A Project report on

Text Condenser using Machine Learning

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

<u>Submitted by</u>

A.KARTHIK
(20H51A0536)

SUMIT CHELLURU
(20H51A05F6)

Y.SANJANA
(20H51A05M5)

Under the esteemed guidance of

Dr. S. KIRUBAKARAN
(Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

EPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase-I report entitled **"Text Condenser using Machine Learning"** being submitted by A.KARTHIK (20H51A0536), SUMIT CHELLURU (20H51A05F6), Y.SANJANA (20H51A05M5) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Dr. S. KIRUBAKARAN
Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Dr. S. Kirubakaran, Professor**, Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

A.KARTHIK	20H51A0536
SUMIT CHELLURU	20H51A05F6
Y.SANJANA	20H51A05M5

TABLE OF CONTENTS

List of Figures

ABSTRACT

This project is dedicated to leveraging natural language processing (NLP) techniques to create an automated text summarization system. The primary objective is to efficiently condense extensive textual content in response to the increasing demand for streamlined information consumption. The investigation explores various summarization methods, culminating in a hybrid approach that harmonizes content extraction with context preservation.

To validate the efficacy of this hybrid approach, advanced NLP techniques, including tokenization, are employed. Additionally, a diverse dataset is meticulously annotated with human-generated summaries, enhancing the study's comprehensiveness. By integrating innovative methodologies, advanced NLP techniques, and a diverse dataset enriched with human insights, this project aims to advance the field of automated text summarization, contributing to more effective information condensation.

CHAPTER 1
INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1.Problem Statement

- Despite the increasing volume of information available, the efficient extraction of key content from large textual datasets remains a significant challenge.

- Existing text summarization tools face challenges in accurately and coherently summarizing large textual datasets, especially in the presence of diverse document structures, languages, and lengths.

- The envisioned text summarization tool aims to be versatile, finding applications in diverse contexts such as document summarization, news aggregation, and content curation.

- The overarching problem is to advance text summarization technology, overcoming current limitations and providing a more robust solution to distill essential information from varied and extensive textual sources.

1.2.Research Objective

To develop a sophisticated and contextually aware text summarization system using the latest natural language processing (NLP) techniques, with an emphasis on improving extractive and abstractive summarization capabilities.

The principal objective is to enhance the effectiveness, precision, and consistency of text summarization in many fields, taking into account variables like document length, linguistic subtleties, and user-specific needs. Furthermore, the study intends to investigate new methods for managing the summarizing of many documents and modifying the summarization model to accommodate different inputs. By tackling issues with automatic summary, the project hopes to advance the field of natural language processing (NLP) and improve the efficiency and accessibility of information condensation across a range of applications, such as news aggregation and document summarization.

1.3. Project Scope and Limitations

Scope:

- Explore various possibilities of diverse datasets for text summarizer to learn form Implementation of various algorithms to provide both extractive as well as abstractive abilities to the tool.

- The tool's scope encompasses user-centric features, allowing customization to meet individual preferences or specific requirements. This involves incorporating user feedback mechanisms and parameters to tailor the summarization output to user expectations.

- The scope includes making sure the text summary tool may be used in a variety of fields, businesses, and topic areas. This entails building in mechanisms to comprehend and process context-specific jargon and information specific to various topic fields.

CHAPTER 2
BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1. Natural Language Processing (NLP) based Text Summarization

2.1.1.Introduction

Automatic Text Summarization remains a challenging problem in the field of natural language processing. This technique is employed to condense extensive textual documents into concise summaries that capture the main points of the original content. The overarching goal of text summarization is to employ software to reduce the length of a document while preserving its essential information.

Text summarization serves the purpose of distilling the crucial elements of a text, making it a valuable tool for various applications. These summarization processes are typically categorized based on their input and output types.

Input-based classification:

1. Single Document Summarization: This category focuses on summarizing relatively shorter texts, where basic summarization models are typically deployed due to the manageable textual context.

2. Multi-Document Summarization: In contrast, multi-document summarization handles longer inputs, often involving a more complex web of semantic connections as multiple documents are processed simultaneously.

Aim-based classification:

1. Generic Summarization: Here, summarization models approach the input text without any prior bias or domain-specific knowledge, aiming to create neutral, unbiased summaries.

2. Domain-Specific Summarization: This category leverages domain-specific information to generate summaries that are more accurate and tailored to the subject matter, drawing on known facts within that domain.

3. Query-Based Summarization: In this approach, the summary is designed to answer specific natural language questions related to the input text, ensuring that the generated summary only contains known responses.

Output-based classification:

1. Extractive Summarization: Extractive summarization involves selecting and assembling important sentences or sections from the input text to construct the summary. This method simplifies the process but may lack the coherence of human-generated summaries.

2. Abstractive Summarization: Abstractive summarization models go a step further by generating their own phrases and sentences to create a more coherent summary, akin to what a human would produce. However, this method is more intricate and challenging to achieve, as it requires a deeper understanding of language.

While tremendous progress has been made in the field of text summarization, particularly with the advent of neural networks, it is important to note that the creation of abstract summaries, as in abstractive methods, still presents a complex task and remains a work in progress. Recent developments in neural machine translation and sequence models have propelled the capabilities of text summarization closer to human-level understanding.

Applications of text summarization are widespread and include media monitoring, search marketing, streamlining internal document workflows, facilitating financial research, enhancing social media marketing, aiding individuals with disabilities, and much more. This technology continues to find new and innovative applications in diverse fields.

2.1.2. Merits, Demerits and Challenges

Merits:

1) Efficient Information Retrieval: Text summarization helps users quickly obtain the key points of a document, saving time and effort, particularly when dealing with lengthy texts.

2) Content Abstraction: Abstractive summarization can produce summaries that convey the essence of the original text more effectively than extractive methods, making them useful for understanding complex materials.

Demerits:

1) Loss of Nuance: Summarization may result in the loss of nuanced or context-specific information, making it less suitable for applications where fine-grained details are crucial.

2) Quality Variation: The quality of summaries generated by algorithms can vary widely, and it can be challenging to consistently produce accurate and coherent summaries, especially in abstractive summarization.

Challenges:

1) Content Understanding: Ensuring that models can accurately understand and represent the content of documents, especially in abstractive summarization, remains a major challenge.

2) Handling Multilingual Text: Summarizing texts in multiple languages, while maintaining accuracy and coherence, is a complex challenge, as languages have varying grammatical structures and nuances.

2.1.3. Implementation

1) **Unsupervised**

Extractive Unsupervised summarization technique means creating the summary from the given document without using any previous labelled group or classification. There are three ways to do so, firstly graph based, secondly latent variable and lastly, term frequency. These are easy to implement and give satisfactory results. using KMeans Clustering for choosing sentences in extractive text summarization which is a ma jor disadvantage. The first step is to eliminate stop words, hyphens and redundant white spaces. This is called pre processing of the input text. The next step is to select the feature using n-gram and finding out the weights using Boolean Weighting(BOOL), Term Frequency(TF), Inverse Document Frequency(IDF) or TFIDF. The next step is to apply KMeans for sentence clustering. KMeans is an iterative process in which values are plotted to the nearest centroid(mean of all values) and then calculating new centroids. In the proposed method, the first sentence is considered as a baseline and the similarity between the sentences is plotted using Euclidean's distance.After the clustering is done using K clusters, the sentences (also called most representative sentences) nearest to the centroids are selected. The proposed method obtains more favourable results than other state-of-art methods. an unsupervised framework for extractive text summarization of a single document called SummCoder. In SummCoder, after the pre-processing, the sentences are converted to fixed length vectors using the skip thought model. For generating summary, sentence selection is done considering 3 scores: Sentence Content Relevance Metric (scoreContR), Sentence Novelty Metric (scoreNov) and Sentence Position Relevance Metric (scorePosR). After calculating all the scores a final score and relative score is calculated. Finally, the summary can be generated by firstly, according to the descending order of relative rank and secondly, according to their occurrence in the input text.

2)Supervised:

Extractive Supervised summarization strategies diminish the weight of summarization by choosing subsets of sentences. Analysts working with NLP are particularly pulled in to extractive summarization. The fundamental focal point of the current work is to recognize remarkable highlights which would help in making a decision about the significance of a sentence in an article. A Supervised learning approach requires a lot of named information or labelled dataset. Extractive procedures select the top N sentences that best speak to the central issues of the article. Extractive procedures are set up as binary classification problems where the objective is to recognize the article sentences having a place in the summary. A directed methodology requires the presence of a bunch of reference, or gold, summaries. Accordingly, an administered model utilizing a minimal robust set of features is the feature of the beneath strategies.

the Recurrent Neural Network (RNN) framework. Since the sentences are randomly ordered in our dataset, there is no immediately available meaning for each sentence from the surrounding sentences. A set of such features are used for each phrase to provide the context locally and globally, which is described below. 1) Abstract ROGUE It is used for summarization as a feature. It uses the abstract, a pre-existing description, to manipulate the known structure of a paper. Abstract ROUGE 's theory is that sentences that are strong qualitative summaries are often likely to have good summaries of the highlights. 2) Numeric Count This is basically calculated on counting the times of numeric occurrences in a sentence, As sentences containing numbers/math do not contribute to a healthy summary. 3) Title Score Nonstop words in the text which match with those in the title are given more importance in the summary. 4) Keyphrase Score Keywords used or predefined by the author are given more importance in the summary when used in the text. 5) Sentence Length We add as an attribute the length of the phrase, an effort to catch the intuition that short phrases are quite unlikely to be successful summaries because they do not communicate as much data as longer phrases.
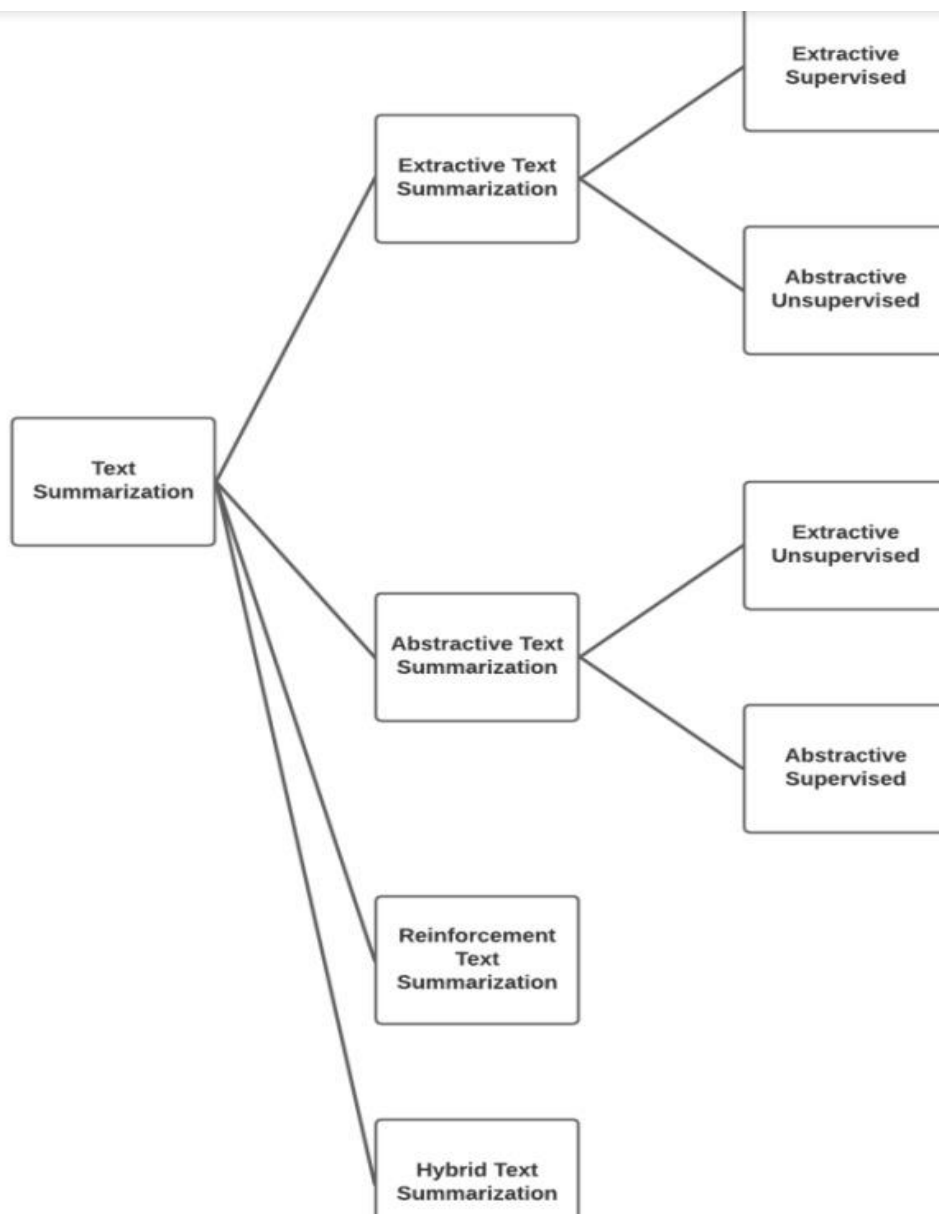
Fig 2.1

2.2. A Survey on NLP based Text Summarization for Summarizing Product Reviews

2.2.1. Introduction

There has been a continuous increase in the number of internet users every year. With an increase in Internet users, comes a great deal of information that gets stored online every second. There is a need for summarizing this data without losing the original meaning of the data. Thus the process of Text Summarization comes into the picture with its benefits spread over different fields such as Machine Learning, Natural Language Processing, Artificial Learning, Semantics etc., Online Shopping has become a common thing these days as a wide variety of products are available at a single place. Everyone refers to the product reviews before buying a product. Then they can conclude which is the best product to buy among the different products available. Suppose a user needs to buy a laptop. Then he must go through different kinds of laptops available at his budget, make a note of different reviews for each product and choose the best among the available laptops. This is a tedious and time taking task. In addition to this, some users' reviews are so long that the user could get the actual meaning of it only after closely going through the review. Thus there is a need for minimizing the review to a shorter representative sentence which depicts the same meaning as the whole content. This is the area where Text Summarization comes into picture with a great deal of benefits that could help us in choosing the best product from the whole lot. Text Summarization methods are broadly categorized into different types as shown in the figure below.
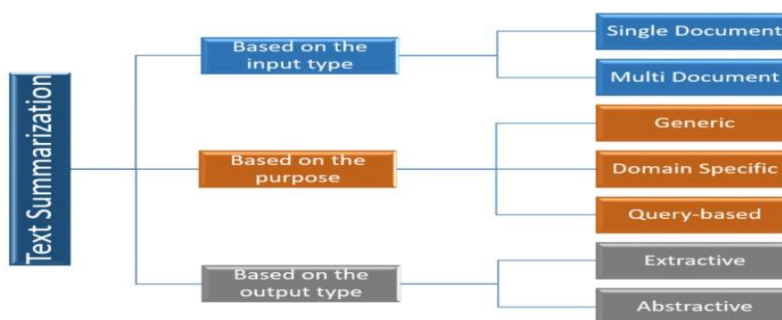

Fig 2.2

2.2.2. Merits, Demerits and Challenges

Merits:

1) Customization: Summarization models can be fine-tuned for domain-specific applications, ensuring the production of tailored and relevant summaries for various industries.

2) Language Translation: Summarization techniques are often utilized to generate concise translations of text in different languages, aiding in global communication.

Demerits:

1) Reference Bias: Extractive methods might favor content near the beginning of a document, potentially overlooking important information located elsewhere.

2) Ambiguity Handling: Handling ambiguous terms or references in text remains a significant challenge for summarization models, which might not always make the correct interpretation.

Challenges:

1) Real-Time Summarization: Generating summaries in real-time, such as during live events or streaming content, requires fast and efficient algorithms that can adapt to evolving information.

2) Fact-Checking and Veracity: Summarization models should be able to verify the accuracy of the information they summarize, as spreading false or misleading information can have serious consequences.

2.2.3. Implementation

Single Document Text Summarization (SDTS): In this type of Summarization, the length of the input is short. There will be only a single document given as the input for Summarization. This was used in the early days of Text Summarization. Multi-Document Text Summarization: This is a process in which the length of the input on a particular topic is too long and therefore multiple documents are provided as an input for a summarization technique. This is often difficult when compared with the SDTS as there is a need to combine the summary of multi documents into a single document. The difficulty here is that there may be diversity in the themes of different documents. An ideal summarization technique often makes condenses the main themes maintaining readability, completeness and without missing the important sentences. BASED ON OUTPUT TYPE: There are two types of techniques Extractive Text Summarization: As the name itself depicts, the extractive text summarization is the process in which the sentences are extracted from the whole text which could depict the similar meaning as the whole text but in a more condensed form. Majority of the text summarization techniques that are being used nowadays are of an extractive type. Abstractive Text Summarization: This is a more advanced type of text summarization which involves the formation of phrases or sentences that are not in the text but reflect the same meaning as the complete text. This method is more captivating but at the same time, it is more difficult for the model to form phrases or sentences that could bring the same meaning. BASED ON PURPOSE: Here it is categorized into the below 3 types. Generic Text Summarization: The method in which the model makes no inferences about the meaning of the text to be summarized or any knowledge of the domain is called Generic Text Summarization. It makes a generic summary of the whole text, documents, photos or video clips. Domain-Specific Text Summarization: In this method of text summarization, the model uses knowledge of a specific domain like scientific documents, medical documents. This increases the accuracy and thereby gives a more meaningful, concise and easily understandable summary of the whole text.

2.3. NPL based Machine learning Approaches for Text Summarization
2.3.1. Introduction

Today's world is centralized on computers and data. Data are our intangible thoughts and imagination. We are producers and consumers of data at the same time. Every little thing in our mundane lives are either a source or receiver of data. For, example when we drive there's data involved, the speed of the car, mileage, distance traveled, etc. Since 20th century, data has been a significant part of our lives, but these days we infer more from data. We store and access them through electronic and wireless systems. Since the advent of the internet, there's an enormous amount of data available today. The Internet is a storehouse of data. Information on news, movies, education, medicine, health, nations, weather, geology, etc. is available on the internet. This could be statistical, numerical, mathematical or text data. Text data is more difficult to interpret due to larger amount of characters. Due to this gigantic amount of information, there must be a system in order to get only the essential parts of the information we access. Text summarization is a way of doing this. Text summarization has been a topic of research and study since decades. Various models have been proposed and tested on different datasets to generate concise summaries. They are compared with different comparison scores. Text summarization can be EXT or ABS, single document or multidocument, and query-based or generic. EXT text summarization is a way of generating summaries by using the same sentences as in the document. ABS is more general and focuses on key concepts of the document. Similarly, single document summarization techniques give summaries of the text of a single document, and multidocument generates summaries of multiple documents. Moreover, these days, there's a need for summarizing text based on queries. Query-based summarization models give summaries of the text based on a specific area as described by the query given by the user, whereas generic summaries are mostly ABS that focus on the general area of the text input. Text summarization has been extensively used in various fields like science, medicine, law, engineering, etc. Researchers have focused on generating summaries of doctor's prescriptions, and that has been proved very useful to patients. Similarly, long news articles have been summarized and this way readers can gain a lot of information on several topics within a short span of time[1].

In this paper, we have discussed the various methods used in text summarization for past five years. The most common methods were found to be Machine Learning(ML), NNs, reinforcement learning, sequence to sequence modeling and fuzzy logic. Similarly, various optimization methods have been used to optimize the proposed objective function for the purpose of text summarization. We can see that various methods were tested on the same dataset, and their accuracy scores were found to be different. Moreover, we can also see that some researchers have combined the different methods and found out the summaries are more accurate than when a single method is used. When NLP processing has been used as a technique to summarize text documents, we see that python libraries such as scikit learn, nltk, spacy, and fastai has been used.

2.3.2. Merits, Demerits and Challenges

Merits:

1) Automated Content Generation: Automated summarization can enhance the productivity of content creators, such as writers and journalists, by providing a quick overview of source material.

2) Data Mining: Summaries can be useful for data mining and analytics, where large volumes of textual data need to be processed efficiently.

Demerits:

1) Ambiguity Handling: Handling ambiguous terms or references in text remains a significant challenge for summarization models, which might not always make the correct interpretation.

2) Training Data Dependency: Many summarization models require substantial amounts of training data, and their performance is often influenced by the quality and diversity of the training dataset.

Challenges:

1) User Customization: Allowing users to specify the level of detail, style, or domain-specific preferences in summaries is a challenge, as it requires fine-grained control over summarization models.

2) Ethical Concerns: The potential for misuse, such as generating misleading or biased summaries, raises ethical concerns that must be addressed in the development and deployment of summarization technologies.

2.3.3. Implementation

Massimo Mauro, et al. have used a sentence extraction method to generate EXT summaries. In this method, the sentences are checked for their relevance and scored accordingly. Similar sentences were then clustered together to discover the most informative sentences and they were selected on the basis of sentence scores[2]. Sarda A.T. et al. have proposed NNs and Rhetorical Structure Theory (RST) to produce EXT summaries of articles. Features are compared and combined to produce a relevant sentence for the summary. The first step is to train the NN so that it learns to select the types of sentences that would be present in the summary[3]. After that, sentences in the document are selected and checked if they fit into the summary. After finding these sentences, they are fed to the rhetorical structure with the help of which linguistic relations are distinguished to form a better summary. Gabriel Silva et al. have performed experiments on CNNCorpus to generate EXT summaries of the documents. . . The sentences of the corpus after removing figures, videos, and tables were assigned with feature vectors for scoring. The dimensionality of feature vectors was reduced using the selection algorithms of WEKA viz. CFS Subset Evaluator, Information Gain Evaluator, and SVM Attribute. Naive Bayes was proven to be the best classifier among 5 classifiers tested using WEKA's platform[4]. Taeho Jo has presented an approach to summarize text by considering similarity between attributes and using KNN algorithm. A paragraph is given as input and it is divided into sentences. Words are represented as vectors. Each sentence is then classified as summary or remaining, and based on the similarity score between them and a human generated summary, it's included in the summary[5].

In this paper, feature is considered as an important factor in order to summarize a text paragraph. This approach can be used in multiple areas like medicine, law, engineering[5]. Similarity between only some features are taken into account, and similarity score is calculated using the KNN method, based on which a paragraph is summarized. Mahsa Afsharizadeh, et al. presented a query-based approach for EXT text summarization. It selects the relevant sentences from the document and includes it in the summary. Eleven query-dependent appropriate features were chosen and used in the paper to find the important sentences[6]. For finding the useful sentences in the document, each sentence is assigned a score by checking on the linear function of its feature values.

CHAPTER 3
RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

- The text summarization project starts by collecting a diverse set of documents and preparing them for analysis. Utilizing a machine learning model, such as a Transformer-based architecture, the system learns to extract key content and generate concise summaries.

- Users interact with the system through a user-friendly application, capturing text input via mobile or PC. The model processes the input and produces abstractive summaries, preserving the essential information from the original text.

- The system incorporates checks for specific summarization features, adapting to different document types and styles. It continuously improves through updates and user feedback, ensuring relevance in the face of evolving language patterns.

- Ethical considerations are paramount, with the system designed to adhere to legal standards and privacy regulations. Quality control mechanisms review generated summaries to confirm accuracy and maintain high standards of output.

CHAPTER 4
CONCLUSION

CHAPTER 4

CONCLUSION

With the huge amount of data available on the internet, extracting the top data as a concept brief would be useful for some users. Since there is a huge amount of data deployed on the internet, there is always a need to find a way to reduce the length of the text and provide a clear summary.

Summarizing the text is still a work in progress in several research areas and needs to be further researched and developed in the summary of the texts. Due to the huge increase in the amount of data available online, extracting the top information as a conceptual summary will be useful for some researchers.

In this project we have drawn the reader's attention to the latest and main data problems as well as the need to summarize the texts and explained how short texts can be useful while keeping the original texts intact.

The optimization way can also be used to solve different problems.

REFERENCES

REFERENCES

- **Ahmad T. Al-Taani, "Automatic Text Summarization Approaches",** *International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017).*

- **Neelima Bhatia and Arunima Jaiswal, "Automatic Text Summarization: Single and Multiple Summarizations",** *International Journal of Computer Applications..*

- **Pankaj Gupta, Ritu Tiwari and Nirmal Robert, "Sentiment Analysis and Text Summarization of Online Reviews: A Survey",** *International Conzatiference on Communication and Signal Processing,* **August 2013.**

- **Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques",** *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE,* **vol. 2, no. 3, AUGUST 2010.**

<u>GitHub Link</u>
1.