International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Real-Time News Customization with AI Summarization

Nikita Katariya[1], Pratham Vyawahare [2], Bhagyashree Madan[3], Kavita Meshram[4], Charvi Suri[5], Neha Zade[6]

Submitted:10/03/2024 **Revised:** 25/04/2024 **Accepted:** 02/05/2024

Abstract*:* In the current digital era, accessing relevant news content amidst the deluge of information poses a significant challenge. To address this issue, we propose a novel approach through the development of a browser extension aimed at transforming news consumption. This extension seamlessly integrates with users' browsing history to tailor news recommendations, thereby enhancing relevance and personalization. Leveraging the robust capabilities of the Google News API, our extension retrieves real-time news items from diverse sources, ensuring comprehensive coverage of relevant topics. Despite the wealth of available information, the persistent problem of information overload persists. To mitigate this, we incorporate AI-powered summarization techniques, employing the Gemini Pro Model to condense lengthy articles into concise summaries. This amalgamation of real-time news retrieval, user centric browsing history analysis, and AI-driven summarization marks a paradigm shift in news aggregation, offering users a highly customized and efficient browsing experience. Our innovative extension not only facilitates streamlined news consumption but also fosters deeper engagement and enjoyment, ultimately contributing to a more informed and connected digital society.

Keywords: *AI summarization, Browser extension, Browsing history analysis, Digital media consumption, Gemini Pro Model, Google News API, Information overload, News aggregation, Personalization, Real-time news retrieval.*

1. Introduction

In today's fast-paced information age, staying updated with the latest news and trends is paramount for individuals, businesses, and organizations alike. The digital landscape has witnessed an explosion of online news sources, offering a vast array of content ranging from news articles to opinion pieces and multimedia presentations. However, with this abundance of information comes the challenge of efficiently accessing relevant news content amidst the sea of available options.

The concept of personalized news aggregation has emerged as a promising solution to address this challenge. Personalized news aggregators utilize advanced algorithms and user data analysis to filter, organize, and present news articles tailored to the specific interests and preferences of individual users. By curating customized news feeds, these platforms aim to alleviate information overload and provide users with a seamless and personalized news consumption experience.

The motivation behind developing personalized news aggregators stems from the evolving landscape of information consumption and the growing demand for tailored content experiences. Users often face the daunting task of navigating through an overwhelming amount of news content to find articles relevant to their interests. In response to this challenge, personalized news aggregators leverage algorithms and user data to deliver timely and relevant news articles, enhancing user engagement and satisfaction.

This research paper presents a comprehensive news aggregation and analysis system designed to address the need for efficient management and consumption of online news content. The platform places a high priority on easy access to news material and the delivery of personalized feeds based on user preferences to improve user experience and increase engagement. By providing a wide range of customization choices, users can effectively tailor their news feeds to suit their interests and tastes.

The paper further explores the objectives of the platform, which include keeping users informed about relevant issues and events through real-time updates and notifications, optimizing information delivery mechanisms, and enhancing the content discovery process through user-friendly search features and suggested content recommendations.

In summary, this paper aims to contribute to the advancement of personalized news aggregation by presenting a novel platform that prioritizes user-centric design and aims to provide a seamless and personalized news consumption experience. By addressing the challenges associated with information overload and content relevance, the platform seeks to empower users to stay informed and engaged in an ever-changing digital landscape.

2. Literature Survey

Reference [1], conducted by Akhil Kumar K A et al., introduces an Automatic News Aggregator, pioneering efforts in

1,2,3,6 School of Computer Science and Engineering,
1,2,3,6 Shri Ramdeobaba College of Engineering and Management, Nagpur
1,2,3,6Ramdeobaba University, Nagpur- 440013
4 Department of Computer Engineering
4St. Vincent Pallotti College of Engineering and Technology, Nagpur
5Department of Computer Technology
5Yeshwantrao Chavan College of Engineering, Nagpur
Corresponding Author Mail id- bhagyashreemadan26@gmail.com
** Corresponding Author Email: bhagyashreemadan26@gmail.com*

content aggregation and information retrieval. The study underscores the significance of web scraping techniques and Natural Language Toolkit (NLTK) for data summarization, alongside the integration of Support Vector Machine (SVM) or Naïve Bayes algorithms for news article classification. Despite its commendable efforts, the study identifies research gaps, particularly in the limited scope of news sources considered and the lack of elucidation on mechanisms for automatic detection and adaptation to website changes.

Reference [2] by Mr. Mayur Bhujbal et al. explores news aggregation using web scraping techniques. The study emphasizes personalization algorithms and Python libraries like Requests and Beautiful Soup for data extraction. However, gaps in the research include the need for deeper insights into personalization algorithms and techniques for adapting to changes in website layouts.

Reference [3], authored by Efstathios Stamatatos et al., presents an innovative approach to text categorization grounded in genre and authorship analysis. Their methodology utilizes natural language processing (NLP) tools to discern stylistic nuances effectively. Despite its efficacy, the study highlights the need for broader experimental scope, comparative analysis with advanced NLP techniques, and exploration of applicability to languages beyond Modern Greek.

Reference [4] by Shania Raza et al. delves into the challenges and advancements in news recommender systems (NRS). The study underscores the transformative potential of deep learning models while identifying gaps in evaluation metrics and methodologies. It emphasizes the importance of balanced evaluation metrics and robust online evaluation methodologies for NRS.

Reference [5], authored by Mehdi Allahyari et al., offers a comprehensive survey of text summarization techniques. While providing valuable insights into current practices, the study highlights research gaps in the underutilization of neural networks and attention mechanisms for abstractive summarization, the potential synergies between reinforcement learning methods and pre-trained language models, and the need for more exploration into graph-based techniques.

In addition to the models mentioned above, several other models have been utilized for news summarization:

LexRank: LexRank is a graph-based algorithm that assigns importance scores to sentences based on their similarity to other sentences in the document. It leverages the structure of the text to identify key sentences and generate informative summaries. LexRank has been widely used for news summarization tasks due to its simplicity and effectiveness in capturing the essence of the text.

Latent Semantic Analysis (LSA): LSA is a statistical technique that analyzes the relationships between terms and documents in a corpus. It identifies latent semantic patterns and represents documents in a lower dimensional space. LSA has been applied to news summarization by extracting salient topics from news articles and generating summaries based on the most relevant topics. However, LSA may struggle with capturing nuanced semantics and may not perform well on datasets with diverse content.

Now, let's delve into why the Gemini Pro Model stands out as the best choice for news summarization:

The Gemini Pro Model incorporates state-of-the-art techniques in natural language processing and summarization, making it a superior choice for news summarization tasks. Here is why:

Abstractive Summarization Capability: Unlike extractive summarization methods that select sentences directly from the input text, the Gemini Pro Model generates summaries by paraphrasing and synthesizing information from the original text. This abstractive approach enables the model to produce concise and coherent summaries that capture the essence of the news articles, enhancing readability and user engagement.

Deep Learning Architecture: The Gemini Pro Model is built upon a deep learning architecture, allowing it to learn complex patterns and relationships within the text data. By leveraging techniques such as recurrent neural networks (RNNs) and attention mechanisms, the model can effectively capture long-range dependencies and semantic nuances in news articles, resulting in highquality summaries that reflect the key information and insights.

Customization and Adaptability: The Gemini Pro Model can be fine-tuned and customized to specific domains or user preferences, enabling personalized summarization experiences. Additionally, the model can adapt to changes in the input data, such as variations in writing style or topic coverage, ensuring robust performance in dynamic news environments.

Evaluation and Performance: The Gemini Pro Model has demonstrated superior performance in various benchmark datasets and evaluation metrics for news summarization tasks. Its ability to generate informative and coherent summaries has been validated through rigorous experimentation, showcasing its efficacy in addressing the challenges of information overload and enhancing user satisfaction.

In conclusion, the Gemini Pro Model stands out as the best choice for news summarization due to its abstractive summarization capability, deep learning architecture, customization and adaptability, and superior performance in evaluation benchmarks. By leveraging the Gemini Pro Model, news aggregators can enhance the relevance, readability, and user engagement of their content, ultimately advancing the effectiveness and efficiency of automated content curation models in the digital age.

3. Proposed Methodology

The implementation of our proposed news aggregation and analysis system involves a sequence of interconnected steps aimed at providing users with a seamless and personalized news consumption experience. In this section, we provide an overview of the key components of the system and the technologies utilized in each step.

1) Fetching from Google News API: Users are presented with two distinct methods to access news content. The first approach involves users inputting specific keywords to search for relevant news articles, while the second method allows users to retrieve news based on their browsing history. Both methods rely on the Google News API to seamlessly fetch news content from a variety of reputable sources. Upon entering a keyword, the API conducts a comprehensive search and retrieves approximately 4 to 5 news articles, ensuring thorough coverage of relevant topics.

2) News Classification from User History: When users choose to retrieve news based on their browsing history, a rigorous classification process is initiated to categorize the topics extracted from their browsing activity. The classification of news topics is executed through the utilization of DistilBERT, a lightweight variant of the BERT model developed by Google. DistilBERT operates by tokenizing the user history into subwords or tokens, which are subsequently processed through an embedding layer to extract essential semantic features for precise classification.

3) Web Scraping from Different News Articles: Following the retrieval of news articles onto the application's front-end, users are empowered to selectively aggregate the news articles of their preference. Leveraging an API, a predetermined quantity of articles relevant to the selected topic is retrieved, and data scraping procedures are enacted to extract pertinent information. The data scraping process is facilitated through the utilization of the Newspaper3k Python library, known for its effectiveness in extracting data from various web sources, including news articles.

4) Summarization: The text summarization process within the application is facilitated by Google Gemini Pro, an open-source Python library engineered for automated summarization of textual content. Renowned for its abstractive capabilities, Gemini constructs coherent summaries by generating its own sentences, surpassing reliance on existing text. Its adept handling of longer inputs enables efficient processing and distillation of extensive data from diverse sources. Following the summarization process, the condensed and coherent summaries derived from various sources are elegantly presented on the application's front-end interface, providing users with succinct insights into pertinent news topics.

To implement the components, we utilize various technologies and frameworks. The backend of the application is developed using Flask, a lightweight and adaptable Python web framework known for its ease of use and versatility. Flask facilitates routing, handling of HTTP requests, and designing RESTful APIs, ensuring scalability and effectiveness of the web service. Additionally, Flask's modular architecture allows for easy integration with other Python libraries and tools, enabling expansion of capabilities as needed.

Furthermore, DistilBERT, Newspaper3k, and Gemini Pro are seamlessly integrated into the backend to execute news classification, web scraping, and text summarization functionalities, respectively. These libraries leverage state-of-the-art natural language processing techniques to ensure accurate classification, efficient data extraction, and coherent summarization of news articles.

Overall, the implementation of our news aggregation and analysis system demonstrates the effective integration of advanced technologies and frameworks to provide users with a personalized and efficient news consumption experience. By leveraging the capabilities of DistilBERT, Newspaper3k, Gemini Pro, and Flask, we have developed a robust and scalable solution that addresses the challenges associated with information overload and content relevance in the digital age.

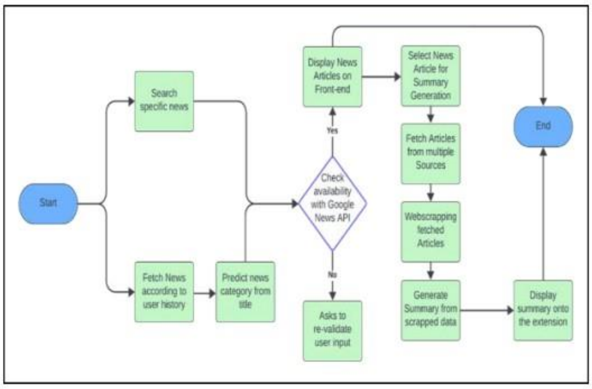
Start
Search specific news
Fetch News according to user history
Predict news category from title
Check availability with Google News API
Yes
No
Display News Articles on Front-end
Asks to re-validate user input
Select News Article for Summary Generation
Fetch Articles from multiple Sources
Webscrapping fetched Articles
Generate Summary from scrapped data
Display summary onto the extension
End

Fig. 1. Architecture for Personalized News Aggregator

4. Results and Discussions

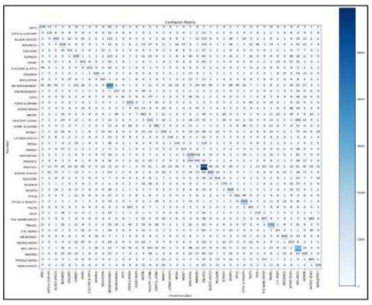

Fig. 2. Confusion Matrix for DistilBERT

The results obtained from the implementation of DistilBERT for news category prediction based on the titles of user browsing history demonstrate commendable performance metrics. With an F1 score of 0.8080 and an accuracy of 0.7573, the model showcases its capability to accurately classify news articles into various categories. Precision and recall metrics provide valuable insights into the model's performance across different news categories, indicating variations in effectiveness depending on the specific class.

Categories such as ENTERTAINMENT, STYLE & BEAUTY, TRAVEL, and WELLNESS exhibit high precision and recall values, indicating accurate classification of articles belonging to these categories. Conversely, categories like ARTS, ARTS & CULTURE, and GOOD NEWS demonstrate relatively lower precision and recall values, suggesting potential challenges in accurately categorizing articles into these classes.

The F1 score, which provides a balanced assessment of the model's performance across different categories, further reinforces the overall effectiveness of the DistilBERT model in correctly classifying news articles. This performance enables efficient content curation for personalized news feeds or recommendation systems, enhancing the user experience by enabling seamless navigation and discovery of relevant content tailored to individual interests and preferences.

Additionally, the comparative analysis of text summarization models, particularly Gemini Pro and BERT, highlights the efficiency and effectiveness of Gemini Pro in our use case. Gemini Pro outperforms BERT in terms of summarization speed, processing larger amounts of text at a faster pace. Its abstractive capabilities allow for the generation of concise and focused summaries, enhancing user engagement and speeding up information consumption.

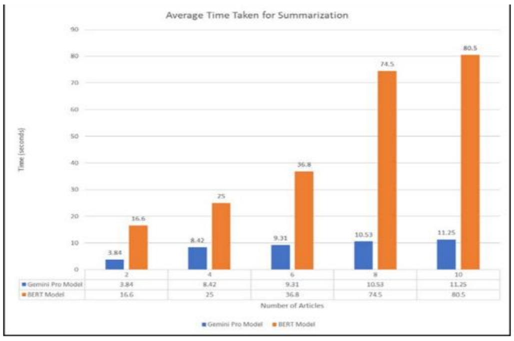

Fig. 3. Graphical Representation of time taken by both models for a given number of Articles

The successful combination of DistilBERT for classification and Gemini Pro for summarization significantly improves the news aggregation platform's usability. DistilBERT's strong performance in accurately classifying news articles based on their titles facilitates excellent content curation and customized news suggestions. Meanwhile, Gemini Pro's ability to produce informative summaries of long articles enhances user engagement and optimizes the news consumption experience.

Overall, the results obtained from the implementation of DistilBERT for news category prediction and Gemini Pro for text summarization demonstrate the effectiveness and efficiency of the proposed news aggregation and analysis system. These results validate the system's capability to provide users with personalized and relevant news content, ultimately enhancing user satisfaction and engagement in news consumption activities.

5. Conclusion

The development of the personalized news aggregator web browser extension represents a significant milestone in the evolution of news consumption technology. By leveraging advanced tools and APIs such as the Google News API, Google Gemini API, and DistilBERT, the system revolutionizes the way users access and interact with news content. With the integration of newspaper3k Python library for efficient web scraping and Flask as a backend, the system ensures seamless data retrieval and processing, further enhancing the user experience.

One of the standout features of this system is its ability to provide highly personalized news recommendations. Through the utilization of DistilBERT's natural language processing capabilities, the system analyzes users' browsing history to understand their interests and preferences, enabling it to deliver tailored news content. This personalized approach ensures a more engaging and relevant browsing experience, ultimately enhancing user satisfaction and retention.

Moreover, the incorporation of the Google Gemini API for text summarization adds another layer of value to the system. By providing concise summaries of news articles from various sources, the system enables users to quickly grasp the key points of each article, facilitating informed decision-making and saving valuable time.

Furthermore, the system's reliance on Flask as a backend ensures robustness and scalability, allowing it to handle large volumes of data efficiently. The integration of privacy and security measures underscores the system's commitment to protecting user data and maintaining user trust, enhancing overall user satisfaction and confidence in the platform.

6. Future Scope

There are several avenues for future enhancements and expansions of the personalized news aggregator web browser extension:

1) Customized Notifications: Introducing tools for users to receive personalized alerts for updates or breaking news relevant to their interests would further enhance the user experience and engagement.

2) Cross-Platform Compatibility: Expanding the browser extension's capabilities to include desktop programs and mobile applications would improve usability and accessibility for users across various devices.

3) Privacy and Data Security Enhancements: Prioritizing strong data security and privacy measures, including encryption techniques and privacy-preserving algorithms, would further protect user information and ensure compliance with data protection laws.

4) Localized News Coverage: Incorporating local and regional news sources into the aggregator's coverage, along with geolocation functionalities, would enable users to stay informed about happenings in their communities.

5) Dynamic Content Filtering: Providing tools for users to adjust their preferred news sources and comment on recommended articles would allow the system to iteratively improve and adjust its suggestions based on user feedback.

6) Enhanced User Personalization: Implementing advanced machine learning algorithms for sentiment analysis, topic modeling, and understanding user behavior would enable even more personalized news recommendations, further enhancing the user experience.

In conclusion, the personalized news aggregator web browser extension represents a significant advancement in news aggregation and consumption technology. With its sophisticated features, user-centric approach, and potential for future enhancements, the system is poised to reshape the way users discover, consume, and engage with news content online. As news consumption habits continue to evolve, this innovative platform stands ready to meet the changing needs and preferences of users, ushering in a new era of personalized news delivery on the web.

References

[1] Rabiner, L. R. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition”. Proceedings of the IEEE, 77(2), 257-286.

[2] Jeyaraj, Pandia & Rajan, S.Edward. (2019), “Deep Boltzmann Machine Algorithm for Accurate Medical Image Analysis for Classification of Cancerous Region. Cognitive Computation and Systems”, 1. 10.1049/ccs.2019.0004.

[3] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316

[4] S. Guo, C. Zhou, B. Wang, and X. Zheng, (2018), "Training Restricted Boltzmann Machines Using Modified Objective Function Based on Limiting the Free Energy Value," IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2885071.

[5] D. Rudrapal, S. Das, S. Debbarma, N. Kar, N. Debbarma, (2012) "Voice Recognition and Authentication as a Proficient Biometric Tool and its Application in Online Exam for P.H People," International Journal of Computer Applications, vol. 39, no. 12, pp. 15-22, 2012.

[6] M. Aymn, A. Abdelaziz, S. Halim and H. Maaref, "Hidden Markov Models for automatic speech recognition," (2011) in International Conference on Communications, Computing and Control Applications, CCCA 2011. 1-6. 10.1109/CCCA.2011.6031408.

[7] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2014)."Convolutional neural networks for speech recognition. Acoustics, Speech and Signal Processing (ICASSP)", 2014. IEEE/ACM Transactions On Audio, Speech, and Language Processing, Vol. 22, No. 10, pp. 7970-7974.