**TEXT SUMMARIZER USING DEEP LEARNING**

PROJECT SYNOPSIS

**Project Lab (IAI-851)**

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

|  |  |
| --- | --- |
| Project Guide: | Submitted By: |
| **Mr. Sudhanshu Kumar** | **Anshika Gupta (TCA1960003)**  **Shivansh Sharma (TCA1960001)** |
|  |  |

FEBRUARY ,2023



Teerthanker Mahaveer University, Moradabad

Faculty of Engineering and Computing Sciences

Table of Contents

[1 Project Title 3](#_Toc31139962)

[2 Domain 3](#_Toc31139963)

[3 Problem Statement 4](#_Toc31139964)

[4 Project Description 5](#_Toc31139965)

[4.1 Project Modules 6](#_Toc31139967)

[5 Implementation Methodology](#_Toc31139968) 8

[6 Technologies to be used](#_Toc31139969) 9

[6.1 Software Platform](#_Toc31139970) 9

[6.2 Hardware Platform](#_Toc31139971) 9

[6.3 Tools](#_Toc31139972) 10

[7 Future Scope and further enhancement of the Project](#_Toc31139974) 10

[8 Team Details](#_Toc31139975) 11

[9 Conclusion](#_Toc31139976) 11

[10 References](#_Toc31139977) 12

### 1.Project Title

**TEXT SUMMARIZER USING DEEP LEARNING**

### 2.Domain

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

### 

### 

### 3. Problem Statement

### 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.

**Objectives**

• Summaries reduce reading time.

• When researching documents, summaries make the selection process easier.

• Automatic summarization improves the effectiveness of indexing.

• Automatic summarization algorithms are less biased than human summarizers.

• Personalized summaries are useful in question-answering systems as they provide personalized information.

• Using automatic or semi-automatic summarization systems enables commercial abstract services to - increase the number of text documents they are able to process.

**4. 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.

**4.1 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)

### 5.Implementation Methodology

In our project, the House Prediction dataset is

imported from Kaggle in Comma Separated

Values (csv) format. The dataset is analyzed with

the help of pandas, numpy and scikit-learn.

Tableau is used as a data visualization tool. After

drawing insights from the dataset with the help of

Tableau, we identify the important factors i.e.

factors majorly affecting the change in prices. The

factors adding insignificant values to the overall

result are omitted. The dataset is divided into two

parts - training set and testing set. The various

machine learning models are trained with the help

of the training set. The testing set is then used to

check the performance of all the machine learning

models. Accuracy score is calculated. Root Mean

Square Error of all the models is calculated. In the

final step the model with the highest accuracy

score and the least RMSE (Root Mean Square

Error) value is used for predicting house prices.

In our project, the House Prediction dataset is

imported from Kaggle in Comma Separated

Values (csv) format. The dataset is analyzed with

the help of pandas, numpy and scikit-learn.

Tableau is used as a data visualization tool. After

drawing insights from the dataset with the help of

Tableau, we identify the important factors i.e.

factors majorly affecting the change in prices. The

factors adding insignificant values to the overall

result are omitted. The dataset is divided into two

parts - training set and testing set. The various

machine learning models are trained with the help

of the training set. The testing set is then used to

check the performance of all the machine learning

models. Accuracy score is calculated. Root Mean

Square Error of all the models is calculated. In the

final step the model with the highest accuracy

score and the least RMSE (Root Mean Square

Error) value is used for predicting house prices.

### Applied Sciences | Free Full-Text | A Novel Approach for Semantic Extractive Text Summarization

### Text Summarization In NLP

### 6.Technologies to be used

**6.1.Software Platform**

Python 3.6.8

Jupyter Notebook from Anaconda 3 (IDE)

Operating System (Windows 7,8,10,11)

**6.2.Hardware Platform**

RAM – Minimum 4GB

Hard Disk – Minimum 32GB

Processor i-5, AMD 3 upwards

**6.3. Tools / Libraries**

Pandas

Numpy

Scikit-Learn

Matplotlib

GridSearch CV

Lasso Regression

Linear Regression

Decision Tree Regressor

**7.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.

**8.Team details**

| **Project Name & ID** | **Course Name** | **Student ID** | **Student Name** | **Role** | **Signature** |
| --- | --- | --- | --- | --- | --- |
| TEXT SUMMARIZER USING DEEP LEARNING | Project Lab (IAI-851) | TCA1960001 | Shivansh Sharma | Analysis. |  |
| TCA1960003 | Anshika Gupta | Implementation |  |

**9.Conclusion**

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.

**10.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://www.kaggle.com/datasets/snap/amazon-fine-food-reviews>