Yahoo Finance News Analysis Project

Overview

This project is a comprehensive pipeline for scraping, analyzing, and visualizing financial news from Yahoo Finance. It involves collecting news articles, storing them in a MongoDB database, performing Exploratory Data Analysis (EDA), extracting topics using BERTopic, and conducting sentiment analysis using FinBERT.

The project is divided into 5 main steps:

1. **Setup MongoDB**: Database configuration.
2. **Scrape Data**: Automated scraping of Yahoo Finance news.
3. **Data Cleaning & EDA**: Data preprocessing and exploratory analysis.
4. **Topic Modeling**: Discovering latent topics in news articles.
5. **Sentiment Analysis**: Analyzing the sentiment of financial news.

Prerequisites

- Python 3.10+
- MongoDB (Atlas or Local)
- Chrome Browser (for Selenium)

Setup

1. **Clone the repository:**

```
git clone
cd YahooFinance
```

2. **Install Dependencies:** You can install the required packages using `pip`. There are two requirement files, one for the scraper and one for the full analysis.

```
pip install -r requirement.txt
```

3. **Environment Variables:** Create a `.env` file in the root directory and add your MongoDB credentials:

```
db_username=your_username
db_password=your_password
db_host=your_cluster_address
MONGODB_CONNECTION_STRING=your_full_connection_string
```

Step 1: Setup MongoDB

This project uses MongoDB to store scraped news articles.

1. Create a MongoDB cluster (e.g., using MongoDB Atlas).
2. Create a database named `finance_news_db`.
3. Create a collection named `numerous_articles`.
4. Ensure your IP is whitelisted in the MongoDB network access settings.
5. Update the `.env` file with your connection details.

Step 2: Scrape Data from Yahoo Finance

The scraping logic is implemented in `daily_scraper.py` (and `scraper.ipynb`). It uses **Selenium** to handle dynamic content loading (infinite scroll) and **BeautifulSoup** for parsing HTML.

What the code does:

- Navigates to `https://finance.yahoo.com/topic/latest-news/`.
- Scrolls down automatically to load more articles.
- Extracts article details: Title, Publisher, Tickers, Link, Author, Time Published, and Content.
- Inserts new articles into the MongoDB collection `numerous_articles`, avoiding duplicates.

How to run:

```
python daily_scraper.py
```

Note: The script runs in headless mode by default.

Automation

A GitHub Actions workflow (`.github/workflows/daily_scrape.yml`) is set up to run the scraper automatically every 4 hours. It checks out the code, installs dependencies, and runs `daily_scraper.py` using secrets configured in the GitHub repository.

Step 3: Data Cleaning and EDA

The `Data_Cleaning_&_EDA.ipynb` notebook handles data preprocessing and exploration.

What the code does:

- **Data Cleaning**: Connects to MongoDB, loads data into a Pandas DataFrame, strips whitespace, and converts `time_published` to datetime objects.
- **EDA**:
  - Analyzes missing values.
  - Visualizes the distribution of **Publishers** and **Authors**.
  - Analyzes the time distribution of articles.
  - Examines the length of titles and content.

Result: Insights into the dataset structure, such as the most active publishers (e.g., Barchart, Business Insider) and peak publication times.

Example of cleaned data

DataFrame after stripping spaces from string attributes:

	_id	title	publisher	tickers	link	authors	time_published	content
0	692af2e04a7fefe22f224c5f	Rising star catch-up: Chad Tredway is back at ...	Business Insider	[{'symbol': 'JPM', 'change': None}]	https://finance.yahoo.com/news/rising-star-cat...	dgeiger@businessinsider.com (Daniel Geiger)	2025-11-29 19:20:00-07:00	Chad Tredway was a rising star at JPMorgan Cha...
1	692af2e14a7fefe22f224c60	Want Steady Income in Retirement? These Overlo...	Investopedia	[]	https://finance.yahoo.com/news/want-steady-inc...	Jonathan Ponciano	2025-11-29 19:08:00-07:00	Klaus Vedfelt / Getty Images\n93% of workers w...
2	692af2e24a7fefe22f224c61	The hottest new AI company is...Google?	CNN Business	[]	https://finance.yahoo.com/news/hottest-ai-comp...	Analysis by Lisa Eadicicco, CNN	2025-11-29 19:01:00-07:00	Google just threw another twist in the fast-ch...
3	692af2e34a7fefe22f224c62	Medicare Advantage woos seniors with plan perk...	Yahoo Finance	[{'symbol': 'HUM', 'change': None}]	https://finance.yahoo.com/news/medicare-advant...	Kerry Hannon · Senior Columnist	2025-11-29 18:57:00-07:00	Seniors have embraced Medicare Advantage plans...
4	692af2e44a7fefe22f224c63	Why Waiting for a Housing Crash Could Be Costi...	Investopedia	[]	https://finance.yahoo.com/news/why-waiting-hou...	Isabel O'Brien	2025-11-29 18:39:00-07:00	Fact checked by Suzanne Kvilhaug\nDrs Producoe...

EDA Visualisation

Step 4: Topic Modeling

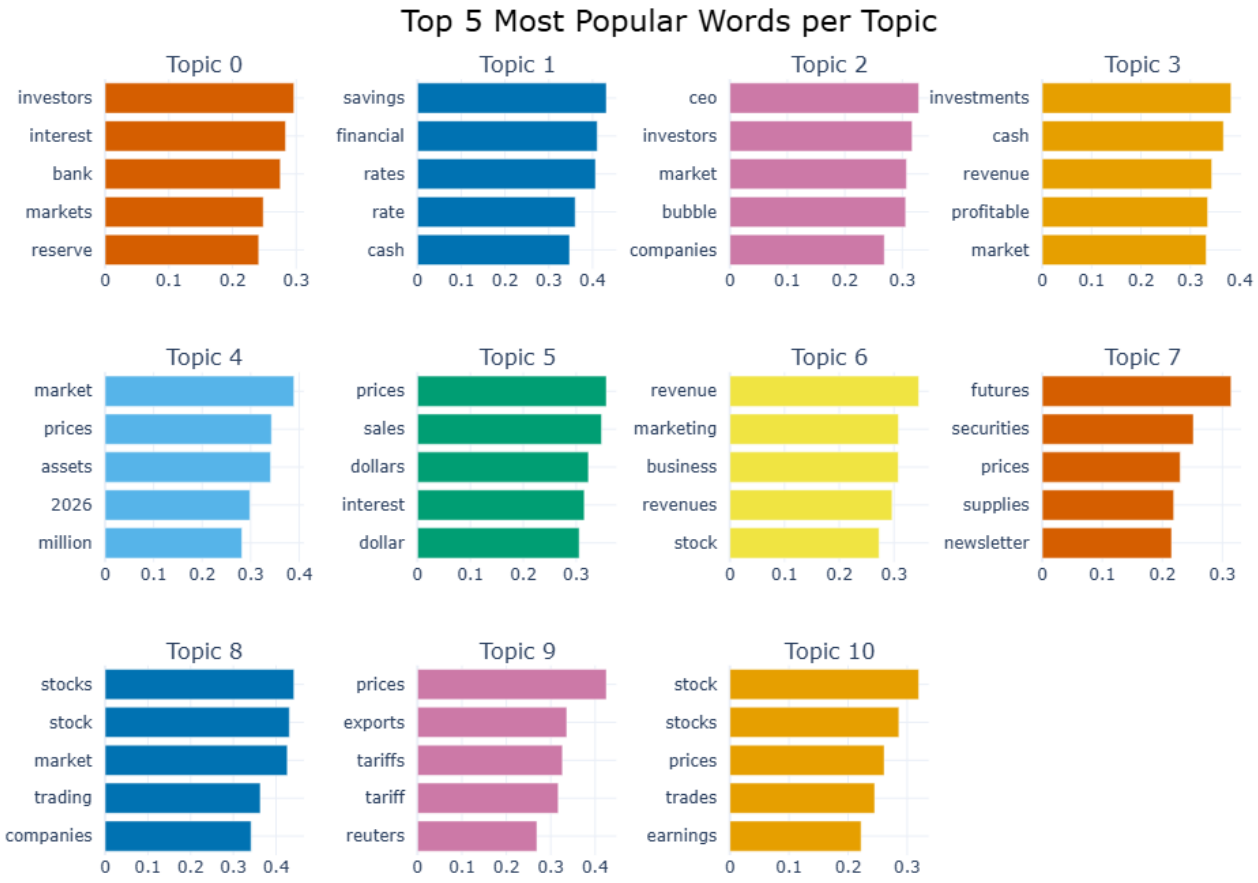
The `Topic_Modeling.ipynb` notebook uses **BERTopic** to discover underlying themes in the news articles.

What the code does:

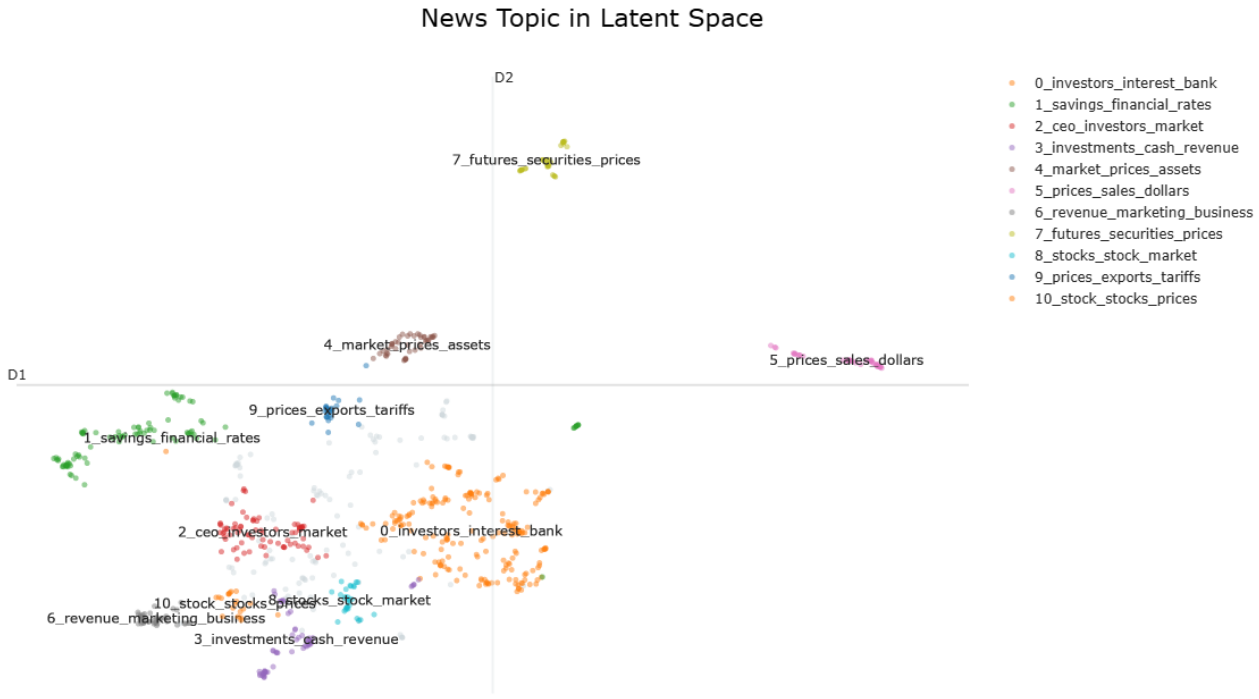
- Loads cleaned data from MongoDB.
- Preprocesses text (removes newlines).
- Uses **UMAP** for dimensionality reduction and **HDBSCAN** for clustering.
- Trains a **BERTopic** model to extract topics.
- Visualizes the most popular words per topic and the document clusters.

Result: Identification of key topics discussed in the financial news, represented by clusters of related keywords.

The most popular words per topic Visualizes



Document Clusters



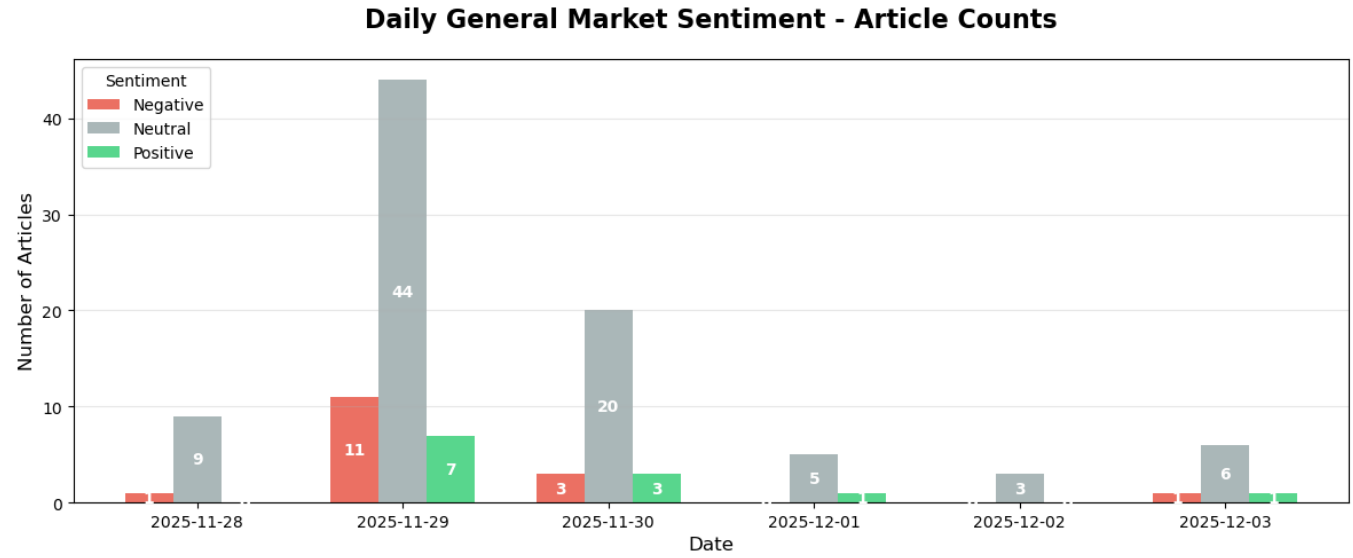
Step 5: Sentiment Analysis

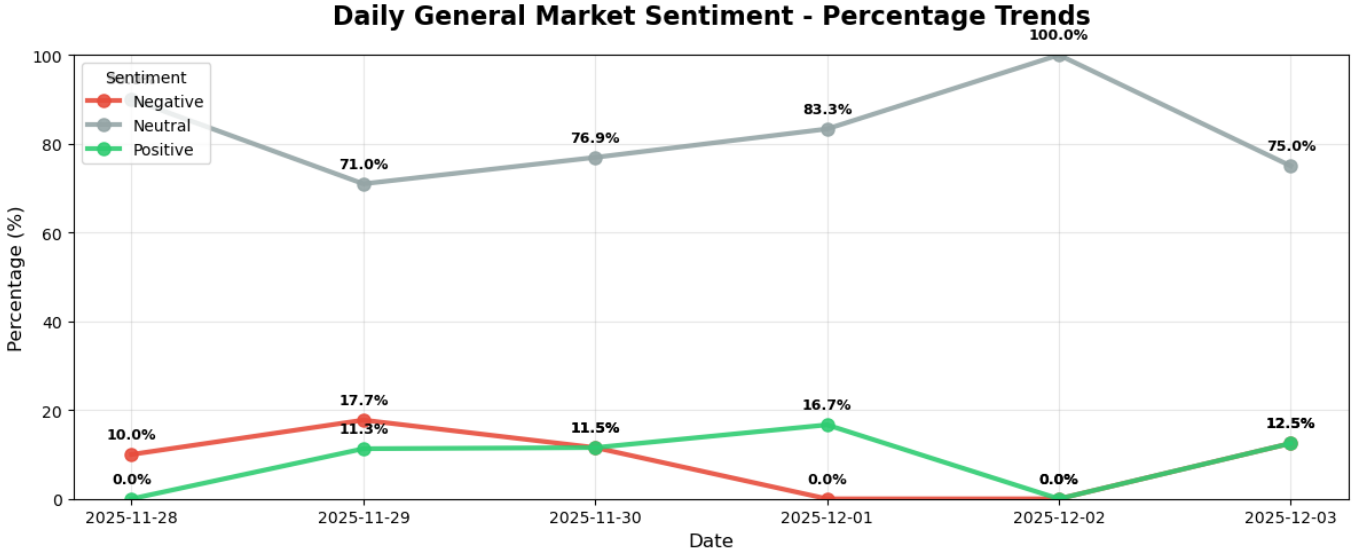
The `sentimental_analysis.ipynb` notebook performs sentiment analysis using **FinBERT**, a BERT model fine-tuned for financial text.

What the code does:

- Loads the **FinBERT** model (`yiyanghkust/finbert-tone`).
- Analyzes the sentiment (Positive, Neutral, Negative) of each article's title and content.
- Performs **Ticker-based Analysis**: Aggregates sentiment for specific stock tickers.
- Performs **Temporal Analysis**: Visualizes sentiment trends over time.

Result: A detailed breakdown of market sentiment, identifying how much the general market are receiving positive or negative coverage and how general market sentiment evolves daily.


Daily General Market Sentiment - Article Counts
Sentiment
Negative
Neutral
Positive
Number of Articles
0
10
20
30
40
9
11
44
7
3
20
3
5
3
6
2025-11-28
2025-11-29
2025-11-30
2025-12-01
2025-12-02
2025-12-03
Date


Daily General Market Sentiment - Percentage Trends
Sentiment
Negative
Neutral
Positive
Percentage (%)
0
20
40
60
80
100
10.0%
71.0%
76.9%
83.3%
100.0%
75.0%
0.0%
17.7%
11.3%
11.5%
0.0%
16.7%
0.0%
12.5%
2025-11-28
2025-11-29
2025-11-30
2025-12-01
2025-12-02
2025-12-03
Date