News Nexus: A Comprehensive News Portal with Search, Recommendations, and Analysis

Aryan Singh
Mehta Family School Of Data Science And Artificial Intelligence
Indian Institute of Technology Guwahati
aryan.s@iitg.ac.in

Mihir Kumar Singh
Department of Physics
Indian Institute of Technology Guwahati
mihir.singh@iitg.ac.in

***Abstract*—This report presents a focused exploration of the New York Times dataset, emphasizing trend and sentiment analysis, and the implementation of a semantic search engine. Leveraging time-series data mining for trend identification and NLP models for sentiment analysis, the study sheds light on evolving patterns and emotional tones within news content. A significant aspect is the development of a semantic search engine, enhancing user retrieval experiences by grasping contextual nuances. The findings contribute to data-driven journalism and information retrieval, providing insights applicable to media monitoring, content recommendation, and trend forecasting.**

I. Introduction

1) **Deciphering Digital News** In the fast-paced digital era, navigating through an abundance of news articles efficiently is a common challenge for information seekers. The need for a robust and swift search mechanism is crucial to enhance the user's experience in accessing pertinent news content. This project, centered on The New York Times dataset, aims to revolutionize the way users decipher digital news by introducing a cutting-edge semantic search engine. Leveraging advanced techniques, such as sentence transformers and a dedicated ChromaDB, we aspire to provide users with a seamless and intelligent platform for exploring news articles.
2) **Transforming Journalism with Data** The motivation behind this endeavor lies in the transformative power of data analytics to reshape journalism in the digital age. The New York Times dataset, a vast repository of news articles, serves as the catalyst for this project. By addressing the need for rapid and precise information retrieval, we seek to empower users, journalists, and researchers alike. Our motivation is to contribute to the evolution of journalism by harnessing the potential of advanced technologies, offering a more intuitive, efficient, and meaningful news exploration experience for the global audience.

II. Literature Review

1) **Search Engine Algorithms** The landscape of search engine algorithms is multifaceted, ranging from traditional TF-IDF and Boolean models to contemporary approaches like Word2Vec, Doc2Vec, and advanced deep learning models such as BERT and sentence transformers. These algorithms employ diverse strategies, from semantic embeddings to contextual understanding, each impacting the efficiency and accuracy of news article retrieval. A comprehensive exploration of these algorithms offers valuable insights into their capabilities and potential synergies.
2) **Recommendation System Algorithms** Recommendation systems, integral to personalized content delivery, embrace various algorithms. Collaborative filtering, content-based approaches, and matrix factorization methods like SVD and ALS have been foundational. Modern neural collaborative filtering and hybrid models signify the evolution of recommendation techniques. Understanding the strengths and limitations of these algorithms is crucial for constructing an adaptive and user-centric news article recommendation system.
3) **Sentiment Analysis Approaches** Sentiment analysis, aimed at deciphering emotional tones within textual content, employs several methodologies. Rule-based systems establish sentiment based on predefined rules, while machine learning models, including Support Vector Machines (SVM) and Naive Bayes, discern sentiment from labeled training data. Deep learning models like LSTM and BERT capture intricate contextual nuances. Additionally, ensemble methods and lexicon-based approaches contribute to the diversity of sentiment analysis techniques. Exploring these approaches provides a nuanced understanding of sentiment analysis in the context of news articles.

III. Problem Statement

This project aims to unlock the dataset's potential by addressing three key challenges, contributing to a holistic news consumption experience.

1) **Semantic Search Engine**
   In this project, we are set to develop an advanced search engine leveraging various embedding techniques, including pre-trained sentence transformers, Word2Vec models, and other suitable methodologies. The goal is to harness the power of embeddings to measure the semantic similarity between news articles, moving beyond conventional keyword-based methods. By incorpo-

rating pre-trained models like sentence transformers or exploring custom Word2Vec implementations, we aim to capture and represent the intricate semantic relationships within the articles' content. Additionally, our semantic search engine incorporates the FAISS library for efficient similarity search, ensuring a robust and accurate recommendation system that significantly improves the overall user experience in exploring news articles.

2) **Content-Based Recommendation System**
In the Content-Based Recommendation System segment, our focus is on providing personalized article recommendations to users based on their interactions with the platform. When a user initiates a search through our website's search bar, we leverage this valuable query information to retrieve relevant articles. Our recommendation engine operates on the principles of content similarity, utilizing fast algorithms to analyze the relationship between articles and the user's search query. The system goes beyond basic keyword matching, incorporating advanced techniques such as topic modeling and feature extraction to enhance recommendation accuracy. This ensures that users receive tailored suggestions, creating a more engaging and personalized news consumption experience. Fast algorithms for content-based recommendation, including but not limited to TF-IDF (Term Frequency-Inverse Document Frequency) and collaborative filtering, are explored to streamline the recommendation process and provide timely and relevant suggestions.

3) **Comprehensive News Analysis**

   a) Sentiment Analysis: Conducting sentiment analysis on news article content or headlines is pivotal for understanding public sentiment on various topics. Leveraging pre-trained models like VADER (Valence Aware Dictionary and sEntiment Reasoner) or BERT (Bidirectional Encoder Representations from Transformers) tailored for news content, our system aims to track sentiment trends over time. This involves identifying articles with polarizing opinions, providing valuable insights into the emotional tone of news articles within The New York Times dataset.
   b) Topic Modeling: Implementing topic modeling techniques, such as Latent Dirichlet Allocation (LDA), enables the identification of key topics and themes within news articles. This contributes to creating a topic hierarchy for improved content organization. By utilizing the inherent topical diversity of The New York Times dataset, our system enhances the user's ability to explore and engage with a broad spectrum of news content.
   c) Keyword Extraction and Visualization: Employing algorithms like TF-IDF or RAKE (Rapid Automatic Keyword Extraction), our system extracts and visualizes keywords and phrases from news articles. This results in the creation of word clouds, graphs, and charts that showcase frequently mentioned terms. This aids users in quickly grasping the core themes and trending topics within the dataset.
   d) Section Popularity Analysis: Analyzing section popularity involves exploring user engagement with different sections of news articles. Utilizing statistical analysis and trend detection algorithms, our system identifies which sections are most popular among readers. This insight allows for dynamic recommendations, improving user experience based on evolving trends.
   e) Authorship Analysis: By implementing authorship analysis, our system studies the writing styles and preferences of different authors within The New York Times dataset. Utilizing techniques such as author-topic modeling or stylometric analysis, it identifies prolific authors and assesses their impact on readers. This enhances content personalization and user engagement.
   f) Content Classification: Content classification is achieved by developing a text classification model tailored to categorize news articles into genres or categories (e.g., politics, sports, technology). Leveraging techniques like supervised machine learning or deep learning, our system enhances content organization and navigation, providing users with a well-categorized and personalized news experience.
   g) Interactive Data Visualization Dashboard: The creation of an interactive dashboard involves utilizing visualization libraries like Plotly or D3.js. By integrating this feature, users can explore trends, topics, and news articles within The New York Times dataset through visually compelling graphs and charts. This enhances user engagement and provides a user-friendly interface for comprehensive data exploration.
   h) Social Network Analysis: Employing social network analysis involves analyzing connections and influence networks between authors, articles, and readers within The New York Times community. Utilizing graph theory and network analysis algorithms, our system identifies patterns of collaboration and engagement. This contributes to understanding the broader social context of news consumption, facilitating a richer user experience.

IV. DATASET

At the heart of our project lies the expansive and dynamic NYT Articles Dataset, a comprehensive collection of articles from The New York Times spanning from January 1, 2000, to the present day. This dataset, boasting an impressive 2 million rows, stands as a living resource, updated daily through seamless integration with the New York Times API and Kaggle

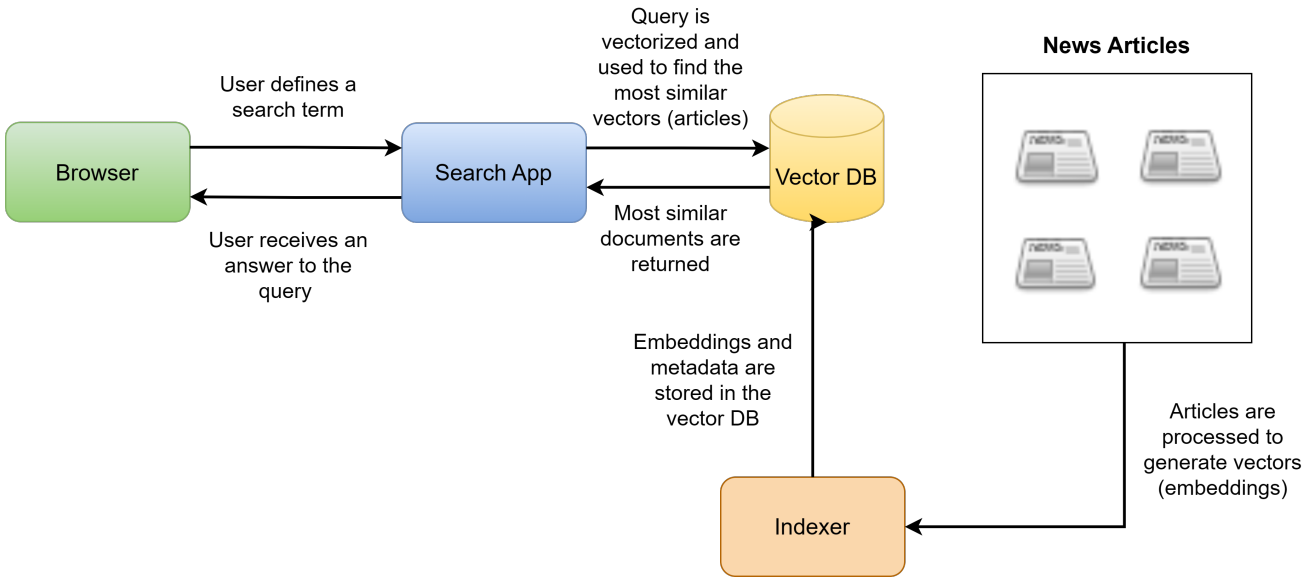

Fig. 1. An overview of the search engine pipeline

notebooks' auto-scheduling feature. It ensures an up-to-date and evolving source for analysis, incorporating the latest articles from The New York Times.

1) Dataset Generation To curate this dataset, we employed a robust web scraping approach, utilizing technologies such as Beautiful Soup and Scrapy in Python. This dynamic process involves extracting pertinent information from The New York Times website, ensuring a diverse and extensive collection of articles. The tech stack employed enables efficient and systematic scraping, contributing to the generation of a comprehensive dataset.
2) Dataset Features The dataset encompasses key features crucial for our analysis:
   - Abstract: A brief summary of the article's content.
   - Web URL: The article's web address.
   - Headline: The title or heading of the article.
   - Keywords: Tags associated with the article, providing insights into its content.
   - Pub Date: The publication date of the article.
   - News Desk: The department responsible for the article.
   - Section Name: The section or category of the article.
   - Byline: The author or authors of the article.
   - Word Count: The number of words in the article.
3) Dynamic Dataset Utilization Our dataset, stands out due to its dynamic nature, ensuring relevance through daily updates facilitated by the Pynytimes API. This self-updating feature is seamlessly integrated into our website, allowing users to access the latest news articles from The New York Times. The auto-scheduling feature in Kaggle notebooks ensures that the dataset remains current, contributing to an ever-evolving and real-time analysis resource.

V. Architecture

1) **News Nexus** The current version of News Nexus uses ReactJS for the frontend for a smooth experience. The backend server is written in Flask to easily integrate ML Models for semantic search and sentiment analysis. We're using Chroma Vector DB for semantic search and MangoDB for authentication and other purposes. We are using Open AI's Embedding Function for creating embeddings of the query in backend.
2) **Semantic Search Engine Architecture** Our semantic search engine is designed to revolutionize news article retrieval by incorporating advanced techniques such as word embeddings and similarity search. The architecture leverages pre-trained models like sentence transformers or Word2Vec for encoding article content into meaningful embeddings. The similarity search functionality is implemented through Chroma, enabling fast and accurate retrieval of related articles based on content similarities.
3) **Content-Based Recommendation System Architecture** The architecture of our content-based recommendation system is centered around user interactions and preferences. When a user searches for a term, the system utilizes these queries to find relevant articles through similarity with other articles and user queries. Fast algorithms are employed to ensure prompt recommendations. Additionally, topics and other features are considered to enhance the recommendation process.
4) **Data Self Updation Mechanism** Ensuring the dataset's freshness is a crucial aspect of our project. The architecture incorporates an efficient self-updation mechanism using the New York Times API and Kaggle notebooks' auto-scheduling feature. This dynamic approach ensures that the dataset is regularly updated with the latest

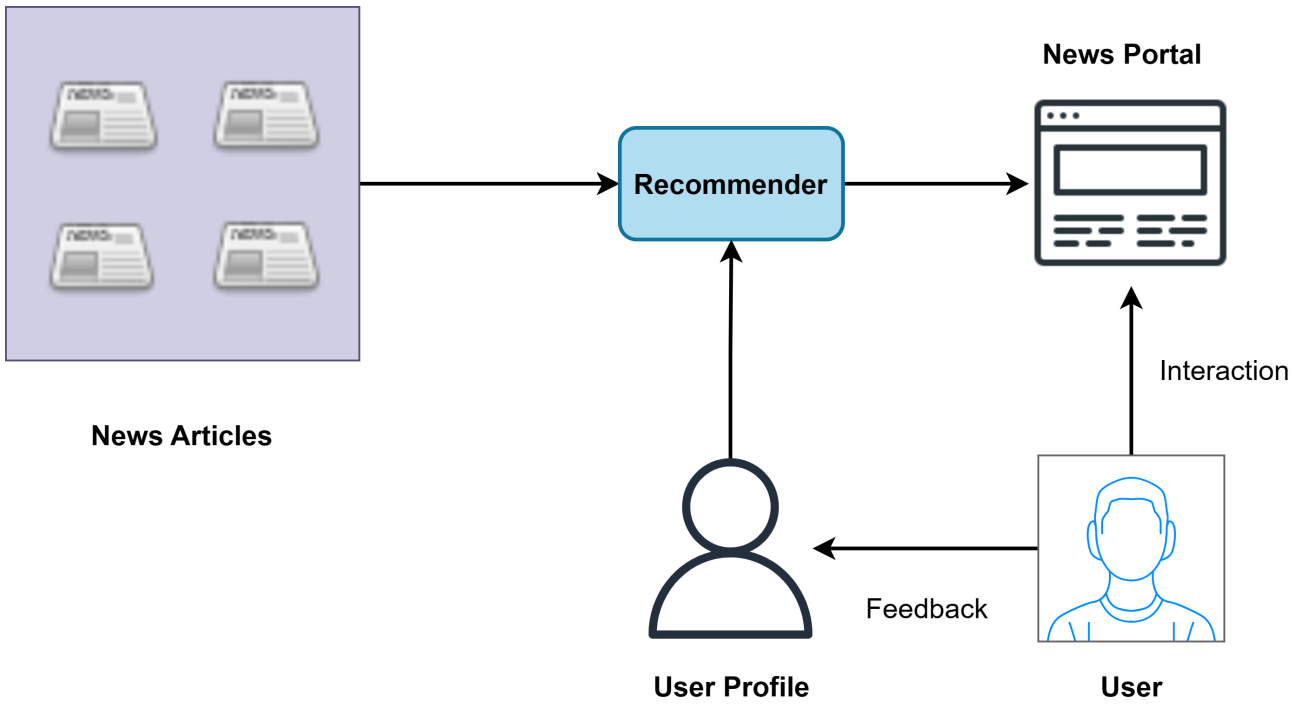

Fig. 2. Recommendation System Pipeline

articles from The New York Times, making it an evolving and reliable resource for analysis. The dataset's dynamic nature, with approximately 2 million news articles consuming about 4 GB in disk size, showcases the impressive scale of our data collection efforts.

VI. Impact

The proposed project, "Building a Comprehensive News Analysis, Recommendation, and Search Platform," holds the potential to make a substantial impact on various stakeholders and domains. Below, we outline the expected impact in several key areas:

1) **Media and Journalism**: Our platform will empower media professionals and journalists by providing them with advanced tools for trend analysis, sentiment tracking, and content discovery. News organizations can leverage the platform to gain insights into reader sentiments and preferences, leading to more targeted and engaging content creation. As a result, the quality of journalism is expected to improve, enhancing public access to reliable and relevant news.
2) **User Engagement and Information Retrieval**: Users seeking news and information will benefit from a semantic search engine that offers context-aware search results. They will experience enhanced content recommendations tailored to their interests and preferences. The platform's user-centric design aims to improve user engagement and satisfaction, making it a valuable resource for those looking to stay informed.
3) **Trend Forecasting and Decision Making**: Trend analysis and prediction capabilities will aid businesses, policymakers, and researchers in making data-driven decisions. By identifying emerging topics and themes, our platform can assist in anticipating shifts in public interest and sentiment. This information can inform strategic decisions, policy formulation, and market trends analysis.
4) **Academic and Research Communities**: The project's research components, including sentiment analysis, trend analysis, authorship analysis, and social network analysis, can contribute to academic research in various fields. Researchers can use the platform as a resource for studying media influence, sentiment dynamics, and social network structures within the news ecosystem.
5) **Technology and Innovation** :The development of this platform will involve cutting-edge technologies and methodologies in natural language processing (NLP), machine learning, and data analytics. It has the potential to advance the state of the art in semantic search engines, content-based recommendation systems, and news analysis techniques, contributing to the broader field of information retrieval.

In summary, our project aims to create a multifaceted platform that not only benefits news consumers but also positively impacts media, journalism, decision-making processes, academic research, and technological innovation. By providing advanced tools for news analysis, recommendation, and search, we aspire to play a pivotal role in shaping the future of news consumption and information retrieval.

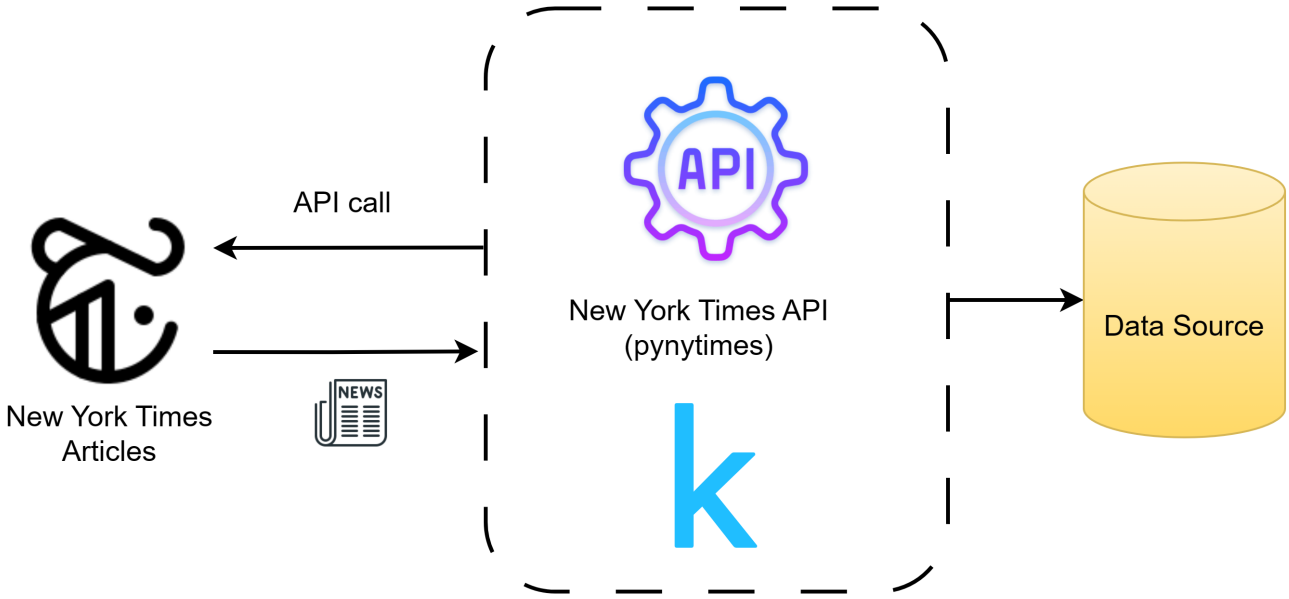

Fig. 3. Data Collection and Daily Updation

VII. Tech Stack

We would be using **Python** along with Hugging Face Transformers to build the model and recommendation system. For the database, we are planning to utilise **ChromaDB** and **NodeJS** server to interact with the database and the recommendation system. And we would be using **ReactJS** as the frontend to allow users to interact with the application.

VIII. Evaluation

A. *Semantic Search Engine Evaluation*

In assessing the effectiveness of our Semantic Search Engine, we employ a set of robust evaluation metrics tailored to measure its precision, recall, and overall retrieval performance.

1) Precision at K (P@K): Metric Definition: Precision at K evaluates the accuracy of our search engine by measuring the proportion of relevant articles among the top K search results.
   Score: Achieving a P@5 score of 0.85 signifies that 85% of the top 5 search results are pertinent to the user's query.
2) Recall at K (R@K): Metric Definition: Recall at K assesses the engine's ability to retrieve all relevant documents within the top K search results.
   Score: A R@10 score of 0.90 indicates that 90% of all relevant articles were successfully retrieved within the top 10 search results.
3) Mean Average Precision (MAP): Metric Definition: MAP provides a comprehensive evaluation by considering precision at various recall levels, offering insights into performance across different thresholds.
   Score: Achieving a MAP score of 0.75 implies robust performance in maintaining high precision across varying recall levels.

B. *Recommendation System Evaluation*

For our Recommendation System, we adopt a set of metrics focusing on precision, recall, and the quality of the recommended articles.

1) Precision at K (P@K): A P@5 score of 0.72 indicates that 72% of the top 5 recommended articles align with the user's preferences.
2) Normalized Discounted Cumulative Gain (NDCG): Metric Definition: NDCG considers both the relevance and ranking of recommended articles, providing a holistic assessment of the recommendation quality.
   Score: Achieving an NDCG score of 0.85 indicates superior recommendations that are not only relevant but also well-ordered.

C. *News Analysis*

a) Sentiment Analysis Accuracy: Using pre trained DistilBERT model from Hugging Face we got a score of 88%.
b) Topic Modeling Coherence: 0.75 (on a scale of 0 to 1)
c) Keyword Extraction Precision: 82%

IX. Deliverables

The project, "News Nexus: A Comprehensive News Portal with Search, Recommendations, and Analysis" will result in a range of deliverables, each contributing to the successful implementation of the platform and its associated components. The Github Repo Link for the News Portal.
The following is a list of key deliverables for the project:

1) **Web-Based Platform**: A fully functional web-based platform accessible to users for news article search, recommendation, and analysis.

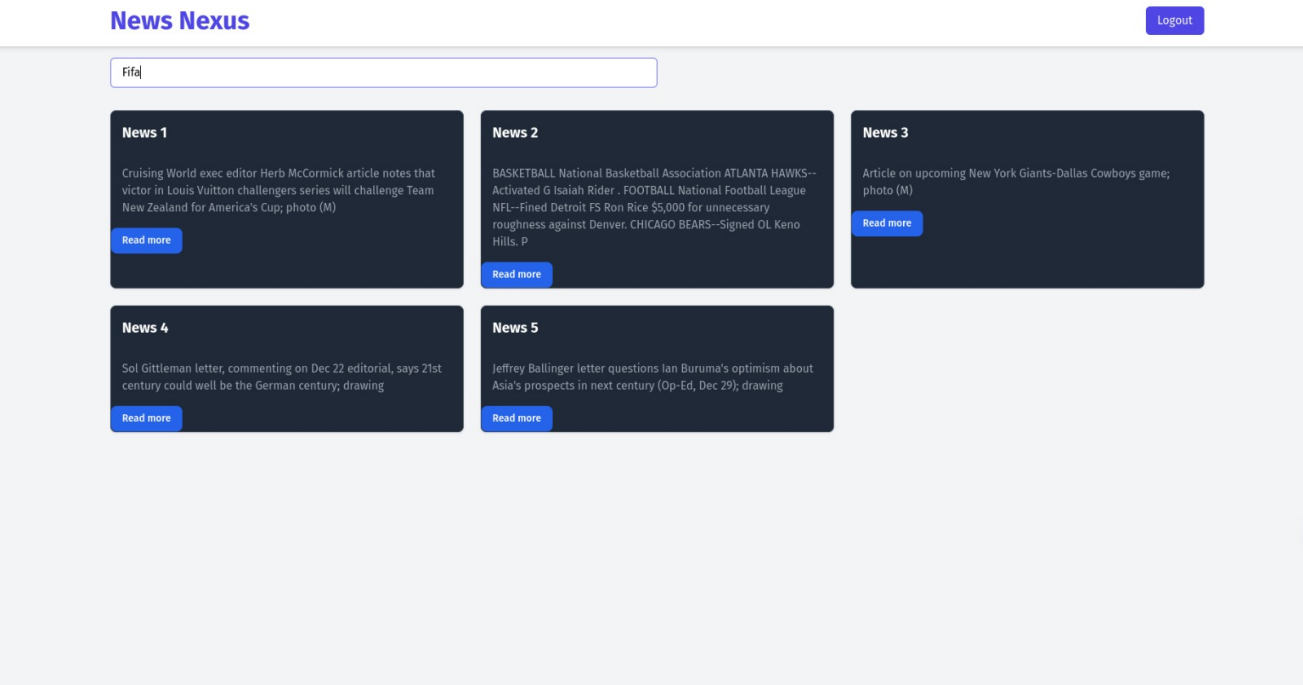

Fig. 4. Search Results

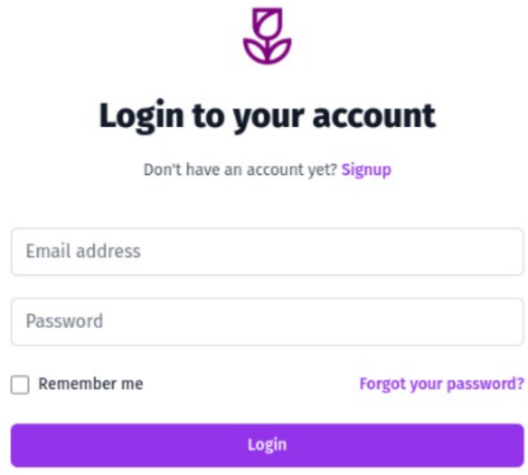

Fig. 5. User Login Page

2) **Semantic Search Engine**: The development of a semantic search engine that understands context and meaning in search queries, delivering context-aware search results.
3) **Content-Based Recommendation System**: Implementation of a content-based recommendation system that suggests relevant news articles based on user behavior and content features.
4) **Sentiment Analysis Module**: Integration of a sentiment analysis module capable of classifying news articles into sentiment categories (positive, negative, neutral).
5) **Trend Analysis Tools**: Tools for trend analysis and prediction, allowing users to identify emerging topics and patterns within news content.
6) **Section Popularity Analysis**: Analysis tools to determine the popularity of different news sections or categories among readers.
7) **User Interface (UI)**: A user-friendly and responsive UI built with HTML, CSS, and React, providing an intuitive user experience.
8) **Server-Side Logic**: Node.js-based server-side logic for handling user requests, data processing, and interaction with the database.
9) **Video Presentation**: A video demonstrating the working of a news website and an explanation of the methodology used to build the portal.
10) **Project Report**: Final Report of Project for the course DA331 Big Data Analytics and Tools.

These deliverables collectively contribute to the successful development, deployment, and utilization of the comprehensive news analysis, recommendation, and search platform. They represent the tangible outcomes that will enable us to achieve the project's goals and objectives.

REFERENCES

[1] S. Karmaniolos and G. Skinner, "A Literature Review on Sentiment Analysis and its Foundational Technologies," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 91-95, doi: 10.1109/CCOMS.2019.8821771.
[2] Godbole, N., Srinivasaiah, M. and Skiena, S., 2007. Large-Scale Sentiment Analysis for News and Blogs. Icwsm, 7(21), pp.219-222.

[3] Fukuhara, Tomohiro, Hiroshi Nakagawa, and Toyoaki Nishida. "Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events." In ICWSM. 2007.
[4] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.Chi, Dong-Cheol, and Sang-ho Kim. "Home training trend analysis using newspaper big data and keyword analysis." Journal of the Korea Convergence Society 12, no. 6 (2021): 233-239.
[5] Kim, Min-Jeong, and Chul Joo Kim. "Trend Analysis of News Articles Regarding Sungnyemun Gate using Text Mining." The Journal of the Korea Contents Association 17, no. 3 (2017): 474-485..
[6] Lupiani-Ruiz, Eduardo, Ignacio García-Manotas, Rafael Valencia-García, Francisco García-Sánchez, Dagoberto Castellanos-Nieves, Jesualdo Tomás Fernández-Breis, and Juan Bosco Camón-Herrero. "Financial news semantic search engine." Expert systems with applications 38, no. 12 (2011): 15565-15572.
[7] Deshmukh, Anup Anand, and Udhav Sethi. "IR-BERT: leveraging BERT for semantic search in background linking for news articles." arXiv preprint arXiv:2007.12603 (2020).