B.E. Project Synopsis on

Automated Text Summarization using NLP

By students of

Computer Engineering Department

Members:

Anjali Masur (18102007)
Harshita Jain (18102049)
Kevin Khimasia (18102029)
Sejal Khedekar (18102010)

Project Guide:

Dr. Pravin Adivrekar

CONTENTS

Page No.

1.Introduction:

As of late, there has been a blast in the measure of text data from an assortment of sources. This volume of text is a priceless source of information and knowledge, which should be effectively summarized to be useful. In this problem, the main objective is to automate text summarization. This expanding availability of documents has demanded exhaustive research in automatic text summarization. Because of increasing information on the internet, these kinds of research are gaining more and more attention among the researchers. The whole concept is to reduce or minimize the valuable information present in the documents.This is commonly used by several websites and applications to create news feed and article summaries. It has become very essential for us due to our busy schedules. We prefer short summaries with all the important points over reading a whole report and summarizing it ourselves. So, several attempts had been made to automate the summarizing process

2. Project Concept:

2.1 Abstract:

Text summarization approaches can be split into two groups: *extractive summarization* and *abstractive summarization*. The objective of extractive and abstractive summarization is to produce a generalized summary, which conveys information in a precise way that generally requires advanced language generation and compression techniques. The extractive model aims at selecting sentences from the passage or text and picking up the main, essential or relevant information from it as it is without any modification. Thus, it has some limitations. To overcome this and come up with a more concise summary, abstractive text summarization comes into the picture. Abstractive text summarization has a neural network called RNN and CNN which has many layers and shortens the summary generated in extractive text summarization.

2.2 Objectives:

- To generate a summary of a text/paragraph by providing the input in the form of a paragraph or an image.
- To generate bullet/key points of a paragraph that would cover only the important points and won't be as long as a summary.
- Summarization of a text/paragraph would be achieved by extractive and abstractive approaches.

2.3 Literature Review:

Extractive Text Summarization is the method of extracting content from the document and combining it to form a text smaller in size. This ensures that only the words having relevance in the document are selected for the summarization.

Whereas, Abstractive Text Summarization is capable of depicting information by creating new sentences. It can be divided into Structured and Semantic approaches, each of which can be subdivided into subcategories based on various methods.
The methods are:

1.In tree-based approaches, clustering from many documents occurs. This clustering is done with the help of the order and significance of these documents. Linearization is used for the formation of sentences using tree traversal

2.Ontology-based approach is used for creating domain-related summaries. A domain or a dataset is fed to the system in advance, and based on this dataset, the system generates summaries with relevant text in the summary

3.Rule-based method is based on random forest classification and feature scoring. The scoring is based on the constraints laid down by the user. The rules can be set in many ways, such as: using verbs and nouns which are related to each other; keywords and syntactic constraints; domain constraints

4.Graph-Based Approach uses the graph data structure for language representation. Here, every word unit is represented by a node, and the structure of the sentences is determined by directed edges. These edges represent the relationship between any two words. The underlying feature of this method is that it uses the shortest path algorithm to find the smallest sentences with a considerable amount of information. The sentence formation is subjected to constraints such as it is mandatory to have a subject, verb, and predicate in it.

2.4. Problem definition:

The need for text summarization is continuously increasing as today's world is getting flooded with a growing number of articles and links to choose from with the expansion of the internet. Human beings tend to read the whole document to develop an understanding of it and generate a summary by keeping the main points in mind. It is getting extremely difficult to obtain the required information from this pool of words and sentences in a short period. Going through all the documents, articles, and different forms of information to manually summarize is extremely time-consuming and exhausting for humans. Summarization helps in saving valuable time and conveys the main essence from which the reader can decide if they want to dig deeper.

2.5. Scope:

The aim of this project is to achieve automation of generating a summary for the given set of data by generating a summarized text of fixed word length by extractive summarization techniques. The model designed in the project will be trained such that it will choose important words and sentences from the input text and arrange them to formulate meaningful sentences. To broaden the scope, we will be implementing abstractive summarization as well where the ambiguity of sentences in the summary will be reduced as this approach generates a summary by framing new sentences that serve the purpose. The existing summarization tools have a restriction on the word length for input text so we will be working on this aspect, and try to remove such barriers.

2.6. Technology stack:

- Python 3.8
- Pandas
- Numpy
- Gensim
- Natural Language Processing (NLP)
- Natural Language Toolkit(NLTK)
- Recurrent Neural Networks
- LSTM (Long Short Term Memory) Networks
- Cosine similarity

2.7. Benefits for Society:

1. Summaries reduce reading time.
2. When researching documents, summaries make the selection process easier.
3. Automatic summarization improves the effectiveness of indexing.
4. Personalized summaries are useful in question-answering systems as they provide personalized information.
5. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.