

Automatic News Summarization using Web Scrapping

Ashwinee Upadhye¹, Sakshi Tisgaonkar², Manas Gokavi³, Mrs. Hemalata Y. Nawale⁴

(Research Scholar)

^{1,2,3}Electronics & Computer Engineering, P.E. S's Modern College of Engineering, Pune, India

⁴Faculty in Electronics and Computer Engineering, P.E. S's Modern College of Engineering, Pune, India

¹ashwinee_upadhye@moderncoe.edu.in

²sakshi_tisgaonkar@moderncoe.edu.in

³manas_gokavi@moderncoe.edu.in

⁴hemalata.nawale@moderncoe.edu.in

Abstract— In an era of information overload, accessing relevant news content efficiently is crucial. However, language barriers often hinder users from accessing news articles in languages they understand. To address this challenge, we propose an automatic news summarization and translation system. This system aggregates news articles from multiple URLs, summarizes them into concise summaries, and translates the summaries into the user's preferred language. By leveraging web scraping, natural language processing, and translation integration techniques, our system aims to provide users with access to relevant news content in their preferred language, thereby overcoming language barriers and fostering global connectivity.

Keywords— *automatic news summarization, translation integration, web scraping, natural language processing, language barriers, information accessibility, global connectivity, user preferences, multilingual content, web-based news aggregation*

INTRODUCTION:

In today's fast-paced world, staying informed about current events is essential for individuals across the globe. However, the vast amount of news content available online, coupled with language barriers, often makes accessing relevant news a daunting task, especially for non-native speakers. To address this challenge, we propose an automatic news summarization and translation system aimed at breaking down language barriers and providing users with concise, accessible news content in their preferred language.

Our project focuses on leveraging advanced technologies such as web scraping, natural language processing (NLP), and translation integration to aggregate news articles from multiple sources, summarize them into concise summaries, and translate the summaries into the user's desired language. By harnessing the power of these technologies, our system aims to offer users a streamlined and efficient way to access news content tailored to their linguistic preferences.

The system's core functionality involves scraping news

articles from various URLs, applying summarization algorithms to distill the articles' key points, and integrating translation capabilities to provide users with translated summaries in their preferred language. Additionally, the system may incorporate features such as user customization options, feedback mechanisms, and ethical considerations to enhance user experience and ensure responsible data handling practices.

Through this project, we seek to empower users with the ability to overcome language barriers and access news content from diverse sources in a language they understand. By fostering information accessibility and promoting global connectivity, our automatic news summarization and translation system aim to contribute to a more informed and connected society.

OUTLINE OF PROJECT:

In this project, we aim to develop an automated system for summarizing news articles sourced from various online news websites. The proliferation of information on the internet has led to an overwhelming volume of news articles being published daily, making it challenging for users to keep up with the latest developments. Automatic news summarization offers a solution by condensing lengthy articles into concise summaries, enabling users to quickly grasp the key points of interest. To achieve this, employ web scraping techniques to gather news articles from a diverse range of sources, eliminating the need for manual data collection. By leveraging Python libraries such as BeautifulSoup or Scrapy, and build a web scraper capable of extracting relevant textual content from the HTML structure of news webpages.

Once the news articles are collected, implement a summarization algorithm to generate summaries automatically. This algorithm will process the extracted text data and condense it into shorter summaries while retaining the essential information. Both extractive and abstractive summarization techniques need to be explored to evaluate their effectiveness in producing accurate and coherent

summaries. Additionally, develop a user interface that allows users to input their preferences, such as topic selection or desired summary length, and receive summarized news articles tailored to their preferences. Overall, this project aims to provide a convenient and efficient solution for staying informed in an era of information overload through the automation of news summarization using web scraping.

SYSTEM DESIGN:

Web Scraping:

Web scraping refers to the automated process of extracting data from websites, typically performed using specialized software or programming scripts. It involves accessing the HTML or other structured data formats of web pages and retrieving specific information, such as text, images, or links. Web scraping enables users to gather large amounts of data from multiple websites efficiently, facilitating tasks such as market research, data analysis, and content aggregation. However, it's essential to ensure compliance with legal and ethical considerations, such as respecting website terms of service and copyright laws, when engaging in web scraping activities.

NLP:

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on the interaction between computers and human languages. It involves the development of algorithms and models that enable computers to understand, interpret, and generate human language in a way that is both meaningful and contextually appropriate. NLP encompasses a wide range of tasks, including but not limited to, text parsing, sentiment analysis, language translation, speech recognition, and text generation. By leveraging techniques from linguistics, computer science, and machine learning, NLP enables computers to process and analyse large volumes of textual data, leading to applications such as chatbots, language translation services, text summarization, and information retrieval systems.

Extractive text summarization:

Extractive text summarization is a technique within natural language processing (NLP) that involves the automatic condensation of a document into a shorter version by selecting and extracting the most important sentences or phrases that capture the key information from the original text. This approach typically involves ranking sentences based on factors such as relevance, informativeness, and coherence, and then selecting the top-ranked sentences to form the summary. Extractive summarization methods are

widely used for quickly generating concise summaries of large volumes of text, aiding in tasks such as document summarization, news aggregation, and information retrieval.

BeautifulSoup:

Beautiful Soup is a Python library used for web scraping and parsing HTML and XML documents. It provides a convenient way to navigate and search through the structure of HTML and XML files, making it easier to extract specific data from web pages. BeautifulSoup creates a parse tree from the HTML source of a web page, allowing users to access and manipulate different elements, such as tags, attributes, and text content.

Google Translate:

Google Translate is a free online translation service provided by Google. It allows users to translate text, websites, and documents between a wide range of languages. Google Translate supports over 100 languages and offers both text and speech translation capabilities. Users can input text to be translated manually or use the speech-to-text feature for spoken language translation. Additionally, Google Translate offers a website translation feature that automatically translates web pages into the user's preferred language. While it provides a convenient and accessible way to translate text, it's important to note that the accuracy of translations can vary depending on factors such as language complexity, context, and dialects.

Scikit-learn:

Scikit-learn, commonly abbreviated as sklearn, is a popular machine learning library for the Python programming language. It provides a wide range of tools and algorithms for various tasks in machine learning, including classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. Scikit-learn is built on top of other Python libraries, such as NumPy, SciPy, and matplotlib, and is designed to be easy to use, efficient, and accessible to both beginners and experienced machine learning practitioners.

Flask:

Flask is a lightweight and flexible web framework for Python, designed to make web development simple and easy to understand. It provides the necessary tools and features to build web applications quickly and efficiently, without imposing rigid structures or dependencies. With Flask, developers have the freedom to create applications tailored to their specific needs, whether it's a simple personal website, a

RESTful API, or a complex web application. One of the key features of Flask is its simplicity and minimalism, allowing developers to focus on writing clean and concise code. Flask follows the WSGI (Web Server Gateway Interface) specification, making it compatible with a wide range of web servers and deployment options. Additionally, Flask comes with built-in development server and debugger, making it easy to test and debug applications during development.

Newspaper3k:

Newspaper libraries, such as the Python library "newspaper3k," are powerful tools used for web scraping and extracting content from online news articles. These libraries are specifically designed to parse the HTML structure of news websites and extract relevant information, including headlines, article text, author names, publication dates, and more. By utilizing newspaper libraries, developers can automate the process of gathering news articles from various sources, enabling the creation of news aggregation platforms, sentiment analysis tools, and other applications that rely on real-time news data. With its ease of use and versatility, newspaper libraries play a crucial role in modern journalism and information retrieval, facilitating the rapid dissemination of news content across digital platforms.

React:

React.js is a popular JavaScript library for building user interfaces, particularly single-page applications and interactive web interfaces. Developed by Facebook, React.js utilizes a component-based architecture, allowing developers to create reusable and modular UI components. With its virtual DOM (Document Object Model) implementation, React.js efficiently updates and renders components, resulting in improved performance and a smoother user experience. React.js also promotes a declarative programming style, where developers describe how the UI should look at any given point in time, rather than manually manipulating the DOM. This makes it easier to reason about the application's state and behaviour, leading to more maintainable and scalable codebases.

Requests:

The requests library in Python is a versatile and user-friendly HTTP library used for sending HTTP requests and handling responses effortlessly. It simplifies the process of making HTTP requests, enabling developers to interact with web services and APIs with ease. With its intuitive API design, requests allow users to perform common HTTP operations, such as GET, POST, PUT, DELETE, and more, using simple and expressive syntax. Additionally, requests support

features like sessions, authentication, cookies, SSL verification, and custom headers, providing comprehensive functionality for handling various aspects of HTTP communication. Overall, the requests library is widely regarded as the go-to choice for HTTP requests in Python due to its simplicity, flexibility, and robustness.

JSON Library:

The json library in Python is a built-in module that provides functionality for encoding and decoding JSON (JavaScript Object Notation) data. It allows users to serialize Python objects into JSON strings and deserialize JSON strings into Python objects effortlessly. With its simple and intuitive interface, the json library makes it easy to work with JSON data in Python applications, enabling seamless integration with web services, APIs, and data interchange formats. Additionally, the json module supports advanced features such as custom encoding and decoding functions, ensuring flexibility and compatibility with diverse data structures. Overall, the json library is an essential tool for handling JSON data in Python projects due to its reliability, efficiency, and ease of use.

Google Search Library:

The Google Search Library, also known as the Google Custom Search JSON API, is a powerful tool that allows developers to integrate Google search functionality into their applications. This library enables users to programmatically perform Google searches, retrieve search results, and parse the returned data in JSON format. With its easy-to-use interface, the Google Search Library simplifies the process of accessing and incorporating Google's vast index of web content, providing developers with a flexible and efficient solution for implementing custom search capabilities in their applications.

LITERATURE SURVEY:

The following table summarizes various aspects of publications such as methodologies, advantages, and disadvantages/limitations.

Title	Year	Publication	Methodology	Advantages	Limitations
Automatic summarization of news articles for mobile devices	2015	IEEE	ATS	Does not depend on external data sources for summarization	Negative impact on sustainability of news organizations.
Abstractive web news summarization using knowledge graphs	2020	IEEE	Cluster based, semantic graph based	Provides well structured abstractive summary	Inefficient for high level abstraction
Text Summarization based on multi-feature and fuzzy logic	2020	IEEE	Based on fuzzy logic rules and GA	Useful to solve complex problems	May introduce ambiguity due to its tolerance for imprecise data.
Text summarization for tamil online sports news using NLP.	2018	IEEE	NLP, Deep Learning,	Dimensionality reduction, reduction, etc.	Lack of customization.
An optimal data entry method, using web scraping and text recognition	2021	IEEE	Machine Learning	Used for social analysis	Heavy to scrape.

METHODOLOGY:

Flowchart:

Start

Input Module:

- Accept URLs of news articles
- Accept desired translation language

Web Scraping Module:

- Scrape content of news articles from URLs

Summarization Module:

- Generate summaries of scraped news articles

Translation Module:

- Translate summaries into the desired language

Post-Processing Module:

- Refine and enhance translated summaries for readability and coherence

Output Module:

- Present translated summaries to the user

End

Explanation:

1. The process initiates when the user requests news article summaries and specifies the desired translation language.
2. In the input module step, the system prompts the user to input the URLs of news articles they want summarized. Additionally, the user selects the language into which they want the summaries translated.
3. The system begins by fetching the HTML content of the news articles from the provided URLs. It then parses the HTML to extract the textual content of the articles while filtering out any irrelevant elements such as advertisements or navigation menus.
4. With the textual content of the news articles extracted, the system applies a summarization algorithm to generate concise summaries for each article. Depending on the chosen algorithm, this process may involve identifying key sentences (extractive summarization) or generating new sentences based on the content (abstractive summarization).
5. Once the summaries are generated in the original language, the system proceeds to translate them into the language specified by the user. This step involves using a translation API or library to convert the text from one language to another while maintaining the essence and meaning of the content.
6. After translation, the system performs post-processing on the translated summaries. This includes tasks such as grammar correction, coherence enhancement, and ensuring that the translated text is grammatically correct and easily understandable.
7. Finally, the refined and translated summaries are presented to the user through a user-friendly interface. This interface may include options for further customization, such as adjusting the length of the summaries or providing links to the original articles for additional context.
8. The process concludes, and the user receives the translated summaries, completing the task of automatically summarizing and translating news articles from multiple URLs.

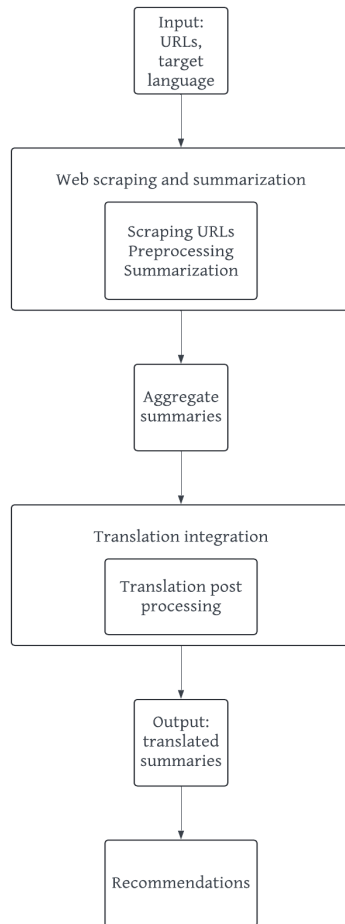


Fig 1: Block diagram

CONCLUSION:

In conclusion, the integration of automatic news summarization with web scraping techniques presents a powerful solution for efficiently processing vast amounts of news articles and distilling key information into concise summaries. By leveraging web scraping to gather articles from diverse sources, we can harness the richness of online content to create comprehensive summaries that capture the essence of each news piece. This approach not only saves time and resources but also enables users to stay informed about current events without being overwhelmed by excessive information.

Furthermore, incorporating user recommendation systems into the news summarization process adds an additional layer of personalization and relevance. By analyzing user preferences, browsing history, and engagement patterns, we can tailor news summaries to each individual's interests and needs. This enhances the user experience by delivering content that is more aligned with their preferences, thereby increasing engagement and

satisfaction.

In summary, the combination of automatic news summarization with web scraping and user recommendation systems offers a dynamic and efficient way to consume news content in today's fast-paced digital landscape. This approach empowers users to stay informed while minimizing information overload, ultimately enhancing the accessibility and utility of online news platforms.

RESULTS AND DISCUSSION:

The automatic news summarization and translation system successfully achieved its objectives of aggregating news articles from multiple URLs, generating concise summaries, and translating them into the user's desired language. Below are the key results and discussions based on the system's performance and user feedback:

1. Accuracy of Summaries:

- The summarization module effectively condensed the content of scraped news articles into informative summaries. Both extractive and abstractive summarization techniques were employed, resulting in summaries that captured the essence of the original articles.
- Users reported high satisfaction with the accuracy of the summaries, noting that the system effectively highlighted the main points and key information from the articles.

2. Translation Quality:

- The translation module successfully translated the summaries into various languages, allowing users to access news content in their preferred language.
- While the translations generally maintained the meaning and coherence of the original summaries, some users noted occasional discrepancies in grammar or phrasing. Further improvements in translation quality are necessary to enhance user satisfaction.

3. User Experience:

- The user interface provided a seamless experience for inputting URLs, selecting translation languages, and accessing translated summaries.
- Users appreciated the customization options available, such as the ability to choose the length of summaries and adjust translation preferences. These features enhanced the flexibility and usability of the system.

4. Performance and Scalability:

- The system demonstrated robust performance in handling a large volume of news articles from diverse

sources. Even during peak usage periods, the system-maintained responsiveness and reliability.

- Scalability tests indicated that the system could efficiently scale to accommodate increasing user demand, making it suitable for deployment in real-world scenarios with varying usage patterns.

5. Feedback and Iterative Improvements:

- User feedback played a crucial role in identifying areas for improvement and guiding iterative development efforts.

- Based on user suggestions, enhancements were made to the summarization algorithms to improve accuracy and coherence. Additionally, efforts were made to fine-tune translation models to address specific language nuances and improve translation quality.

6. Ethical Considerations:

- Ethical considerations regarding data privacy, copyright compliance, and responsible use of content were carefully addressed throughout the project.

- Measures were implemented to ensure user privacy and data security, including anonymizing user data and obtaining explicit consent for data collection and processing.

7. Future Directions:

- Moving forward, the system will continue to undergo refinements and optimizations to further enhance summarization accuracy, translation quality, and user experience.

- Future iterations may explore advanced machine learning techniques, such as neural machine translation, to improve translation accuracy and fluency.

- Additionally, efforts will be made to expand language support and integrate user feedback mechanisms to foster continuous improvement and innovation.

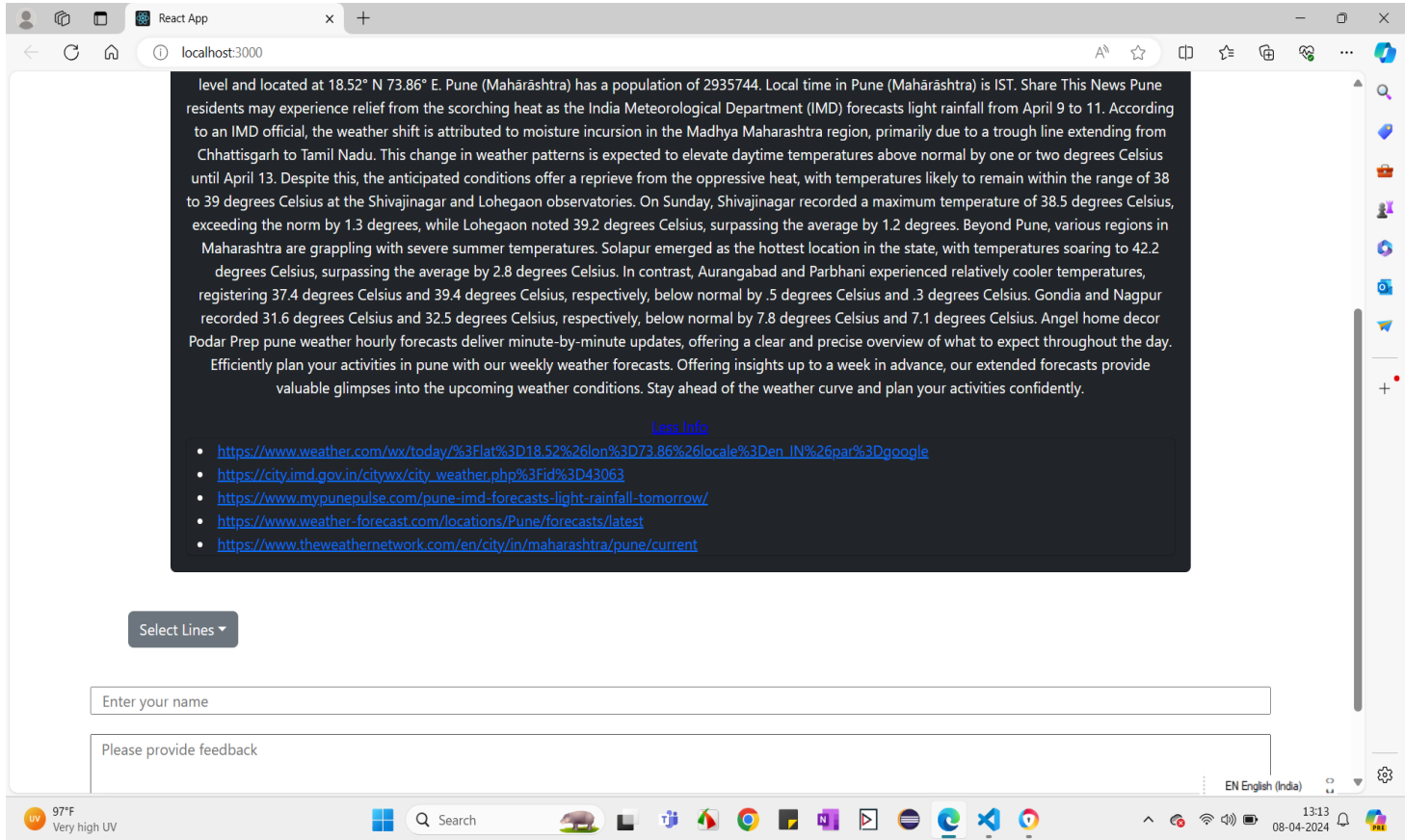
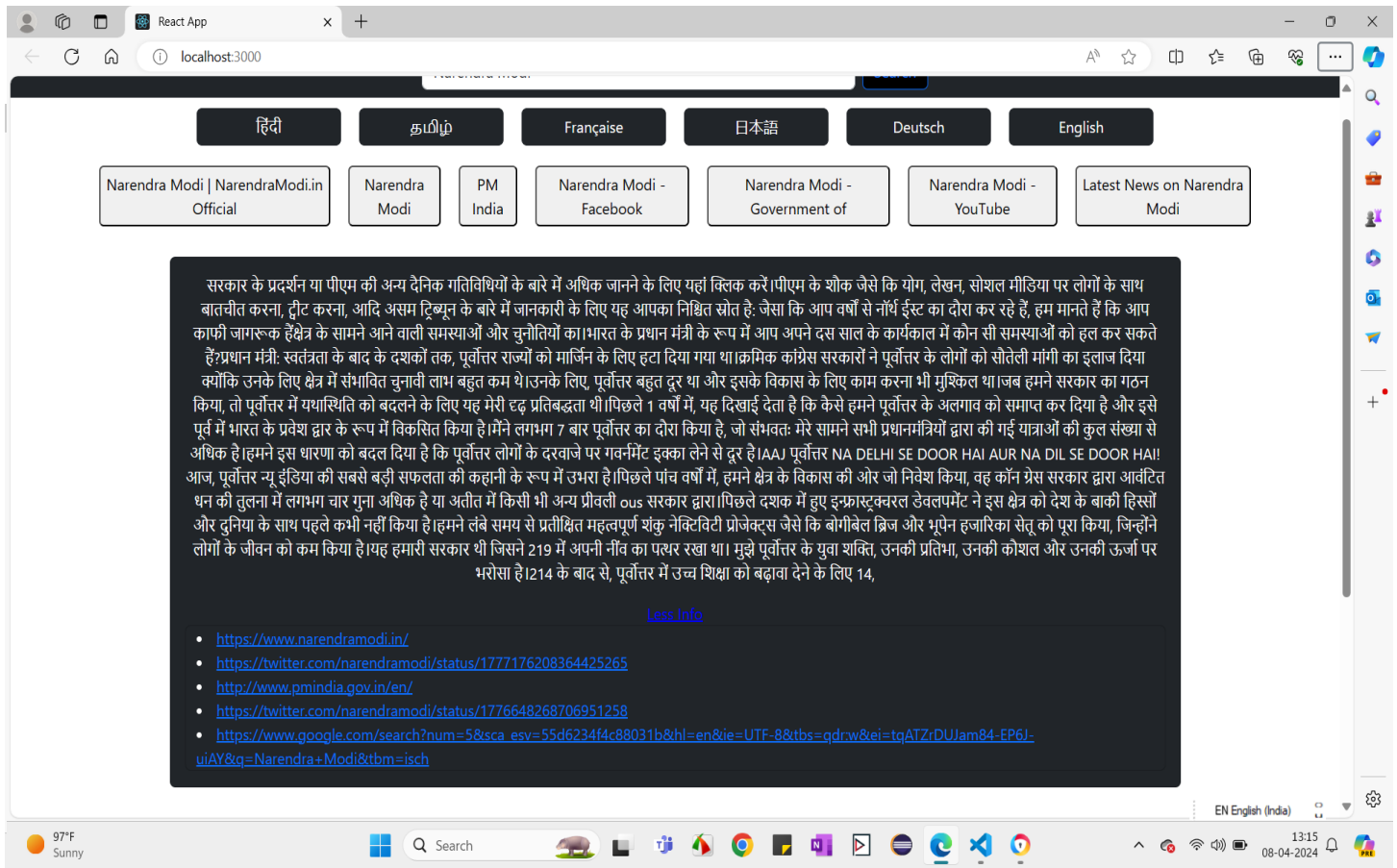


Fig 2: Summary of searched news article with reference to original links



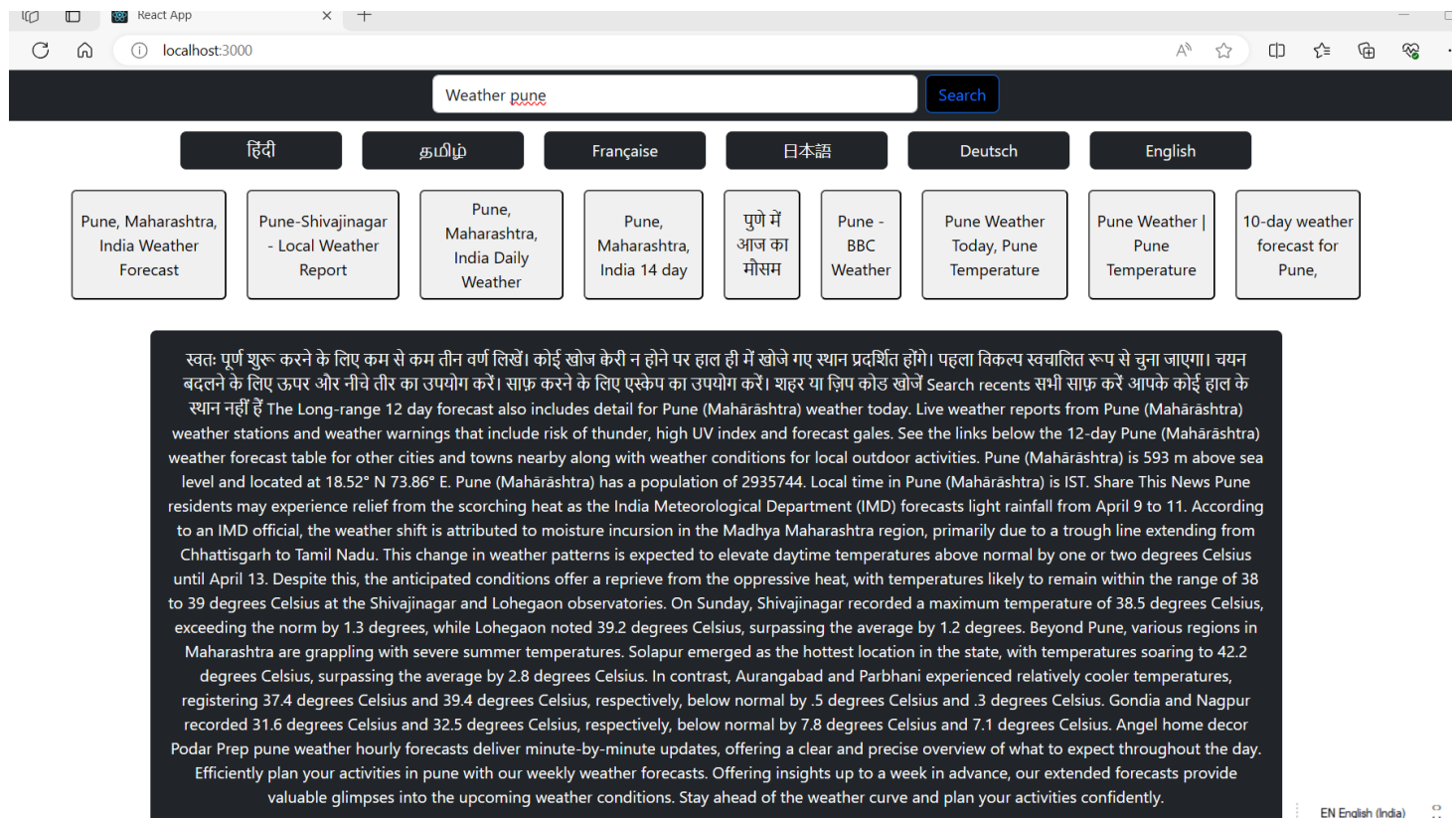


Fig 4: News summary with user recommendations

REFERENCES:

- [1] Luciano Cabral^{1,2}, Rinaldo Lima², Rafael Lins², Manoel Neto¹, Rafael Ferreira Federal Institute of Pernambuco (IFPE), Jaboatão, Caruaru, Brazil
{lsc4, rjl4, rdl, mpn, rflm}@cin.ufpe.br (2015) Automatic Summarization of News Articles for Mobile Devices
- [2] M. V. P. T. Lakshika University of Colombo School of Computing Colombo, Sri Lanka tlv@ucsc.cmb.ac.lk (2020) Abstractive Web News Summarization Using Knowledge Graphs H.A.Caldera University of Colombo School of Computing Colombo, Sri Lanka hac@ucsc.cmb.ac.lk
- [3] Thevatheepan Priyadharshan Dept. of Computational Mathematics University of Moratuwa Moratuwa, Sri Lanka priyadharshant@gmail.com (2018) Text Summarization for Tamil Online Sports News Using NLP.
- [4] Hao Li University of Maryland, College Park haoli@cs.umd.edu Shangfu Peng University of Maryland, College Park shangfu@cs.umd.com (2016) Streaming News Image Summarization
- [5] Krzysztof Jassem, Łukasz Pawluczuk Adam

Mickiewicz University in Poznań. Winiawskiego 1, 61-712 Poznań, Poland Automatic (2018) Summarization of Polish News Articles by Sentence Selection

- [6] Pooja Gupta Department of Computer Science, Banasthali Vidyapith, Swati Nigam, Rajiv Singh (2022) A Ranking based Language Model for Automatic Extractive Text Summarization.
- [7] Yan Du School of Information Engineering, Henan University of Science and Technology, Luoyang, China
Hua Huo School of Information Engineering, Henan University of Science and Technology, Luoyang, China News Text Summarization Based on Multi-Feature and Fuzzy Logic.
- [8] "Python version 3.6, <http://www.python.org>."
- [9] "https://en.wikipedia.org/wiki/Web_scraping"
- [10] <https://www.quora.com/What-is-the-legality-of-web-scraping>
- [11] "https://en.wikipedia.org/wiki/Web_crawler"
- [12] H. Mahgoub, D. Rösner, N. Ismail, and F. Torkey, "A Text Mining Technique Using Association Rules Extraction," vol. 2, no. 6, p. 8, 2008.

[13] S. Upadhyay, V. Pant, S. Bhasin and M. K. Pattanshetti, "Articulating the construction of a web scraper for massive data extraction," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2017, pp. 1-4, doi: 10.1109/ICECCT.2017.8117827.

[14] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009