**News Scraper Application for Sentiment Analysis and Summarization of News Articles**

**- Rahul Tammalla , Clarkson University.**

**1. Executive Summary**

This project is a web scraping application that extracts the latest news articles from multiple sources, performs sentiment analysis, and summarizes the descriptions of the articles. It uses pre-trained machine learning models to analyze the text and provide concise summaries and sentiment labels (positive, negative, or neutral). The application is designed to update automatically every 24 hours, ensuring users get fresh and relevant content daily.

**2. Problem Statement**

With the overwhelming amount of daily news articles available, it becomes difficult for users to stay updated or quickly absorb important details. This project aims to:

* Automatically fetch the latest articles.
* Provide summaries of articles to make information consumption faster.
* Analyze the sentiment of the articles to understand public opinion.

**3. Objectives**

* **Scraping news articles**: Automatically fetch the latest top news articles from trusted US sources.
* **Sentiment analysis**: Apply sentiment analysis to classify the article descriptions into categories such as positive, negative, or neutral.
* **Summarization**: Provide concise summaries of news articles using a text summarization model.
* **Automatic daily updates**: The application refreshes every 24 hours at 12:01 AM to ensure up-to-date news.

**4. Technologies Used**

**5.1 Programming Languages**

* **Python**: The main programming language used for building applications.

**5.2 Libraries and Frameworks**

* **Streamlit**: A framework to easily build interactive web apps. Used to display the news articles, sentiment, and summaries on the web.
* **Transformers**: A library from Hugging Face that provides pre-trained models for natural language processing tasks like sentiment analysis and summarization.
* **Pandas**: Used to store and manipulate news articles in a structured format (DataFrame).
* **Requests**: A simple library used to make HTTP requests to the NewsAPI for fetching news articles.
* **NewsAPI**: A service used to fetch the latest news articles from multiple trusted sources.

**5.3 Machine Learning Models**

* **DistilBERT (for Sentiment Analysis)**: A smaller, faster, and more efficient version of the BERT model that is fine-tuned for sentiment classification tasks. It is used to classify the sentiment of news article descriptions into positive, negative, or neutral categories.
* **T5-small (for Summarization)**: A smaller variant of the T5 model, fine-tuned for text summarization tasks. It generates concise summaries of the article descriptions.

**5.4 Deployment Platform**

* **Streamlit**: Used to deploy the app as a web-based application, allowing users to interact with it easily.

**5.5 Other Tools**

* **GitHub**: Used for version control and storing the project code.

**6. Detailed Workflow**

**6.1 Data Collection**

* The app fetches the latest news articles using **NewsAPI**, which aggregates articles from various trusted news sources in the United States.
* The articles are filtered based on the **publish date** to ensure that only news articles from the previous day are fetched.

**6.2 Sentiment Analysis**

* **DistilBERT**, a pre-trained model, is used for sentiment analysis. It processes the description of each article and returns a sentiment label: positive, neutral, or negative.

**6.3 Text Summarization**

* **T5-small**, a text-to-text transformer model, is used for summarizing the article descriptions. The model generates a concise version of the article’s description while maintaining its key points.

**6.4 Web Scraping & Display**

* Once the news data is fetched, the app displays it using **Streamlit**. The user can click a button to scrape news, and the app will show the title, description, sentiment label, and summary of the articles in a structured table format.

**6.5 Timer Functionality**

* A countdown timer is implemented that tracks the time left until the next update at 12:01 AM. This timer ensures that the app updates every 24 hours, ensuring fresh content.

**7. Key Features**

* **Scraping News**: Automatically fetches the latest news articles based on the date.
* **Sentiment Analysis**: Analyzes the sentiment of the news articles to classify them into positive, neutral, or negative categories.
* **Summarization**: Provides a summary of each article to make it easier to digest the content.
* **Timer for Daily Update**: The app shows the time remaining until the next update, which happens every 24 hours at 12:01 AM.

**8. Results**

* The app successfully scrapes news articles, analyzes the sentiment, and generates concise summaries for each article.
* The interface is user-friendly, with clear options to scrape the news for the previous day and display the results in a readable format.

**9. Challenges Faced**

* **Timezone Handling**: Initially, there was an issue with displaying the correct date (local time) due to Streamlit’s UTC-based time handling. The issue was resolved by applying **timezone conversion** to ensure the date and news are fetched in the correct time zone (EST).
* **API Rate Limits**: NewsAPI has rate limits, so the app handles any errors gracefully and displays appropriate messages to the user if the limit is exceeded.

**10. Future Enhancements**

1. **Expand to Other Countries**: Currently, the app fetches news from the US, but in the future, we could expand this to other countries and regions.
2. **Real-Time Scraping**: Implement real-time scraping of articles instead of relying on daily updates.
3. **News Categorization**: Categorize news articles into topics like technology, politics, sports, etc., to allow users to filter based on their interests.
4. **Improved Summarization**: Fine-tune the models specifically for news articles to improve summarization quality.

**11. Conclusion**

This News Scraper application is a powerful tool that helps users stay updated with the latest news in a quick and easy way. By utilizing **Machine Learning models** for sentiment analysis and summarization, it provides value-added features to standard news scraping tools.