A

Mini Project Report on

**Multilingual News Summarizer using**

# AI Techniques

Submitted in partial fulfilment of the requirements for the degree of

BACHELOR OF ENGINEERING

IN

**Computer Science & Engineering** Artificial Intelligence & Machine Learning by

Ayush Gondhali (22106124)

Harsh Jain (22106117)

Krishna Mishra (22106131)

Tanuj Kokamkar (22106088)

Under the guidance of

# Prof. Tejashree Kolhe



**Department of Computer Science & Engineering**

**(Artificial Intelligence & Machine Learning)**

**A. P. Shah Institute of Technology G. B. Road, Kasarvadavali, Thane (W)-400615 University of Mumbai 2024-2025**

# A. P. SHAH INSTITUTE OF TECHNOLOGY



## CERTIFICATE

This is to certify that the project entitled “**Multilingual News Summariser using AI Techniques”** is a bonafide work of Ayush Gondhali (22106124), Harsh Jain (22106117), Krishna Mishra (22106131), Tanuj Kokamkar (22106088) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Prof. Tejashree Kolhe | Dr. Jaya Gupta |
| Mini Project Guide | Head of Department |

# 



# A. P. SHAH INSTITUTE OF TECHNOLOGY

## Project Report Approval

This Mini project report entitled “**Multilingual News Summariser using AI Techniques*”*** by **Ayush Gondhali, Harsh Jain, Krishna Mishra and Tanuj Kokamkar**is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

External Examiner:

Internal Examiner:

Place: APSIT, Thane Date:

**Declaration**

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Ayush Gondhali Harsh Jain Krishna Mishra Tanuj Kokamkar

(22106124) (22106117) (22106131) (22106088)

**ABSTRACT**

The rapid growth of digital news content in multiple languages has made it challenging for users to stay informed efficiently. This project, Multilingual News Summariser using AI Techniques, aims to address this issue by developing an intelligent system that can automatically summarize news articles in various languages while preserving key information. The system leverages Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyze and generate concise summaries. It utilizes text preprocessing methods such as tokenization, stop word removal, and stemming, followed by summarization algorithms like Extractive and AbstractiveSummarization. Deep learning models, including Transformers, are integrated to enhance the summarization quality across different languages. Additionally, multilingual embedding models, ensure effective cross-language understanding.The proposed system collects news data from various online sources, processes the text, and provides users with short, meaningful summaries. By reducing reading time and overcoming language barriers, this AI-powered summarizer enhances information accessibility for global users. Future improvements may include real-time news updates to further enrich user experience. Additionally, a Text-to-Speech (TTS) feature has been integrated to allow users to listen to the generated summaries, improving accessibility for individuals with visual impairments or those preferring audio content.

This project demonstrates the power of AI in automated text summarization and highlights its potential for use in journalism, research, and personalized content delivery.

**Keywords: Multilingual News Summariser, Text Preprocessing Methods, Machine Learning, Natural Language Processing (NLP), Text-to-Speech (TTS).**

## Index

|  |  |  |  |
| --- | --- | --- | --- |
| Index | | | Page no. |
| Chapter-1 | | |  |
|  | Introduction | | 1 |
|  |  |  |  |
| Chapter-2 | | |  |
|  | Literature Survey | | 3 |
|  | 2.1 | History | 4 |
|  | 2.1 | Review | 5 |
|  |  |  |  |
| Chapter-3 | | |  |
|  | Problem Statement | | 8 |
|  |  |  |  |
| Chapter-4 | | |  |
|  | Experimental Setup | | 11 |
|  | 4.1 | Hardware setup | 12 |
|  | 4.2 | Software Setup | 12 |
|  |  |  |  |
| Chapter-5 | | |  |
|  | Proposed system and Implementation | | 13 |
|  | 5.1 | Block Diagram of proposed system | 14 |
|  | 5.2 | Description of Block diagram | 14 |
|  | 5.3 | Implementation | 16 |
|  | 5.4 | Advantages | 18 |
| Chapter-6 | | |  |
|  | Conclusion | | 19 |
|  |  |  |  |
| References | | | 21 |
|  |  |  |  |

**CHAPTER 1**

**INTRODUCTION**

### 1. INTRODUCTION

In today’s digital era, a vast amount of news content is generated in multiple languages every second. However, accessing and understanding news in different languages can be a major challenge for users, especially when they need quick insights. Traditional translation methods often fail to preserve the essence of news articles, and manually reading through long reports is time-consuming. To address these challenges, AI-driven text summarization techniques can be employed to extract key information from multilingual news sources and present it in a concise, easy-to-read format.

This project focuses on developing a Multilingual News Summariser using AI Techniques that automatically generates summaries of news articles written in different languages. By leveraging Natural Language Processing (NLP) and Deep Learning models, the system can effectively analyze, understand, and generate meaningful summaries while maintaining the original context. The use of transformer-based architectures, and multilingual BERT ensures accurate and context-aware summarization across various languages. Additionally, One of the major enhancements of our system is the integration of Text-to-Speech (TTS) technology, allowing users to listen to the generated summaries instead of reading them. This feature is particularly beneficial for individuals with visual impairments, multitasking professionals, and those who prefer audio-based news consumption. By leveraging AI-powered speech synthesis, the system provides natural and clear pronunciation for multilingual summaries, further enhancing accessibility.

With the increasing demand for personalized and real-time news consumption, this AI-powered summarizer can be highly beneficial for individuals, researchers, and media organizations. It allows users to quickly grasp important news highlights without language constraints. The project also lays the groundwork for further enhancements, such as speech-to-text integration, sentiment analysis, and real-time news aggregation, making it a valuable tool in today’s fast-paced information landscape.

# CHAPTER 2

**LITERATURE SURVEY**

**2**

## 2. LITERATURE SURVEY

### 2.1-HISTORY

1. **Early Text Summarization (1950s-1990s)** 
   * The concept of text summarization emerged in the 1950s with early statistical methods.
   * In the **1960s-70s**, basic rule-based and keyword extraction techniques were developed
   * By the **1990s**, extractive summarization models using sentence ranking and frequency analysis were introduced.
2. **Introduction of Natural Language Processing (NLP) (2000s)** 
   * The early 2000s saw advancements in NLP-based extractive summarization, where techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and Latent Semantic Analysis (LSA) were used.
   * Initial multilingual summarization attempts relied on machine translation tools, but they often failed to maintain contextual accuracy.
3. **Evolution of Text-to-Speech (TTS) Technology** 
   * Early TTS systems (1980s-2000s) used rule-based and concatenative synthesis, which sounded robotic.
   * Deep learning models like WaveNet (2016) by Google revolutionized TTS, making it more natural-sounding.
   * Advances in AI-driven TTS led to real-time speech synthesis with high accuracy and multilingual support.
4. **Deep Learning and Transformer Models (Late 2010s - Present)** 
   * Transformer models like BERT (2018), GPT, T5, and BART revolutionized NLP tasks, including summarization.
   * M-BERT (Multilingual BERT), mBART, and mT5 enabled efficient multilingual text processing without explicit translation.
   * AI-driven abstractive summarization gained popularity, allowing models to generate humanlike summaries instead of just extracting key sentences.

3

1. **Integration of Multilingual Summarization and TTS (Present and Future)** 
   * The combination of AI-powered text summarization and TTS provides a seamless news consumption experience.
   * AI models like Gemini, GPT-4, and LLaMA are making multilingual summarization more efficient and accessible.
   * Real-world applications include AI-powered news platforms, voice assistants, and accessibility tools for the visually impaired.
   * Future advancements aim for real-time, highly accurate, and personalized summarization with natural speech output.

### 2.2-LITERATURE REVIEW

**1) "Enhancing Text Summarization and Audio Generation Using Hybrid Model." Engineering Research Express (2025).**

**By Koreddi, Venkatesh, Shaik Chandini, BVT Kalyan Challa, and B. Sai Ram Teja.** In today's fast-paced world, efficiently processing the vast amount of information in lengthy documents poses significant challenges. Traditional methods of reading and understanding multipage documents, often spanning six to seven pages, are time-consuming and cumbersome for students, scholars, and professionals who face constant time constraints. This project introduces a novel approach by integrating advanced text summarization with real-time audio generation. Unlike existing solutions, which either focus on summarization or audio conversion separately, this dual-system solution uniquely combines both, allowing users to quickly grasp key content and consume it in audio form while on the move. The summarization component effectively condenses large documents into concise summaries, preserving the critical information and context. Meanwhile, the audio generation component transforms these summaries into speech, enhancing accessibility for auditory learners and users on the go. This innovative combination not only saves time but also offers flexibility, catering to diverse user preferences and learning styles.

**2) "A survey on extractive text summarization." IEEE Access 10 (2022):**

**By Moratanch, N., and S. Chitrakala.**

Text Summarization is the process of obtaining salient information from an authentic text document. In this technique, the extracted information is achieved as a summarized report and conferred as a concise summary to the user. It is very crucial for humans to understand and to describe the content of the text. Text Summarization techniques are classified into abstractive and extractive

4

summarization. The extractive summarization technique focuses on choosing how paragraphs, important sentences, etc produces the original documents in precise form. The implication of sentences is determined based on linguistic and statistical features. In this work, a comprehensive review of extractive text summarization process methods has been ascertained. This paper interprets extractive text summarization methods with a less redundant summary, highly adhesive, coherent and depth information

1. **"** **Performance study on extractive text summarization using BERT models." International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE) 4, no. 11 (2021).**

**By Abdel-Salam, Shehab, and Ahmed Rafea**

The task of summarization can be categorized into two methods, extractive and abstractive. Extractive summarization selects the salient sentences from the original document to form a summary while abstractive summarization interprets the original document and generates the summary in its own words. Deep learning has achieved promising performances in comparison to the classical approaches, and with the advancement of different neural architectures such as the attention network (commonly known as the transformer), there are potential areas of improvement for the summarization task. The

introduction of transformer architecture and its encoder model “BERT” produced an improved performance in downstream tasks in NLP.. The objective of this paper is to produce a study on the performance of variants of BERT-based models on text summarization through a series of experiments, and propose “SqueezeBERTSum”, a trained summarization model fine-tuned with the SqueezeBERT encoder variant, which achieved competitive ROUGE scores retaining the BERTSum baseline model performance by 98%, with 49% fewer trainable parameters.

1. **"** **Review of automatic text summarization techniques & methods."**

**Journal of King Saud University-Computer and Information Sciences 34, no. 4 (2022)**

**By Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik and Edi Noersasongko**.

This paper provides a broad and systematic review of research in the field of text summarization published from 2008 to 2019. There are 85 journal and conference publications which are the results of the extraction of selected studies for identification and analysis to describe research topics/trends, datasets, preprocessing, features, techniques, methods, evaluations, and problems in this field of research. The results of the analysis provide an in depth explanation of the topics/trends that are the focus of their research in the field of text summarization; provide references to public datasets, preprocessing and features.

5

**CHAPTER 3**

**Problem Statement**

6

## 3. Problem Statement

In today’s digital age, vast amounts of news articles are published across the globe in multiple languages, making it difficult for individuals to access and comprehend information beyond their native language. Traditional news aggregation systems either rely on direct translation, which often loses context and meaning, or require users to manually sift through lengthy articles. Additionally, the increasing volume of online news makes it challenging to quickly extract key insights.

The main challenges includes:

1. **Language Barrier in News Consumption:**

Many individuals cannot access global news due to language limitations. News articles are often published in regional languages, restricting their reach. Machine translation tools provide literal translations but fail to convey the full context. Misinterpretation of translated news can lead to misinformation and biased perspectives. A multilingual news summarization system can bridge this gap by generating easy-to-understand summaries in different languages

1. **Information Overload:**

The digital age has led to an overwhelming influx of news articles from multiple sources. Users do not have the time to read through lengthy articles to extract key insights. Manual summarization is impractical, especially for professionals needing quick updates. Irrelevant or redundant information often clutters news feeds, making it hard to focus on essential details.

**4.Challenges in Context Preservation:**

Extractive summarization methods often pull out key sentences but fail to create a coherent flow.Direct translations can alter the meaning, especially for complex sentences or idiomatic expressions.News articles often have implicit meanings, requiring contextual awareness for accurate summarization.

**5. Need for AI-Powered Automation:**

Manual summarization and translation are time-consuming and error-prone. AI models like BERT, mBERT, and T5 can process large volumes of text efficiently. Deep learning techniques improve the quality of summaries by ensuring fluency and relevance. Real-time summarization can benefit journalists, researchers, and global readers by providing instant insights.

7

**CHAPTER 4**

**Experimental Setup**

8

## 4.Experimental Setup

### 4.1 Hardware Setup

* **Memory:** 4GB RAM
* **Web-Browser:** Google / Opera / Mozilla Firefox / Microsoft Edge
* **Operating System:** Windows / MacOS / Linux distributions / iOS / Android. **• Network:** Wi-Fi / Internet (4G / 5G)

### 4.2 Software Setup

The development of the Multilingual News Summarizer Website involved the use of several software tools and technologies to create a dynamic and responsive web presence. The following is a summary of the software setup used for this project.

* **Code editor:** For this project, we utilized PyCharm, a powerful integrated development environment (IDE) specifically designed for Python programming.
* **HTML and CSS:** HTML and CSS were the core technologies for building the website. HTML5 was used to structure the content and define the layout, while CSS3 was employed for styling and layout design
* **JavaScript:** JavaScript is a client-side scripting language, which means that it runs on the user's browser rather than on a web server. It is used to create interactive and dynamic elements on websites. It allows developers to add interactivity, animations, and other dynamic features to web pages.
* **Libraries:**
* **transformers** – for using pre-trained multilingual summarization models like mBART, mT5, Pegasus, etc.
* **langdetect** – for automatic language detection of news articles.
* **newspaper3k** – for scraping and extracting clean news content from URLs.
* **nltk** – for basic NLP preprocessing like tokenization, stopwords removal, etc.
* **torch** – for running or fine-tuning deep learning models (if using PyTorch-based models).

9

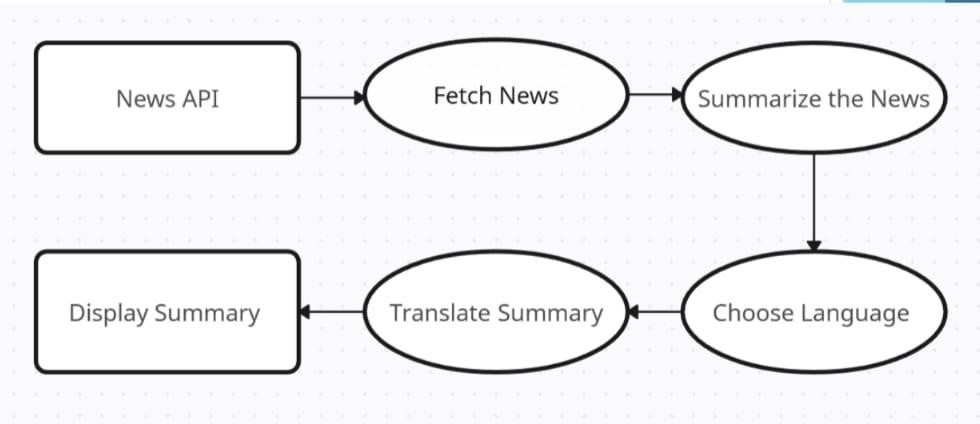
# CHAPTER 5

**Proposed System & Implementation**

10

## 5. Proposed system & Implementation

### 5.1 Block diagram of proposed system



5.1 (a) Block Diagram

### 5.2 Description of block diagram

The Description of block diagram can be interpreted as follows:

**1. News API (Source of News Articles):**

- The system fetches news articles from online sources using a News API. The API retrieves realtime news data from various categories such as politics, sports, technology, etc. The fetched news includes headlines, full article text, and metadata (e.g., source, timestamp) .

**2. Fetch News:**

- The system processes the API response and extracts the relevant news articles. The extracted news data is cleaned by removing unwanted characters, advertisements, and non-textual content.

**3. Summarize the News:**

-AI-based Natural Language Processing (NLP) techniques are used to generate a concise summary. The summarization can be extractive (selecting key sentences) or abstractive (generating new summarized text). Deep learning models like BERT, T5, or mT5 may be used to improve summary quality.

11

**4. Choose Language:**

- The user is given an option to select the target language for translation. A predefined set of languages is provided, allowing multilingual access to the summarized news.

**5. Translate Summary:**

- The summarized news is translated into the user-selected language using AI-based translation models. Transformer-based models like mBERT, MarianMT, or Google T5 are used to ensure accurate translation while maintaining context. The system ensures that the translated summary remains coherent and readable.

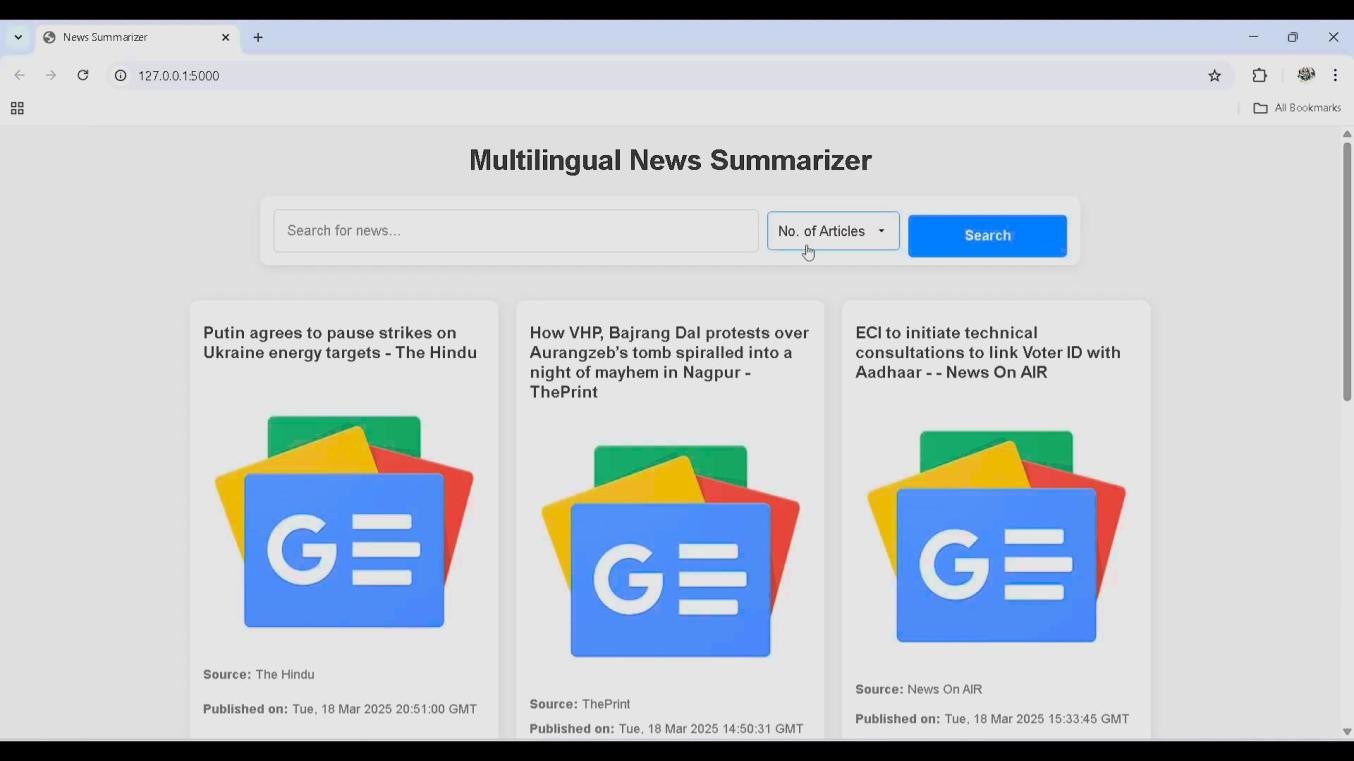
**6. Display Summary:**

- The final translated summary is displayed to the user in a simple and readable format. The user can read the news in their preferred language without losing important information. dditional features like text-to-speech (TTS) can be integrated to allow users to listen to the summarized news.

This workflow automates the process of Multilingual News Summarizer, helping user to read Summarize news in their preferred language, removing language barriers.

12

5.3 Implementation

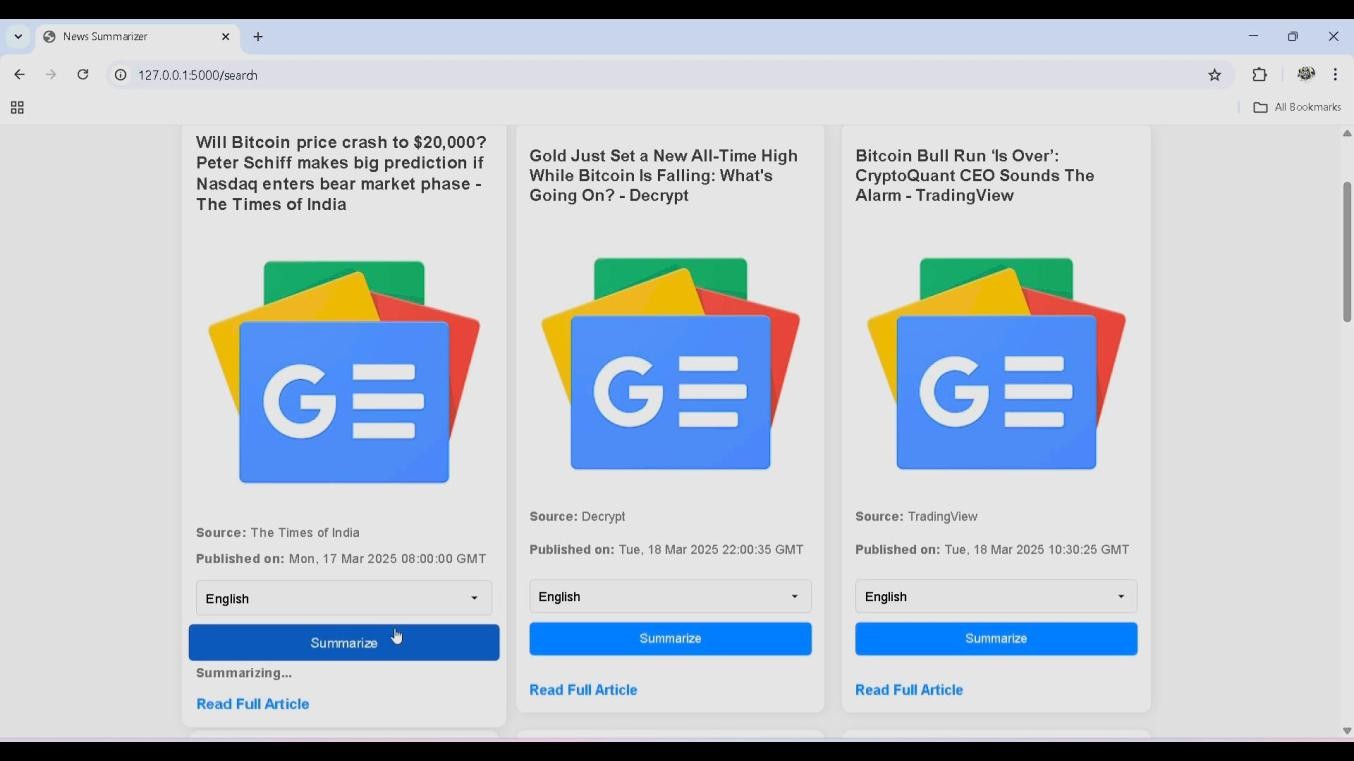


5.3 (A)

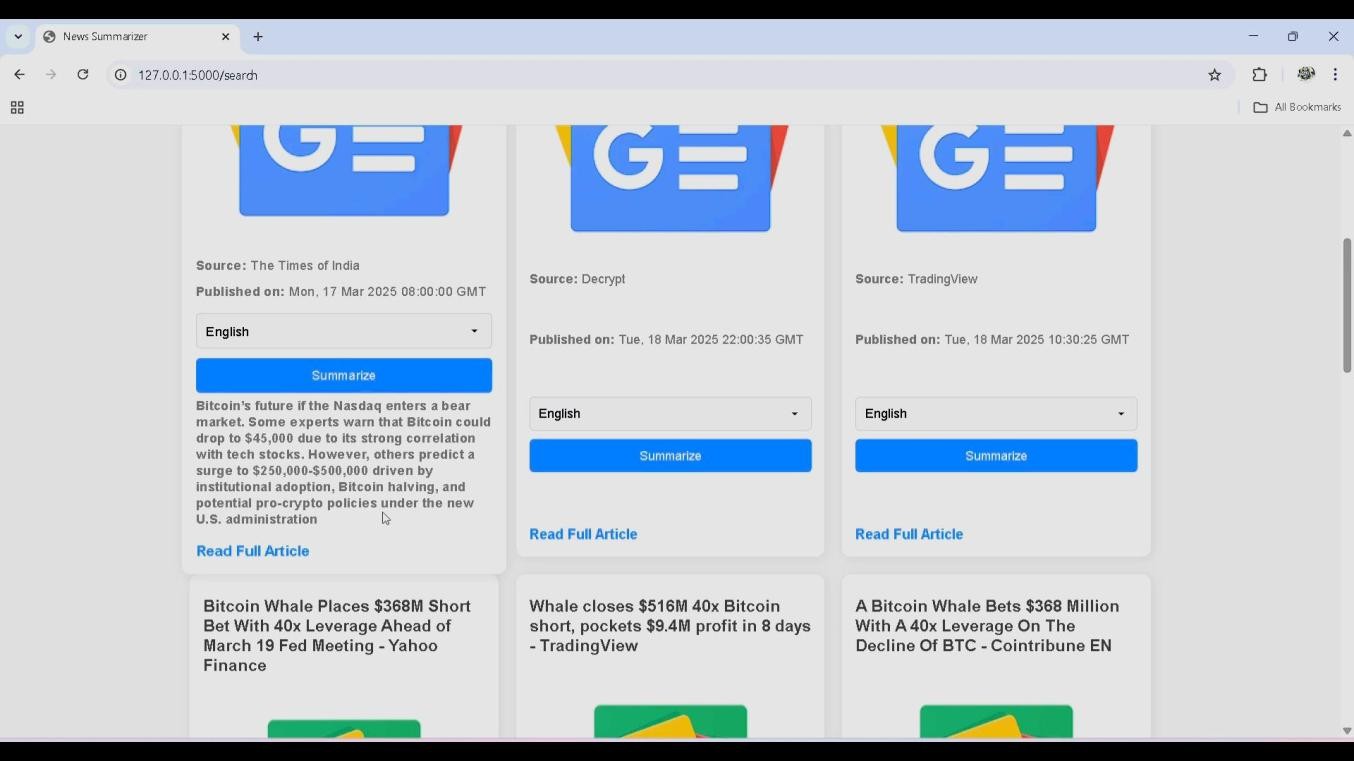


5.3 (B)

13



3(C)



5.3 (D)

14

**5.4 Advantages:**

Multilingual News Summarizer offers several advantages:

1. **Global Accessibility:** Enables users to read news in their preferred language, removing language barriers. Helps non-English speakers access international news effortlessly. Supports multilingual audiences, making news available to a diverse population.

1. **Time-Efficient Information Consumption:** Provides concise and relevant summaries, saving users from reading lengthy articles. Reduces information overload by extracting key insights from news articles. Helps professionals and researchers get quick updates on important topics.
2. **Improves News Understanding and Awareness:** Helps users grasp the essence of news quickly without language constraints. Reduces bias and misinformation by providing diverse language perspectives. Encourages cross-cultural understanding by making global news more accessible.

1. **Potential for Integration with Other Applications:** An be integrated into news platforms, mobile apps, and voice assistants. Text-to-Speech (TTS) can be added for listening to summarized news in multiple languages. Chatbots and smart assistants can use this system to deliver real-time news summaries.

These advantages collectively enhance enhances accessibility, efficiency, and engagement by providing concise, language-adaptive news summaries

15

**CHAPTER 6**

**Conclusion**

16

**6. Conclusion**

​The Multilingual News Summarizer using AI Techniques is an innovative solution designed to tackle information overload by delivering concise and accurate news summaries. Employing advanced Natural Language Processing (NLP) models, it distills essential information from diverse news sources, ensuring users receive relevant content without losing the original context. The system's translation feature allows individuals from various linguistic backgrounds to access news in their preferred languages, thereby breaking down language barriers and promoting global accessibility. Additionally, the integration of Text-to-Speech (TTS) technology enhances usability by enabling users to listen to summaries, catering to those with visual impairments or those who favor audio consumption. This project exemplifies the transformative power of artificial intelligence in automating and personalizing news consumption, with potential future enhancements including support for additional languages, real-time news analysis, sentiment detection, and personalized news recommendations.

17

### References

**Research paper**

[1 Koreddi, Venkatesh, Shaik Chandini, BVT Kalyan Challa, and B. Sai Ram Teja. "Enhancing Text Summarization and Audio Generation Using Hybrid Model." Engineering Research Express (2025). [2] Moratanch, N., and S. Chitrakala. "A survey on extractive text summarization." In 2017 international conference on computer, communication and signal processing (ICCCSP), pp. 1-6. IEEE, 2017.

1. Abdel-Salam, Shehab, and Ahmed Rafea. "Performance study on extractive text summarization using BERT models." Information 13, no. 2 (2022): 67.
2. Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, and Affandy Affandy. "Review of automatic text summarization techniques & methods." Journal of King Saud University-Computer and Information Sciences 34, no. 4 (2022).

18