**Industry Oriented Mini Project Report**

**on**

# Auto Title Craft

Submitted in partial fulfillment of the requirements
for the award of degree of

BACHELOR OF TECHNOLOGY

in

Information Technology

by

**A.Shivani (20WH1A1262)**

**B.Sneha (20WH1A12A3)**

**K.Shivani (20WH1A12C0)**

**Under the esteemed guidance of**

**Dr. P Kayal**

**Associate Professor**


Department of Information Technology
BVRIT HYDERABAD College of Engineering for Women
Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyd–500090
(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)
(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE & IT)

December 2023


# BVRIT HYDERABAD

## College of Engineering for Women

**Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090**
**(Affiliated to Jawaharlal Nehru Technological University Hyderabad)**
**(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE & IT)**

## CERTIFICATE

This is to certify that the Project report on **" Auto Title Craft "** is a bonafide work carried out by **A.Shivani (20WH1A1262), B.Sneha (20WH1A12A3) and K.Shivani (20WH1A12C0)** in the partial fulfillment for the award of B.Tech degree in **Information Technology , BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad** affiliated to Jawaharlal Nehru Technological University, Hyderabad, under my guidance and supervision. The results embodied in the project work have not been submitted to any other university or institute for the award of any degree or diploma.

**Internal Guide**
**Dr.P Kayal**
**Associate Professor**
**Department of IT**

**Head of the Department**
**Dr. Aruna Rao S L**
**Professor & HoD**
**Department of IT**

**External Examiner**

# DECLARATION

We hereby declare that the work presented in this project entitled**"Auto Title Craft"** submitted towards completion in IV year I sem of B.Tech IT at "BVRIT HYDERABAD College of Engineering for Women",Hyderabad is an authentic record of our original work carried out under the esteemed guidance of **Dr. P. Kayal, Associate Professor**, Department of Information Technology.

A.Shivani (20WH1A1262)

B.Sneha (20WH1A12A3)

K.Shivani (20WH1A12C0)

# ACKNOWLEDGMENTS

We would like to express our profound gratitude and thanks to **Dr. K. V. N. Sunitha, Principal, BVRIT HYDERABAD College of Engineering for Women** for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. Aruna Rao S L, Professor & Head, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for all the timely support, constant guidance and valuable suggestions during the period of our project.

We are extremely thankful and indebted to our internal guide, **Dr.P Kayal, Associate Professor, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for her constant guidance, encouragement and moral support throughout the project.

Finally, we would also like to thank our Project Coordinators **Mr Ch. Anil Kumar, Assistant Professor and Mr N.Anand, Assistant Professor**, all the faculty and staff of Department of IT who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

A.Shivani (20WH1A1262)

B.Sneha (20WH1A12A3)

K.Shivani (20WH1A12C0)

# ABSTRACT

In the context of Natural Language Processing (NLP), a title generator is a tool or model that generates titles or headlines for text-based content, such as articles, blog posts, news stories, or other documents. These generators use various NLP techniques, including language modeling and text summarization, to create concise and engaging titles that capture the essence of the content. Title generators can be useful for content creators, journalists, and marketers to automate the process of coming up with catchy and informative titles. They often take the input text or content and analyze it to produce a headline that is both relevant and attention-grabbing. This can save time and effort in the content creation process and help improve the visibility and content is influenced as per page rank. Title generators can be implemented using different NLP models and algorithms, such as Recurrent Neural Networks (RNNs), transformers, or rule-based approaches, depending on the specific requirements and available data. The proposed project will identify a suitable title for the document given by the user using efficient algorithms and will be evaluated using the evaluation metrics ROUGE (Recall-Oriented Understudy for Gisting Evaluation).

***Keywords***: ROUGE, Auto Title,NLP

# Contents

# List of Figures

# Chapter 1

# Introduction

A title generator is a tool or model that generates titles or headlines for text-based content.Title generators can be useful for content creators, journalists, and marketers to come up with catchy and informative titles. It not only entices readers but also encapsulates the core message or theme, providing a glimpse into the substance that lies within.In the realm of digital communication, where attention spans are often fleeting, the art of title generation becomes even more critical. A compelling title not only grabs attention but also stimulates curiosity, prompting readers to delve deeper into the content. It serves as a marketing tool, influencing the perceived value of the material and encouraging engagement.The process of title creation involves a delicate balance of creativity, clarity, and relevance. It requires an understanding of the target audience, the subject matter, and the desired tone. Whether aiming for humor, intrigue, or straightforward informativeness, the title sets the tone for the entire piece, acting as a guiding beacon for both the writer and the reader.In this era of information overload, where countless pieces of content vie for recognition, mastering the skill of title generation is a valuable asset for any writer or communicator. It's the first impression, the handshake, and the invitation—all rolled into a few carefully chosen words. As such, delving into the nuances of title creation can unlock a world of possibilities for effective communication and engagement.

## 1.1 Motivation

A title is the first thing your audience encounters. It's the gateway to your content, and a compelling title can leave a lasting impression, drawing readers in from the very beginning. In a world filled with information overload, capturing attention is a precious commodity. A well-crafted title acts as a magnet, enticing readers and prompting them to explore what lies beneath.A strong title sets expectations for

the quality of the content. Readers are more likely to engage with material they perceive as valuable, and the title plays a key role in shaping that perception.A well-crafted title is a call to action, inviting readers to explore, discover, and engage with your ideas. It's the catalyst for shares, the driver of curiosity, and the foundation of a lasting connection with your audience.

## 1.2 Objective

The model's primary objective is to automate the creation of concise and coherent summaries from user-provided text, condensing essential information while preserving key points. Subsequently, it extracts keywords from these summaries to identify crucial terms, aiding in understanding the main focus of the content. Ensuring user engagement, the model prompts users to input text for summarization and allows customization through parameter adjustments, such as the summarization ratio. Additionally, it evaluates performance by assessing the quality of summaries and extracted keywords, aiming to condense information effectively. The model handles potential warnings and errors, enhancing the user experience, and integrates external packages like BERT Extractive Summarizer and YAKE for state-of-the-art efficiency and accuracy in text summarization and keyword extraction. In achieving these objectives, the model significantly enhances the overall efficiency and effectiveness of its summarization, extraction, and title generation processes.

## 1.3 Problem Definition

Developing an automated solution for generating accurate and contextually relevant titles in the automotive domain using Natural Language Processing (NLP) is crucial due to the current challenges in manual title creation. The absence of a sophisticated NLP model results in time-consuming and error-prone processes, leading to suboptimal titles that may not attract the intended audience. This deficiency impacts the efficiency, consistency, and coherence of content creation across various automotive platforms, affecting user engagement and discoverability. Additionally, the lack of standardized and SEO-friendly titles hampers the visibility of automotive content online, limiting its potential reach. Addressing these issues through an advanced NLP-based auto title generation system is essential for enhancing content effectiveness, meeting the demands of modern platforms, and adapting to dynamic industry trends, ultimately improving user experience and fostering innovation in the automotive sector.

# Chapter 2

# Literature Survey

This paper introduces [1] an innovative approach to Automatic Title Generation using a pre-trained GPT-2 Transformer Language Model. The model employs a three-module pipeline – Generation, Selection, and Refinement – along with a Scoring function. Despite a limited corpus, the model leverages GPT-2's natural language generation capabilities, ensuring accurate titles. The Selection and Refinement modules further enhance semantic and syntactic accuracy. Trained on arXiv abstracts and evaluated on three test sets, the proposed pipeline demonstrates promising results using ROUGE and BLEU metrics. Human evaluation validates the effectiveness of the approach, showcasing its adaptability to various domains with sufficient relevant data.

This paper presents [2] a novel approach for automatic title generation for Hindi stories, articles, or passages. The system utilizes various techniques, including noun and adjective combinations, keywords associated with proverbs, and direct use of proverbs. The algorithm encompasses Parts of Speech Tagging, Heading Generation, and Natural Language Processing. Tested on 10 randomly selected Hindi short stories, the system demonstrates satisfactory results, generating accurate and relevant titles. The proposed tool has potential applications in educational institutes for Hindi tutors, novel or story writers, contributing to the promotion and awareness of the Hindi language in computer systems and society.

This research [3] focuses on keyword extraction from scientific articles in Bahasa Indonesia using the TextRank algorithm. The process involves pre-processing, TextRank-based keyword extraction, and post-processing stages. The study emphasizes the importance of the number of assigned keywords in determining the recall value. Results indicate an increase in recall from 38.46% to 61.54% with 5 to 15 keywords. The research suggests potential improvements by considering multiword expression candidates in the pre-processing stage. The study con-

tributes insights into TextRank's performance for Bahasa Indonesia scientific article keyword extraction, demonstrating its applicability across languages.

This paper [4] explores automatic text summarization, focusing on the extractive method TextRank applied to Serbian-language news corpora. With the increasing volume of textual data on the internet, the need for efficient algorithms that provide concise and accurate summaries has grown. The study compares TextRank results on Serbo-Croatian corpora with an implementation in the Gensim project. The complexities of Slavic languages, marked by high inflection, are considered, and unsuccessful attempts with a deep learning method are briefly discussed. The research aims to contribute to effective text summarization in non-English languages.

The article [5] introduces YAKE!, an unsupervised keyword extraction method for automatic summarization of large text collections. YAKE! relies on statistical text features, including term co-occurrence and frequencies, to identify the most relevant keywords within a single document. Unlike supervised approaches, YAKE! does not require large annotated corpora, making it versatile across languages and domains. The system demonstrates superiority over ten unsupervised and one supervised method across diverse datasets. Its simplicity, independence from training corpora, and effectiveness in various scenarios position YAKE! as a valuable solution for automated keyword extraction.

This paper [6] introduces a system for generating Topic Level Summaries, leveraging abstractive summarization with Bidirectional Encoder Representation from Transformer (BERT). The system validates user input, scrapes relevant information from Wikipedia and 'The Hindu', and employs a two-step summarization process to drown out irrelevancies. The first step generates summaries for Wikipedia and news articles separately, and the second step combines these summaries to create a precise Topic Level Summary. The implemented model achieves promising ROUGE scores, emphasizing its effectiveness in providing concise and informative summaries for user-entered topics.

# Chapter 3

# System Design

## 3.1 Architecture

The operational framework commences with the acquisition of user data, traversing a series of meticulously designed stages encompassing text summarization and keyword extraction. The journey begins with the diligent collection of user data, which then undergoes a systematic procession. Employing advanced text summarization techniques, the system distills the amassed information into succinct and meaningful summaries. This condensed representation serves as a foundational step for subsequent analysis. Moving forward, the process incorporates keyword extraction and ranking mechanisms. The system meticulously dissects the text, identifying key terms and phrases, and subsequently ranks them based on their contextual relevance, frequency, and significance to the overall content. This evaluation serves as a critical metric for determining the content's relevance to the user's intent or inquiry. The system further refines the output by selecting an appropriate title that encapsulates the core essence of the content. This title becomes a pivotal component in presenting the refined information to the user effectively. Following this, the architecture delves into the creation of a comprehensive dataset. This compilation integrates the original user data, the distilled text summaries, extracted keywords, and additional metadata, forming a structured repository of information. The culmination of this intricate process involves an assessment of relevance accuracy. The system rigorously evaluates its performance, ensuring that the distilled content aligns with user expectations and intent. This iterative and comprehensive workflow—from user input to dataset creation—exemplifies the system's proficiency in condensing content cohesively and determining relevance accurately, showcasing its adaptability and efficacy in handling diverse user data scenarios.

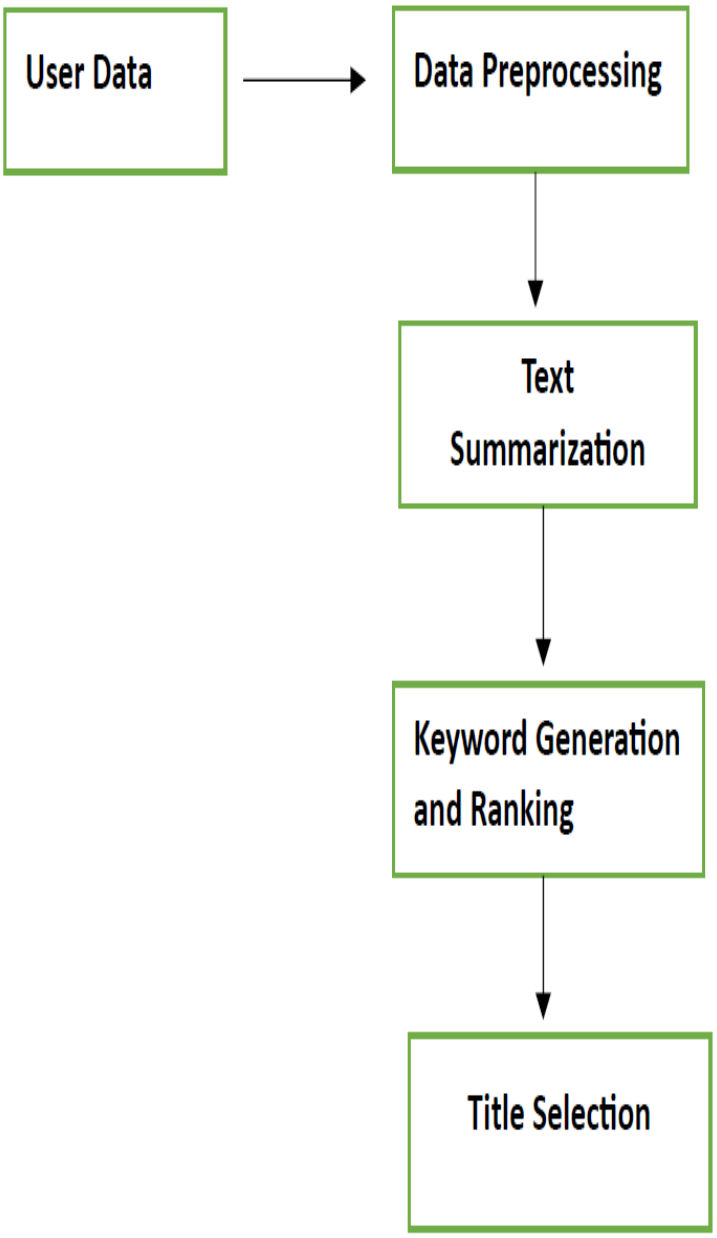



Figure 3.1: Architecture

## 3.2 Tools and Libraries used

The following are the tools and libraries used involved in the project :

- **Google Colab :** While not a specific Python module, Google Colab is a cloud-based environment utilized here to run Python code interactively, providing access to GPUs and facilitating collaborative coding.
- **Pandas :** Pandas is employed for data manipulation and handling, especially for reading and managing structured data from CSV files.
- **NumPy :** NumPy is essential for numerical computations, especially when dealing with arrays and mathematical operations on large datasets.
- **Yake :** YAKE (Yet Another Keyword Extractor) is an automatic keyword extraction algorithm used in Natural Language Processing (NLP). It is designed to identify and extract key terms or keywords from a given text, helping to summarize the main topics or themes within the content.
- **Bert Summarizer :** BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing model. It excels in understanding context and semantics in language, making it a powerful tool for various NLP tasks, such as text classification, named entity recognition, and question answering.
- **Deduplication :** To deduplicate values means to remove or eliminate duplicate instances of the same value within a dataset or a collection, ensuring each value appears only once. This process is commonly performed in data cleaning or preprocessing to maintain data integrity and avoid redundancy.

## 3.3 UML Diagram

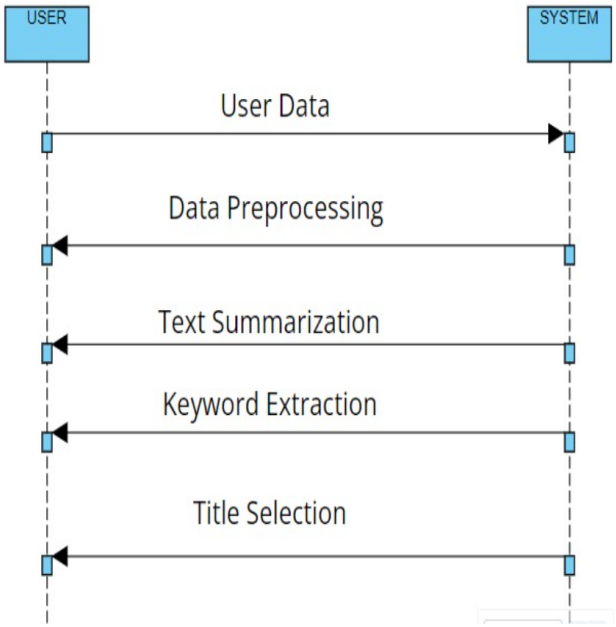



Figure 3.2: Usecase Diagram for Title Generation

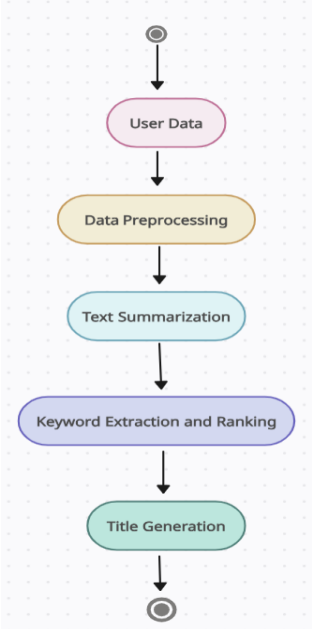



Figure 3.3: Activity Diagram for Title Generation

# Chapter 4

# Modules

## 4.1 Proposed Method

### 4.1.1 Text Summarization

Text summarization is the process of distilling the key information, main ideas, and essential points from a given document or text, while retaining its core meaning. The goal is to create a shorter version, known as a summary, that provides a condensed yet coherent representation of the original content. The Bert-extractive-summarizer stands as an advanced tool for producing succinct text summaries, employing the Bert model to extract pivotal information. This innovative approach significantly streamlines the summarization process by identifying and pulling out key details from the user-provided text. By doing so, it enhances the efficiency and effectiveness of information presentation. The generated summaries serve as valuable aids, offering users a quick and comprehensive understanding of the text. Leveraging state-of-the-art natural language processing, the summarizer ensures a more nuanced extraction of essential content. Overall, the Bert-extractive-summarizer is a robust solution for condensing information, making it an invaluable asset in diverse applications where concise and informative summaries are paramount.

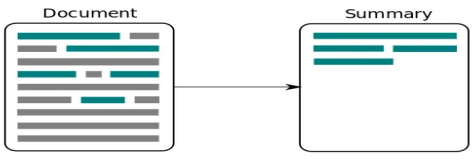



Figure 4.1: Text Summarization

### 4.1.2 Keyword Extraction and Ranking

The process of keyword extraction and ranking plays a pivotal role. Keywords serve as concise representatives of the main themes and critical elements within a text. The YAKE library is instrumental in this process, employing a sophisticated approach to identify and extract keywords from a given summary. YAKE utilizes a deduplication algorithm, as specified by the 'dedupFunc' parameter in the 'Keyword Extractor' instantiation. This algorithm ensures that the extracted keywords are unique and diverse, enhancing the quality and relevance of the identified terms. The library assigns a "score" to each keyword, reflecting its significance within the context of the summary. Efficient keyword extraction is invaluable for content retrieval, search engine optimization, and information categorization. By providing a succinct representation of essential concepts, YAKE contributes to improved content understanding, enabling users to grasp the core ideas and themes with greater precision.

Figure 4.2: Keyword Extraction

### 4.1.3 Title Generation

In title generation, the YAKE library employs a meticulous methodology for extracting keywords that significantly enhances the process. By utilizing a dedicated deduplication algorithm, YAKE ensures that the assigned scores to keywords reflect a nuanced understanding, preventing unnecessary repetition. This sophisticated approach not only contributes to the avoidance of redundancy but also adds a layer of precision in identifying pivotal terms within the content. The deduplication algorithm plays a crucial role in refining the selection of keywords, promoting the generation of titles that encapsulate the essence of the text with greater accuracy. This nuanced strategy distinguishes YAKE in title generation, facilitating the creation of impactful and informative titles for diverse content.

## 4.2 Performance Measure

### 4.2.1 BERT for Text Summarizer

BERT, or Bidirectional Encoder Representations from Transformers, stands as a cutting-edge natural language processing (NLP) model developed by Google. BERT apart is its bidirectional approach to context understanding, enabling it to consider the full context of a word by analyzing both its left and right surroundings within a sentence. BERT learns contextualized representations of words through tasks like predicting missing words in sentences. This contextualization allows BERT to generate embeddings influenced by the context, making it particularly effective for various NLP applications. Its state-of-the-art performance on diverse benchmarks has positioned BERT as a go-to model for tasks such as text classification and question-answering. BERT has computational intensity and resource requirements, posing challenges for deployment in resource-constrained environments.

### 4.2.2 YAKE for Keyword Extraction

YAKE, or Yet Another Keyword Extractor, is an innovative automatic keyword extraction algorithm designed for text analysis. Unlike conventional methods, YAKE employs a data-driven strategy by considering both statistical features, such as term frequency, and linguistic features, like part-of-speech patterns. Notably, YAKE focuses on extracting key phrases rather than individual keywords, providing a more contextually relevant representation of important terms within the text. The algorithm constructs a term-document matrix to capture relationships between terms and documents, facilitating the extraction of key phrases based on their significance in the corpus. YAKE is content-independent, applicable to various types of text data without requiring domain-specific adaptations. Users can customize the exclusiveness of extracted key phrases, offering a flexible balance between specificity and generality. Open-source and versatile, YAKE finds applications in document summarization, clustering, and information retrieval, making it a valuable tool for researchers and practitioners in natural language processing.

### 4.2.3 Randomforest Classifier for Accuracy

RandomForest classifier proves to be an effective strategy. RandomForest is an ensemble learning method that combines the predictions of multiple decision trees, offering robustness and accuracy. Each decision tree in the forest is trained on a random subset of the data, and their collective output helps mitigate overfitting and enhances generalization. For title generation, the RandomForest clas-

sifier can be trained on a dataset with labeled titles, learning patterns and relationships between input features and corresponding titles. The model's ability to handle non-linearity and capture complex relationships makes it well-suited for tasks where the quality of generated titles depends on various factors. By leveraging the diversity and independence of multiple decision trees, the RandomForest classifier holds promise in enhancing the precision and generalization of title generation in the project. Fine-tuning the model parameters and optimizing feature selection can further tailor its performance to the specific requirements of the title generation task.

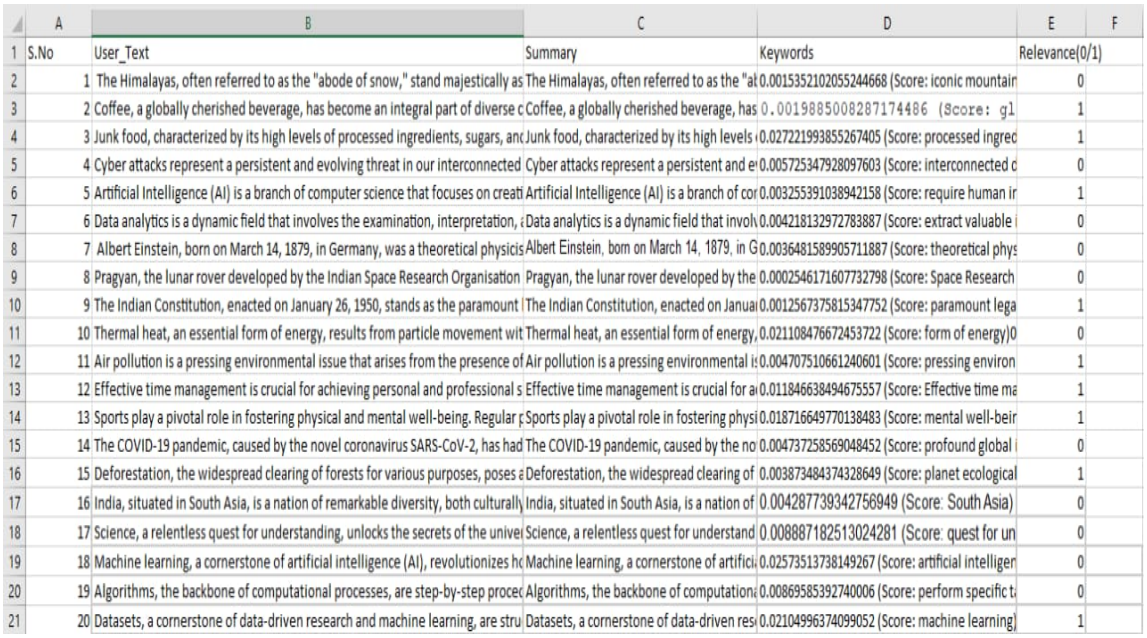
# Chapter 5

# Implementation

As a part of implementation of this project, We implemented and built 3 features, Text Summarization, Keyword extraction and Ranking, Title Generation. For building the Text Summarization module, We used BERT. BERT stands for Bidirectional Encoder Representations from Transformers, it is a type of recurrent neural network(RNN). BERT known for its bidirectional context understanding and exceptional performance across diverse language understanding tasks. For building the keyword extraction and ranking module, We have used YAKE library deduplicate algorithm. YAKE stands for Yet Another Keyword Extractor. Deduplication in the context of YAKE (Yet Another Keyword Extractor) is crucial to enhance the quality and relevance of the extracted keywords.

## 5.1 Dataset

The dataset contains approximately 200 rows and 5 columns consists of S.No, User Text, Summary, Keywords, Relevancy(0/1). Relevance is calculated manually either 0 or 1 based on the keywords extracted from the summary .If 0 is relevance then keywords are not suitable and if its 1 then keywords are related to summary .



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | S.No | User_Text | Summary | Keywords | Relevance(0/1) | |
| 2 | 1 | The Himalayas, often referred to as the "abode of snow," stand majestically as | The Himalayas, often referred to as the "al | 0.0015352102055244668 (Score: iconic mountain | 0 | |
| 3 | 2 | Coffee, a globally cherished beverage, has become an integral part of diverse c | Coffee, a globally cherished beverage, has | 0.0019885008287174486 (Score: gl | 1 | |
| 4 | 3 | Junk food, characterized by its high levels of processed ingredients, sugars, an | Junk food, characterized by its high levels | 0.027221993855267405 (Score: processed ingred | 1 | |
| 5 | 4 | Cyber attacks represent a persistent and evolving threat in our interconnected | Cyber attacks represent a persistent and e | 0.005725347928097603 (Score: interconnected d | 0 | |
| 6 | 5 | Artificial Intelligence (AI) is a branch of computer science that focuses on creati | Artificial Intelligence (AI) is a branch of cor | 0.003255391038942158 (Score: require human ir | 1 | |
| 7 | 6 | Data analytics is a dynamic field that involves the examination, interpretation, a | Data analytics is a dynamic field that involv | 0.004218132972783887 (Score: extract valuable i | 0 | |
| 8 | 7 | Albert Einstein, born on March 14, 1879, in Germany, was a theoretical physicis | Albert Einstein, born on March 14, 1879, in G | 0.0036481589905711887 (Score: theoretical phys | 0 | |
| 9 | 8 | Pragyan, the lunar rover developed by the Indian Space Research Organisation | Pragyan, the lunar rover developed by the | 0.0002546171607732798 (Score: Space Research | 0 | |
| 10 | 9 | The Indian Constitution, enacted on January 26, 1950, stands as the paramount | The Indian Constitution, enacted on Januai | 0.0012567375815347752 (Score: paramount lega | 1 | |
| 11 | 10 | Thermal heat, an essential form of energy, results from particle movement wit | Thermal heat, an essential form of energy, | 0.021108476672453722 (Score: form of energy)0 | 0 | |
| 12 | 11 | Air pollution is a pressing environmental issue that arises from the presence of | Air pollution is a pressing environmental is | 0.004707510661240601 (Score: pressing environ | 1 | |
| 13 | 12 | Effective time management is crucial for achieving personal and professional s | Effective time management is crucial for a | 0.011846638494675557 (Score: Effective time ma | 1 | |
| 14 | 13 | Sports play a pivotal role in fostering physical and mental well-being. Regular p | Sports play a pivotal role in fostering physi | 0.018716649770138483 (Score: mental well-beir | 1 | |
| 15 | 14 | The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has had | The COVID-19 pandemic, caused by the no | 0.004737258569048452 (Score: profound global i | 0 | |
| 16 | 15 | Deforestation, the widespread clearing of forests for various purposes, poses a | Deforestation, the widespread clearing of | 0.003873484374328649 (Score: planet ecological | 1 | |
| 17 | 16 | India, situated in South Asia, is a nation of remarkable diversity, both culturally | India, situated in South Asia, is a nation of | 0.004287739342756949 (Score: South Asia) | 0 | |
| 18 | 17 | Science, a relentless quest for understanding, unlocks the secrets of the univer | Science, a relentless quest for understand | 0.008887182513024281 (Score: quest for un | 0 | |
| 19 | 18 | Machine learning, a cornerstone of artificial intelligence (AI), revolutionizes hc | Machine learning, a cornerstone of artifici | 0.02573513738149267 (Score: artificial intelliger | 0 | |
| 20 | 19 | Algorithms, the backbone of computational processes, are step-by-step proce | Algorithms, the backbone of computationa | 0.00869585392740006 (Score: perform specific t | 0 | |
| 21 | 20 | Datasets, a cornerstone of data-driven research and machine learning, are stru | Datasets, a cornerstone of data-driven res | 0.02104996374099052 (Score: machine learning) | 1 | |

Figure 5.1: Dataset

## 5.2 Code

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
```

Figure 5.2: Importing Libraries

```
import pprint
# Take input from the user for the text
user_text = input(" ")
# Print the user-provided text with a line break
pprint.pprint("TEXT:{}".format(user_text))
```

Figure 5.3: Data Preprocessing

```
pip install bert-extractive-summarizer

Collecting bert-extractive-summarizer
  Downloading bert_extractive_summarizer-0.10.1-py3-none-any.whl (25 kB)
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (from bert-extractive-summarizer) (4.35.2)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (from bert-extractive-summarizer) (1.2.2)
Requirement already satisfied: spacy in /usr/local/lib/python3.10/dist-packages (from bert-extractive-summarizer) (3.6.1)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->bert-extractive-summarizer) (1.23.5)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->bert-extractive-summarizer) (1.11.3)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->bert-extractive-summarizer) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->bert-extractive-summarizer) (3.2.0)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy->bert-extractive-summarizer) (3.0.12
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy->bert-extractive-summarizer) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy->bert-extractive-summarizer) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy->bert-extractive-summarizer) (2.0.8)
```

Figure 5.4: Installing BERT summarizer

```
from summarizer import Summarizer
def generate_summary(user_text):
    model = Summarizer()
    summary = model(user_text, ratio=0.5)
    sentences = summary.split('. ')
    summary_3_lines = '. '.join(sentences[:3])
    return summary_3_lines
summary = generate_summary(user_text)
```

Figure 5.5: Generating Summary

```python
from rouge import Rouge
def calculate_rouge_scores(hypothesis, reference):
    rouge = Rouge()
    scores = rouge.get_scores(hypothesis, reference)
    return scores
# Example usage
if __name__ == "__main__":
    # Sample reference and hypothesis summaries
    reference_summary = "Junk food, characterized by its high levels of processed ingredients, sugars, and unhealthy fats, is a prevalent aspect of modern diet
    generated_summary = "Junk food, characterized by its high levels of processed ingredients, sugars, and unhealthy fats, is a prevalent aspect of modern diet
    # Calculate ROUGE scores
    rouge_scores = calculate_rouge_scores(generated_summary, reference_summary)
    # Access individual scores
    rouge_n_score = rouge_scores[0]['rouge-1']['f']
    rouge_l_score = rouge_scores[0]['rouge-1']['f']
    # Print the scores
    print(f"ROUGE-N Score: {rouge_n_score}")
    print(f"ROUGE-L Score: {rouge_l_score}")
```

Figure 5.6: Rouge Score for Summary

```python
import warnings
warnings.filterwarnings("ignore")
warnings.resetwarnings()
```

Figure 5.7: Ignoring Warnings

```
import yake

def get_keywords_yake(text):
    # Create a YAKE keyword extractor
    keyword_extractor = yake.KeywordExtractor(lan='en',  # language
                                              n=3,      # n-gram size
                                              dedupLim=0.9,  # deduplication threshold
                                              dedupFunc='seqm',  # deduplication algorithm
                                              windowsSize=1,
                                              top=10)    # number of keywords to extract
    # Extract keywords
    keywords = keyword_extractor.extract_keywords(text)
    return keywords
def print_results(keywords):
    print("Keywords:")
    for score, keyword in keywords:
        try:
            score = float(score)
            print("{} ({:.2f})".format(keyword, score))
        except ValueError:
            print("{} (Score: {})".format(keyword, score))
keywords = get_keywords_yake(summary)
print_results(keywords)
```

Figure 5.8: Importing YAKE Library

# Chapter 6

# Results and Discussions

### 6.1 Experimental

The proposed text processing pipeline involves multiple steps: input text is summarized using the BERT Summarizer, and then YAKE extracts keywords from the generated summary. The highest-scoring keyword serves as the title. A manually labeled dataset is created, associating each keyword with a relevance score (0 or 1). Subsequently, a Random Forest classifier is trained on this dataset to predict keyword relevance. The accuracy of the model in determining keyword relevance is assessed. This comprehensive approach combines advanced natural language processing techniques, machine learning, and manual evaluation, offering a nuanced strategy for automating title generation with potential applications in information retrieval systems.

```
Coffee, often hailed as the elixir of wakefulness, holds a special place in the hearts an
('TEXT:Coffee, often hailed as the elixir of wakefulness, holds a special '
 'place in the hearts and cups of people around the world. This aromatic '
 'beverage, derived from the roasted seeds of the Coffea plant, weaves a rich '
 'tapestry of culture, history, and daily rituals. As the sun rises, so does '
 'the steam from coffee mugs, signaling the start of a new day.  In the '
 "highlands of Ethiopia, where the legend of coffee's discovery began, locals "
 'speak of a goat herder named Kaldi. According to folklore, Kaldi noticed his '
 'goats becoming unusually energetic after munching on red berries. Intrigued, '
 'he sampled the berries himself, and the rest is history. This mythical tale '
 'encapsulates the essence of coffee – a stimulant that ignites vitality and '
 'fuels tales of discovery.  As coffee cultivation spread across the '
 'continents, it evolved into a global phenomenon. The coffee plantations of '
 'Latin America, the robusta fields of Africa, and the aromatic Arabica farms '
 'in Asia all contribute to the diverse flavor profiles found in our daily '
 'brews. Each sip encapsulates the geographical nuances, altitude, and '
 'climate, creating a sensory journey with every cup.  Beyond its geographical '
 'roots, coffee has embedded itself in societal rituals. It is the companion '
 'to morning rituals, the catalyst for business meetings, and the muse for '
 'countless artists. The bustling café culture, with the hiss of the espresso '
✓ 0s  completed at 6:30 AM
```

Figure 6.1: Data Preprocessing

```
Enter the number of texts you want to summarize: 1
Enter text 1:The Himalayas, often referred to as the "abode of snow," represent a majestic and formidable mountain range in South Asia, spanning five countries:
Summary for Text 1:
The Himalayas, often referred to as the "abode of snow," represent a majestic and formidable mountain range in South Asia, spanning five countries: India, Nepal,
```

Figure 6.2: Generated Summary

```
print(f"ROUGE-L Score: {rouge_l_score}")

ROUGE-N Score: 0.6338797770909852
ROUGE-L Score: 0.6338797770909852
```

Figure 6.3: Rouge Score for Summary

```
Keywords for Text 1:
Keywords:
0.003782293066447377 (Score: South Asia)
0.0065435388788183555 (Score: formidable mountain range)
0.011829620965203699 (Score: abode of snow)
0.011829620965203699 (Score: spanning five countries)
0.015553393367982867 (Score: represent a majestic)
0.01828768303803474 (Score: colossal mountain range)
0.01828768303803474 (Score: mountain range plays)
0.02047306640806586 (Score: majestic and formidable)
0.021984705787557363 (Score: mountain range)
0.029938739240021932 (Score: formidable mountain)
```

Figure 6.4: Extracted Keywords

```python
from sklearn.ensemble import RandomForestClassifier
# instantiate the classifier
rfc = RandomForestClassifier(random_state=0)
# fit the model
rfc.fit(X_train, y_train)
# Predict the Test set results
y_pred = rfc.predict(X_test)
# Check accuracy score
from sklearn.metrics import accuracy_score

print('Model accuracy : {0:0.4f}'. format(accuracy_score(y_test, y_pred)))
```

```
Model accuracy : 0.6066
```

Figure 6.5: Model Accuracy

# Chapter 7

# Conclusions and Future Scope

### 7.1 Conclusion

In this text processing workflow, a user provides input text, and a summary is generated from that text. Subsequently, keywords are extracted from the generated summary, and the word with the highest score is chosen as the title. This approach simplifies complex texts, condensing them into a concise summary and distilling key information into a single-word title. However, the effectiveness of this method relies on the accuracy of both the summarization and keyword extraction processes. Continuous refinement of these algorithms and user feedback can enhance the overall performance, making the system a valuable tool for efficiently extracting and highlighting essential content from user-provided texts.

### 7.2 Future Scope

The future scope for the given text summarization and keyword extraction approach is promising. Enhancements can be made to improve the title selection process. Implementing advanced natural language processing techniques and machine learning models can refine keyword scoring, considering semantic relationships and context. Integration with domain-specific ontologies and knowledge graphs can enhance keyword relevance. Additionally, exploring multi-document summarization and abstractive summarization methods can broaden applicability. Collaboration with sentiment analysis tools can enable title generation reflecting user sentiments. Further, deploying the system in real-time applications, such as news aggregation or content recommendation, can provide valuable insights and enhance user experience. Continuous updates and adaptability to evolving language patterns will be crucial for sustained effectiveness.

# Bibliography


[1] C. D. P.Mishra, "Automatic title generation for text with pre-trained transformer language model," 2021.

[2] L. Jain and P. Agarwal, "Title generation tool for hindi short stories," 2018.

[3] Gunawan and Dani, "Keyword extraction from scientific articles in bahasa indonesia using textrank algorithm," *Telecomminication and Computer Engineering(ELTICOM)*, 2020.

[4] Kosmajac, Dijana, and V. Keselj, "Automatic text summarization of news articles in serbian language," *18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019.

[5] Campos and Ricardo, "Yake! keyword extraction from single documents using multiple local features.," *Information Sciences 509*, pp. 257–289, 2020.

[6] M.Ramina, N.Darnay, C.Ludbe, and A.Dhruv, "Topic level summary generation using bert induced abstractive summarization model," *IEEE Access*, pp. 747–752, 2020.