"NEWS SCRAPING AND SUMMARIZATION USING NATURAL LANGUAGE PROCESSING TO TRANSFORM THE INFORMATION ACCESS"

[1]Mr.Suhas Bharadwaj R [2]Amith H D [3]Anand J [4]Bhargav M Gowda [5] Chinmaya Y N
[1] Dept.of IS&E, Assistant Professor, NIE Mysuru[2]B.E, Information Science & Engineering, NIE Mysuru [3]B.E, Information Science & Engineering, NIE Mysuru [4]B.E, Information Science & Engineering, NIE Mysuru [5] B.E, Information Science & Engineering, NIE Mysuru

Abstract

The project titled "News Summarizer and Ranking App" is an ambitious endeavor aimed at revolutionizing news consumption by offering a comprehensive solution for news retrieval and summarization. Integrated seamlessly with a state-of-the-art mobile application developed using Flutter, this project sets the stage for an unparalleled user experience. Powered by the versatility of Dart, the backend of the application showcases remarkable efficiency in aggregating diverse news articles into a unified platform.

At its core, the project boasts a sophisticated summarization module driven by Natural Language Processing (NLP), enabling the distillation of intricate content into concise and accessible summaries. Within the Flutter-based mobile app, users are greeted with an intuitive interface facilitating effortless browsing, interactive engagement through features like likes and comments, and real-time updates on the latest news.

Beyond its technical efficiency, this comprehensive system exemplifies user-centric design principles, offering not just streamlined functionality but also delivering an immersive and visually captivating experience. As users navigate through succinctly summarized news content, they are immersed in a holistic ecosystem that seamlessly blends technological innovation with a user-friendly interface, setting new benchmarks in the realm of news consumption applications.

I. INTRODUCTION

In the ever-evolving landscape of digital news consumption, the "News Summarizer and Ranking App" project aspires to redefine user interaction by seamlessly amalgamating efficient news retrieval, user engagement through advanced summarization techniques, and visually compelling presentation. The endeavor entails the development of a robust RESTful API, facilitating the streamlined aggregation of news articles from diverse sources while ensuring real-time updates and integration with various news outlets. Leveraging the power of Natural Language Processing (NLP), the project endeavors to distill lengthy news articles into concise and informative summaries, thereby fostering heightened user engagement. A pivotal aspect of the project involves the implementation of a dynamic ranking system, fueled by user interactions such as likes and comments, to prioritize and showcase articles with enhanced engagement levels. Additionally, an innovative image generation module complements this effort by creating captivating visuals based on news headlines. The mobile app, built on Flutter and designed with an intuitive interface, enables seamless navigation, empowering users to effortlessly browse, react to, and engage with summarized news content. Ultimately, the overarching goal of the project is to offer a comprehensive solution that reshapes how individuals interact with and stay informed about the latest news in the digital age.

II. RELATED WORK

In this research, we explored several state-of-the-art models and techniques across different domains. For ranking, we examined models such as BERT and its derivatives, which excel in understanding context for improved search relevance. In text-to-image generation, we utilized Stable Diffusion, known for its ability to generate high-quality images from textual descriptions. For summarization, we reviewed methods includng T5 model trainer and LLaMA LLM, both of which provide concise and coherent summaries. Additionally, our news scraping leveraged frameworks like BeautifulSoup and Scrapy to efficiently extract and process information from various sources. This comprehensive approach ensures our system is robust, accurate, and versatile.

III. PROBLEM STATEMENT

"News Summarizer and Ranking App" project addresses the challenges associated with traditional news consumption methods in the digital era. Despite the abundance of news content available online, users often struggle with information overload, making it difficult to stay informed efficiently. Existing news aggregation platforms lack sophisticated summarization techniques, leading to lengthy articles that may deter users from engaging with the content fully. Additionally, the absence of dynamic content ranking systems results in a lack of personalization, making it challenging for users to discover relevant and engaging news articles. These issues underscore the need for a comprehensive solution that integrates efficient news retrieval, sophisticated summarization algorithms, and dynamic content ranking mechanisms to redefine the digital news consumption experience.

IV. OBJECTIVES

The core objectives of our project, the "News Summarizer and Ranking App," are tailored to revolutionize the digital news consumption experience. Through the seamless integration of efficient news retrieval, sophisticated summarization techniques, and dynamic user engagement strategies, our aim is to redefine how users interact with news content. Specifically, we aim to develop a robust RESTful API for efficient news retrieval from diverse sources in real-time, ensuring continuous updates and seamless integration with various news outlets. Leveraging advanced Natural Language Processing (NLP) techniques, we intend to condense lengthy news articles into concise and informative summaries, thereby enhancing user engagement and comprehension. Furthermore, the implementation of a dynamic ranking system, fueled by user interactions such as likes and comments, will prioritize articles with increased engagement and relevance. Additionally, we will introduce an innovative image generation module to augment the visual appeal of the app, creating captivating images based on news headlines. Complementing these features, our user-friendly Flutter-based mobile application with an intuitive interface will facilitate effortless navigation and interaction with summarized news content. By accomplishing these objectives, our project aims to provide a comprehensive solution that redefines the digital news consumption landscape, empowering users with efficient, personalized, and engaging news experiences.

V.IMPLEMENTATION
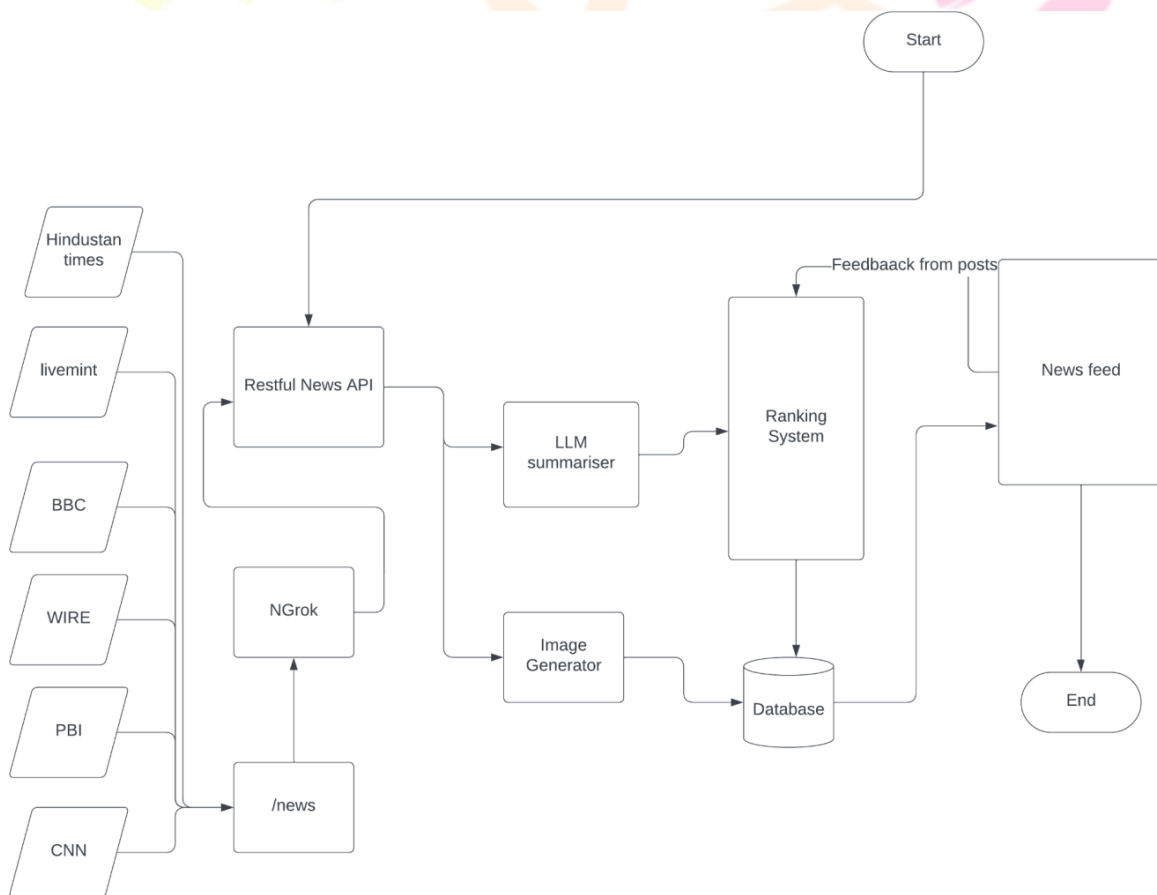
a) Block Diagram

Figure1: Block Diagram

News Scraping

The first step involves collecting news articles from three prominent sources: Hindustan Times, CNN, and Viacom. This process includes accessing their respective websites, parsing the HTML content, and extracting relevant news articles using web scraping techniques. By fetching data from multiple sources, the system ensures a diverse range of news topics for summarization.

Summarization Using LLM

Once the news articles are gathered, they undergo summarization using Large Language Models (LLMs) such as GPT-3.5. These models employ advanced natural language processing techniques to generate concise summaries while preserving the key information and context of the original articles. Summarization is essential to condense lengthy articles into digestible snippets for efficient consumption by users.

Ranking Algorithm

After summarization, the system implements a ranking algorithm to prioritize the summarized news articles based on their relevance, importance, and novelty. Various factors are considered in the ranking process, including the recency of the news, user preferences, trending topics, and the credibility of the news sources. This ensures that users receive a curated selection of top-ranked news articles tailored to their interests.

News Feed to User

Finally, the ranked and summarized news articles are delivered to the end-users through a personalized news feed. The news feed presents the articles in a user-friendly format, allowing users to quickly browse through the headlines and summaries. Additionally, the system may offer customization options for users to refine their news preferences and receive updates on specific topics or categories of interest. By delivering relevant and timely news content, the system aims to enhance the user experience and keep users informed about current events from diverse perspectives.

a) Hardware Specifications:

Core i5 processor from Intel
Memory: 8 GB
500GB HDD
Operating System: Windows 7 or later

b) Necessary Software:

1. Firebase: Purpose: Backend infrastructure User: Real-time database for storage and retrieval, Secure user authentication, Analytics for user behavior tracking, Cloud functions for serverless processing

2. Ngrok: Purpose: Create secure tunnels to localhost Uses: Facilitates testing of webhooks and API integrations, Enables remote access to local applications for collaboration and testing.

3. Android Studio: Purpose: Primary IDE for Android development Uses: Comprehensive tools for building, testing, and debugging, Emulator for testing on virtual devices, Seamless integration with Firebase, Robust UI design tools.

4. Kaggle: Purpose: Data analysis and machine learning Uses: Data processing with vast datasets, Machine learning for sentiment analysis and summarization, Collaboration and code sharing within the Kaggle community.

VI. CONCLUSION AND FUTURE ENHANCEMENT

a) CONCLUSION

In conclusion, the News Summarizer and Ranking App project represents a significant advancement in the realm of digital news consumption. While existing systems rely on conventional news aggregation platforms and traditional summarization techniques, our proposed system introduces cutting-edge approaches to enhance the depth, breadth, and user engagement in the news consumption process.

By leveraging advanced summarization techniques, particularly the BERT model for abstractive summarization, our system aims to provide users with more contextually rich and coherent news summaries. The inclusion of web scraping techniques further ensures a comprehensive collection of news articles from diverse sources, offering users a more holistic view of current events.

The dynamic content ranking system, driven by user interactions such as likes, shares, and comments, sets our proposed system apart from existing ones. This personalized approach to content ranking ensures that the most engaging and relevant news articles rise to the top, tailoring the news feed to individual user preferences.

Additionally, the integration of the Flutter mobile application enhances the user experience, providing a visually appealing and user-friendly interface across different platforms. Addressing the challenge of information overload, the system aims to strike a balance between depth and conciseness in news summaries, offering users an efficient and personalized news consumption experience.

Through these innovations, the News Summarizer and Ranking App project strives to redefine the way individuals interact with news content, providing a comprehensive and user- centric platform in the evolving landscape of digital news consumption. The project's ultimate goal is to contribute to a more informed, engaged, and personalized news-consuming community.

b) FUTURE ENHANCEMENT

Multilingual Translation
1. Purpose:
    To break down language barriers, improve user experience, and cater to a global audience by translating news articles into various languages.

2. Implementation:
   - Translation APIs: Integrate APIs like Google Translate or DeepL for reliable translations.
   - Backend Processing: Translate scraped articles and store them alongside originals.
   - User Interface: Include options for users to select their preferred language.
   - Quality Assurance: Implement a review system for users to provide feedback on translation quality.

3. Benefits:
   - Accessibility: Enables non-native speakers to understand content.
   - Engagement: Attracts a diverse user base, enhancing engagement.
   - Global Reach: Positions the application as a global platform.

Public Release on Google Play Store

1. Purpose:
    To increase visibility and accessibility by reaching millions of potential users through the Google Play Store.

2. Preparation:
   - Optimization: Ensure compatibility with various Android devices.
   - Compliance: Adhere to Google Play Store policies and guidelines.
   - User Testing: Conduct extensive beta testing and incorporate user feedback.
   - Marketing: Develop a comprehensive marketing strategy for promotion.

VII. REFERENCES

[1] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[2] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.

[3] P. N. Varalakshmi K and J. S. Kallimani, "Survey on Extractive Text Summarization Methods with Multi Document Datasets," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 2113-2119, doi: 10.1109/ICACCI.2018.8554768.

[4] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006

[5] Richardson, L. (2007). Beautiful soup documentation April.

[6] Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Veselin Stoyanov and Luke Zettlemoyer (2019). BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. CoRR, abs/1910.13461.

[7] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin (2017). Attention Is Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin (2017). Attention Is All You Need. CoRR, abs/1706.03762. Lukasz Kaiser and Illia Polosukhin (2017). Attention Is All You Need. CoRR, abs/1706.03762.

[8] Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Jamie Brew (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing. CoRR, abs/1910.03771.

[9] Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu (2019).Exploring the Limits of Transfer Learning with a UnifiedText-to-Text Transformer. CoRR, abs/1910.10683.

[10] Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. ACL 2004.