Building a Comprehensive News Analysis, Recommendation, and Search Platform

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 comprehensive study on harnessing the power of the New York Times dataset for trend analysis, sentiment analysis, and the development of a semantic search engine. The New York Times dataset, with its vast collection of articles spanning diverse topics, offers a rich source of information for uncovering trends and sentiments within news content and enabling efficient semantic searches.**

In this research, we detail the methodologies employed for data collection, preprocessing, and analysis. For trend analysis, time-series data mining techniques were utilized to identify emerging patterns and topics of interest over time. Sentiment analysis, powered by natural language processing (NLP) models, provided insights into the emotional tone of news articles.

Moreover, we describe the architecture and implementation of a semantic search engine that enhances user information retrieval experiences by understanding context and meaning. Leveraging a knowledge base built from the dataset, the search engine offers precise and context-aware search results.

Results and findings from these analyses are presented, shedding light on significant trends, sentiment fluctuations, and the effectiveness of the semantic search engine. This research contributes to the domain of data-driven journalism and information retrieval systems, offering valuable insights for various applications, including media monitoring, content recommendation, and trend forecasting.

***Index Terms*—component, formatting, style, styling, insert**

I. Problem Statement

In an era of vast digital media content, the New York Times dataset stands as a treasure trove of information encompassing diverse topics, opinions, and sentiments. The project aims to harness this dataset's potential by addressing three key challenges, each contributing to a holistic news consumption experience.

1) Search Engine for News Articles
    a) **Word Embeddings and Similarity Search** : Developing an advanced search engine for news articles leveraging word embeddings (e.g., Word2Vec) to measure the similarity between articles. Implementing a similarity search engine that recommends related articles based on content similarities.
2) Content-Based Recommendation System
    a) **Content Filtering for User Preferences**: Building a content-based recommendation system that suggests news articles to users based on their preferences and interactions along with content filtering mechanisms that allow users to customize their news feeds based on specific keywords, topics, or sections they are interested in.
3) Comprehensive News Analysis
    a) **Sentiment Analysis**: Performing sentiment analysis on news article content or headlines to get public sentiment on various topics. Tracking sentiment trends over time and identifying articles with polarizing opinions.
    b) **Topic Modeling**: Employing topic modeling techniques (e.g., Latent Dirichlet Allocation) to identify key topics and themes within news articles. Creating a topic hierarchy for improved content organization.
    c) **Keyword Extraction and Visualization**: Extracting and visualizing keywords and phrases from news articles to create word clouds, graphs, and charts that showcase the most frequently mentioned terms.
    d) **Section Popularity Analysis**: Analyzing which sections of news articles are most popular among readers. Identifying trends in section popularity over time to improve recommendations.
    e) **Authorship Analysis**: Studying the writing styles and preferences of different authors by analyzing the news articles they've written. Identifying prolific authors and their impact on readers.
    f) **Content Classification**: Developing a text classification model to categorize news articles into genres or categories (e.g., politics, sports, technology) to improve content organization and navigation.
    g) **Interactive Data Visualization Dashboard**: Creating an interactive dashboard that allows users to explore trends, topics, and news articles within the dataset through visually compelling graphs and charts.
    h) **Social Network Analysis**: Analyzing connections and influence networks between authors, articles, and readers within The New York Times community. Identifying patterns of collaboration and engagement.

The project aims to revolutionize the way users access, discover, and understand news articles by addressing these challenges comprehensively.

II. Dataset

We're utilizing the NYT Articles Dataset . This dataset contains a comprehensive collection of articles from The New York Times, spanning from January 1, 2000, to the present day. It is important to note that this dataset is updated daily, ensuring that the latest articles from The New York Times are included, providing an up-to-date and evolving resource for analysis.
What sets this dataset apart is its dynamic nature, as it is updated daily through the seamless integration of the New York Times API and Kaggle notebooks' auto-scheduling feature.
The dataset includes key features:

- 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.

III. 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:

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.

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.

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.

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.

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.

IV. 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 **MongoDB** 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.

V. Deliverables

The project, "Building a Comprehensive News Analysis, Recommendation, and Search Platform," will result in a range of deliverables, each contributing to the successful implementation of the platform and its associated components. The following is a list of key deliverables: **Web-Based Platform**: A fully functional web-based platform accessible to users for news article search, recommendation, and analysis.

Semantic Search Engine: The development of a semantic search engine that understands context and meaning in search queries, delivering context-aware search results.

Content-Based Recommendation System: Implementation of a content-based recommendation system that suggests relevant news articles based on user behavior and content features.

Sentiment Analysis Module: Integration of a sentiment analysis module capable of classifying news articles into sentiment categories (positive, negative, neutral).

Trend Analysis Tools: Tools for trend analysis and prediction, allowing users to identify emerging topics and patterns within news content.

Section Popularity Analysis: Analysis tools to determine the popularity of different news sections or categories among readers.

Authorship and Social Network Analysis: Tools and visualizations for authorship analysis and social network analysis within the news ecosystem.

Database Implementation: Deployment of a MongoDB database to store and manage news article metadata and user interactions.

User Interface (UI): A user-friendly and responsive UI built with HTML, CSS, and React, providing an intuitive user experience.

Server-Side Logic: Node.js-based server-side logic for handling user requests, data processing, and interaction with the database.

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.

VI. Literature Review

Paper 1: "A Literature Review on Sentiment Analysis and its Foundational Technologies"

Sentiment Analysis, a prominent field at the intersection of Natural Language Processing (NLP) and Machine Learning (ML), has emerged as a pivotal tool for computational opinion assessment. This computational treatment of expressed sentiments enables the quantification and interpretation of opinions within textual data, making it a valuable asset for decision-making in various domains.

Paper 2: "Trend Analysis"

In parallel, the realm of Trend Analysis addresses the critical task of monitoring and managing the evolution of reliability and efficiency in development activities. This domain, exemplified by works like "Trend Analysis" (Kanoun and Laprie, 1996), underscores the importance of reliability trend analyses in aiding project managers.

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).