Image-based News Aggregator Using OCR and NLP for Summarization

Santosh Reddy P, S S Sanjan†, Spandhana K Devadiga*, Vinati Thakkar†

[1,2,3,4*]Dept. of Computer Science and Engineering, BNMIT Institute of Technology, Bangalore, Karnataka, India.

Contributing authors:
santhoshreddy@bnmit.in;
sanjanass2702@gmail.com;
spandhana050604@gmail.com;
vinati.thakkar1@gmail.com;
†These authors contributed equally to this work.

Abstract: The development of digital information raises the demand for insight extraction from large data in the shortest possible time. Users suffer from inability to keep updated due to growing online news content and time limits. There-fore, this solution combines text extraction using OCR (Tesseract) in newspaper images, real-time news using NewsAPI.org, and abstractive summarization. This approach condenses articles into compact, easily understandable form. Sum-maries can be turned into audio using TTS tools like gTTS or pyttsx3 to increase accessibility. These combined technologies provide news faster, personalized, and easily digestible to the users without actually reading full articles.

Keywords: —OCR, NLP, Summarization, Image Processing, News Aggregator, Text Extraction, Tesseract, Transformers

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_40

1 Introduction

In this increasingly expanding digital landscape, the amount of data on the internet is also increasing. But accessing the correct data efficiently and quickly can be challeng-ing. Hence the ability to automatically extract and process textual the required image has become increasingly important. And in today's fast phased world, people don't have enough time t oread an entire newspaper or manually search through multiple websites to find accurate and relevant news article only to end up not getting the cor-rect information. With the overwhelming amount of data online, locating trustworthy content quickly has become a major challenge.

Hence summarization can be helpful by selecting only the significant information from the article which sufficient information. Using technologies like Optical Character Recognition (OCR) combined with NewsAPI.org API, along with text summarization techniques and Text-to-Speech (TTS) feature that converts the summarized news text into audio, offers an ideal solution to this problem. By automating the process of extracting news content directly from the newspaper images, based on the title collecting information through NewsAPI.org API and summarizing the information collected.

Optical Character Recognition (OCR) is a technology allows machine to recognize text from images, like scanned document or photos. Various types of OCR exist, includ-ing printed text recognition, handwritten text recognition, and intelligent character recognition (ICR), each suited to different input complexities.

Application Programming Interface (API) is a set of rules and tools that facil-itates communication between to software application like exchanging data, service and functionalities. APIs define the methods and data formats that applications can use to request and exchange information. The API allows users to search for articles based on specific keywords, topics, or even sources, making it a highly flexible tool for fetching news content. API is helpful in streamline the process of gathering and integrating news content into the system.

Text Summarization is the process of reducing a large amount of data and retain-ing only the significant and relevant information. Instead of spending a lot of time on reading the whole content, get the relevant information in minutes saving a lot of time. There are two types of test summarization tech niques Extractive and Abstractive. In Extractive summarization which works by directly selecting important sentences, phrases or section from the original text and adding them to form summary. In Abstractive summarization involves generating new sentences that convey the core information of the original data.

Text-to-Speech technology helps in converting the summa rized text into audio, which increases the accessibility and usability of the proposed solution. By integrating TTS, the system enhances user accessibility, broadens its usability, and ensures that news content is available in both text and audio forms for greater convenience.

2 Related Work

Liu [1] proposed a BERT-based extractive summarization model that treats sentence selection as a sequence classification task. By fine-tuning pre-trained BERT with a summarization-specific layer, the model achieved high ROUGE scores on CNN/-DailyMail datasets, demonstrating the capability of contextual embeddings to retain semantic relations and key information without redundancy. Extending this idea, Liu and Lapata [2] introduced a unified framework combining extractive and abstractive summarization using transformer-based language models. Their method fine-tuned BERT-like encoders to improve coherence and informativeness and performed well even on noisy input, such as OCR-extracted headlines.

Attention-based approaches have also been explored. Gambhir and Gupta [3] developed a word-level attention-based deep learning model for extractive summariza-tion, where the attention mechanism enhanced sentence selection based on semantic salience, proving suitable for large-scale tasks involving noisy OCR text. Khatri et al. [4] proposed a hybrid summarization framework using document context vec-tors and RNNs to generate coherent summaries, particularly effective for processing context-rich headlines from visual news sources.

Methods based on embeddings have also been widely investigated regarding deal-ing with robustness related to noisy or incomplete text input data. An extractive summarization technique that uses sentence embeddings and scoring functions to cap-ture salient content has been shown to work well when tested on noisy OCR text input by Sinha et al. [5]. Another technique that tackles robustness related to contex-tual features has been proposed by Dhakras et al. in their BoWLer model in [6] when it comes to partially captured headline information. In another research, the author Hennig [7] used Probabilistic Latent Semantic Analysis for Multi-document Summa-rization, which allowed for the identification of latent themes even in cases where only a small set of annotated data was available. In addition, the work done by Abdel-Salam and Rafea in [8] further evaluated BERT-based models for extractive summarization with a focus on coherence and readability in the evaluation criteria based on ROUGE measures, for low-quality OCR results.

Apart from the embedding approaches, there have been attempts on hybrid approaches as well as topic-driven approaches. This is because, in addition to using embeddings, a hybrid approach can also be designed, especially if there is a need to adapt to irregular data that may be encountered in the real world, as in the case of news headlines captured through an image. This was pro-posed by author Pai in his study [9]. Another study by Hidayat et al. in [10] combined latent Dirichlet allocation. In addition, Gupta and Lehal proposed an elaborate sur-vey on the methods for performing extractive summarization that included rule-based methods, graph-based methods, and machine learning-based methods, thus providing crucial insights into the ways and means for choosing suitable approaches for headline-based summarization. Recently, graph-related models have emerged prominently in extractive

summarization tasks. Jia et al. [12] proposed a hierarchical attention heterogeneous graph network that incorporates both syntactic and semantic structural information to reinforce the learning of representations, which is particularly impor-tant in dealing with noisy structural data, like OCR-extracted texts. Similar efforts were seen with Jing et al. [13], who proposed a multiplex graph-based neural network that could effectively leverage multiple semantics, which was valuable in producing coherent summaries from short and fragmented headline segments, usually seen in text extraction from images.Methodology

The proposed newspaper summarization system follows a modular and structured design, integrating multiple technologies including Optical Character Recognition (OCR), NewsAPI-driven content retrieval, Natural Language Processing (NLP), and Text-to-Speech (TTS). This framework enables the conversion of image-based news-paper content into concise and meaningful summaries, thereby improving accessibility, efficiency, and overall user experience. The system architecture is organized into six distinct functional modules, each responsible for a specific stage in the processing pipeline, as described in the following sections.

2.1 Input Acquisition and Preprocessing

The system begins with the acquisition of input data, which may be a newspaper image, scanned document, or plain text f ile. Users can upload PDFs or image formats. Preprocessing includes grayscale conversion, noise removal, thresholding, and resizing to enhance OCR accuracy, especially in docu ments with complex layouts and fonts.

2.2 Text Retrieval using OCR

The preprocessed image is processed using Tesseract OCR, which extracts textual data from visual media. It captures headlines, subheadings, and article bodies across multiple columns and font types. The raw text output may contain one or multiple articles or headlines.

2.3 Title Detection and Contextual Content Enrichment

To improve context for summarization, key titles and key words are extracted from the OCR output. These are used as queries to NewsAPI.org, which returns the most recent and relevant news content. The combined dataset—OCR results plus API-enhanced articles—forms a robust input for summa rization.

2.4 Text Summarization

The main theme of this project is text summarization as shown in Fig. 1 .The goal is to summarize lengthy articles or several news sources into a concise, informative summary keeping the most important information. The method employed is a

hybrid summarization method that integrates both extractive and abstractive techniques to produce results of better quality.

2.4.1 Extractive Summarization

This involves word-for word copying of important sentences, phrases, or pas sages from the original work. Various approaches are attempted: Graph-based and ranking methods such as TextRank, LSA, and Latent Dirichlet Allocation (LDA) have been used very widely [10], [11], [7]. Recent graph neural networks (GNNs), i.e., hierarchical or multiplex GNNs [12], [13] and syntactic compression techniques [18], significantly

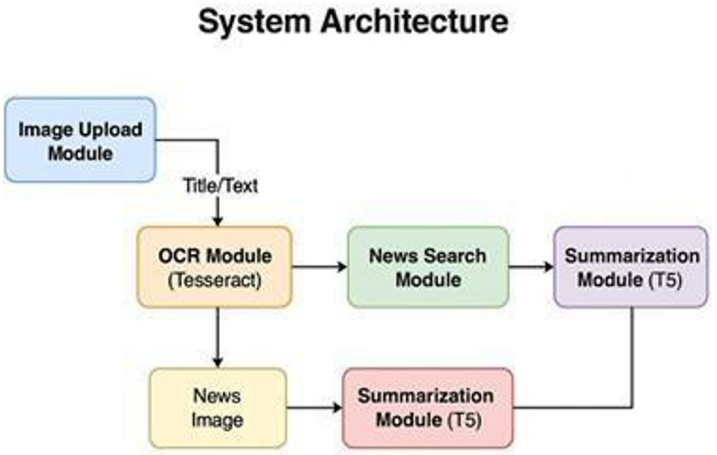

Fig. 1 Interface of the News Aggregator

improved extractive sum marization content quality. Rank fusion and submodular optimization and language models further refine the extraction process [14], [15].

These models guarantee that the summary holds valuable information without modifying the author's original word choice, although they can lack fluency and readability

2.4.2 Abstractive Summarization

Techniques of abstractive summarization involve the use of deep learning approaches for the generation of summaries by reformulating the source text in new manually readable sentences by constructing these sentences using the sourced text by deep learning approaches like BART, PEGASUS, or T5 [2], [16], [19]. It uses the long contextual dependencies of the sentences for better syntactical reconstruction, hence better quality

of generated summarizations but at the cost of heavy computations required, requiring large training data [19].

Recent approaches involving the use of cascade forests [20] or the extraction of the latent features of the text for better results in the generated summary [17] are still being explored for usage in hybrid approaches for summarization.

2.5 Text-to-Speech (TTS) Conversion

After the text summary has been generated, the output is then passed to the Text-to-Speech module for the text to be presented in an audio form. This adds to the usefulness of the system, especially for people with visual impairments who can now be able to read the text by hearing the audio. The deployment of the system can dictate the type of module that can be used for the text-to-speech function, whereby for online use, one can use the gTTS module (Google's Text-to-Speech), while for offline applications, one can consider the pyttsx3 module. Both of these text-to-speech modules are capable of being used for several lan-guages, have the ability to control the rate of the voice, and also have different accents. This audio output gets displayed together with the text summary in the user interface.

2.6 Output Generation and Display Interface

The final result of this system includes both text and audio forms of the summarized news content. The design intended for this purpose allows easier understanding by distinguishing both text and the result summary. Users are also availed with an inter-active layout that allows them to read the news summary and then listen to the audio result when needed.

3 Results And Discussion

The proposed newspaper summarization framework effectively automates news extrac-tion, classification, summarization, and the generation of audio. By embedding into one system all the tools like OCR, content recovery based on NewsAPI, hybrid sum-marization techniques, and TTS modules, the system turns out to be comprehensive and accessible. Experimental evaluation using both scanned newspaper images and online news articles demonstrated reliable performance with consistency for different input formats.

For image-based inputs, the Tesseract OCR module had an average text recognition accuracy of about 92% under conditions of adequate image quality. The system was successfully able to extract headlines and body texts from varied layout newspapers. In contrast, news content directly via NewsAPI.org bypassed the OCR stage, reducing the processing time and computation involved.

Fig. 2 Interface of the News Aggregator

The further component was the summarization part which adopted a hybrid approach that incorporated both abstractive and extractive techniques. The former was achieved through the use of abstractive techniques like the transformer models T5 that produced linguistically fluent summaries through paraphrasing, with the end result being summaries that were coherent and concise as shown in Fig. 2.
Figure 3 showcases a comparison of the original news text extracted into the system with its summarized form. The news text extracted into the system is represented by a well-defined upper portion of the image, whereas the summarized result appears in the larger portion at the end of the figure. The audio play option for the summarized news text appears at the bottom of the interface of the system.

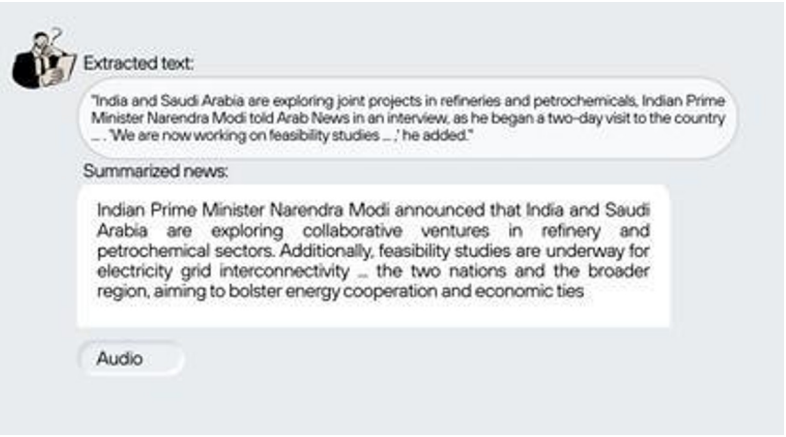

Fig. 3 Interface of the News Aggregator

4 Conclusion

This work proposes an integrated system for automated newspaper summarization to address information overload. The proposed system is capable of converting printed newspaper contents into machine-readable text with the aid of Optical Charac-ter Recognition and hence digitizes news from offline sources. NewsAPI.org further enhances the contextual understanding by fetching articles relevant to detected key-words and headlines. Its hybrid summarization strategy then summarizes long news articles with the aid of extractive and abstractive methods in a condensed form without losing their essence. This is further supported by adding Text-to-Speech functionality, which improves the accessibility for visually impaired users and supports the consump-tion of contents in the audio format. In short, the proposed framework represents a streamlined pipeline for the acquisition, processing, summarizing, and presenting of news content both in textual and spoken formats and is thus well-suited for modern time-sensitive information environments.

References

1. Liu, Y.: Fine-tune BERT for Extractive Summarization. arXiv preprint arXiv:1903.10318 (2019)
2. Liu, Y., Lapata, M.: Text Summarization with Pretrained Encoders. arXiv preprint arXiv:1908.08345 (2019
3. Gambhir, M., Gupta, V.: Deep learning-based extractive text summarization

with word-level attention mechanism. Multimedia Tools and Applications 81, 20829–20852 (2022). https://doi.org/10.1007/s11042-022-12729-y

4. Khatri, C., Singh, G., Parikh, N.: Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks. arXiv preprint arXiv:1807.08000 (2018)
5. Sinha, A., Yadav, A., Gahlot, A.: Extractive Text Summarization using Neural Networks. arXiv preprint arXiv:1802.10137 (2018)
6. Dhakras, P., Shrivastava, M.: BoWLer: A neural approach to extractive text sum-marization. Proc. 32nd Pacific Asia Conference on Language, Information and Computation, 1–3 (2018)
7. Hennig, L.: Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis. Proc. International Conference RANLP-2009, 144–149 (2009)
8. Abdel-Salam, S. Rafea, A.: Performance Study on Extractive Text Summarization Using BERT Models. Information 13(2), 67 (2022). https://doi.org/10.3390/info13020067
9. Pai, M.: Text Summarizer Using Abstractive and Extractive Method. International Journal of Engineering Research & Technology (IJERT) 3(5) (2014).
10. Hidayat, S., et al.: Automatic text summarization using latent Dirichlet alloca-tion (LDA) for document clustering. International Journal of Advanced Intelligent Informatics 1(3), 132–139 (2015).
11. Gupta, V., Lehal, G.S.: A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence 2(3), 258–268 (2010).
12. Jia, R., et al.: Neural extractive summarization with hierarchical attentive hetero-geneous graph network. Proc. EMNLP 2020, 3622–3631 (2020)
13. Jing, B., et al.: Multiplex graph neural network for extractive text summarization. arXiv preprint arXiv:2108.12870 (2021)
14. Joshi, A., et al.: RankSum—An unsupervised extractive text summarization based on rank fusion. Expert Systems with Applications 200 (2022)
15. Ghadimi, A., Beigy, H.: SGCSumm: An extractive multi-document summariza-tion method based on pre-trained language model, submodularity, and graph convolutional neural networks. Expert Systems with Applications 215 (2023)
16. Zhang, X., et al.: Neural latent extractive document summarization. Proc. EMNLP 2018, 779–784 (2018)
17. Xu, J., Durrett, G.: Neural extractive text summarization with syntactic compres-sion. Proc. EMNLP 2019, 3292–3303 (2019)
18. Yang, K., et al.: EcForest: extractive document summarization through enhanced sentence embedding and cascade forest. Concurrency and Computation: Practice and Experience 31(17), e5206 (2019)
19. Yin, W., Pei, Y.: Optimizing sentence modelling and selection for document summarization. Proc. IJCAI 2015, 1383–1389 (2015)
20. Yasunaga, M., et al.: Graph-based neural multi-document summarization. Proc. CoNLL 2017, 452–462 (2017)