AI Based News Aggregator and Summarizer

Bhavya Thakkar[1], Raj Shah[2], Vandit Shah[3], Prof. Sweedle Machado[4]

[1,2,3,4] *Dwarkadas J Sanghvi College of Engineering, Information Technology Department*

[1]*thakkarbhavya4@gmail.com*, [2]*rajs55996@gmail.com*, [3]*shahvanditt@gmail.com*, [4]*sweedle.machado@djsce.ac.in*

Abstract: For many centuries, newspapers have been a reliable source of news and information for us. They were previously offered in the form of newspapers, but they are now also accessible online. There are multiple news sources that provide news on various topics. The users desire to read them in their own comfort, like not having to type or navigate to different websites. The system suggested in this study makes it simpler for the user to read news articles. The articles are compiled from many sources and provided in a condensed format. The software is available in Hindi, Gujarati, and English. Voice assistant helps the user to control the flow and delivers a better user experience. To keep the user interested, the recommendation system makes article suggestions.

Keywords: newspaper, voice assistant, recommendation system, text summarization

1. Introduction

In this modern world, news is the important start of the day because it provides access to information and a wide range of expertise. Reading the news is a good habit and a staple of modern life. The user will gain a wider perspective and deeper knowledge by reading the news. Nowadays, people read news online rather than in newspapers, as they once did. Reading a news story requires the user to visit multiple distinct news websites, which takes a lot of time, due to the large amount of news websites today and the variety of topics they cover.

The proposed system aims to develop a news aggregator, summarizer, recommendation system and voice assistant system. A news aggregator is a very useful programme because it gathers information from various websites. Most news stories, on the other hand, are extremely long, making it difficult for readers to finish them. Thus, news summarization is extremely beneficial because it shortens a lengthy article by emphasising key points in multiple languages (most notably English, Hindi, and Gujarati) and allows the reader to comprehend the entire piece by reading a summary of it. In addition, similar news articles are recommended to users based on their previous reading preferences, and the voice assistant feature assists the user by reading summarised news articles for them.

2. Literature Survey

This section covers the literature survey related to existing systems, methodology/algorithms and tool/framework used.

2.1 Literature Related to Existing Systems

2.1.1 A News Aggregator and Efficient Summarization System [1]: This paper focused on developing a news aggregator by compiling pertinent articles based on a certain input term or keyphrase and a news summarizer using the Textrank algorithm.

2.1.2 The News Reader Application and Recommendation System [3]: Popular recommendation algorithms including content-based and collaborative recommender systems are examined and discussed in this study. Additionally, it suggests a hybrid approach that is more effective because it combines the finest aspects of the aforementioned algorithms while striving to eliminate any potential drawbacks.

2.1.3 Matrix-Based News Aggregation System [4]: Users can gain a thorough understanding of the news with the help of Matrix-based News Analysis (MNA). The usage of MNA by a news aggregator to identify multiple perspectives on international news topics is also illustrated. The outcomes of a case study demonstrate that this approach broadens the user's understanding of the news while also providing news aggregation features that are equivalent to those of well-known systems.

2.2 Literature Related to Methodology/Approaches/Algorithms

2.2.1 News Aggregator

2.2.1.1 Matrix Based News Analysis (MNA) [4]: The two-dimensional matrix that MNA uses to organise articles' elements shows what entity I (row) has to say about entity j. This approach presents a variety of viewpoints in the news (column). MNA organises news items into rows and columns of a user-made matrix. Documents are things that have been given a cell assignment. The articles' subjects are then succinctly summarised by MNA in each cell.

2.2.2 News Summarization

2.2.2.1 Extractive and Abstractive based summarization [12]: The essential phrases and sections from the documents are chosen using the extractive technique. The important lines are then combined to make the summary. Abstractive text summarising aims to condense the major ideas of the original text into a concise, understandable summary. In order to maintain the integrity of the ideas, it substitutes new words and phrases for those in the original text. Thus, Abstractive is considerably more difficult than Extractive.

2.2.3 Recommendation System

2.2.3.1 Content Based and Collaborative-Based Filtering Approach [3]: Content-based filtering, as its name suggests, is largely focused on the content of the objects in a collection. It functions by identifying connections among several items in a batch of data. One disadvantage of this strategy is how severely confined the coverage of content-based suggestions is when there is not much user history. The technique to recommend systems that has been used most commonly is Collaborative-based filtering, which is based on user history. It can be applied to a wide range of situations. In collaborative-based filtering, the item-based strategy takes into account the similarities between the items utilised by customers. The cosine similarity is employed to provide suggestions for comparable products used by the user's K-neighbours.

2.3 Literature Related to Technology/Tools/Framework

2.3.1 Language Translation (Translation Models) [14]

2.3.1.1 Indic NLP:The Indic NLP Library supports the majority of the common text processing and NLP features for Indian languages. It offers many functionalities, including tokenization, transliteration, translation, and text normalization.

2.3.1.2 IndicTrans:The Samanantar dataset, which at the time of writing was the biggest publicly accessible parallel corpus collection for Indic languages, was used to train the Transformer-4x (434M) multilingual NMT model known as IndicTrans (14 April 2021).

3. Proposed Methodology

This section outlines the system's fundamental design, which can combine online news from many sources and summarise its content to cut down on user reading time. Also, visitors will be given recommendations for related content based on their reading interests.

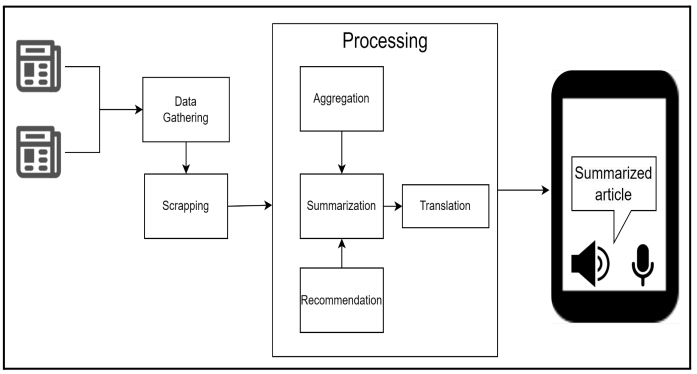

Fig 1: Architecture of News Aggregator

The main phases of the suggested system are depicted in Fig. 1. It has six primary stages, which are listed below.

3.1 Data Gathering[13]

Data gathering involves collecting information from several sources and storing it in one place so that it may be conveniently accessed for future processing. This stage involves collecting the **Really Simple Syndication** (RSS) feed from various news websites and storing them in the database. RSS feeds are freely available on all news websites which is an XML file and it is semi-structured in nature. It consists of basic details and links of all articles which are updated frequently.

3.2 Data Extracting

This stage entails parsing the RSS feed, which is stored in the database, to extract the link for each article. After extracting the link, each article link is scraped to obtain the title, description of the news, published date, author, and image URL of the news articles.

3.3 Aggregation

Using TF-IDF, extracted articles from multiple websites are vectorized. This phrase combines the terms "term frequency" (TF) and "inverse document frequency" (IDF). TF is used to count the number of times a word appears in a document[6]. The algorithm treats all terms equally, regardless of whether they contain stop words like "if," which is erroneous. Each keyword has a

different level of significance. IDF, or Inverse Document Frequency, is employed in this [5]. It gives words that occur frequently a lower weight and uncommon terms a higher weight. The data is additionally clustered using DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise. Since the data is unlabelled, and the number of clusters is unknown, DBSCAN is preferred. To cluster the articles, cosine similarity between vectors is calculated first, and then similar articles are grouped using cosine similarity [9]. Two parameters are required by DBSCAN: "minpts" and "eps". The term "minpts" refers to the smallest number of points in a cluster, and the term "eps" defines the distance between two points necessary for them to be included in the same cluster. [8]. As a result, there are several sets of articles on the same subject. Then PageRank algorithm is used which ranks the output articles, and is applied to these categories [7]. The most popular article will be summarised in greater depth.

3.4 Summarization

In this stage, the highly ranked article that was obtained during the aggregation stage will be summarised into three different languages (i.e. English, Hindi and Gujarati) using extractive-based summarization. This method's main objective is to identify the most important data, which is then extracted and structured to produce a brief summary. Pre-processing is done on the news articles (i.e. raw text). After eliminating all commas and punctuation, the article is tokenized into sentences. The pre-processed text is used for word embedding. Word embedding turns a word's syntactic and semantic data into a vector. Word2Vec can be used to extract a single word embedding from a corpus of text. This can be used to compare the cosine similarity between two vectors. The cosine similarity metric can be used to compare two vectors. Cosine similarity has a range of 0 to 1. If the value is 0, the vectors are completely distinct. A value of +1 similarly denotes that the vectors being evaluated are the same. The vectors were ranked using the TextRank algorithm in the end. To identify the most crucial sentences inside a text, a text processing graph-based ranking system called a TextRank is used. The fundamental tenet of TextRank is to rank phrases by importance before giving each phrase a score [10].

3.5 Recommendation

A recommender system is a tool that can identify the top items for a user's future preference among a group of items. Collaborative-based Filtering, as its name suggests, is used in the proposed methodology and is mostly focused on the content of the objects in a collection. It functions by identifying commonalities among various database objects. Other users' preferences or activities are not taken into account. The recommendation system remembers a user's interest when they read news in a particular category and suggests related items based on their past behaviour and interests.

3.6 Summarised Articles

After completing all of the preceding steps, the summarised article is displayed to the end user. In addition to the summarised article, the user has the option of reading the entire article. Simultaneously, similar articles will be recommended to the end user.

3.7 Voice Assistant[2]

The suggested methodology also has a voice assistant function that reads article summaries and news headlines to the user. This feature allows the user to utilise the application at their convenience by allowing them to regulate the application's flow. Alan AI is used to implement this feature. It includes a complete serverless environment for building sophisticated and reliable voice assistants. A fantastic tool for seamless voice recognition is Alan Studio. Making spoken language models, honing speech recognition software, launching, and hosting voice parts are not necessary.

4. Implementation Details

The hardware and software specifications of the suggested system are described in this section. Additionally, it discusses the walkthrough of each webpage and provides a snapshot of the webpage for the proposed system.

4.1 Hardware and Software Requirements

4.1.1 User End: The proposed system will require a working smartphone/laptop/pc/tablet or any such device with a working internet connection and browsers such as Chrome, Firefox,Microsoft Edge. The users should also have the latest version of the browser installed and operating system version above Windows 7/ Android 7/ ios 10.
4.1.2 Developer End: The proposed system will also require tools such as Git and a code editor such as Visual Studio, jupyter notebook, MongoDB Atlas.

4.2 User Interface.

Fig 2: Home Page

The home page of the suggested system is depicted in Fig.2. This webpage provides instructions for using voice assistant commands based on news categories and keywords to get news articles.

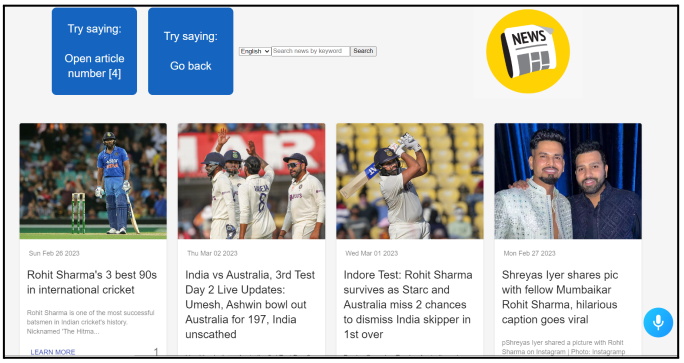

Fig 3: Aggregated Article Page

Fig. 3 shows the suggested system's Aggregated Article page. This webpage offers aggregated content in response to the user's request on the home page. The voice assistant function on this webpage assists users by reading the headlines of each news article.

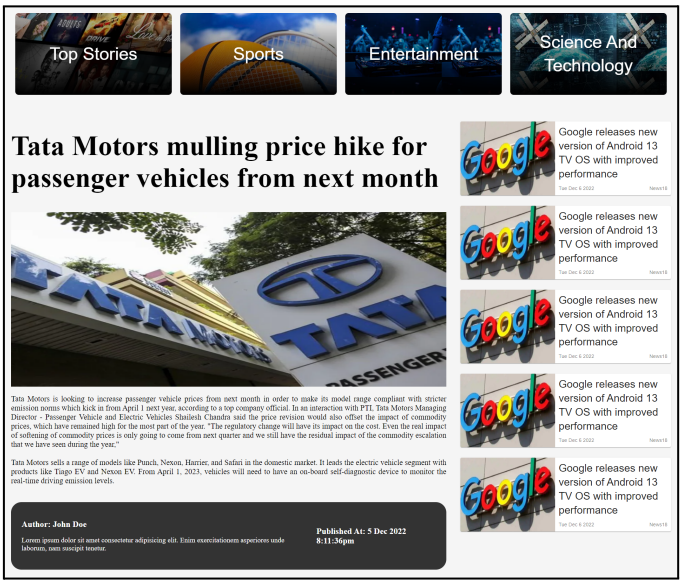

Fig 4: Detail Article Page

The Detail Article Page of the proposed system is seen in Fig 4. This page displays a thorough summary of the relevant news item, complete with photographs, the author, and other details. Also, this website offers cards of different news categories so that users may easily browse through articles in a certain topic. Recommendations for articles are also displayed on this website.

5. Conclusion

A news aggregator and summarizer was developed in this study to aggregate news articles from various websites and display the most relevant summarised article to the user. Everyone can use this system to read the news and keep up with current events around the world. On the basis of previous user behaviour and interests, the algorithm also makes news article suggestions. When the end user is unable to read the news on their own, the proposed system includes a voice assistant feature that reads the summarised news articles for them.

Future development of the suggested system could add news verification in an effort to spot fake news by estimating its likelihood of being untrue. The system can also be improved by adding voice assistant functions that are available in a variety of languages.

6. References

[1] *A. Mohamed, M. Ibrahim, M. Yasser, M. Ayman, M. Gamil, W. H. El Ashmawi, "News Aggregator and Efficient Summarization System". International Journal of Advanced Computer Science and Applications 11. 636-641. 10.14569/IJACSA.2020.0110677.*

[2] *Alan AI | Conversational Voice AI Platform." https://alan.app/ (accessed May 13, 2022).*

[3] *Athalye, Shweta, "Recommendation System for News Reader" (2013). Master's Projects. 294. DOI:https://doi.org/10.31979/etd.xn48-6q4j.Available:https://scholarworks.sjsu.edu/etd_projects/294.*

[4] *F. Hamborg, N. Meuschke, and B. Gipp, "Matrix-based news aggregation: exploring different news perspectives," in Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. IEEE Press, 2017, pp. 69–78.*

[5] *S. Qaiser, & R. Ali. (2018). "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". International Journal of Computer Applications (0975 – 8887). 181. 10.5120/ijca2018917395. Volume 181 – No.1, July 2018.*

[6] *Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology, (ICITEE), 2014.*

[7] *Trajkovski, I. (2014). "Pagerank-Like Algorithm for Ranking News Stories and News Portals". in ICT Innovations 2013. Advances in Intelligent Systems and Computing, vol 231. Springer, Heidelberg. ICT Innovations 2013 pp 87–96.*

[8] *D. Deng, "DBSCAN Clustering Algorithm Based on Density," in 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.*

[9] *A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.*

[10] *A. K. Yadav, M. Kumar, A. Pathre, "Implemented Text Rank based Automatic Text Summarization using Keyword Extraction" International Research Journal of Innovations in Engineering and Technology - IRJIET, Volume 4, Issue 11, pp 20-25, November 2020.*

[11] *S. Alhojerry and J. Kalita, "Recent Progress on Text Summarization," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 1503-1509, doi: 10.1109/CSCI51800.2020.00278.*

[12] *FeedSpot. (2017, Dec, 19). Top 100 Indian News RSS Feeds[Online]. Available: https://blog.feedspot.com/indian_news_rss_feeds/.*

[13] *R. Gowtham & S. Doddapaneni, A. Bheemaraj, M. Jobanputra, & R. AK, A. Sharma, S. Sahoo, H. Diddee, J. Mahalakshmi D. Kakwani, N. Kumar A. Pradeep, D. Kumar, V. Raghavan, A. Kunchukuttan, P. Kumar, M. Khapra. (2021). "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages".*