News Recommendation System

Shubhi Agarwal
Data Science and Artificial Intelligence
Indian Institute of Technology, Guwahati
Guwahati, India
a.shubhi@iitg.ac.in

***Abstract*—In today's digital age, personalized news consumption is paramount. This project introduces an innovative News Recommendation System (NRS) that transcends conventional platforms, offering a comprehensive and enriching news consumption experience. The NRS comprises core functionalities, including User Registration and Profiles, Content Aggregation, Topic and Category Selection, Personalized News Feed, Article Recommendation, and Search Functionality.**

Going beyond traditional NRS features, this project pioneers the integration of groundbreaking elements such as Emotion-Based Feedback, Fake News Detection, Bookmarking, Content Moderation, Sentiment Analysis, Historical Context in Search and Privacy and Security of accounts. Emotion-Based Feedback enables users to express sentiments towards articles, enhancing personalization. Fake News Detection safeguards against misinformation, fostering media literacy. Bookmarking and Content Moderation empower users while ensuring a secure environment. Sentiment Analysis optimizes search results based on user moods, and Historical Context in Search enriches the user experience.

This NRS leverages various machine learning algorithms, web technologies for frontend development, advanced page ranking techniques, and big data frameworks, including Hadoop and Spark for seamless implementation. In a digital landscape teeming with information and varying emotional states, this NRS not only elevates user engagement and satisfaction but also promotes responsible news consumption. It serves as a valuable addition to evolving news recommendation platforms, addressing the multifaceted needs of today's news consumers.

***Index Terms*—news recommendation, Big Data, machine learning, TF IDF, cosine similarity, Flask, ensemble models.**

I. Introduction & Motivation

News recommendation systems are vital in the digital age as they help users navigate the overwhelming volume of information. They provide personalized news feeds tailored to individual interests, saving time and effort in finding relevant content. These systems also expose users to diverse perspectives, promoting a more well-rounded understanding of current events. Additionally, they play a crucial role in combating misinformation by guiding users towards credible sources. News recommendation systems not only enhance user convenience but also foster media literacy and responsible news consumption, making them indispensable tools in today's information-rich world.

II. Problem Statement

In the digital age, news consumption has undergone a seismic shift, transitioning from traditional sources to online platforms. News recommendation systems (NRS) have emerged as the linchpin in facilitating personalized news discovery for users. Existing news recommendation systems play a pivotal role in today's digital landscape by offering personalized news content to users. They leverage algorithms that analyze user behavior, providing convenience and a wide range of news sources. However, these systems face challenges. They can inadvertently create filter bubbles and echo chambers, limiting exposure to diverse perspectives. Ensuring the quality and credibility of news remains a concern, with the occasional promotion of sensationalized or inaccurate content. Privacy issues arise from data collection, raising concerns about user data protection. Additionally, contextual understanding and historical context are areas that require improvement

In this project I have created a news recommendation, which recommends news based on topics a user select. The searching is based on hybrid recommendation techniques. Here, basically I have worked on 2 parts: fake news detection and enhanced personalized news recommendation.

III. Architecture Detail

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Term Frequency (TF): This component measures how frequently a term appears in a document. It is calculated as the ratio of the number of occurrences of a term to the total number of words in the document. The idea is to emphasize terms that appear frequently within a document.
  - Inverse Document Frequency (IDF): This component evaluates the importance of a term across a collection of documents. It is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term. Terms that are common across many documents receive a lower IDF score.
  - TF-IDF Score: The TF-IDF score of a term in a document is the product of its TF and IDF scores. This score reflects how important a term is to a specific document in the context of a larger document collection.

- Purpose: TF-IDF is used to highlight terms that are both frequent within a document and distinctive across the entire document collection. It helps in transforming a collection of documents into a numerical vector space, facilitating various natural language processing tasks such as text classification and clustering.

- **Cosine Similarity** Cosine Similarity is a metric used to determine the cosine of the angle between two non-zero vectors in an inner product space. In the context of NLP, these vectors represent the TF-IDF representations of documents. - Calculation: For two vectors A and B, the cosine similarity (cos) is computed as the dot product of A and B divided by the product of their magnitudes. The result ranges from -1 (completely dissimilar) to 1 (completely similar).
  - Purpose: In information retrieval and document similarity analysis, cosine similarity is employed to measure how similar two documents are in terms of their content. It is widely used in tasks like document clustering, recommendation systems, and search engines.
  - Advantages: Cosine similarity is robust to document length and is particularly useful when dealing with high-dimensional data, as it focuses on the orientation rather than the magnitude of vectors.

- **Text Processing:** Before analysis, the news article texts undergo preprocessing. This includes converting text to lowercase, removing square brackets, eliminating URLs and HTML tags, and replacing non-word characters and punctuation. These steps ensure a standardized and clean corpus for subsequent analysis.

- **Machine Learning Model:** The application incorporates a machine learning model,for fake news detection. Logistic Regression is a classification algorithm suitable for binary outcomes. It is trained on labeled data to predict whether a given news article is fake or not.

- **Recommendation System:** The recommendation system suggests articles similar to a given one. It leverages both cosine similarity and category matching to provide relevant recommendations. This system enhances user engagement and information discovery.

- **User Interface::** The Flask web application offers a user-friendly interface. It includes search functionality, a fake news detection feature, and a display for recommended articles. The UI is designed to be intuitive, allowing users to interact seamlessly with the recommendation system.

IV. Methodology and Experiment

Workflow: Dataset creation and pre-processing Model creation Feature extraction using TF-IDF and cosine similarities Extracting recommendations and results Website display The project appears to be a News Recommendation System with elements of Fake News Detection. Let's break down the workflow:

1. Data Loading and Preprocessing: - The datasets are loaded and cleaned and preprocessed by handling missing values in the "Title" and "Abstract" columns.

2. Text Vectorization: - The TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is used to convert the text data into numerical vectors. - The combined text from the "Title" and "Abstract" columns is used for vectorization.

3. Cosine Similarity Calculation: - Cosine similarity is calculated between the articles based on the TF-IDF vectors. This similarity measure helps identify articles with similar content.

4. Fake News Detection Model Loading: - A machine learning model for fake news detection is loaded from the 'fakenewsmodel.pkl' file. - The TF-IDF vectorizer used for training the model is also loaded from 'tfidfvectorizer.pkl'.

5. Flask Web Application: - A Flask web application is created with three main routes: - **Home (''/'):** Displays recommended articles. - **Fake News Detection (''/detect-fakenews''):** Accepts user input, processes it, and returns the fake news prediction. - **Search (''/search''):** Accepts search queries, finds relevant articles, and displays recommendations.

6. User Interaction: - Users can interact with the system through a web interface ('index.html'). - The system provides recommended articles on the home page. - Users can input news text for fake news detection, and the system returns a prediction.

7. Fake News Detection: - The user-input news text is preprocessed using the 'wordproc' function. - The loaded machine learning model ('loadedLR') and TF-IDF vectorizer are used to predict whether the news is fake or not.

8. Search Functionality: - Users can search for articles based on a query that matches the title, category, subcategory, or abstract. - The system returns a list of recommended articles based on the search results.

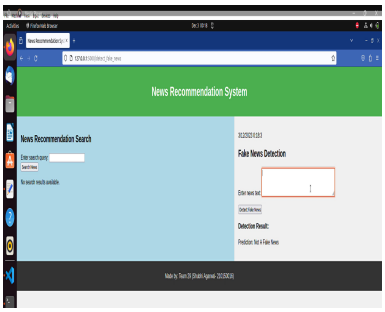
9. Web Application Execution: - The Flask application is run locally, and users can access the functionality through a web browser.

10. Output: - The web interface displays the recommended articles, search results, and fake news detection results.

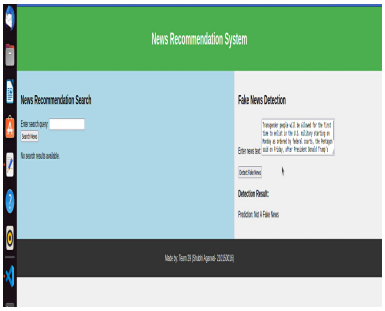
This workflow demonstrates a comprehensive News Recommendation System with an added feature of Fake News Detection, providing users with personalized recommendations

and tools to identify potentially misleading news articles.

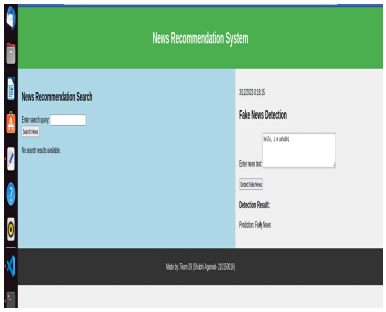
DEMO:



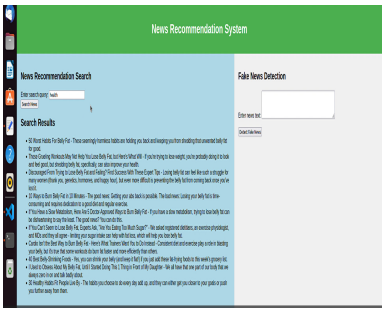
This demonstrates the start when none search operation is performed



This demonstrates identification of a fake news.



This demonstrates identification of an authentic news.



This demonstrates news recommendation based on search by any keyword.

V. Result and Conclusion

In conclusion, the implemented News Recommendation System coupled with Fake News Detection offers a multi-faceted solution for users seeking reliable and personalized news content. Through a user-friendly web interface, individuals can effortlessly navigate recommended articles, engage in topic-specific searches, and employ a robust tool for discerning the authenticity of news stories. The incorporation of TF-IDF vectorization and cosine similarity enables the system to provide tailored recommendations based on content similarity, enhancing the user experience. Moreover, the integration of a machine learning model for fake news detection further fortifies the system's utility by empowering users to make informed judgments about the veracity of news articles. The combination of these features fosters a dynamic and interactive platform that addresses the contemporary challenges of information credibility and content personalization in the realm of digital news consumption. Overall, the project demonstrates the efficacy of leveraging advanced technologies to create a comprehensive and intelligent news recommendation system.

Though there were more features like historical context research, emotion based feedbackwhich I wanted to implement as discussed earlier in the research proposed but could not be implemented because of lack of time.

References

[1] Ahmed H, Traore I, Saad S., "Detecting opinion spams and fake news using text classification," Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[2] X. Han, W. Shang and S. Feng, "The design and implementation of personalized news recommendation system", 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), Las Vegas, NV, USA, 2015, pp. 551-554, doi: 10.1109/ICIS.2015.7166653.

[3] https://www.kaggle.com/datasets/achintyatripathi/news-dataset-18920.
[4] https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
[5] MIND Dataset : https://msnews.github.io/