**TEXT SUMMARIZER USING DEEP LEARNING**

PROJECT REPORT

**Project Lab (IAI-851)**

**Bachelor of Technology CSE-AI (I-Nurture)**

|  |  |
| --- | --- |
| PROJECT GUIDE:  **Mr. Sudhanshu Kumar** | SUBMITTED BY:  **Anshika Gupta (TCA1960003)**  **Shivansh Sharma (TCA1960001)** |

FEBRUARY ,2023



**FACULTY OF ENGINEERING & COMPUTING SCIENCES**

**TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD**

**DECLARATION**

We hereby declare that this Project Report titled **TEXT SUMMARIZER USING DEEP LEARNING** submitted by us and approved by our project guide, Faculty of Engineering & Computing Sciences. Teerthanker Mahaveer University, Moradabad, is a bonafide work undertaken by us and it is not submitted to any other University or Institution for the award of any degree diploma / certificate or published any time before.

|  |  |  |
| --- | --- | --- |
| **Project ID :** | **IAI-851** | |
| **Student Name:** | **Shivansh Sharma** | Signature |
| **Student Name:** | **Anshika Gupta** | Signature |
|  |  |  |
| **Project Guide :** | **Mr. Sudhanshu Kumar** | Signature |

Table of Contents

[1 Project Title 4](#_Toc31392321)

[2 Problem Statement 4](#_Toc31392322)

[3 Project Description 5](#_Toc31392323)

[3.1 Scope of the Work 6](#_Toc31392324)

[3.2 Project Modules 7](#_Toc31392325)

[3.3 Context Diagram (High Level) 7](#_Toc31392326)

[4 Implementation Methodology 10](#_Toc31392327)

[5 Technologies to be used 11](#_Toc31392328)

[5.1 Software Platform 11](#_Toc31392329)

[5.2 Hardware Platform 11](#_Toc31392330)

[5.3 Tools, if any 11](#_Toc31392331)

[6 Advantages of this Project 12](#_Toc31392332)

[7 Future Scope and further enhancement of the Project 13](#_Toc31392334)

[8 Project Repository Location 14](#_Toc31392335)

[9 Conclusion 15](#_Toc31392337)

[10 References 16](#_Toc31392338)

**Appendix**

**A: Data Flow Diagram (DFD)**

**B: Entity Relationship Diagram (ERD)**

**C: Use Case Diagram (UCD)**

**D: Screen Shots**

# Project Title

**TEXT SUMMARIZER USING DEEP LEARNING**

# Problem Statement

### In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it.In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.

### Customer reviews can often be long and descriptive. Analyzing these reviews manually, as you can imagine, is really time-consuming. This is where the brilliance of Natural Language Processing can be applied to generate a summary for long reviews. We will be working on a really cool dataset. Our objective here is to generate a summary for the Amazon Fine Food reviews using the abstraction-based approach we learned about above. You can download the dataset from Kaggle.

To create a text summarizer which summarizes the text or the content of the paragraph in minimum words without changing its meaning. This system is made using NLP and deep learning based model which is branch of machine learning. This text summarizer also summarizes text from the weblinks and also summarizes text from PDF document.

# Project Description

In the modern Internet age, textual data is ever increasing. Need some way to condense this data while preserving the information and meaning. We need to summarize textual data for that. Text summarization is the process of automatically generating natural language summaries from an input document while retaining the important points. It would help in easy and fast retrieval of information.

There are two prominent types of summarization algorithms.

• Extractive summarization systems form summaries by copying parts of the source text through some measure of importance and then combine those part/sentences together to render a summary. Importance of sentence is based on linguistic and statistical features.

• Abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding in short possibly using new words and phrases. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating real world knowledge. Majority of the work has traditionally focused on extractive approaches due to the easy of defining hard-coded rules to select important sentences than generate new ones. Also, it promises grammatically correct and coherent summary. But they often don’t summarize long and complex texts well as they are very restrictive.

## Scope of the Work

## The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. There are two different groups of text summarization : indicative and informative. Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text .

## Project Modules

**Module 1**

* **Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection.

* **Data Cleaning**

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn’t the fanciest part of machine learning and at the same time, there aren’t any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning.

### Module 2

### Feature Selection

### Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

### It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.

### Model Training

### A machine learning training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from.

### ML models can be trained to benefit manufacturing processes in several ways. The ability of ML models to process large volumes of data can help manufacturers identify anomalies and test correlations while searching for patterns across the data feed. It can equip manufacturers with predictive maintenance capabilities and minimize planned and unplanned downtime.

### Module 3

### Model Evaluation

### Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

### To understand if your model(s) is working well with new data, you can leverage a number of evaluation metrics.

### Accuracy

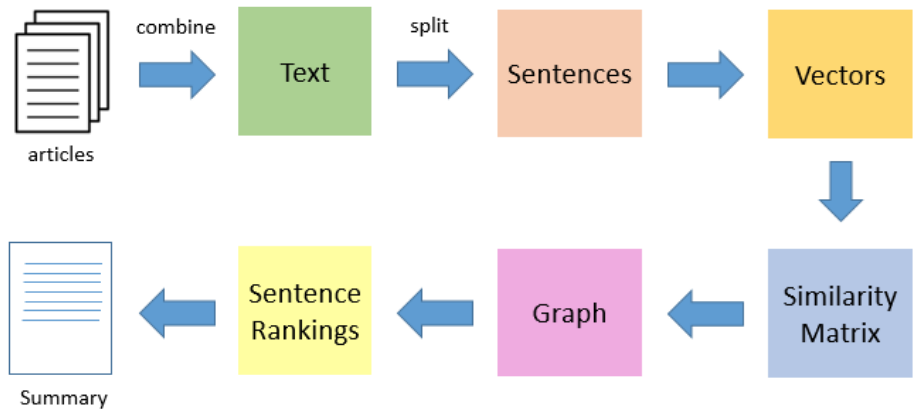
### Precision

### Confusion Matrix

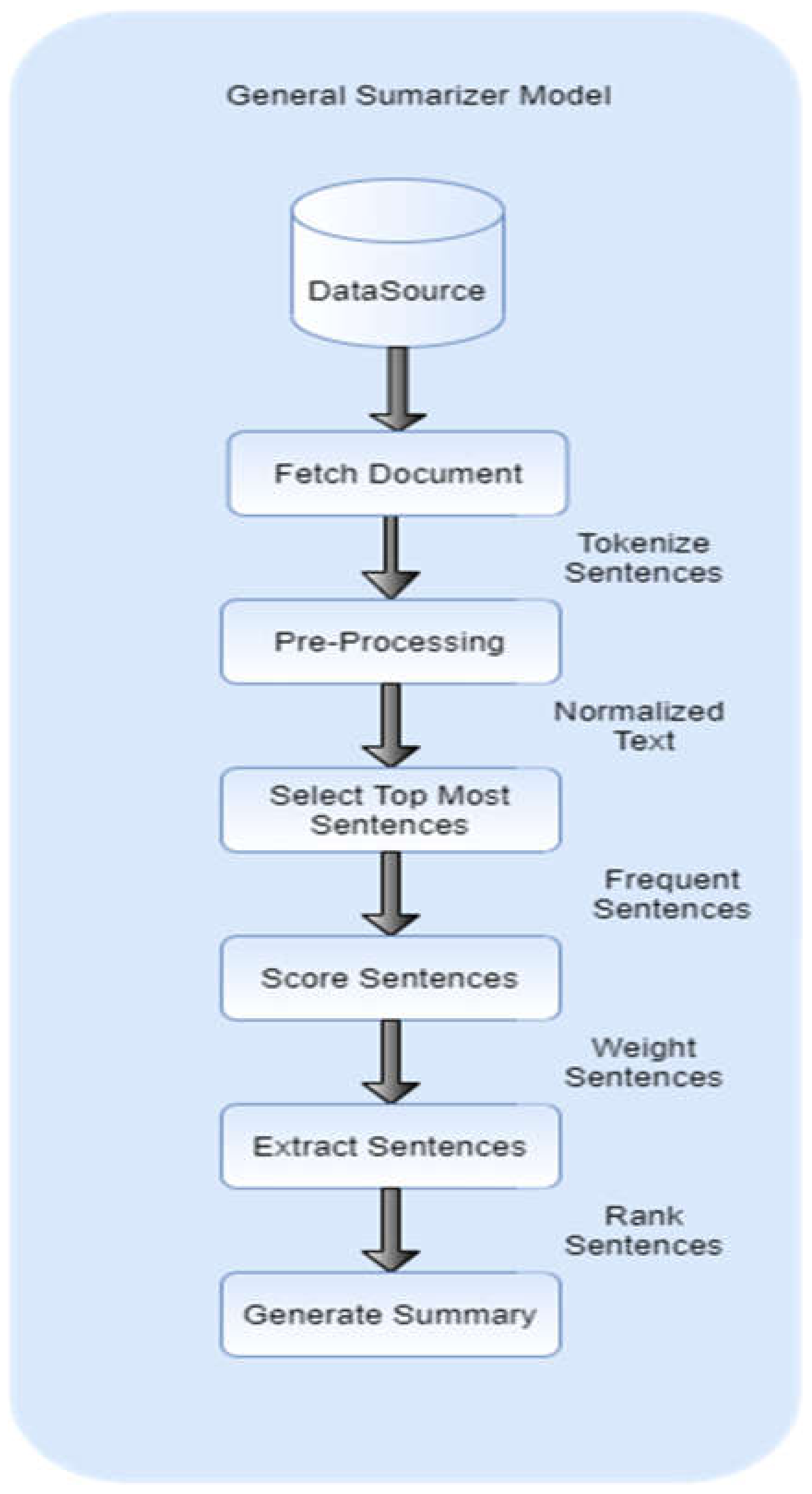
### Log-Loss

### AUC(Area under Curve)

## Context Diagram (High Level)



# Implementation Methodology



# Technologies to be used

## Software Platform

* Python 3.6.8
* PyCharm (IDE)
* Operating System (Windows 7,8,10,11)

## Hardware Platform

* RAM – Minimum 4GB
* Hard Disk – Minimum 32GB
* Processor i-5, AMD 3 upwards

## Tools, if any

* Pandas
* Numpy
* Scikit-Learn
* Matplotlib
* GridSearch CV
* Lasso Regression
* Linear Regression
* Decision Tree Regressor

# Advantages of this Project

* Summarizing reduces perusing time
* While investigating reports, outlines make the determination procedure simpler
* Summarization improves the adequacy of ordering
* Summarization calculations are less one-sided than human summarizers
* Personalized summaries are useful in question-answering systems as they provide personalized information
* Utilizing programmed or Summarization frameworks empower business theoretical administrations to build the number of content archives they can process
* Text summarization can help users quickly grasp the essence and context of the data, and identify the most important or interesting aspects.
* This can improve the efficiency and productivity of data analysis, as well as the communication and presentation of the results.

# Future Scope and further enhancement of the Project

The future study is to build a robust, domain and language independent extractive text summarization that works well with multi-documents. Similarly, because the quality evaluation of the summary is done manually by experienced assessors, it is highly subjective. There are specific quality assessment criteria, such as grammaticality and coherence, but the results are different when two experts evaluate the same summary.

# Project Repository Location

| **S#** | **Project Artifacts (softcopy)** | **Location** (Mention Lab-ID, Server ID, Folder Name etc.) | **Verified by Project Guide** | **Verified by Lab In-Charge** |
| --- | --- | --- | --- | --- |
|  | Project Synopsis Report (Final Version) | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Project Progress updates | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Project Requirement specifications | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Project Report (Final Version) | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Test Repository | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Project Source Code (final version) with executable | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |
|  | Any other document | https://github.com/shiva-3/Final-year-project | Name and Signature | Name and Signature |

# Conclusion

Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document .Text summarization approaches based on Neural Network, Graph Theoretic, Fuzzy and Cluster have, to an extent, succeeded in making an effective summary of a document.Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods. Abstractive method is similar to summaries made by humans. Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate into the domain specific areas.

Text summarization is an interesting machine learning field that is increasingly gaining attraction. As research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. Hereby, We can say we have successfully completed text summarization using NLP as per problem statement with efficiency. By this project we have solved the problem by the summaries of the text to gain information. We have tried our best to make these summaries as important as possible in the aspect of text intention. We can add various features to our web applications like we can take input of almost any text format like(.doc and .docx,.rtf) by uploading it directly in our input box for text summarization. We can also integrate features like the voice text acceptance for the text summarization. Example, someone reads out loud the text paragraph from the newspaper or passage from novel which is difficult to understand and needs to be summarized. We have certain limitation while dealing with punctuation marks and spaces so in future we will try to make it as proper as possible.

# References

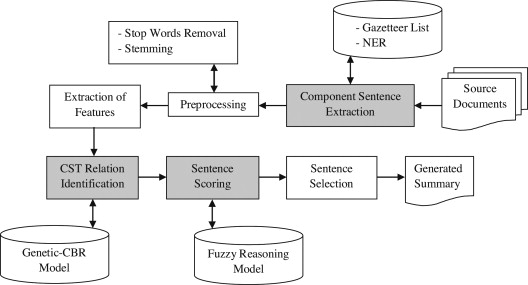
* <https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>
* <https://data-flair.training/blogs/machine-learning-text-summarization/>
* <https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html>
* [https://www.google.com/url?sa=i&url=https%3%2F%2Fwww.topcoder.com%2Fthrive%2Farticles%2Ftext-summarization in nlp&psig=AOvVaw2Bjnb9Ups2KLcLZEM5PtfT&ust=1677141229227000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCPCoy4vcqP0CFQAAAAAdAAAAABAE](https://www.google.com/url?sa=i&url=https%253%2F%2Fwww.topcoder.com%2Fthrive%2Farticles%2Ftext-summarization%20in%20nlp&psig=AOvVaw2Bjnb9Ups2KLcLZEM5PtfT&ust=1677141229227000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCPCoy4vcqP0CFQAAAAAdAAAAABAE)
* <https://research.ijcaonline.org/volume97/number3/pxc3897087.pdf>
* <https://app.genmymodel.com/api/repository/luciferdevil/TextSummarization>
* https://www.sciencedirect.com/science/article/abs/pii/S1568494614001598

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S#** | **Reference Details** | **Owner** | **Version** | **Date** |
|  | Project Synopsis | IAI-851 | 1.0 | 06-05-2023 |
|  | Project Requirements | IAI-851 |  |  |
|  |  |  |  |  |

**Annexure A**

**Data Flow Diagram (DFD)**

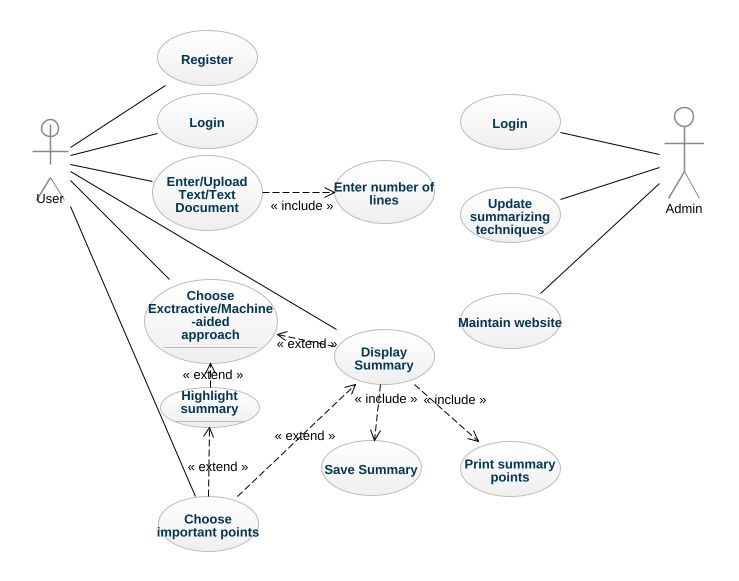
**(Mandatory)**

****

**Annexure B**

**Entity-Relationship Diagram (ERD)**

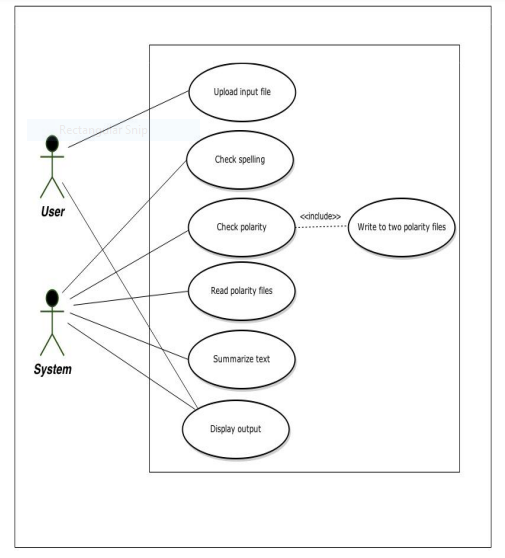
**(Mandatory)**

****

**Annexure C**

**Use-Case Diagram (UCD)**

**(Optional)**



**Annexure D**

**Screen Shots**

