval (Antoniou and Bassiliades, 2022). Recently, QA systems have been developed as chatbots to increase user experience.

2. 2. 1. QA bot

A chatbot is a computer program that understands customer questions and automates responses, simulating human conversation. Chatbots are referred to as bots as well. QA bot is a combination of chatbot features and a QA system that provides users with answers in a friendly manner.

3. Problem Definition

Currently, there is no proper architecture developed to search semantically within a raw document. The only available search mechanism is to search for the exact word throughout the document. This search method is inefficient for a user who wants to find a particular information, because the user will be directed to various places within the document or may not find the relevant information. Therefore, the user will invest a considerable time and power in searching for the relevant information within the document.

It is ideal to use user queries to find information. Therefore, allowing user queries to search within a taw document will be beneficial. Moreover, semantic search for those queries will be productive and question answering system approach will improve the user experience.

3. 1. Problem Statement

It is difficult to retrieve information relevant to the user’s intention within a document that is not encoded nor feature processed.

4. Research Motivation

In 2018 google released a system called "Talk to Books.". In that system, users can ask questions from books, and then the system will give an output of texts from books which answers user's question. This system primarily works based on sentence embeddings. It uses Universal Sentence Encoder to encode sentences (Cer, Yang, Kong, Hua, Limtiaco, John, et al., 2018). However, the sentences from the books are pre-encoded.

The motivation of this research is to find out if it is possible to ask questions from a user's book and get the answers. For example, a school student will be studying science. That student needs an answer to "What are the types of rocks?". Even though the student has a science book, the student will not be able to find the answers quickly. Manually, the student has to search for the index, then find the relevant page and get the answer.

Therefore, this project is expected to add value to the NLP domain and users who primarily use documents to learn or work. The project idea started from the page “For Developers - Semantic Experiences”.

5. Related Work

Table 1: Related Works

Citation	Brief	Contribution	Limitations
Lee et al., 2020	Develop a QA system that provides information to users about covid. This system primarily uses BEST for entity recognition, the DenSpi model, and the SPARC model for QA.	Effective in getting answers from unstructured data. A logical approach with less latency QA model.	Although the threshold for the scores of answers, the distribution of scores was very inconsistent across different questions
Yoo et al., 2021	This study analyzes the similarity of Korean phrases by combining deep	Very accurate result since two ways of similarities, word level and sentence level	Performance can be further improved by considering the

	learning methodologies with a method that considers lexical associations.	similarities calculated and the average of those are obtained to make decisions.	information in the order/ part of speech
Liu, 2022	A topic modeling semantic search approach to find relevant court judgments.	This paper has used topic modeling to classify and retrieve documents semantically rather than a keyword match. Abstractive and extractive summarizations are included to justify the displayed document's relevance to the user.	No specified limitations
Arif, Latif and Latif, 2021	Answer type classification is a key step in automating question-answer systems. Maintaining high accuracy for answer type classification by identifying the correct answer types can result in precise answers for a question. Here the questions are classified into predefined categories according to the	A novel approach, Question Sentence Embedding (QSE), was introduced. This approach uses Universal Sentence Encoder-based Transformer architecture for feature extraction.	The system is tested only with a specific dataset which is the COVID-Q dataset.

	domain, which narrows down the answer selection process.		
Chen, Cheng and Heh, 2021	Chatbot programs process and stimulate human-to-human dialogue using natural language. Chatbots can be adjusted according to the learning speed of a user. Education systems make use of chatbots to create interactive learning experiences for students.	Question Answering System (QAS) using Line Bot data retrieval. The system consists of three modules sentiment analysis, data retrieval, and answer retrieval.	The research has taken place only considering the educational domain and other domains can be tested using the system.

Many research suggests QA systems that only use knowledge bases as the primary data source (Antoniou and Bassiliades, 2022). Retrieving information from already existing data will be inefficient as time passes. Mechanisms that can extract raw data from the internet will increase the recency of the system, which will be more reliable as well.

Research that used text similarity for QA systems has more room for improvement. Google has released many papers on natural language understanding, improving the semantic experiences in NLP. They have developed encoders to encode sentences (Cer, Yang, Kong, Hua, Limtiaco, St. John, et al., 2018), which are getting more popular gradually. Above mentioned work by Arif Latif and Latif uses a sentence encoder but does not incorporate word similarity verification for questions.

Research done by Yoo et al. in 2021 clearly shows the performance improvement when word representation models are combined with universal sentence encoder. Research done by Yoo et al. in 2021 clearly shows performance improvement when word representation models are combined with universal sentence encoder. Therefore, improving the word similarity and combining it with sentence embeddings will be productive.

6. Research Gap

There has been relatively little research in semantic question answering Systems when compared to question answering systems (Antoniou and Bassiliades, 2022). No attempt has been made to search for answers semantically within a simple document. If such a system is developed, the answerability of questions for a large amount of unstructured text should also be considered (Lee et al., 2020). This project proposes a bot that aids users in searching for information within a document provided by themselves.

This project focuses on empirical and theoretical gaps in the question answering domain and Performance gaps in Text Similarity. Having pre-encoded knowledge will not be sufficient in the future. Data are collected rapidly in large volumes. Therefore, combining different models to form a proper architecture will be beneficial when retrieving data without prior preparation.

7. Contribution to Body of Knowledge

This project will be contributing a novel architecture combining a variety of NLP architectures to search semantically within a document. A word space will be created respective to the application scope. And encoders will be fine-tuned to assign vectors relevant to the domain.

This project mainly focuses in text similarity and QA systems in achieving the gap. QA systems will be the problem domain as it is used as a technology to help users get information. Text similarity will be this project's research domain, where it acts as a semantic calculator. It is hypothesized that this novel approach combining QA and text similarity domain models will fill the research gap.

7. 1. Contribution to Problem Domain

Finding information for users is the primary aim of this project. This research focuses on developing a QA bot to make a user-friendly interface for the user to get information. A couple of QA techniques will be used to shortlist the sentence candidates. QA techniques will reduce the search latency, have high relevancy, and improves the user experience. This work is expected be useful in information extraction domain as well.

7. 2. Contribution to Research Domain

Text similarity is used to investigate the relevance between user queries and the information. Therefore, the accuracy of calculating the similarity will be increased. Which will also, give accurate output to users. Recent research in text similarity has been significantly excellent. On top of those, this project will contribute to natural language understanding.

8. Research Challenge

Retrieving information from raw documents has always been a challenge. Data is processed to become meaningful information. Therefore, it is essential to identify suitable ways to pre-process the document in the proposed question-answering system.

Parsing a document into usable inputs is also not easy because a document may have footers, headers, and tables with sentences, which may or may not be necessary to the user's queries. Identifying which data to engineer and give to the machine learning model will be tricky. Furthermore, after parsing the document and getting a list of sentences, it is still not ideal to encode all the sentences because a document may have more than 1000 sentences. Encoding everything will take hours and obtaining an output will take even more time. Latency is a significant aspect of question-answering systems. Therefore, tackling the difficulty of encoding sentences is important.

Processing the user queries is important as it will help shortlist the parsed document sentence list. Since the proposed QA system works semantically, user queries must be processed in a way that could shortlist the sentences logically. When shortlisting, the intention of the user's queries should be considered. Therefore, having a word space will come in handy when analyzing the query.

9. Research Questions

RQ1: How can a system help the user search semantically within a non-pre-encoded document?

RQ2: How to calculate the similarity between sentences? (What encoders can be used? What is the best method /formula to calculate the similarity between two vectors?)

RQ3: What are the important aspects to consider when developing a QA system?

10. Research Aim

The aim of this research is to design, develop and evaluate a novel architecture that will find the relevant information to the user from non-pre-encoded document by considering the context of user's queries.

This research project will produce a question-answer system that can retrieve answers/ information for users based on user queries within a raw document produced by the user. The system will semantically search for the information when retrieving the answers. Moreover, the question-answering system should be better in terms of user experience.

Several NLP techniques/will be used to achieve precise information retrieval. The questions-answer system will act more like a chatbot to improve the user experience. Therefore, it will be interactive and have a short-term memory to store the history of retrieved information/ answers. The threshold score will be high to obtain accurate answers, and a couple of techniques will be used when calculating the similarities.

The knowledge to achieve the aim will be studied, researched, and documented. The proposed system will be developed, and the performance will be evaluated to justify the hypothesis. The developed system will be hosted for public use, and this research's source code and documents will be publicly available for further research. A research paper about the proposed system and the outcomes of the findings will be published.

11. Research Objectives

Table 2 : Research Objectives

Objective	Description	Learning Outcomes	Research Questions
Literature Survey	Investigate previous work to collate relevant data on related work and critically evaluate them. **RO1:** Conduct a preliminary study on existing QA systems & Architectures. **RO2:** Analyze the document parsing methods that can be used.	LO4, LO2, LO5	RQ1, RQ2, RQ3

	RO3: Conduct a preliminary study on Text similarity. **RO4:** Analyze various models and conclusion methods that is considered when calculating the text similarities.		
Requirement Elicitation	Documenting the requirements of the project using appropriate techniques and tools to meet the expected research gaps & challenges to be addressed based on previous related research and any domain-specific sources of knowledge. **RO1:** Gather information about requirements of basic QA understand end-user expectations. **RO2:** Gather the requirements to parse document. **RO3:** Gather insights & opinions from technology & domain experts to build a suitable system.	LO1, LO2, LO5, LO7	RQ1, RQ2, RQ3
Design	Designing architecture and a system that can solve the identified problems with recommended techniques. **RO2:** Design an architecture find the relevant information from documents. **RO2:** Design a QA Bot with low latency and high relevancy. **RO3:** Design a data-preprocessing pipeline to parse the document and feature engineer the user query for further processing.	LO1, LO3	RQ1, RQ2, RQ3

	RO4: Design a system that finds the similar sentences for user query from pre-processed data.		
Development	Implementing a system that addresses the gaps in this project that were aimed to be solved. **RO1:** Develop a QA Bot that can find relevant text for user's information seeking questions. **RO2:** Include an algorithm with preprocessing models that improves the speed and accuracy of the system. **RO3:** Develop a model that will classify the question and find the exact intention of the question. **RO4:** Documents uploaded by the user should not be stored in the database. If it is stored then proper encryption techniques should be used to secure the user's activities.	LO1, LO5, LO6	RQ1, RQ2, RQ3
Testing and Evaluation	Testing the developed system & deep learning model with appropriate data and evaluating them with baseline techniques. **RO1:** Create a test plan and perform unit, integration and functional testing for the developed system. **RO2:** Evaluate the performance by f1 score, compared against baseline models.	LO4	RQ1, RQ2, RQ3

12. Project Scope

The project's objectives, an evaluation of comparable items, and the time allotted for this research project are used to establish the scope, which is as follows.

12. 1. In-scope

The following is a list of the project’s scope:

- A question answering system that takes a document from the user and retrieve answers user’s queries from the provided document.
- Users can only upload documents that are in pdf format.
- Users can search information with user queries that can be a question, unstructured text, or structured text.
- If the system identifies couple of answers, then only five answers that the system identifies will be displayed to the user.
- Creation of Question Answering system that will work as a chatbot.
- Graphical user interface that allows users to upload the document, ask question and view the output.
- Chat between the user and the system will store so that the users can view the history until the session gets over.
- Users can only interact in English.

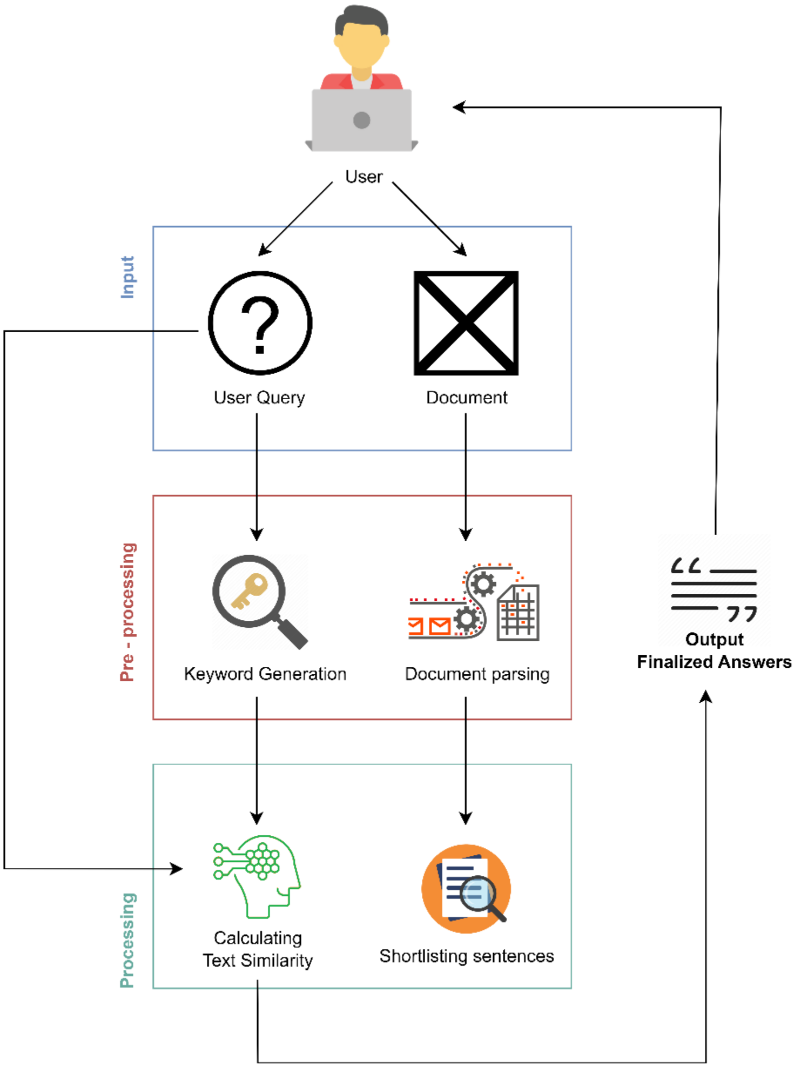
12. 2. Out-scope

The following are the parts that will not be covered by the project

- Allowing to upload documents other than pdfs.
- Displaying more than five answers.
- Retrieving answers from external source if the system could not find the answers within the provided document.
- History of chats with previously uploaded documents will be available.

12. 3. Prototype Diagram

Figure 1: Prototype diagram



13. Methodology

13. 1. Research Methodology

The quality of every project is determined by three critical factors: cost, time, and scope, all of which must be appropriately handled during the project's duration. Consequently, research methodologies are necessary.

The Saunders Research Onion Model (Saunders, Lewis, and Thornhill, 2003) was used to determine the methodologies. The acceptable methodologies for the project are indicated in the table below.

Table 3 : Research Methodology

Research philosophy	The philosophy of research affects data collection and processing because it relates to the investigated reality. Positivism, interpretivism & pragmatism are possible philosophical approaches for this research. For this research, **positivism** was chosen since the research outcomes can be **evaluated** using metrics.
Research approach	Deductive and Inductive approaches are two approaches that researchers can use to conduct research. Since this project is expected to be evaluated to prove the **hypothesis**, the **deductive** approach has been chosen.
Research strategy	The strategy focuses on the techniques for gathering data that will be used to address the research issues. The methods to meet the research plan are **interviews, documents, research analysis, experiments,** and **surveys.**
Research choice	The choice of approach reveals whether the research is concerned with qualitative or quantitative characteristics. Although quantitative results are the key focus, it has been determined that the qualitative nature of the data will significantly impact the quantitative outcomes of the research. Therefore, **multi-method** was chosen.
Time horizon	Data collection will be collected at one point in the evaluation phase of the research. Therefore, the **cross-sectional** method was selected.
Procedure	Techniques for data collecting and analysis are discussed here. Surveys, annotated, question answer data will be used.

13. 2. Development Methodology

13. 2. 1. Life cycle model

Agile software development life cycle was chosen as the research development method since the project development time is limited. This approach will iteratively improve the system while brainstorming the requirements based on time and importance.

13. 2. 2. Design Methodology

The author selected Object-Oriented Analysis and Design as the Design Approach to enable an incremental methodology that can be used to grow the system with the ability to reuse system components.

13. 2. 3. Evaluation Methodology

Accuracy will be the metric that will be used to evaluate the model as proof of concept. The accuracy equation is given by,

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ predictions}$$

However, the accuracy metric will not be good when evaluating the proposed model. Because the dataset is obtained from public sources and not self-imposed. The data can be imbalanced. Therefore, to evaluate the proposed model with other baseline models, the F1 score will be used. Precision and recall are the foundation metrics of F1 score. The equation of precision and recall are given below,

- Precision is a measure of true positives (the text accurately matched by the model) predicted by the model divided by the total number of true positives and true negatives predicted by the model.

$$Precision = \frac{number\ of\ true\ positive}{number\ of\ true\ positive + number\ of\ false\ positives}$$

- The recall is the ratio of text that belongs to the positive category and correctly matched texts by the model by the total number of sentences that belongs to the positive categories.

$$Recall = \frac{number\ of\ true\ positive}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

The F1 score is defined as the harmonic mean of precision and recall. The F1 score formulae is as follows,

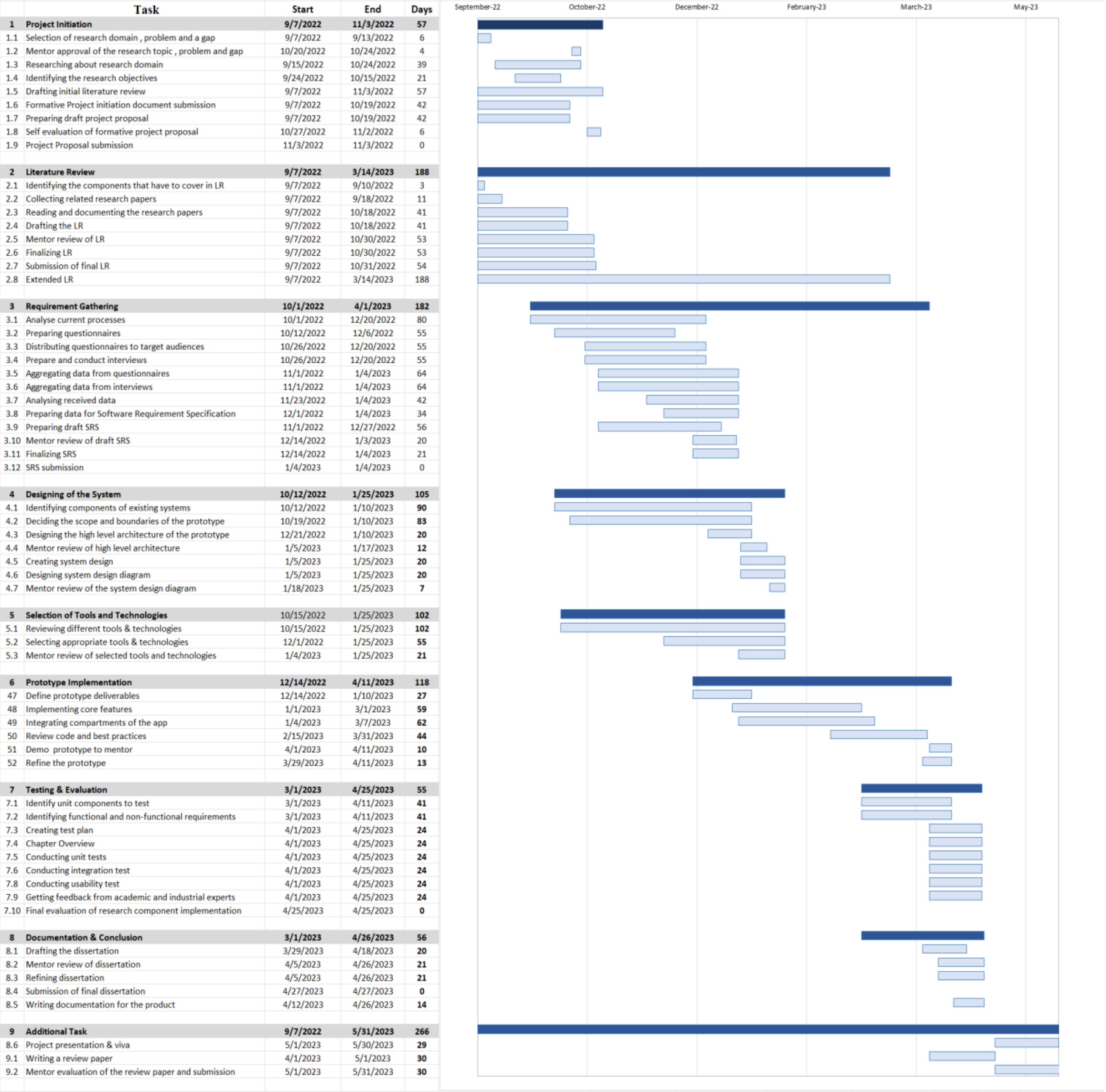
$$F1\ score = 2 \times \frac{precision \ \times recall}{precision + recall}$$

13. 3. Project Management Methodology

The **Prince2 Agile** project management methodology was selected. It enables the author work flexibly. Also, it enables to focus on short-term goals and progress which makes it more suitable methodology for this project.

13. 3. 1. Gannt Chart

Figure 2 : Gantt chart



	Task	Start	End	Days
1	**Project Initiation**	**9/7/2022**	**11/3/2022**	**57**
1.1	Selection of research domain , problem and a gap	9/7/2022	9/13/2022	6
1.2	Mentor approval of the research topic , problem and gap	10/20/2022	10/24/2022	4
1.3	Researching about research domain	9/15/2022	10/24/2022	39
1.4	Identifying the research objectives	9/24/2022	10/15/2022	21
1.5	Drafting initial literature review	9/7/2022	11/3/2022	57
1.6	Formative Project initiation document submission	9/7/2022	10/19/2022	42
1.7	Preparing draft project proposal	9/7/2022	10/19/2022	42
1.8	Self evaluation of formative project proposal	10/27/2022	11/2/2022	6
1.9	Project Proposal submission	11/3/2022	11/3/2022	0
2	**Literature Review**	**9/7/2022**	**3/14/2023**	**188**
2.1	Identifying the components that have to cover in LR	9/7/2022	9/10/2022	3
2.2	Collecting related research papers	9/7/2022	9/18/2022	11
2.3	Reading and documenting the research papers	9/7/2022	10/18/2022	41
2.4	Drafting the LR	9/7/2022	10/18/2022	41
2.5	Mentor review of LR	9/7/2022	10/30/2022	53
2.6	Finalizing LR	9/7/2022	10/30/2022	53
2.7	Submission of final LR	9/7/2022	10/31/2022	54
2.8	Extended LR	9/7/2022	3/14/2023	188
3	**Requirement Gathering**	**10/1/2022**	**4/1/2023**	**182**
3.1	Analyse current processes	10/1/2022	12/20/2022	80
3.2	Preparing questionnaires	10/12/2022	12/6/2022	55
3.3	Distributing questionnaires to target audiences	10/26/2022	12/20/2022	55
3.4	Prepare and conduct interviews	10/26/2022	12/20/2022	55
3.5	Aggregating data from questionnaires	11/1/2022	1/4/2023	64
3.6	Aggregating data from interviews	11/1/2022	1/4/2023	64
3.7	Analysing received data	11/23/2022	1/4/2023	42
3.8	Preparing data for Software Requirement Specification	12/1/2022	1/4/2023	34
3.9	Preparing draft SRS	11/1/2022	12/27/2022	56
3.10	Mentor review of draft SRS	12/14/2022	1/3/2023	20
3.11	Finalizing SRS	12/14/2022	1/4/2023	21
3.12	SRS submission	1/4/2023	1/4/2023	0
4	**Designing of the System**	**10/12/2022**	**1/25/2023**	**105**
4.1	Identifying components of existing systems	10/12/2022	1/10/2023	**90**
4.2	Deciding the scope and boundaries of the prototype	10/19/2022	1/10/2023	**83**
4.3	Designing the high level architecture of the prototype	12/21/2022	1/10/2023	**20**
4.4	Mentor review of high level architecture	1/5/2023	1/17/2023	**12**
4.5	Creating system design	1/5/2023	1/25/2023	**20**
4.6	Designing system design diagram	1/5/2023	1/25/2023	**20**
4.7	Mentor review of the system design diagram	1/18/2023	1/25/2023	**7**
5	**Selection of Tools and Technologies**	10/15/2022	1/25/2023	**102**
5.1	Reviewing different tools & technologies	10/15/2022	1/25/2023	**102**
5.2	Selecting appropriate tools & technologies	12/1/2022	1/25/2023	**55**
5.3	Mentor review of selected tools and technologies	1/4/2023	1/25/2023	**21**
6	**Prototype Implementation**	**12/14/2022**	**4/11/2023**	**118**
47	Define prototype deliverables	12/14/2022	1/10/2023	**27**
48	Implementing core features	1/1/2023	3/1/2023	**59**
49	Integrating compartments of the app	1/4/2023	3/7/2023	**62**
50	Review code and best practices	2/15/2023	3/31/2023	**44**
51	Demo prototype to mentor	4/1/2023	4/11/2023	**10**
52	Refine the prototype	3/29/2023	4/11/2023	**13**
7	**Testing & Evaluation**	**3/1/2023**	**4/25/2023**	**55**
7.1	Identify unit components to test	3/1/2023	4/11/2023	**41**
7.2	Identifying functional and non-functional requirements	3/1/2023	4/11/2023	**41**
7.3	Creating test plan	4/1/2023	4/25/2023	**24**
7.4	Chapter Overview	4/1/2023	4/25/2023	**24**
7.5	Conducting unit tests	4/1/2023	4/25/2023	**24**
7.6	Conducting integration test	4/1/2023	4/25/2023	**24**
7.8	Conducting usability test	4/1/2023	4/25/2023	**24**
7.9	Getting feedback from academic and industrial experts	4/1/2023	4/25/2023	**24**
7.10	Final evaluation of research component implementation	4/25/2023	4/25/2023	**0**
8	**Documentation & Conclusion**	**3/1/2023**	**4/26/2023**	**56**
8.1	Drafting the dissertation	3/29/2023	4/18/2023	**20**
8.2	Mentor review of dissertation	4/5/2023	4/26/2023	**21**
8.3	Refining dissertation	4/5/2023	4/26/2023	**21**
8.4	Submission of final dissertation	4/27/2023	4/27/2023	**0**
8.5	Writing documentation for the product	4/12/2023	4/26/2023	**14**
9	**Additional Task**	**9/7/2022**	**5/31/2023**	**266**
8.6	Project presentation & viva	5/1/2023	5/30/2023	**29**
9.1	Writing a review paper	4/1/2023	5/1/2023	**30**
9.2	Mentor evaluation of the review paper and submission	5/1/2023	5/31/2023	**30**

13. 3. 2. Deliverables

Table 4 : Deliverables

Deliverable	Date
Project Proposal Document	9th of November 2022
Literature Review Document	12th of December 2022
Software Requirement Specification	24th of November 2022
System Design Document	17th of December 2022
Project Specifications Design and Prototype	2nd of February 2023
Prototype	27th of April 2023
Thesis	27th of April 2023
Final Research Paper	1st of May 2023

13. 3. 3. Resource Requirements

The necessary project resources are identified based on the project's objectives, expected solutions, and deliverables. The following are the necessary software, hardware, and data resources.

Software Requirements

- **Operating System (Windows) -** Windows will be the default option for development since it can handle all the necessary tools for the project.
- **Python** - The programming language to develop Machine Learning and Deep Learning models. Python is an all-purpose language in many applications that integrate with data science.
- **TensorFlow/ Scikit learn Python packages** - Libraries used to facilitate model construction, training, and testing.
- **NodeJS** - The API communicates between the ML back and front. Node.js will allow the program to support concurrency. Moreover, connecting the ML backend will be easy since Node.js supports TensorFlow.
- **JavaScript (React)** - The application's frontend, where the output will be displayed. The aspect of the project since it will be the point of interaction between users and the system.

- **PyCharm/ VSCode-** Development environments that facilitate the project's development.
- **Google Colab** - Cloud development environment to build, train & test ML & Deep Learning models.
- **Zotero** - A research management application for storing and backing up research items and managing references.
- **MS Office/ Google Docs/ Canva/ Figma** - Tools to create reports, figures & documentation.
- **Google Drive/ One drive/ GitHub** - To back up project-related files and code.

Hardware Requirements

- **Core i7x processor of the tenth generation or higher** - To perform high resource-intensive tasks.
- **2GB VGA or above** - To manage training processes of data science models.
- **16GB RAM or above** - To manage data sets & development environments.
- **Disk space of 40GB or above** - To store data & application code.

Data Requirements

• **QA data** – SQUAD data set.

Skill Requirements

• Ability to create question answering bot.

• Ability to create optimized Machine Learning & Deep Learning models to encode sentences to calculate the similarity.

• Research writing skills.

14. Risk Management

The project was identified with the following risks, along with suggested mitigation strategies.

Table 5 : Risk mitigations

Risk Item	Risk Level	Frequency	Mitigation Plan
Lack of knowledge in the domains of research	5	4	Read research papers, self-learning through books, Online learnings

Corruption of documentation	5	5	Save the document in cloud and enable the auto save
Work overlord	4	5	Have a proper task break down list
Limited hardware resources	5	5	Use cloud services
Inability to explain the research	4	2	Prepare a document with explanations and a recording of demonstration

References

Antoniou, C. and Bassiliades, N. (2022). A survey on semantic question answering systems. *The Knowledge Engineering Review*, 37, e2. Available from https://doi.org/10.1017/S0269888921000138.

Arif, N., Latif, S. and Latif, R. (2021). Question Classification Using Universal Sentence Encoder and Deep Contextualized Transformer. *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*. 7 December 2021. Sharjah, United Arab Emirates: IEEE, 206–211. Available from https://doi.org/10.1109/DeSE54285.2021.9719473 [Accessed 17 October 2022].

Asai, A. et al. (2021). XOR QA: Cross-lingual Open-Retrieval Question Answering. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021. Online: Association for Computational Linguistics, 547–564. Available from https://doi.org/10.18653/v1/2021.naacl-main.46 [Accessed 30 October 2022].

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., et al. (2018). Universal Sentence Encoder. Available from http://arxiv.org/abs/1803.11175 [Accessed 8 November 2022].

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., et al. (2018). Universal Sentence Encoder for English. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018. Brussels, Belgium: Association for Computational Linguistics, 169–174. Available from https://doi.org/10.18653/v1/D18-2029 [Accessed 17 October 2022].

Chen, L.E., Cheng, S.Y. and Heh, J.-S. (2021). Chatbot : A Question Answering System for Student. *2021 International Conference on Advanced Learning Technologies (ICALT)*. July 2021. Tartu, Estonia: IEEE, 345–346. Available from https://doi.org/10.1109/ICALT52272.2021.00110 [Accessed 8 November 2022].

For Developers - Semantic Experiences. (no date). *Google AI*. Available from https://research.google.com/semanticexperiences/for-developers.html [Accessed 6 November 2022].

Lee, J. et al. (2020). Answering Questions on COVID-19 in Real-Time. Available from http://arxiv.org/abs/2006.15830 [Accessed 30 October 2022].

Liu, Y. (2022). Court Judgement Labeling Using Topic Modeling and Syntactic Parsing. Available from http://arxiv.org/abs/2208.04225 [Accessed 30 October 2022].

Majumdar, D. (2022). Introduction to Lexical Similarity. Available from https://ladal.edu.au/lexsim.html [Accessed 31 October 2022].

Sarkar, D. (2016). *Text Analytics with Python*. Berkeley, CA: Apress. Available from https://doi.org/10.1007/978-1-4842-2388-8 [Accessed 31 October 2022].

Semantic Similarity of Two Phrases | Baeldung on Computer Science. (2021). Available from https://www.baeldung.com/cs/semantic-similarity-of-two-phrases [Accessed 31 October 2022].

Wang, J. and Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information*, 11 (9), 421. Available from https://doi.org/10.3390/info11090421.

Yoo, Y. et al. (2021). A Novel Hybrid Methodology of Measuring Sentence Similarity. *Symmetry*, 13 (8), 1442. Available from https://doi.org/10.3390/sym13081442.