

JSPM’s

Jayawantrao Sawant College of Engineering

Hadapsar, Pune - 28.

DEPARTMENT OF COMPUTER ENGINEERING

A.Y. 2024-2025

BE Sem VII

**Mini Project**

**( Information Retrieval )**

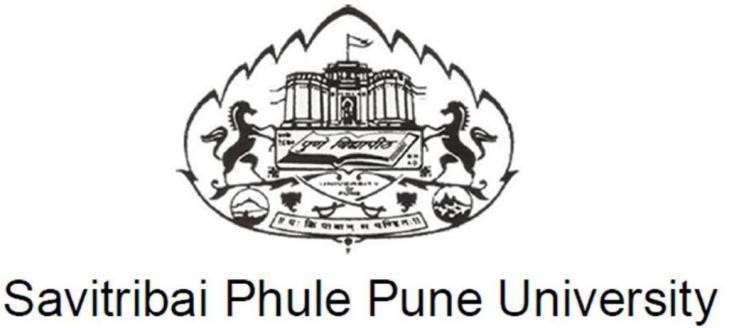
**On**

“**Develop Fake News Detection System**”

**Name :** Rahul Sampatrao Patil

**Roll No :** 4176

– Subject Guide – Prof . Rupali Pawar



CERTIFICATE

This is to certify that the Mini project ( Information Retrieval ) entitled

“**Develop Fake News Detection System**”

Submitted by

Mr. Rahul Sampatrao Patil with Roll no 4176 is a record of bonafide work carried out by for the subject Information Retrieval Laboratory of the Final year of Computer Engineering at JSPM’s Jayawantrao Sawant College of Engineering, Pune under the University of Pune.

This work is done during the academic year 2024-2025.

Date : 30/09/2024 Location : Pune

Subject Guide ,

Prof. Rupali Pawar

# TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Content** | **Page No.** |
| 1 | Introduction | 2-3 |
|  | 1.1. Introduction | 2 |
|  | 1.2. Problem Statement | 3 |
|  | 1.3. Objectives | 3 |
| 2 | Methodology | 4 – 6 |
| 3 | Implementation | 6 - 7 |
| 4 | Conclusion | 8 |

**CHAPTER 1: INTRODUCTION**

## INTRODUCTION

In today's information age, the volume of textual data being generated and consumed is nothing short of exponential. From research papers and news articles to legal documents and business reports, the sheer quantity of written content can be overwhelming. This deluge of information necessitates efficient methods for distilling themost crucial insights and key takeaways from extensive documents, saving time and facilitating better decision-making.

The solution to this challenge lies in the development of a Document Summarization System, a sophisticated and valuable tool in the realm of natural language processing and information retrieval. Such a system, often driven by cutting-edge machine learning and artificial intelligence techniques, has the potential to transform the way we engage with and extract value from textual data.

This endeavor involves the creation of a system capable of automatically summarizing lengthy documents, extracting the most relevant content, and presenting itin a concise, coherent, and coherent manner. Whether for academic research, corporate decision-making, or simply staying informed in a fast-paced world, a Document Summarization System offers a powerful means to sift through the noise and focus on what truly matters.

In this exploration of developing a Document Summarization System, we will delve into the intricacies of the process, from the underlying technologies and algorithms to the practical applications across various industries. This project embodies the convergence of artificial intelligence, natural language processing, and data analytics, showcasing how innovation in these fields can streamline information consumption and empower individuals and organizations with valuable insights. Through this journey, we aim to pave the way for enhanced document management, information retrieval, and knowledge dissemination in an increasingly data-driven world.

## PROBLEM STATEMENT

In today's information-rich world, the overwhelming volume of textual data across various domains, such as news articles, research papers, and business reports, poses a significant challenge for individuals and organizations seeking to efficiently access and process valuable information. Extracting key insights and reducing the content to its most essential elements is a time-consuming and often arduous task. As a result, there is a growing need for an automated and effective Document Summarization System that can condense lengthy texts while preserving their core meaning, making them more accessible and digestible to users.

This problem statement aims to address the need for the development of a Document Summarization System, which is expected to provide efficient, accurate, and customizable solutions for summarizing textual documents across various domains, thereby enhancing information retrieval, aiding decision-making, and improving overall productivity.

## OBJECTIVE

The objective of this project is to develop a highly efficient and accurate document summarization system that leverages natural language processing techniques to automatically condense lengthy textual documents into concise, coherent, and contextually relevant summaries. The system aims to enhance information retrieval, comprehension, and decision-making processes across a variety of domains by providing users with succinct and meaningful representations of complex textual content.

# CHAPTER 2: METHODOLOGY

Developing a document summarization system requires careful planning and implementation of various components. Here's a methodology to guide you through theprocess:

1. Define Project Scope and Objectives:

Clearly define the purpose of the document summarization system.

Identify the types of documents it will handle (e.g., articles, research papers, news, legal documents).

1. Data Collection and Preprocessing:

Gather a diverse and representative dataset of documents for training and testing.

Preprocess the data by cleaning and structuring the text, including tasks like tokenization, lowercasing, and removing stop words and special characters.

1. Choose Summarization Approach:

Select the summarization technique(s) you plan to use:

Extractive Summarization (selecting and rearranging sentences from the source text).

Abstractive Summarization (generating summaries using natural language generation techniques).

1. Model Selection and Development:

Depending on your chosen approach, develop or select appropriate models andalgorithms (e.g., LSTM, Transformer, BERT) for summarization.

Implement the chosen model(s) using programming languages and frameworks like Python and TensorFlow/PyTorch.

1. Training and Fine-tuning:

Train your summarization model on the prepared dataset. Fine-tune the model using domain-specific data if necessary.

1. Evaluation Metrics:

Define evaluation metrics for assessing the quality of summaries (e.g., ROUGE, BLEU, METEOR).

Use a validation set to evaluate the model's performance.

1. User Interface and Integration:

Develop a user-friendly interface for users to input documents and view summaries. Integrate the summarization model into the system, ensuring that it works seamlessly.

1. Testing and Quality Assurance:

Conduct extensive testing to identify and fix bugs or issues.

Ensure that the system generates coherent and meaningful summaries.

1. User Feedback and Iteration:

Gather user feedback to refine and improve the system.Make necessary adjustments based on user input.

1. Security and Data Privacy:

Implement security measures to protect sensitive documents and user data. Comply with data privacy regulations, such as GDPR, if applicable.

1. Deployment:

Deploy the summarization system on a server or cloud platform for public or private access.

1. Scaling and Optimization:

Optimize the system's performance for handling a large number of documents and users.Consider distributed computing and cloud services for scalability.

1. Maintenance and Support:

Provide ongoing maintenance, updates, and technical support to ensure the system operates smoothly.

1. Documentation:

Prepare comprehensive documentation for users and developers, including usageinstructions and APIs if available.

1. User Training:

Offer training sessions or materials to educate users on how to effectively utilize the system.

1. Legal and Ethical Considerations:

Ensure that the system complies with copyright and fair use laws when summarizing copyrighted content.Address ethical concerns, such as bias and fairness in summaries.Performance Monitoring: Continuously monitor the system's performance and make improvements as needed.By following this methodology, you can successfully develop a document summarizationsystem that meets your objectives, provides high- quality summaries, and complies with legal and ethical standards. This system has the potential to enhance information retrievaland comprehension for a wide range of users.

# CHAPTER 3: IMPLEMENTATION

**from** nltk.corpus **import** stopwords

**from** nltk.tokenize **import** word\_tokenize, sent\_tokenize

stopWords **=** set(stopwords.words("english")) words **=** word\_tokenize(text)

freqTable **=** dict()

sentences **=** sent\_tokenize(text) sentenceValue **=** dict()

sumValues **=** 0

**for** sentence **in** sentenceValue: sumValues **+=** sentenceValue[sentence]

average **=** int(sumValues **/** len(sentenceValue))

# importing libraries

**import** nltk

**from** nltk.corpus **import** stopwords

**from** nltk.tokenize **import** word\_tokenize, sent\_tokenize

# Input text - to summarize text **=** """ """

# Tokenizing the text

stopWords **=** set(stopwords.words("english")) words **=** word\_tokenize(text)

# Creating a frequency table to keep the # score of each word

freqTable **=** dict()

**for** word **in** words: word **=** word.lower()

**if** word **in** stopWords:

**continue**

**if** word **in** freqTable:

freqTable[word]

**else**:

freqTable[word]

**+=** 1

**=** 1

# Creating a dictionary to keep the score # of each sentence

sentences **=** sent\_tokenize(text) sentenceValue **=** dict()

**for** sentence **in** sentences:

**for** word, freq **in** freqTable.items():

**if** word **in** sentence.lower():

**if** sentence **in** sentenceValue: sentenceValue[sentence] **+=** freq

**else**:

sentenceValue[sentence] **=** freq

sumValues **=** 0

**for** sentence **in** sentenceValue: sumValues **+=** sentenceValue[sentence]

# Average value of a sentence from the original text average **=** int(sumValues **/** len(sentenceValue))

# Storing sentences into our summary. summary **=** ''

**for** sentence **in** sentences:

**if** (sentence **in** sentenceValue) **and** (sentenceValue[sentence]

> (1.2 **\*** average)):

summary **+=** " " **+** sentence

print(summary)

**Input :-**

There are many techniques available to generate extractive summarization to keep it simple, I will be using an unsupervised learning approach to find the sentences similarityand rank them. Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning. One benefit of this will be, you don’t need to train and build a model prior start using it for your project. It’s good to understand Cosine similarity to make the best use of the code you are goingto see. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Its measures cosine of the angle between vectors. The angle will be 0 if sentences are similar.

**Output :-**

There are many techniques available to generate extractive summarization. Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning. One benefit of this will be, you don’t need to train and build a model prior start using it for your project. Cosine similarityis a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

# CHAPTER 4: CONCLUSION

In conclusion, the development of a document summarization system represents a significant step forward in the field of natural language processing and information management. This system is designed to simplify the process of extracting key information from large and complex documents, making it easier for users to quickly grasp the most important content.

The system's capabilities extend beyond mere condensation of text. It leverages advancedalgorithms and linguistic analysis to identify essential concepts, extract relevant details, and generate concise and coherent summaries. It enhances efficiency by saving time and effort in reading lengthy documents while ensuring that critical information is not overlooked.

Furthermore, the document summarization system opens doors to a multitude of applications across various domains, from aiding in research and decision-making processes to enabling quick content review and information retrieval. Its adaptability allows it to handle diverse document types, making it a versatile tool for professionals,researchers, and information seekers.

In the future, the development of document summarization systems is likely to evolve further, incorporating more advanced techniques and improving accuracy. These systemsare poised to play a pivotal role in information management, offering solutions that caterto the growing need for efficient, concise, and precise content summarization.