WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Samachar - News Aggregator

Rutuja Dighe, Sakshi Sawant, Rhea Sinha, Prof. Mohan Bonde

Usha Mittal Institute of Technology, SNDT Women’s University, Mumbai, India 400049
Emails: rutujadighe8237@gmail.com, saksh2003sawant@gmail.com, rheasinha1234@gmail.com, mohan.bonde@umit.sndt.ac.in

DOI: https://doi.org/10.55248/gengpi.5.0424.0957

ABSTRACT—

Samachar is a web-based news application designed to offer users a convenient and efficient means of staying abreast of current events. Utilizing web scraping techniques, the appli- cation gathers articles from diverse sources, which are then con- cisely summarized using the TextRank algorithm. Users have the flexibility to tailor their news feed by selecting preferred sources and categories, ensuring relevance to their interests. Moreover, Samachar features a robust search functionality, enabling users to locate specific articles with ease. The primary objective of this research is to develop a user-friendly and dependable news aggregator that streamlines the process of accessing information.

Index Terms—: *web scraping, Sqlite, TextRank algorithm, news aggregation, search functionality, user customization*

I. INTRODUCTION

The expanse of the internet has exponentially increased access to vast amounts of unstructured data, presenting both opportunities and challenges in information management. Particularly, the analysis and summarization of unstructured text data poses significant hurdles, necessitating innovative approaches to efficiently organize and retrieve information. Web scraping and summarization techniques offer promising solutions by systematically gathering and condensing data from disparate sources, thereby facilitating streamlined access and analysis.

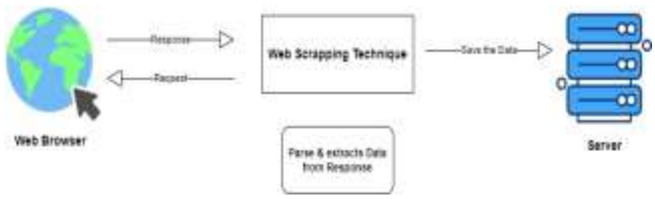


Fig. 1. Web Scrapping Architecture

As shown in Figure 1. It involves making an HTTP request to a website’s server, downloading the page’s HTML and parsing it to extract the desired data [1].

This research project endeavors to develop a comprehensive web scraping and summarization framework tailored for text data sourced from diverse websites, focusing on genre-based summarization. Web scraping is the process of extracting data from a specific web page [2]. Leveraging Python modules such as Beautiful Soup, Requests, and Newspaper, the framework aims to automate the process of data collection and summarization, enabling efficient condensation of website content based on their genre [3].

The emergence of digital media platforms, coupled with the boost of online content sources, has led to an era of information overload, where individuals struggle to navigate the deluge of available information effectively. In response, news aggregation platforms have emerged as a viable solution, offering users a centralized platform for accessing curated news content from various sources conveniently [4]. This project seeks to address the limitations of traditional news consumption methods by providing users with a centralized hub for accessing news articles tailored to their preferences[5]. Central to this endeavor is the development of a user- friendly news aggregation platform equipped with customized features, including personalized news feeds [6], topic-based summarization, and intelligent summarization algorithms. By condensing news articles from diverse sources into concise summaries on a single platform, the project aims to streamline the process of information consumption, thereby empowering users with greater control over their news consumption habits [7].

The research questions underpinning this study are designed to evaluate the efficacy and usability of the news aggregation platform, as well as the underlying technologies and method- ologies employed in its development. Key areas of inquiry include the effectiveness of web scraping techniques in gather- ing news articles from diverse online sources, the performance of extractive summarization algorithms in condensing content while preserving its essence, and user perceptions regarding the platform’s features and functionality.

Furthermore, the project aims to address pertinent issues such as privacy concerns associated with data collection for personalization purposes, potential algorithmic biases in content representation, and the challenges inherent in web scraping and content summarization.By comprehensively ad- dressing these research questions, this study aims to contribute valuable insights to the field of news aggregation systems, informing future research and development efforts in this rapidly evolving domain.

Moreover, the project holds significance in its potential to improve the accuracy and efficiency of content summarization, offering practical implications for market research, content analysis, and website classification. From a societal stand- point, the development of a news aggregation platform has broader implications for promoting media literacy, facilitating informed civic engagement, and combating misinformation,

thereby contributing to the enhancement of public discourse and democratic participation in the digital age.

II. METHODOLOGY

Web scraping, also known as web harvesting or web data extraction, is a technique that extracts data from websites such as news websites, blogs, forums, and social network- ing platforms. Python libraries such as Requests, Beautiful Soup, and Newspaper are utilized for efficient web scraping

. Researchers target specific domains, topics, or regions to collect relevant news articles, retrieving HTML content from web pages, parsing it to extract desired information such as article titles, publication dates, and article text, and storing the extracted data in a structured format for further analysis. Techniques for handling dynamic content and overcoming anti-scraping measures are employed to ensure comprehensive data collection.

keeping crucial informational aspects and the meaning of content, is known as summarization of the news article. This will enable the reader to obtain all relevant news information by simply reading a brief synopsis. As a result, we use natural language processing to make this happen [9]. Natural language processing (NLP) is an artificial intelligence discipline in which computers intelligently analyze, comprehend, and infer meaning from human language.

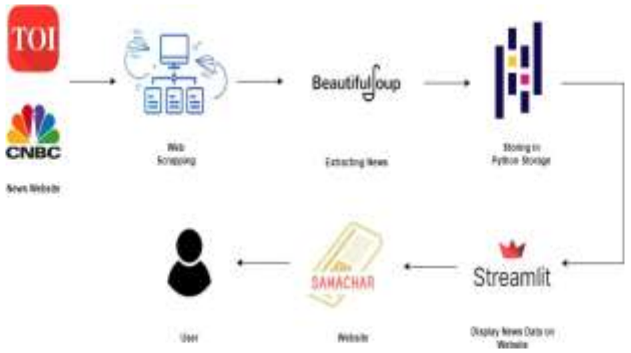


Fig. 2. Workflow of Samachar

Figure 2 provides a comprehensive view of the above process in a diagrammatic manner. It defines the flow of content through the website [8].

Once news articles are scraped, the data is organized and stored in a structured database for efficient retrieval and management. SQLite, a lightweight relational database management system, is utilized for storing scraped articles. The database is structured with separate tables for article metadata, ensuring data integrity and facilitating querying and analysis. Articles are organized into distinct tables based on categories, genres, or sources, enabling targeted analyses and comparisons.

Text summarization is a crucial component of news ag- gregation systems, enabling users to digest large volumes of text content efficiently. The TextRank algorithm, an unsuper- vised graph-based approach, is employed for automatic text summarization. TextRank constructs a graph representation of the text, ranking sentences based on their importance scores to generate concise and informative summaries. The algorithm considers factors such as sentence position, length, and semantic similarity to identify key sentences capturing the essential information of the text.

The goal of compressing the whole news article into a shorter version, lowering the size of the original text while

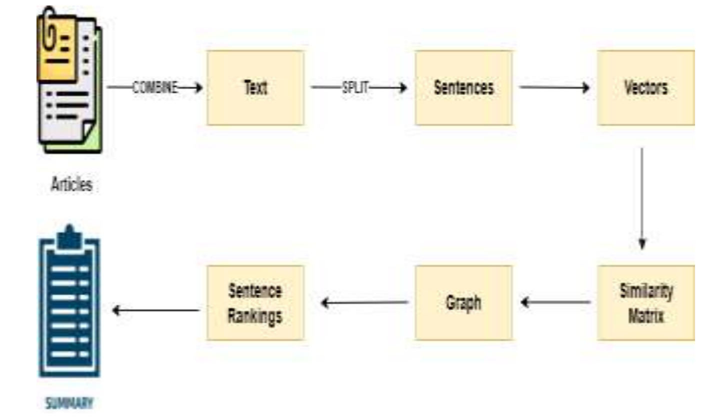


Fig. 3. Working of Textrank Algorithm

Fig. 3. depicts the splitting of text into sentences, con- vert each to a vector then Construct similarity matrix and graph from vectors. After this Rank sentences uses random walk algorithm to select top-ranked sentences to form summary.

Text Preprocessing:

Text preprocessing involves several steps to prepare the input text for summarization. Initially, the input text is tokenized into individual words or sentences, facilitating further analysis at a granular level. Following tokeniza- tion, commonly occurring words that do not contribute significantly to the semantic meaning of the text, known as stopwords, are removed. This step helps to eliminate noise and focus on more meaningful content. Addition- ally, words are often lemmatized or stemmed to reduce them to their base or root form, promoting normalization and reducing redundancy within the text.

Graph Representation:

In the graph representation stage, the preprocessed text is transformed into a structured graph where nodes correspond to individual sentences, and edges denote the similarity between sentences. The similarity between sentences is typically calculated using cosine similarity, based on their TF-IDF (Term Frequency-Inverse Doc- ument Frequency) representations. This representation allows for a comprehensive understanding of the relation- ships between sentences, forming the basis for subsequent ranking and summarization steps. The calculated similar- ity scores serve as edge weights in the graph, capturing the strength of connections between sentences.

Graph Ranking:

Graph ranking involves applying an algorithm, often

inspired by Google’s PageRank algorithm [10], to rank sentences within the constructed graph. The PageRank algorithm assigns importance scores to sentences based on their connectivity and the importance of sentences they are connected to. This iterative calculation process continues until convergence, ensuring that the ranking reflects the relative significance of each sentence within the context of the document. By considering both the local importance of sentences and their global impact on the entire document, the ranking process identifies key sentences that are essential for summarization.

Summary Generation:

Using the ranked sentences, the summary generation stage selects the top-ranked sentences based on their PageRank scores to compose the final summary. The length of the summary can be predetermined or dynami- cally determined based on factors such as the input text length or user preferences. By prioritizing sentences with higher importance scores, the summary captures the most salient information from the original text, providing a concise and informative overview.

Implementation:

TextRank summarization can be implemented using pro- gramming languages such as Python, leveraging libraries like NLTK (Natural Language Toolkit) for text prepro- cessing and scikit-learn for TF-IDF vectorization. Fur- thermore, integration with a database system enables effi- cient processing of large volumes of text data, facilitating real-time summarization tasks in diverse applications.

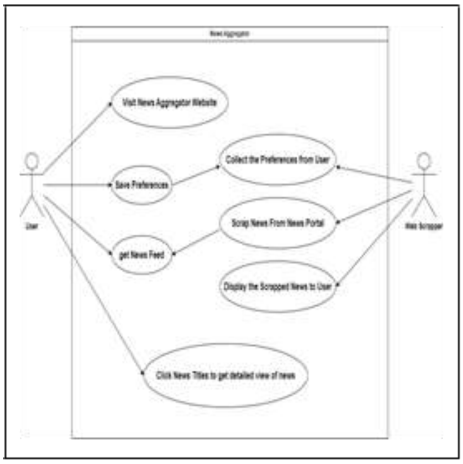

News Aggregator
Visit News Aggregator Website
Collect the Preferences from User
Save Preferences
Scrap News From News Portal
get News Feed
Display the Scrapped News to User
Click News Titles to get detailed view of news
User
Web Scrapper

Fig. 4. Use case Diagram of Website

Figure 4 illustrates the process wherein a user visits a news ag- gregator website, sets preferences including topics of interest and preferred news sources. The aggregator collects news from diverse sources and presents it in a unified feed. Users can click on news titles for detailed views, though this is optional for the aggregator's primary function.

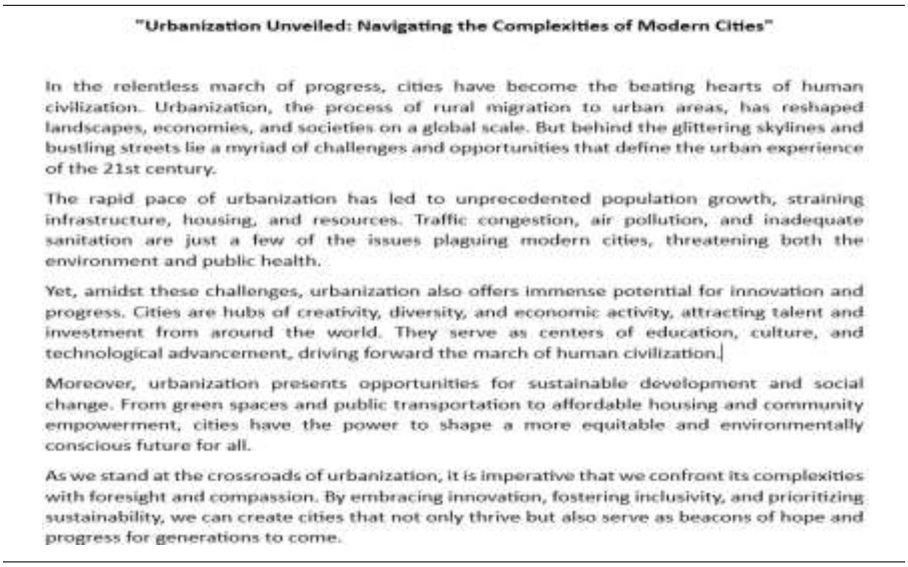
The registration process allows users to create accounts and access the full range of features offered by the news aggrega- tion platform. During registration, users provide personal in- formation, which is securely stored in the platform's database for authentication and account management. Passwords are securely encrypted to ensure account security. Additionally, users specify their preferences and interests, enabling the platform to deliver personalized content recommendations.

The email newsletter feature enables the platform to extend its reach beyond the website by delivering curated news updates directly to users' email inboxes. Users can subscribe to newsletters based on their interests and preferences, allowing for strategic segmentation and targeting . Integration with backend email delivery services and analytics tools facilitates tracking of user interactions and optimization of content delivery strategies.

RESULTS AND ANALYSIS

Our hypothesis suggested that the integration of advanced technologies, including web scraping, text summarization, and personalized content delivery mechanisms, would lead to enhanced user engagement, satisfaction, and information consumption behavior within the news aggregation platform. We anticipated that these features would improve the rele- vance, timeliness, and personalization of news content, thereby fostering a more dynamic and interactive user experience.

Experimental Analysis



"Urbanization Unveiled: Navigating the Complexities of Modern Cities"

In the relentless march of progress, cities have become the beating hearts of human civilization. Urbanization, the process of rural migration to urban areas, has reshaped landscapes, economies, and societies on a global scale. But behind the glittering skylines and bustling streets lie a myriad of challenges and opportunities that define the urban experience of the 21st century.

The rapid pace of urbanization has led to unprecedented population growth, straining infrastructure, housing, and resources. Traffic congestion, air pollution, and inadequate sanitation are just a few of the issues plaguing modern cities, threatening both the environment and public health.

Yet, amidst these challenges, urbanization also offers immense potential for innovation and progress. Cities are hubs of creativity, diversity, and economic activity, attracting talent and investment from around the world. They serve as centers of education, culture, and technological advancement, driving forward the march of human civilization.

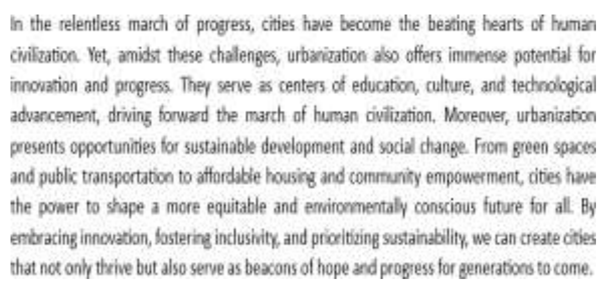
Moreover, urbanization presents opportunities for sustainable development and social change. From green spaces and public transportation to affordable housing and community empowerment, cities have the power to shape a more equitable and environmentally conscious future for all.

As we stand at the crossroads of urbanization, it is imperative that we confront its complexities with foresight and compassion. By embracing innovation, fostering inclusivity, and prioritizing sustainability, we can create cities that not only thrive but also serve as beacons of hope and progress for generations to come.

Fig. 5. Original Article

The survey utilised a comparative analysis of algorithm- generated summaries, presenting participants with two sum- maries each, labeled as Summary 1 and Summary 2 of the above article. Figure 5 contains the original article used as an Input to generate the two summaries. A rating scale ranging from

1 to 5 was utilized, where 1 represented the least preferred and 5 denoted the most preferred summary.[7] Participants were tasked with ranking the summaries accord- ing to their preference level. Additionally, participants were prompted to indicate their overall preferred summary. The survey was conducted via the Google Forms platform and gained responses from 40 participants, ensuring an impartial and unbiased environment for data collection, while also guaranteeing anonymity for participants.

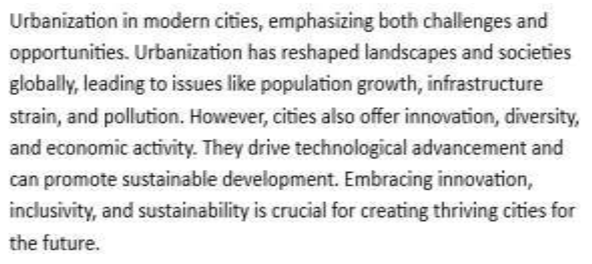


In the relentless march of progress, cities have become the beating hearts of human civilization. Yet, amidst these challenges, urbanization also offers immense potential for innovation and progress. They serve as centers of education, culture, and technological advancement, driving forward the march of human civilization. Moreover, urbanization presents opportunities for sustainable development and social change. From green spaces and public transportation to affordable housing and community empowerment, cities have the power to shape a more equitable and environmentally conscious future for all. By embracing innovation, fostering inclusivity, and prioritizing sustainability, we can create cities that not only thrive but also serve as beacons of hope and progress for generations to come.

Fig. 6. The TextRank Summary

TextRank Algorithm (Summary 1): _

The Figure 6 depicts the summary generated using TextRank Algorithm. It is a graph-based approach to extract key sen- tences from a text document. It identifies the importance of sentences based on their similarity to other sentences in the document. The implementation involves preprocessing the text, constructing a graph representation, ranking sentences using an iterative algorithm, and selecting the top-ranked sentences to form the summary.



Urbanization in modern cities, emphasizing both challenges and opportunities. Urbanization has reshaped landscapes and societies globally, leading to issues like population growth, infrastructure strain, and pollution. However, cities also offer innovation, diversity, and economic activity. They drive technological advancement and can promote sustainable development. Embracing innovation, inclusivity, and sustainability is crucial for creating thriving cities for the future.

Fig. 7. The Word Frequency Summary

Word Frequency Algorithm (Summary 2): _

The figure 7 illustrates the summary generated using Word Frequency algorithm. It generates summaries based on the frequency of words appearing in the text. It identifies the most frequently occurring words or phrases and selects sentences containing these words to compose the summary.

Analysis

The preferences of participants for each summary were calculated by aggregating the rankings assigned by all par- ticipants.

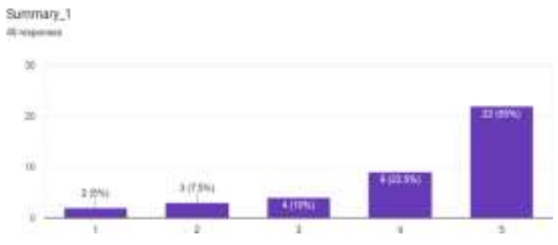


Fig. 8. TextRank Chart

1) *Figure 8 portrays:* Percentage of participants assigning each mark: 2.5% (1 mark), 7.5% (2 marks), 10% (3 marks), 22.5% (4 marks), 55% (5 marks). Overall preference: 72.5% of participants preferred Summary 1.

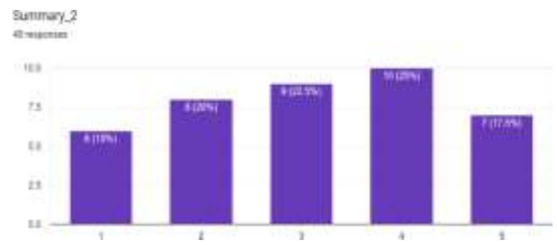


Fig. 9. Word Frequency Chart

2) *Figure 9 portrays:* Percentage of participants assigning each mark: 15% (1 mark), 20% (2 marks), 22.5% (3 marks), 25% (4 marks), 17.5% (5 marks). Overall preference: 27.5% of participants preferred Summary 2.

Inference

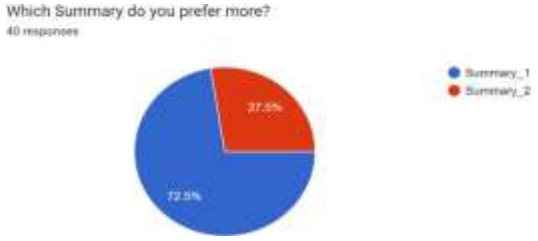


Fig. 10. Summary Preference Chart

The survey results indicate a significant preference for summaries generated using the TextRank algorithm

(Summary 1) over those generated using the Word Frequency algorithm (Summary 2).Summary 1, utilizing the TextRank algorithm, was preferred by 72.5% of participants, refer Figure 10, highlighting its effectiveness in capturing the essence of the text and generating more appealing summaries. These findings suggest that algorithms that incorporate semantic analysis and prioritize sentence importance, such as TextRank, may outperform simpler frequency-based approaches, like Word Frequency, in summarization tasks.

The results confirm the validity of our hypothesis, indicating significant improvements in user experience and satisfaction following the integration of advanced technologies within the news aggregation platform. The systematic collection of diverse news articles from online sources has enriched the platform's content repository, providing users with access to a wide range of information across various topics and interests. The implementation of the TextRank algorithm for auto- matic text summarization has notably enhanced content con- sumption efficiency by condensing lengthy articles into concise summaries [11]. This improvement has led to increased user engagement and satisfaction, as users can quickly grasp the key points of news articles without navigating through extensive texts.

Furthermore, the structured storage of scraped articles in an SQLite database has facilitated efficient data retrieval and management, enabling targeted analyses and personalized con- tent delivery based on user preferences. The integration of user registration and preference selection features has empowered users to customize their news feeds and receive personalized recommendations, resulting in heightened engagement levels. In conclusion, the detailed analysis of the project's results underscores the transformative impact of integrating advanced technologies and personalized features within the news ag- gregation platform. By leveraging web scraping, text sum- marization, and personalized content delivery mechanisms, the platform has significantly enhanced its content richness, accessibility, and user engagement dynamics. These findings signify a significant milestone in the evolution of digital news consumption, positioning the platform as a leading player in the industry.

4. CONCLUSIONS

The conclusion of the Samachar News Aggregator project marks a significant milestone in the successful integration of advanced technologies, including web scraping techniques, text summarization algorithms, and personalized content de- livery mechanisms, within the platform. The platform has aggregated a diverse range of content across numerous themes and interests using sophisticated web scraping techniques. This rich repository of articles serves as the foundation for providing users with a comprehensive and dynamic news experience.

Furthermore, the implementation of the TextRank algorithm for automatic text summarization has played a pivotal role in enhancing content consumption efficiency. By condensing lengthy articles into concise summaries, the platform enables users to quickly grasp the key points of news articles without the need to navigate through extensive texts.

The survey results further indicated a substantial prefer- ence among participants for summaries generated using the TextRank algorithm (Summary 1) over those generated using the Word Frequency algorithm (Summary 2). Notably, Sum- mary 1, employing the TextRank algorithm, garnered preference from 72.5% of participants, underscoring its efficacy in capturing the essence of the text and producing more engaging summaries. These findings imply that algorithms integrating semantic analysis and prioritizing sentence importance, like TextRank, hold an advantage over simpler frequency-based methods such as Word Frequency in the domain of summa- rization tasks.

In summary, the integration of web scraping techniques, text summarization algorithms, and personalized content delivery mechanisms has significantly enhanced the Samachar News Aggregator platform. The project's success validates the initial hypothesis and underscores the platform's commitment to pro- viding users with a more efficient, personalized, and satisfying news consumption experience.

ACKNOWLEDGMENT

We have a great pleasure to express our gratitude to all those who have contributed and motivated during our project work. Special thanks to our guide Prof. Mohan Bonde whose suggestions helped us in writing this paper.

REFERENCES

[1] S. F. F. S. I. H. Prof. Jyothi Patil, Saba Fatima, "News content aggregator using web scraping," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 05, no. 05, pp. 7500–7505, 2023.

[2] Techopedia, "Web scraping." [Online]. Available: https://www. techopedia.com/definition/5212/web-scraping

[3] A. M. K. B. Mr Rakesh Kumar Rai, Dr Isha Singh, "News aggrega- tor web application using django," *International Journal Of Creative Research Thoughts(IJCRT)*, vol. 09, no. 07, pp. 871–876, 2021.

[4] R. K. P. Dilip Suthar, Ajit Kumar Trivedi, "Web scraping news por- tals for the ease of news reading," *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, no. 05, pp. 2402–2405, 2021.

[5] C. Grozea, D.-C. Cercel, C. Onose, and S. Trausan-Matu, "Atlas: News aggregation service," in *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 2017, pp. 1–6.

[6] V. Singrodia, A. Mitra, and S. Paul, "A review on web scrapping and its applications," 01 2019, pp. 1–6.

[7] A. Mohamed, M. Ibrahim, M. Yasser, M. Ayman, M. Gamil, and W. El-Ashmawi, "News aggregator and efficient summarization system," vol. 11, pp. 636–641, 07 2020.

[8] M. Bhujbal, M. Bibawanekar, and P. Deshmukh, "News aggregation using web scraping news portals," *International Journal of Advanced Research in Science, Communication and Technology*, vol. Volume 3, pp. 2581–9429, 07 2023.

[9] P. B. D. R. Akhil Kumar K A, Basavaraj Raga, "Automatic news aggre- gator," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 03, no. 07, pp. 1828–1831, 2021.

[10] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: https://aclanthology.org/W04-3252

[11] I. J. of Engineering Science and A. T. (IJESAT), "Data