DOI: https://doi.org/10.52756/ijerr.2024.v46.017 *Int. J. Exp. Res. Rev.*, Vol. 46: 212-221 (2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

ISSN 2455-4855

Advanced News Archiving System with Machine Learning-Driven Web Scraping and AI-Powered Summarization Using T5, Pegasus, BERT and BART Architectures

Check for updates

Narasimhula L V Venugopal[1], K Visala[2], Sammingi Nirmala[3]*, Ch Vinod Varma[4], Adibabu Triparagiri[5], Athmakuri Satish Kumar[6] and Ch Sekhar[1]

[1]Department of Computer Science and Engineering, GMR Institute of Technology(A), Rajam, Andhra Pradesh, India; [2]Department of Computer Science and Engineering, Vignan Institute of Information Technology(A), Visakhapatnam, Andhra Pradesh, India; [3]Department of CSE (AI & ML, Data Science), Anil Neerukonda Institute of Technology and Sciences (A), Visakhapatnam, Andhra Pradesh, India; [4]Department of Computer Science and Engineering, SRKR Engineering College(A), Andhra Pradesh, India; [5]Department of Computer Science and Engineering, N S Raju Institute of Engineering and Technology, Visakhapatnam, Andhra Pradesh, India; [6]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India

E-mail/Orcid Id:

NLVV, nlvvgopal@gmail.com, https://orcid.org/0009-0004-8011-2009; *KV,* visalasai@gmail.com, https://orcid.org/0009-0003-2870-2258; *SN,* samminginirmala77@gmail.com, https://orcid.org/0009-0009-8358-2002; *CVV,* vinodvarmaaa@gmail.com, https://orcid.org/0000-0002-7229-4234; *AT,* adibabu.5b8@gmail.com, https://orcid.org/0009-0003-5945-156X; *ASK,* satishathmakuri@gmail.com, https://orcid.org/0000-0002-9489-2091; *CHS,* sekhar.ch@gmrit.edu.in, https://orcid.org/0000-0001-5603-1453

Article History:

Received: 30th Jun., 2024
Accepted: 20th Dec., 2024
Published: 30th Dec., 2024

Keywords:
News Summarization, Web Scraping, Structured Archive, News Organization, Information Retrieval, BERT, BART

How to cite this Article:
Narasimhula L V Venugopal, K Visala, Sammingi Nirmala, Ch Vinod Varma, Adibabu Triparagiri, Athmakuri Satish Kumar and Ch Sekhar (2024). Advanced News Archiving System with Machine Learning-Driven Web Scraping and AI-Powered Summarization Using T5, Pegasus, BERT and BART Architectures. *International Journal of Experimental Research and Review*, *46*, 212-221.
DOI: https://doi.org/10.52756/ijerr.2024.v46.017

Abstract: Data plays a crucial role in the contemporary era of technology, as it is a vital element in the publication of news on the internet or a website. Nevertheless, understanding long reports in order to fully comprehend events can be a challenging endeavor, frequently leading to subjective judgments. The application's architecture integrates the categorization of news stories by day, resulting in a well-organized and readily accessible archive. The application employs the web scraping method, which entails pulling pertinent news articles from numerous internet sources. The application employed sophisticated summarizing libraries, including the BERT, BART, T5 model and Google Pegasus, to condense the information into a succinct and comprehensible style. The T5 model performs exceptionally well in text summarization and other natural language processing tasks because of its text-to-text structure; it is also a very customizable language model. Google Pegasus, an expert in abstractive summarizing, uses self-attention mechanisms and rigorous pre-training to generate high-quality, concise news summaries. To summarize, these are the most important parts of our app's process. When it comes to collecting, storing, and summarizing news articles, the system has you covered. In addition, it will offer a straightforward design that makes it simple to browse past news stories and their summaries.

Introduction

The news plays a crucial role in keeping individuals updated about the latest events in our rapidly changing world, and staying informed is of the highest significance. The information presented in the many complicated news stories frequently posted on the internet can be difficult for even the most committed readers to fully understand (Suleiman et al., 2020; Asmitha et al., 2024; Dharrao et al., 2024). Readers may be prone to biased judgments or a lack of context if lengthy papers with complicated facts are difficult to understand. So, with our program's intuitive interface, reading news stories is a breeze.

Our software provides a more user-friendly way to read and understand news articles, helping people deal with overwhelming information. To save users the

***Corresponding Author:** samminginirmala77@gmail.com

trouble of reading long reports, the system compiles news stories from multiple websites and summarizes them in a nutshell. With the help of cutting-edge methods like web scraping, the application swiftly obtains relevant news content so that users may enjoy a personalized selection of articles.

There are a lot of systems out there that disseminate news, however, some of these more conventional platforms may have complicated interfaces and demand subscription fees to access premium services. There may be times when websites don't give enough or relevant information. Not having an organized archive also makes it hard to follow previous news stories and navigate the site, which is bad for users. However, our software lets users choose when they want news stories to be retrieved, shows them a full list of articles with links, headings, and sources, lets them see summary versions of the items they've chosen, and lets them search using headings or sources.

Our new website is here to help close this gap. Allow me to explain it to you: Our technology automatically prompts customers to specify their desired date the moment they land. The first step of our system is to collect news articles from trustworthy online sources using web scraping techniques. After that, high-tech summarizing tools are employed to reduce articles to brief summaries, such as T5 Model and Google Pegasus Model. By utilizing an established archive system, our program is designed to arrange news articles according to date, making them easy to retrieve and refer to in the future. Features for perusing old news and summaries are built into the platform and an easy-to-navigate UI is created for smooth interaction (Zhang et al., 2020; Haque et al., 2022; Dharrao et al., 2023; Gite et al., 2023; Wang et al., 2023).

Our platform's development has made strategic use of a wide range of tools and techniques to provide a pleasant user experience and hassle-free operation. The backbone of our technical architecture is built on strong technologies like Python for processing data and scraping websites and HTML, CSS, and JavaScript for building the front end. Beautiful Soup for web scraping, MongoDB Cluster for database administration, and FastAPI for quick backend development are crucial components that our platform combines. We guarantee top-notch speed and scalability with the help of GPU technology, more especially T4 double-sided.

Building an efficient tool for news summaries is our principal objective in writing this paper. The system will streamline the process of gathering and summarizing news articles from reputable web sources, making it easier to acquire pertinent information. Summaries should be brief (less than 150 characters) for the benefit of users, and we strive to accomplish this goal by summarizing at least 70% of daily news pieces. Daily, we will mechanically scrape around 30-35 articles from different sources to compile our news articles. Utilizing the summarizing models, our program generated summaries that saved users time and decreased the likelihood of biased judgments caused by an absence of sufficient information. Not only does it summarize news stories, but it also has an extensive archive system that sorts them by day for simple retrieval and future reference. Staying informed about the latest happenings has become more convenient than ever with our easily navigable design. We may easily navigate through archived news stories and their synopses.

Literature Survey

Raffel et al. (2020) introduced an extremely successful natural language processing technique that is based on a unified text-to-text transformer paradigm. The effectiveness of their strategy was demonstrated by rigorous testing and evaluations, therefore elevating the standard for the potential of transfer learning. Using our collection of news items, we upgraded pre-trained models in order to render them more effective at summarizing.

Zhang et al. (2020) proposed PEGASUS, an approach that pre-trains with extracted gap sentences to enhance summarization. We can improve the user experience by providing high-quality summaries and efficiently processing lots of news articles using this method.

The capabilities Mohamed et al. (2020) suggested for news synthesis and summary could be quite helpful. During their examination of several text summarization techniques, they always emphasized the necessity of swiftly retrieving data from huge datasets. Their innovative method of picking and condensing news stories from different sources on the internet was developed by integrating cutting-edge web scraping techniques with sharp summary algorithms.

Mastropaolo et al. (2021) developed the Text-To-Text Transfer Transformer (T5) to enhance coding methods. We put T5 through its paces to see how well it handled various forms of code. You can trust the T5-type transformer. Changing the text-to-text model T5 to meet different natural language processing objectives demonstrates its versatility.

Widyassari et al. (2022), after reading a lot of research on the subject, looked into the methods for automatically summarizing text. In fact, they looked at how well a number of different methods worked, some of which

DOI: https://doi.org/10.52756/ijerr.2024.v46.017

were more traditional and some of which were more new. With the help of new technologies like phrase extraction and semantic analysis, we expect to refine our summarization process better and get better results.

The research conducted by Khilji et al. (2021) emphasized the significance of abstract writing in summarizing methodologies and evaluation instruments. In order to improve the uniformity and excellence of the summaries, they engaged in an examination of several models and methodologies. We utilized evaluation techniques gathered from their study to reduce news stories and evaluate our summarizing program.

Abodayeh et al. (2023) conducted a web scraping operation utilising BeautifulSoup, a tool for analysing data sets. A potential method for the researchers to achieve their analytical goals by collecting data from various online sources was the exploration of web scraping techniques. Our technological capabilities enabled us to rapidly obtain news stories, therefore facilitating the process of summarizing and assessing them. The ability to access news articles from a multitude of online sources was enhanced by this proficiency.

Deaf and hard-of-hearing people benefit greatly from sign language, which uses sophisticated hand and body gestures. Sekhar et al. (2024) use a Random Forest model, Media Pipe for gesture sensing, with TensorFlow for machine learning to recognize sentences. On a given dataset, the system can translate hand movements into phrases with 80% accuracy. Real-time phase synthesis using powerful machine learning and natural language processing is possible with ChatGPT integrated to improve sign language-non-user communication.

Wolf et al. (2020) presented modern techniques for NLP using the Transformer model. Several natural language processing tasks have been significantly improved by these models' use of self-attention processes, which enable a more precise comprehension of textual context. Several models that have done exceptionally well on several NLP benchmarks are part of the Transformer architecture. These models are Pegasus, GPT, BERT, and T5. To ensure optimal performance, we exhaustively evaluated and validated the transformer-based models used in our application.

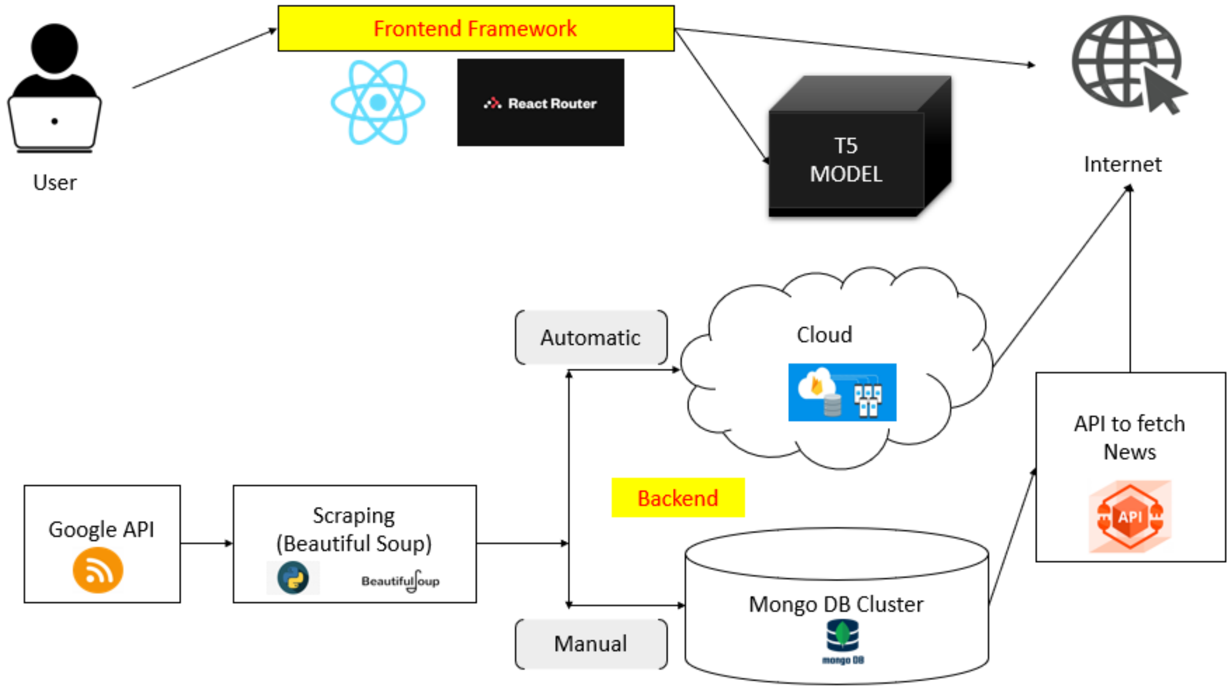


Figure 1. System Architecture.

Reddit and Hacker News were among the resources used by Aniche et al. (2018) to stay abreast of technological developments. Findings from interviews and analyses of these groups point to their contributions to knowledge sharing and involvement and their struggles with moderation. Improvements to developer habits and tool design, as well as research on developer communities, are made possible by these discoveries.

DOI: https://doi.org/10.52756/ijerr.2024.v46.017

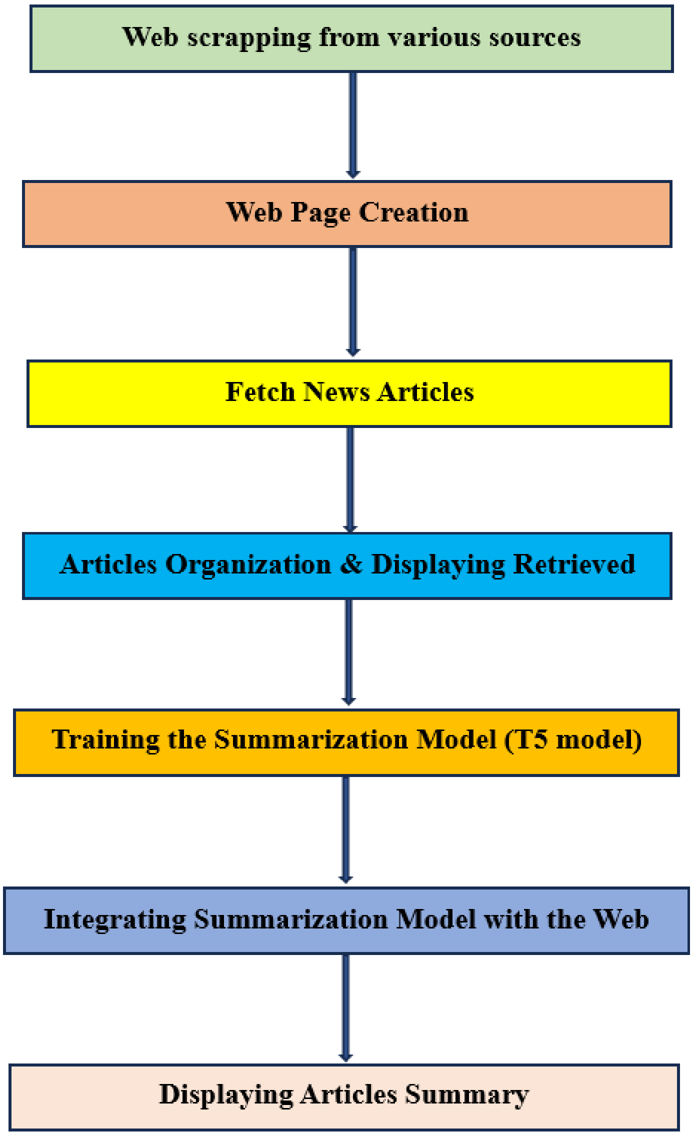

Web scrapping from various sources
Web Page Creation
Fetch News Articles
Articles Organization & Displaying Retrieved
Training the Summarization Model (T5 model)
Integrating Summarization Model with the Web
Displaying Articles Summary

Figure 2. Proposed Work Flow.

One website that rapidly gathers and summarizes news headlines from all across the globe is NewsOne, according to Sundaramoorthy et al. (2017). The software collects news stories from many sources (web scraping and RSS feeds) and sorts them into various categories (e.g., Technology, Health, and Business). Readers can choose news pieces to read based on their interests using this classification system. Users are not charged for using the site, which strives for the fastest news delivery possible. There are rumors that the app is getting regular updates that will make the material more accessible and encourage more user participation.

Methodology

To develop a user-friendly application that simplifies access to news information and summarizes news content efficiently, we have to perform a few important steps:

Figure 1 describes the web application framework, detailing the flow of data and functionalities. At the front end, React Router facilitates the creation of a dynamic user interface, enabling seamless navigation within the application. Users interact with the system through the internet, accessing its features and functionalities as described by Keerthana et al. (2024). Moving to the backend, the system utilizes various components to gather and process data. Google API grants access to Google Services, while web scraping, facilitated by Beautiful Soup, extracts data from websites. An API fetches news articles, potentially from multiple sources. The T5 Model, a machine learning model capable of tasks like text summarization, enhances data processing. Finally, data is stored and managed in a Mongo DB Cluster database, ensuring efficient data storage and retrieval. So, In this way, our website or system works.

Figure 2 describes the paper flow of our application. Our system extracts news articles from various sources through a web scraping mechanism. Upon Scraping, our system creates a web page to present the extracted

DOI: https://doi.org/10.52756/ijerr.2024.v46.017

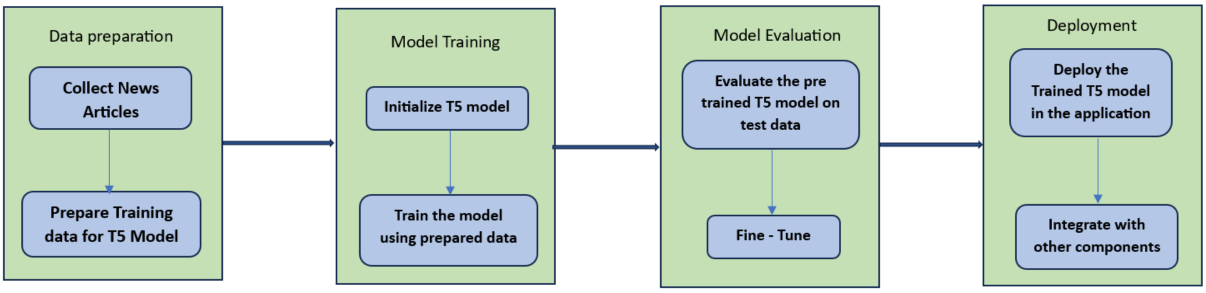

Data preparation
Collect News Articles
Prepare Training data for T5 Model
Model Training
Initialize T5 model
Train the model using prepared data
Model Evaluation
Evaluate the pre trained T5 model on test data
Fine - Tune
Deployment
Deploy the Trained T5 model in the application
Integrate with other components

Figure 3. T5 Model Flow.

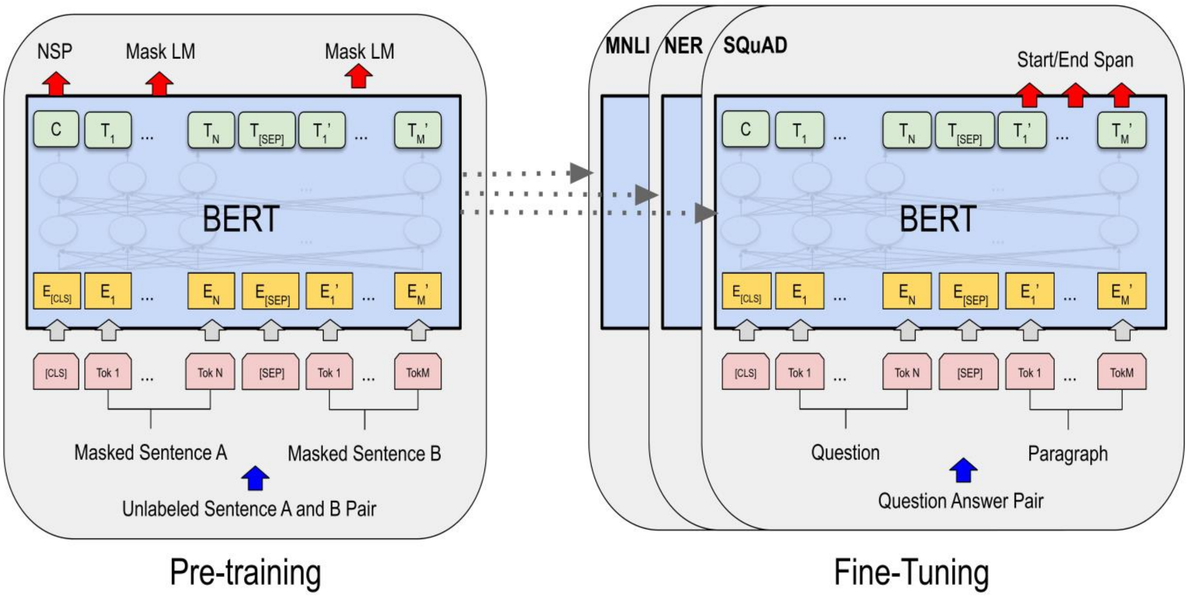


Figure 4. Pre training and Fine - Tuning Architecture.

information. Then, it organizes the fetched news articles and showcases them on the previously created web page. We have particularly used the T5 model to train on a dataset. T5 is a text-to-text generative model widely used for summarization and question-answering tasks. The trained T5 model is integrated into the web page to facilitate summarization. Finally, our system displays summaries of the news articles on the user-friendly web page generated by the integrated T5 model.

Figure 3 describes the t5 model flow of our application. The Flow T5 model involves data collection, model initialization, training, evaluation, and deployment stages, resulting in a user-friendly system for effectively summarizing news articles.

BERT – Bidirectional Encoder Representations Form Transformer Machine Learning Algorithm

Google developed a ground-breaking language representation model called BERT, which is especially efficient at interpreting the context of words in a sentence. It uses transformer architecture with a bidirectional approach, which means that examining the words that come before and after a word takes into account the whole context. Because of these features, BERT is very useful for tasks like named entity recognition, sentiment analysis, and question-answering. It can also capture complex relationships and meanings inside text. To make sure that the most important elements are emphasized in the summaries, you can use BERT in the context of your news summarizing application to extract important sentences and pertinent information from lengthy articles. The application can improve its comprehension and processing of the data by integrating BERT.

BERT consists of two steps: fine-tuning and pre-training. The model is trained on unlabelled data across several pre-training tasks during pre-training. The pre-trained parameters are used to initialize the BERT model for fine-tuning, and labelled data from the downstream jobs is used to adjust each and every parameter. Despite being started with identical pre-trained parameters, each downstream task has its own fine-tuned models.

BART – Bidirectional and Auto-Regressive Transformer Machine Learning Algorithm:

BART is an effective model that combines the best features of autoregressive and bidirectional transformers, especially for text production tasks. It works especially effectively for jobs involving translation, summarization, and other aspects of natural language processing. In order to teach the model to create coherent and fluid text from

DOI: https://doi.org/10.52756/ijerr.2024.v46.017

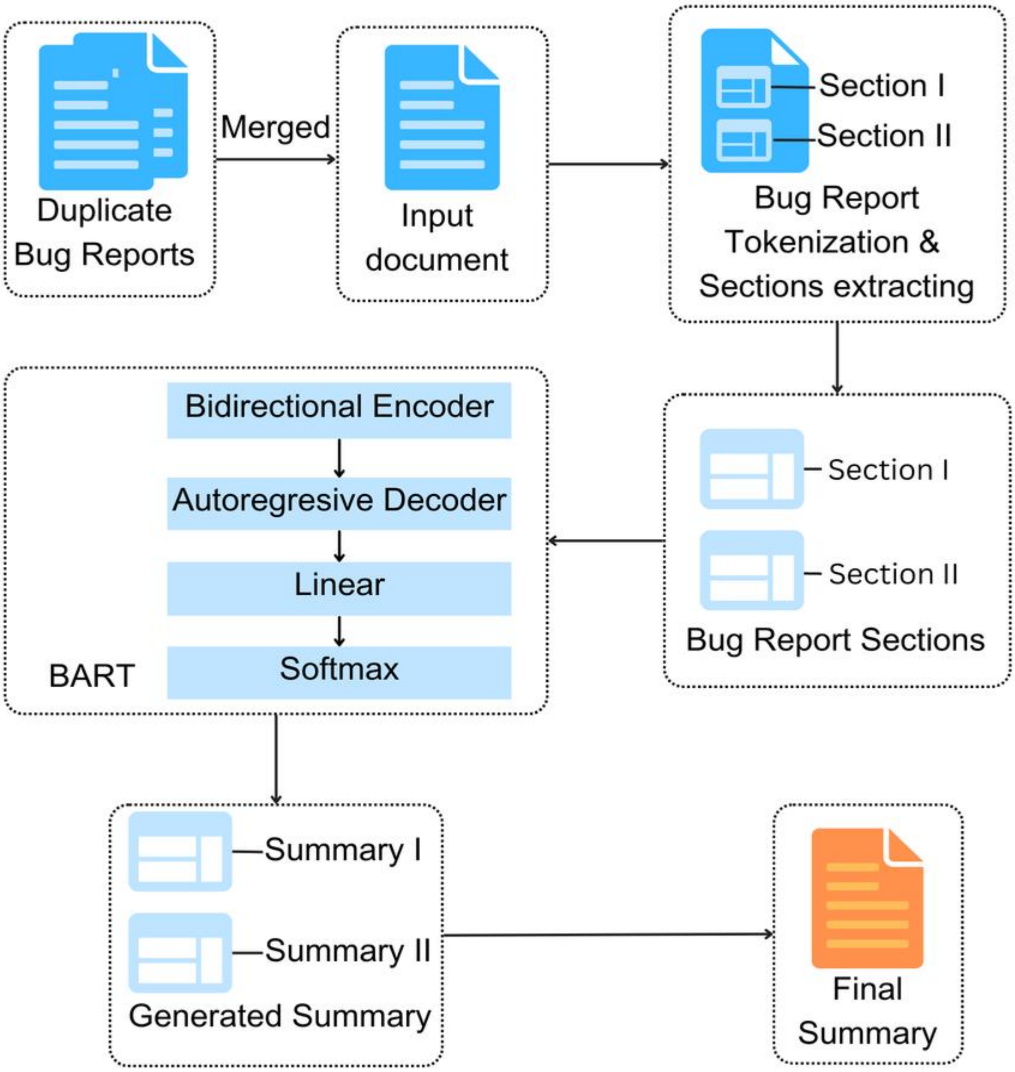


Figure 5. BART Architecture.

Incomplete input, BART corrupts text and then trains it to reconstruct it. This makes BART a great option for your application since it can generate outstanding abstractive summaries that capture the essential points of news stories in a readable manner. The program may produce succinct, context-aware summaries that users find engaging by integrating BART into your summarization process. This guarantees that users get the most pertinent information from the news stories they are interested in.

Results and Discussion

Deploying our paper on port 10000 and using MongoDB database makes hosting our frontend and backend components easier. This setup helped us manage everything smoothly, from the React.js frontend to the Node.js backend, simplifying database management.

Figure 6 describes the home page of our website. Here you will find the scrolling news articles on a specific date and you will also find a calendar option.

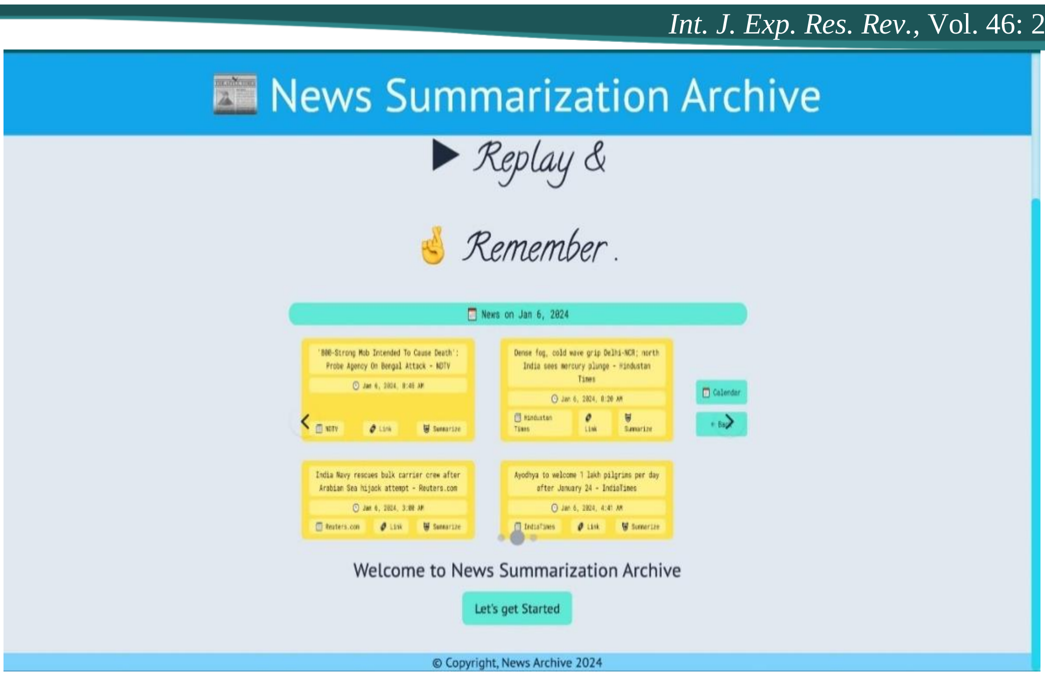
The Figure 7 shows the calendar view of a news summarization archive. The calendar is used to navigate you and access archived summaries. As you can see, a "Choose the Date" option at the top of the calendar will likely allow users to select a specific date to view archived summaries.

Figure 8 describes the collection of news articles on a particular day. The user selects a specific date and a list of articles appears on the screen which contains a news headline, date and time, relevant links and summarize buttons. The archive also allows users to search for news articles or sources by entering keywords in the search bar.

Figure 9 displays the summary of the news articles. The model processes the news articles and presents the summary for user consumption.

The Figure 10 and Figure 11 displays the model evaluation metrics. We got these results ROUGE-1: Achieving an F-score of approximately 0.44, indicating 44% unigram overlap with reference summaries. ROUGE-2: F-score of approximately 0.23 indicates moderate bigram overlap. Important for coherence and fluency in summaries. ROUGE-L: F-score around 0.41, indicating common subsequence between summaries. Important for news summarization's coherence and key information.

DOI: https://doi.org/10.52756/ijerr.2024.v46.017



News Summarization Archive

▶ Replay &

🤞 Remember .

News on Jan 6, 2024

'800-Strong Mob Intended To Cause Death': Probe Agency On Bengal Attack - NDTV
Jan 6, 2024, 8:45 AM
NDTV | Link | Summarize

Dense fog, cold wave grip Delhi-NCR; north India sees mercury plunge - Hindustan Times
Jan 6, 2024, 8:20 AM
Hindustan Times | Link | Summarize

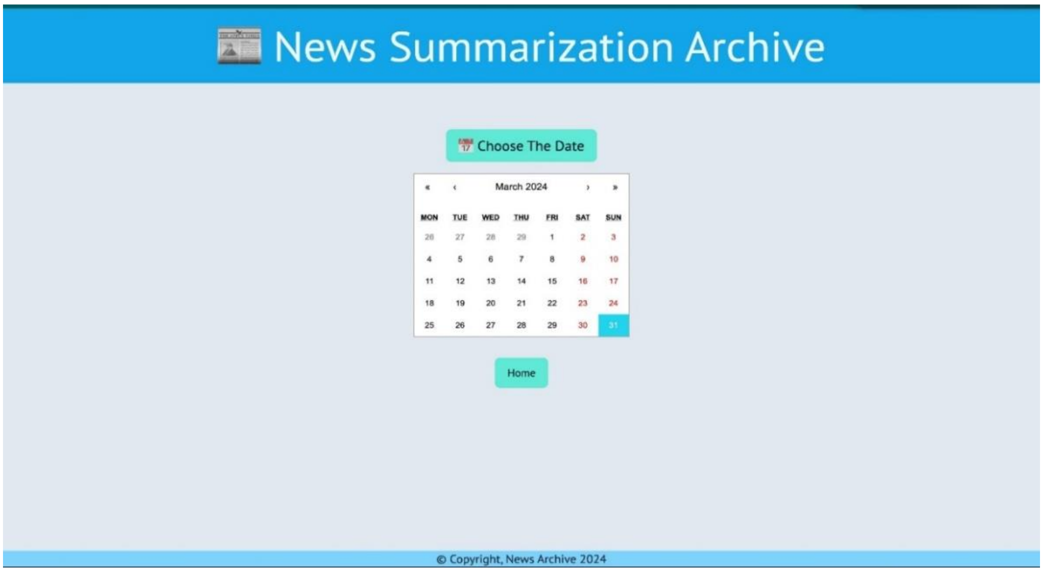
India Navy rescues bulk carrier crew after Arabian Sea hijack attempt - Reuters.com
Jan 6, 2024, 3:00 AM
Reuters.com | Link | Summarize

Ayodhya to welcome 1 lakh pilgrims per day after January 24 - IndiaTimes
Jan 6, 2024, 4:41 AM
IndiaTimes | Link | Summarize

Welcome to News Summarization Archive

© Copyright, News Archive 2024

Figure 6. News Summarization Home Page.

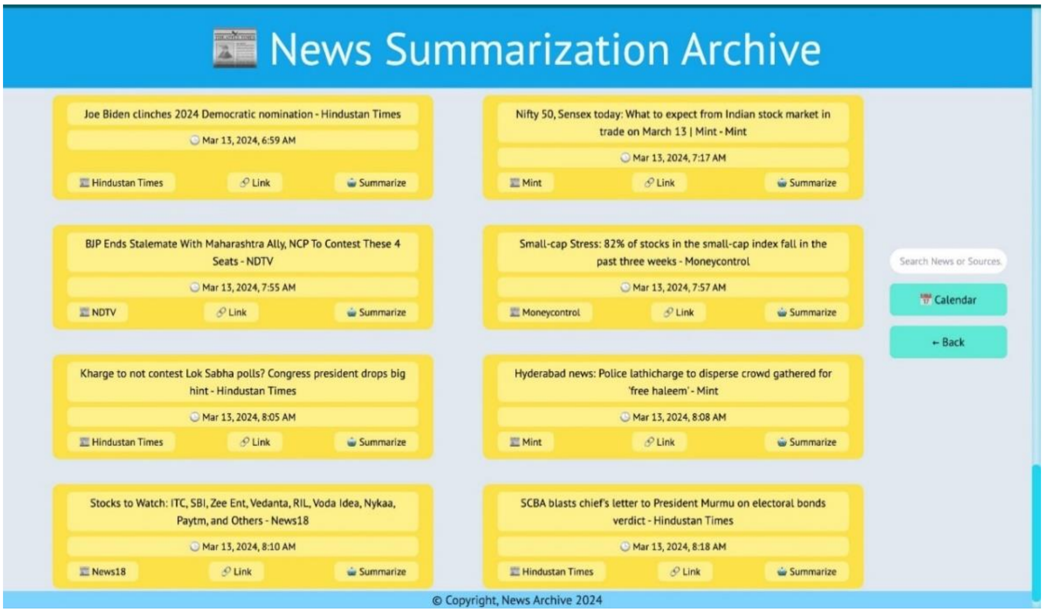


News Summarization Archive

MON	TUE	WED	THU	FRI	SAT	SUN
26	27	28	29	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

© Copyright, News Archive 2024

Figure 7. News Archive Calendar.



News Summarization Archive

Joe Biden clinches 2024 Democratic nomination - Hindustan Times
Mar 13, 2024, 6:59 AM
Hindustan Times | Link | Summarize

Nifty 50, Sensex today: What to expect from Indian stock market in trade on March 13 | Mint - Mint
Mar 13, 2024, 7:17 AM
Mint | Link | Summarize

BJP Ends Stalemate With Maharashtra Ally, NCP To Contest These 4 Seats - NDTV
Mar 13, 2024, 7:55 AM
NDTV | Link | Summarize

Small-cap Stress: 82% of stocks in the small-cap index fall in the past three weeks - Moneycontrol
Mar 13, 2024, 7:57 AM
Moneycontrol | Link | Summarize

Kharge to not contest Lok Sabha polls? Congress president drops big hint - Hindustan Times
Mar 13, 2024, 8:05 AM
Hindustan Times | Link | Summarize

Hyderabad news: Police lathicharge to disperse crowd gathered for 'free haleem' - Mint
Mar 13, 2024, 8:08 AM
Mint | Link | Summarize

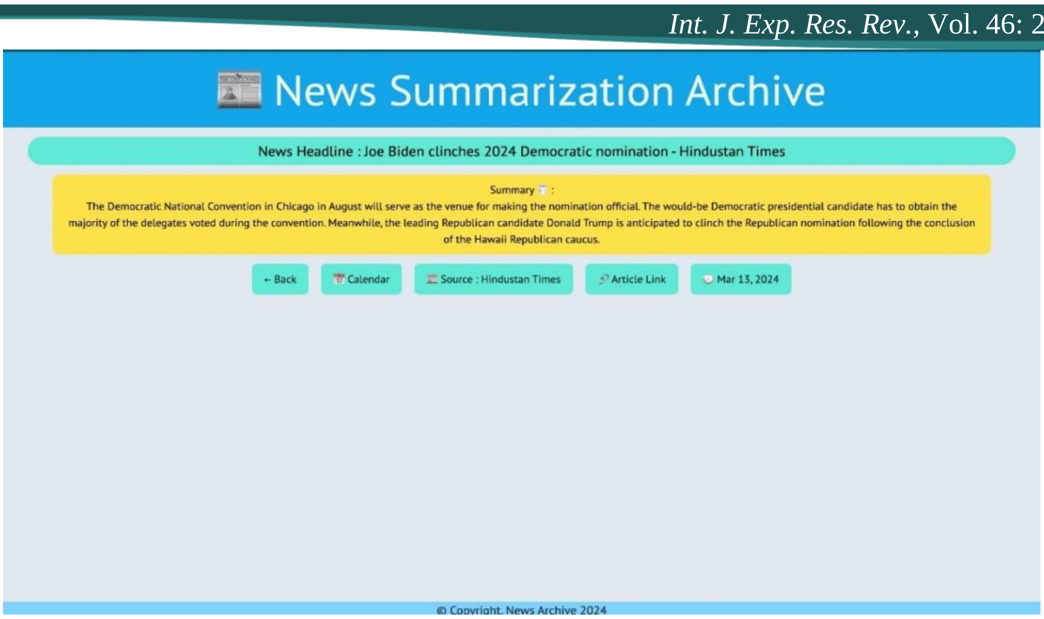
Stocks to Watch: ITC, SBI, Zee Ent, Vedanta, RIL, Voda Idea, Nykaa, Paytm, and Others - News18
Mar 13, 2024, 8:10 AM
News18 | Link | Summarize

SCBA blasts chief's letter to President Murmu on electoral bonds verdict - Hindustan Times
Mar 13, 2024, 8:18 AM
Hindustan Times | Link | Summarize

© Copyright, News Archive 2024

Figure 8. News Archive.

DOI: https://doi.org/10.52756/ijerr.2024.v46.017



News Summarization Archive

News Headline : Joe Biden clinches 2024 Democratic nomination - Hindustan Times

Summary :

The Democratic National Convention in Chicago in August will serve as the venue for making the nomination official. The would-be Democratic presidential candidate has to obtain the majority of the delegates voted during the convention. Meanwhile, the leading Republican candidate Donald Trump is anticipated to clinch the Republican nomination following the conclusion of the Hawaii Republican caucus.

Figure 9. News Summary.

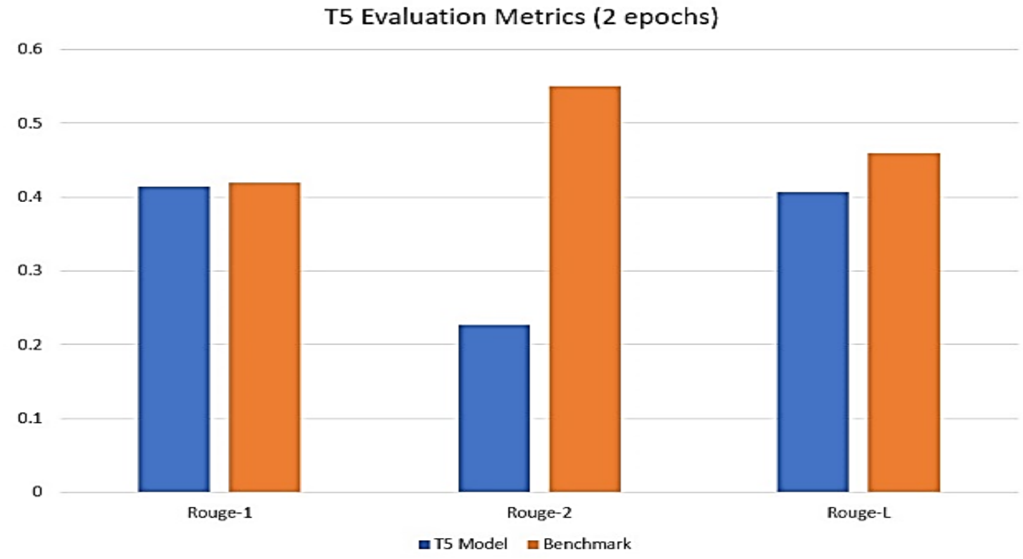


Figure 10. T5 Model Evaluation.

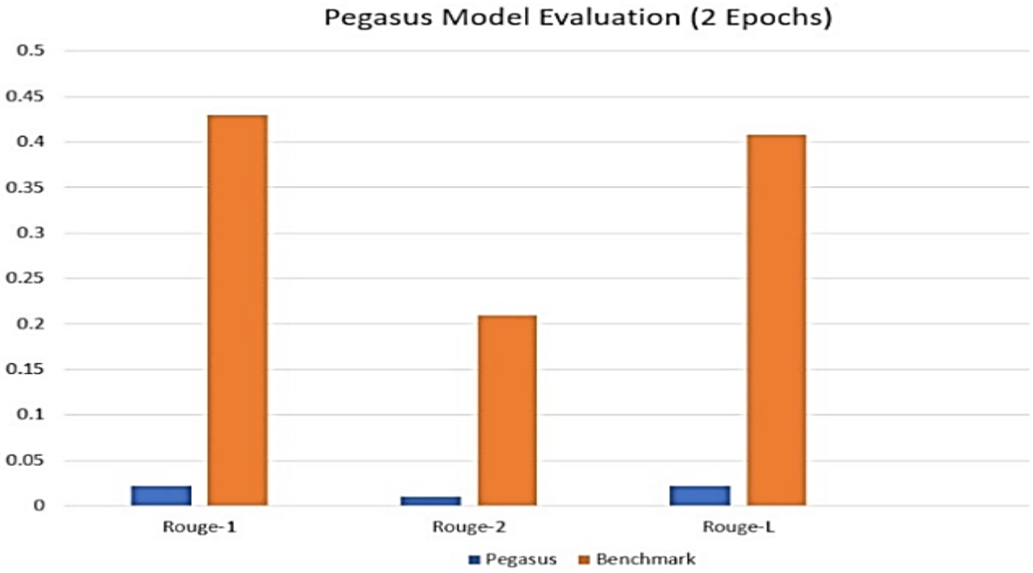


Figure 11. Pegasus Model Evaluation.

DOI: https://doi.org/10.52756/ijerr.2024.v46.017

Metric	F- Score	Description
ROUGE – 1	0.44	44% unigram overlap with reference summaries
ROUGE – 2	0.23	Moderate bigram overlap; important for coherence and fluency
ROUGE – L	0.41	Common subsequence, important for coherence and key information in news summarization

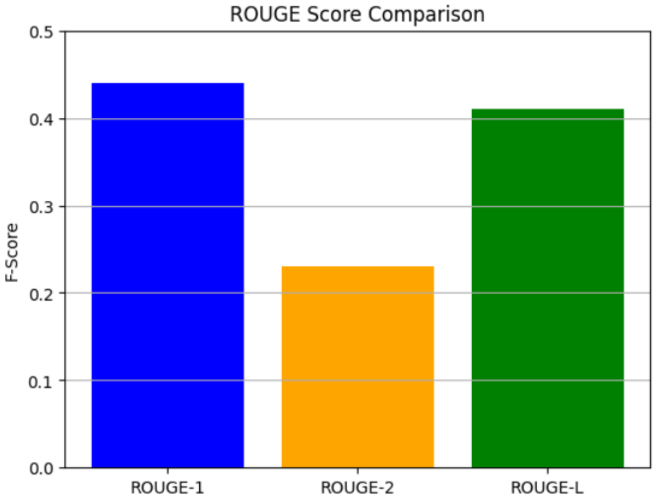


Figure 12. Comparison Graph of ROGUE Scores.

Conclusion

In our effort to make news reading quicker and more efficient, we utilised the T5 model to generate summarised news articles. This way, users can get the gist of the news without having to read the entire article, saving time. However, we encountered challenges when integrating the T5 model with our user interface (UI). This made it difficult for users to access the summarised articles seamlessly. To address this, we worked on improving the integration process to ensure a smoother user experience. Despite the hurdles, our goal remains clear: to provide users with a convenient way to stay informed by quickly scanning through summarised news articles. With continued efforts, we aim to overcome these challenges and offer a user-friendly platform where everyone can easily access and benefit from summarised news content.

Conflicts of Interest

According to the authors, there is no conflict of interest.

References

Abodayeh, A., Hejazi, R., Najjar, W., Shihadeh, L., & Latif, R. (2023). Web Scraping for Data Analytics: A BeautifulSoup Implementation. *2023 Sixth International Conference of Women in Data Science at Prince Sultan University* (WiDS PSU), pp. 65–69. https://doi.org/10.1109/wids-psu57071.2023.00025

Aniche, M., Treude, C., Steinmacher, I., Wiese, I., Pinto, G., Storey, M.-A., & Gerosa, M. A. (2018). How modern news aggregators help development communities shape and share knowledge. *Proceedings of the 40th International Conference on Software Engineering*. https://doi.org/10.1145/3180155.3180180

Asmitha, M., Kavitha, C.R., & Radha D. (2024). Summarizing News: Unleashing the Power of BART, GPT-2, T5, and Pegasus Models in Text Summarization. *2024 4th International Conference on Intelligent Technologies* (CONIT), Karnataka, India. *2024*, 1-6.

Dharrao, D., Mishra, M., Kazi, A., Pangavhane, M., Pise, P., & Bongale, A.M. (2024). Summarizing business news: Evaluating BART, T5, and PEGASUS for effective information extraction. *Revue d'Intelligence Artificielle*, *38*(3), 847-855. https://doi.org/10.18280/ria.380311

Dharrao, D., Bongale, A.M., Kadalaskar, V., Singh, U., & Singharoy, T. (2023). Patients' medical history summarizer using NLP. *In 2023 International Conference on Advances in Intelligent Computing and Applications* (AICAPS), Kochi, India, pp. 1-6. https://doi.org/10.1109/AICAPS57044.2023.10074336

Gite, S., Patil, S., Dharrao, D., Yadav, M., Basak, S., Rajendran, A., & Kotecha, K. (2023). Textual feature extraction using ant colony optimization for hate speech classification. *Big Data and Cognitive Computing*, *7*(1), 45. https://doi.org/10.3390/bdcc7010045

Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic similarity metrics for evaluating source code summarization. *In Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, Pittsburgh, PA, USA, pp. 36-47. https://doi.org/10.1145/3524610.3527909

Keerthana, B., Vamsinath, J., Kumari, C. S., Appaji, S. V. S., Rani, P. P., & Chilukuri, S. (2024). Machine Learning Techniques for Medicinal Leaf Prediction and Disease Identification. *International Journal of Experimental Research and Review*, *42*, 320–327. https://doi.org/10.52756/ijerr.2024.v42.028

Khilji, A. F. U. R., Sinha, U., Singh, P., Ali, A., & Pakray, P. (2021). Abstractive Text Summarization Approaches with Analysis of Evaluation Techniques. In *Communications in Computer and Information Science*, pp. 243–258. https://doi.org/10.1007/978-3-030-75529-4_19

Mastropaolo, A., Scalabrino, S., Cooper, N., Nader Palacio, D., Poshyvanyk, D., Oliveto, R., & Bavota, G. (2021). Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks. *2021 IEEE/ACM 43rd International Conference on Software Engineering* (ICSE), 336–347. https://doi.org/10.1109/icse43902.2021.00041

Mohamed, A., Ibrahim, M., Yasser, M., Ayman, M., Gamil, M., & Hassan, W. (2020). News aggregator and efficient summarization system. *International Journal of Advanced Computer Science and Applications*, *11*(6). https://doi.org/10.14569/ijacsa.2020.0110677

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. https://jmlr.org/papers/volume21/20-074/20-074.pdf

Sekhar, C., Devi, J., Kumar, M., Swathi, K., Ratnam, P., & Rao, M. (2024). Enhancing Sign Language Understanding through Machine Learning at the Sentence Level. *International Journal of Experimental Research and Review*, *41*(Spl Vol), 11-18. https://doi.org/10.52756/ijerr.2024.v41spl.002

Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. *Mathematical Problems in Engineering*, *2020*, 1-29. https://doi.org/10.1155/2020/9365340

Sundaramoorthy, K., Durga, R., & Nagadarshini, S. (2017). NewsOne — An Aggregation System for News Using Web Scraping Method. *2017 International Conference on Technical Advancements in Computers and Communications* (ICTACC), pp. 136–140. https://doi.org/10.1109/ictacc.2017.43

Wang, M., Xie, P., Du, Y., & Hu, X. (2023). T5-Based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, *13*(12), 7111. https://doi.org/10.3390/app13127111

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning. *PMLR*, pp. 11328-11339. https://doi.org/10.48550/arXiv.1912.08777

How to cite this Article:

Narasimhula L V Venugopal, K Visala, Sammingi Nirmala, Ch Vinod Varma, Adibabu Triparagiri and Athmakuri Satish Kumar (2024). Advanced News Archiving System with Machine Learning-Driven Web Scraping and AI-Powered Summarization Using T5, Pegasus, BERT and BART Architectures. *International Journal of Experimental Research and Review*, *46*, 212-221.

DOI : https://doi.org/10.52756/ijerr.2024.v46.017