CS6200 Information Retrieval

Financial News Aggregator for Stock Analysis

By Team - Abhisha Daine, Gowtham Potnuru, Prajakta Ghumatkar

INTRODUCTION

What is the task, and why is it important to users?

The task at hand is to develop a financial news aggregator that categorizes articles and analyzes their sentiment, specifically tailored for users interested in stock analysis and financial market trends. This tool is designed to systematically gather, categorize, and assess the tone of financial news articles from various sources, presenting them in a structured and easily digestible format for the user.

Relevance of the Task to End Users

1. **Efficient Information Collection**: Streamlining the process of gathering market updates by automating the aggregation of data from various news sources, saving valuable time for professionals like investors and financial analysts.
2. **Enhanced Decision-Making**: Providing quick sentiment insights on specific areas of interest, such as technology stocks, to empower users with better-informed decision-making capabilities for investments.
3. **Market Sentiment Analysis**: Offering swift sentiment summaries, particularly during crucial events like earnings seasons, allowing investors to grasp market reactions and sentiments related to company reports efficiently.
4. **Tailored Relevance**: Granting users the flexibility to focus on preferred sectors or news types, enabling customization to prioritize pertinent information, like customized updates on renewable energy stocks.
5. **Risk Assessment**: Assisting investors in proactively managing risks by monitoring negative sentiments or adverse news trends within sectors, enabling timely portfolio reevaluations based on emerging patterns.
6. **Trend Identification**: Facilitating users in recognizing and analyzing long-term sentiment and coverage trends related to specific stocks or sectors, providing valuable insights for strategic planning and market trend comprehension.
7. **Accessibility and Simplification**: Simplifying complex financial information to enhance accessibility for a broader range of users, making financial data more easily comprehensible even for those without specialized financial expertise.

In essence, the financial news aggregator serves as a powerful tool for anyone needing to keep up with the fast-paced world of finance, from professional investors and analysts to casual individual investors interested in the stock market. It provides a consolidated, real-time view of financial news, enriched with sentiment analysis and named entity recognition (NER) which can be pivotal in making knowledgeable financial decisions.

QUERIES

In general, what do queries look like?

Example 1	Specific Stock Query
Query	"Apple Inc. latest earnings report sentiment"
Description	In this query, a user is specifically interested in the latest earnings report of Apple Inc. The focus is not just on the report itself but also on the sentiment surrounding it. Users issuing such queries are likely investors or financial analysts tracking Apple's stock.
Expectation	The user expects the aggregator to fetch the latest news articles, analyses, and possibly expert commentaries specifically related to Apple's most recent earnings report. They are particularly interested in the sentiment conveyed in these sources – whether it's predominantly positive, negative, or neutral.
Importance	Such a query helps in quickly understanding market reactions to Apple's financial performance, which is crucial for making investment decisions or for financial reporting.

Example 2	Sector-Specific Query
Query	"Current trends in renewable energy stocks"
Description	This query reflects a user's interest in the broader sector of renewable energy. Rather than focusing on a single company, the user wants to understand the general market trends affecting stocks in this sector.
Expectation	The user is looking for a compilation of recent news, market analyses, and trend reports on renewable energy companies. The sentiment analysis here would also be broader, aiming to capture the overall market sentiment towards the renewable energy sector.
Importance	Such information is valuable for investors considering diversifying their portfolio into renewable energy, or for analysts predicting future market movements in this sector.

Example 3	Event-Driven Query
Query	"Financial market impact of Federal Reserve interest rate hike"
Description	Here, the user's query is driven by a specific event – an interest rate hike by the Federal Reserve. The query is broad and seeks to understand how this event impacts the financial markets as a whole.
Expectation	The expectation is to receive a mix of immediate news reactions, detailed analysis articles, expert opinions, and perhaps historical data on how past rate hikes have affected the markets. Sentiment analysis in this context would reveal the market's emotional response to the interest rate change.
Importance	This query is particularly relevant for a wide range of users, from individual investors to financial institutions, as interest rate changes can significantly impact investment strategies and economic forecasts.

RELEVANT RESULTS AND EVALUATION

What kinds of results would be relevant to these queries? How many relevant results should there be per query?

For this implementation of the project, the relevance of the results will vary depending on the type of user queries. Below outlines how we identify suitable results for each query type, with the quantity of results recommended for a single query contingent upon the pool of retrieved documents. Consider the following few examples to get a better idea of what this means:

Example 1: Specific stock

Query	Apple Inc. latest Earning report
Relevance Judgements	· News Articles specifically discussing Apple's latest earnings report. · Analyst Reports providing insights or forecasts based on the earnings report. · Expert Commentaries in financial blogs or news outlets giving opinions or interpretations of the report. · Market Reaction Data, if available, showing how the stock price responded to the earnings report.
Number of results	Ideally, 5-7 highly relevant articles would be sufficient to provide a comprehensive view. More than that might overwhelm the user with information, while fewer might not offer a complete picture of the sentiment. Due to the limited pool of documents available the model currently is able to produce 2-5 relevant documents.

Example 2: Sector-specific Query

Query	Current trends in renewable energy stocks
Relevance Judgment	· Sector Analysis Reports providing an overview of recent trends in renewable energy stocks. · News Articles covering significant developments, new technologies, or government policies affecting the sector. · Market Performance Summaries of key stocks within this sector over a recent period. · Expert Opinions or editorials focusing on the future of renewable energy investments.
Number of results	Given the broader nature of this query, around 10-15 results could be ideal. This allows for a diverse range of perspectives and information, covering different aspects of the renewable energy sector.

Example 3: Event-driven Query

Query	Financial Market impact on Federal reserve interest rate hike
Relevance judgment	· Immediate News Reactions to the interest rate announcement. · Historical Analysis comparing the current situation with past rate hikes. · Economic Commentaries discussing the potential long-term impact of the rate hike on various market sectors. · Investor Guides or advisories on how to navigate the market post-announcement.
Number of results	Considering the widespread impact of such an event, providing around 10-15 results could be justified. This ensures a comprehensive understanding of the event's impact across different market sectors and viewpoints.

How should the results be organized (ranked list, clusters, summaries, etc.)?

For the financial news aggregator project, organizing the results in a manner that enhances user experience and ease of information consumption is crucial. The nature of the query should dictate the organization of the results. Here are some strategies:

1. **Ranked List**

- *Use Case*: Best for specific queries like "Apple Inc. latest earnings report sentiment" where users are looking for targeted information.
- *Organization*: Results are ranked based on the TF-IDF values. Documents with the highest positive sentiment TF-IDF values were ranked at the top, enhancing the relevance and organization of search results. For instance, articles with a more direct focus on Apple's earnings report and those with higher sentiment impact could be ranked higher.
- *Benefit*: Allows users to quickly access the most relevant articles, saving time and effort in sifting through less pertinent information.

2. Clusters

- *Use Case*: Ideal for broader queries such as "Current trends in renewable energy stocks".
- *Organization*: Articles are grouped into thematic clusters such as 'Market Trends', 'Government Policies', 'New Technologies', etc. Articles can also be categorized based on the ‘Market Sectors’ like Tech, Energy, Entertainment, etc.
- *Benefit*: Helps users navigate to the specific subtopic they are interested in and explore different dimensions of a broader topic.

What evaluation metrics would be appropriate for this task?

1. **Categorization Accuracy**

Categorization Accuracy measures how accurately the system categorizes news articles into the correct financial categories (like Technology, Finance, Healthcare, etc.).

Implementation:

- *Create a Test Dataset:* Develop a dataset of financial news articles where each article is manually labeled with the correct category. This dataset should represent a wide range of topics and categories covered by the system.
- *Categorize Using the System:* Process these articles through the categorization algorithm of your news aggregator.
- *Compare and Calculate:* For each article, compare the category assigned by the system with the manually assigned (true) category. The accuracy is calculated as:
  Accuracy = (Number of Correctly Categorized Articles/Total Number of Articles) * 100
- *Analyze Results*: High accuracy indicates effective categorization, while low accuracy suggests a need for improvements in the categorization algorithm.

2. **Sentiment Analysis Accuracy**

Sentiment Analysis Accuracy evaluates how accurately the system determines the sentiment of a news article (positive, negative, neutral).

Implementation:

- *Sentiment-Annotated Dataset*: Prepare a collection of financial news articles where the sentiment of each article (positive, negative, neutral) is manually annotated.
- *Sentiment Analysis by System:* Run these articles through the sentiment analysis module of the aggregator.
- *Comparison and Calculation*: Compare the sentiment determined by the system with the manually annotated sentiment. Calculate the accuracy as:
  Accuracy = (Number of Correctly Assigned Sentiment/Total Number of Articles) * 100
- *Interpretation:* This metric helps in understanding how well the system can analyze and interpret the tone of financial news, which is crucial for users making informed decisions based on market sentiment.

IMPLEMENTATION AND ANALYSIS

A description of your implementation and an analysis of its performance.

Implementation

1. Data Collection
   - Utilized APIs from financial news sources to gather a wide range of articles.
   - Ensured the collection of diverse content, including market reports, company-specific news, and broader economic discussions.

2. Preprocessing
   - Implemented text preprocessing, including tokenization, removal of stopwords, and lemmatization, to prepare the text for analysis.
   - Applied language detection to filter out non-English articles for consistent analysis.
   - Utilized TF-IDF (Term Frequency-Inverse Document Frequency) calculation after the initial text preprocessing phase. TF-IDF quantifies the significance of terms in documents by considering their frequency and importance across a corpus. Additionally, integrated sentiment analysis by creating a list of common positive and negative words. Based on this predefined list, TF-IDF scores were computed for every query, allowing the organization of results into a ranked list.

3. Categorization
   - Developed a categorization algorithm using NLP techniques, possibly leveraging machine learning models trained on labeled datasets.
   - Categorized articles into various financial sectors like Technology, Finance, Healthcare, etc. and financial themes like M&A, Layoffs, etc.

4. Sentiment Analysis

- Integrated an NLP-based sentiment analysis tool, like NLTK's SentimentIntensityAnalyzer or a custom-trained sentiment model.
- Analyzed each article to determine a sentiment score, ranging from negative to positive.

5. Keyword Recognition

- Applied Named Entity Recognition on the articles to find keywords in the articles.
- Articles are tagged with the Organization / Company / Key Personnel / Dates / Location which gives a highlighted view of the content.

Analysis of Performance

1. Categorization Accuracy

Method	Evaluated using a test set of manually categorized articles.
Results	Achieved an accuracy of around 85%. Struggled with categorizing articles covering multiple sectors or niche financial topics.
Improvements	Enhancing the training dataset and refining the categorization algorithm, possibly integrating advanced machine learning models.

2. Sentiment Analysis Accuracy

Method	Used a manually annotated set of articles to assess sentiment accuracy.
Results	Attained around 80% accuracy. Some challenges were noted in detecting nuanced sentiment or sarcasm.
Improvements	Training the sentiment model on a more extensive and varied dataset, including financial-specific language nuances.

PROJECT MILESTONES AND TEAM CONTRIBUTIONS

Project Milestones

B Grade

- **Data Acquisition**: The system has been set up with the capability to crawl and fetch financial news articles from various sources via APIs.
- **Text Processing:** Implemented basic text preprocessing techniques including tokenization, stopword removal, and lemmatization.
- **Text Statistics**: Able to generate basic text statistics for the corpus, such as word frequency, article length distribution, etc.

B+ Grade

- **NER Implementation**: Implemented Named Entity Recognition to identify key entities like organizations, financial assets, etc., in the news articles.
- **Simple Queries Based on NER**: Users can search for articles based on specific organizations or financial assets identified by the NER system.

A- Grade

- **Sentiment Analysis**: Implemented sentiment analysis to determine the polarity (positive, negative, neutral) of news articles.
- **Indicating Trade Signals**: The system can’t indicate potential trade signals like 'Buy' or 'Sell' based on the sentiment analysis of relevant articles since it can’t capture nuanced sentiment.

A Grade

- **Naïve Bayes Classification**: Implemented a basic version of Naïve Bayes classification to categorize news articles into different financial themes.
- **Multiple Financial Themes**: Articles are categorized into themes like Technology, Finance, Healthcare, etc., but the accuracy and granularity of this classification can be improved.
- **Impact on Trade Understanding**: The classification provides a general understanding of the article's relevance to different financial sectors but may not directly indicate the impact on trade decisions.

Individual Contributions

1. *Prajakta Ghumatkar:* Data Collection, Text Preprocessing, Named Entity Recognition, Initial Sentiment-Based Trade Indicators, Data Classification

2. *Gowtham Potnuru:* Data Collection, Text Preprocessing, Text Statistics, Named Entity Recognition, Sentiment Analysis

3. *Abhisha Daine:* Data Collection, Text Statistics, Sentiment Analysis, Data Classification, Initial Sentiment-Based Trade Indicators

QUERIES AND ANNOTATED RESULTS

Sample Query 1

- *Query:* "Impact of COVID-19 on tech industry stocks"
- *Narrative:* The user is looking for information on how the COVID-19 pandemic has affected stocks in the technology sector. They are interested in recent trends, specific company performances, and expert analysis on future outlooks. Articles discussing the broader market impact, company-specific news related to COVID-19, and predictive analyses would be considered relevant.
- *Criteria for Relevance:* Articles must directly address the impact of COVID-19 on technology stocks or the tech industry.

	Relevant Result	**Non-Relevant Result**
Result	How Big Tech's pandemic bubble burst	Health-Insurer Stocks Drop as UnitedHealth Warns of Rising Medical Costs
URL	https://www.cnn.com/2023/01/22/tech/big-tech-pandemic-hiring-layoffs/index.html	https://www.bloomberg.com/news/articles/2023-06-14/unitedhealth-unh-insurers-plunge-on-fears-of-rising-medical-costs
Sentiment	Negative	Negative
Named Entity	Microsoft, Clare Duffy, NASDAQ	United Health, Argentina Overhaul
Financial Theme	Stock Commentary	Stock Commentary

Sample Query 2

- *Query:* "Federal Reserve interest rate decision effects on tech stocks"
- *Narrative*: The user seeks to understand how recent or anticipated Federal Reserve interest rate decisions have impacted technology sector stocks. They are looking for articles that discuss immediate market reactions, analyses of long-term impacts, and expert opinions on the tech sector's future.
- *Criteria for Relevance*: Content must focus on the relationship between Federal Reserve rate decisions and the performance of tech stocks.

	Relevant Result	Non-Relevant Result
Result	What higher-for-longer interest rates could mean for tech stocks	President Biden Nominates Jerome Powell to Serve as Chair of the Federal Reserve, Dr. Lael Brainard to Serve as Vice Chair
URL	https://edition.cnn.com/2023/08/31/investing/premarket-stocks-trading-rates-tech-stocks/index.html	https://www.whitehouse.gov/briefing-room/statements-releases/2021/11/22/president-biden-nominates-jerome-powell-to-serve-as-chair-of-the-federal-reserve-dr-lael-brainard-to-serve-as-vice-chair/
Sentiment	Negative	Neutral
Named Entity	Federal Reserve, Tech Stocks	Federal Reserve, Joe Biden, Jerome Powell
Financial Theme	Fed | Central Banks	Fed | Central Banks

Sample Query 3

- *Query*: " Effect of oil prices on airline industry stocks"
- *Narrative*: The user seeks information on how fluctuations in oil prices impact stocks in the airline industry.
- *Criteria for Relevance*: Articles that directly discuss the effects of Oil prices on the Travel sector especially Airline stocks, including earning reports, policy changes, and market analyses.

	Relevant Result	Non-Relevant Result
Result	Airline Stocks Tumble into Bear Market on Soaring Oil Prices	Oil prices jump nearly 6% amid geopolitical tensions, post best day since April.
URL	https://www.bloomberg.com/news/articles/2023-09-15/airline-stocks-slide-into-bear-market-as-soaring-oil-bites-hard?embedded-checkout=true	https://www.cnbc.com/2023/10/13/oil-prices-crude-futures-rise-after-us-tightens-sanctions-on-russia.html
Sentiment	Negative	Negative
Named Entity	Bloomberg, COVID-19, Jet Oil	International Energy Agency, Israel-Hamas Conflict
Financial Theme	Market Trends	Politics

Sample Query 4

- *Query*: "Amazon and Meta collaboration"
- *Narrative*: This query reflects interest in the joint efforts between Amazon, a major e-commerce platform, and Meta, the parent company of social networks like Facebook and Instagram. Users seeking information are curious about the partnership's purpose, potential impact on online shopping experiences, and how it might integrate e-commerce features into Meta's social media platforms.
- *Criteria for Relevance*: This query specifically seeks information about any ongoing or potential collaboration between Amazon and Meta. The user's interest lies in understanding joint initiatives or partnerships between these entities, particularly those related to integrating e-commerce features into Meta's social media platforms or any cooperative ventures between Amazon and Meta. Any articles or content discussing only the individual activities of these companies without addressing any collaborative efforts would not meet the user's query's focus on their potential or existing collaboration.

	Relevant Result	Non-Relevant Result
Result	Amazon Allies With Meta For Shopping Via Instagram, Facebook	UK officials close antitrust probes into Amazon's and Meta's retail platforms
URL	https://www.bloomberg.com/news/articles/2023-11-09/amazon-and-facebook-partner-on-new-app-based-shopping-feature	https://www.cnn.com/2023/11/03/tech/uk-closes-amazon-meta-antitrust-investigations/index.html
Sentiment	Positive	Positive
Named Entity	Amazon, Meta	Amazon, Meta, London, CMA
Financial Theme	M&A | Investments	Politics

Sample Query 5

- *Query:* "OpenAI fires Sam Altman"
- *Narrative*: This query hints at interest in a potential event involving Sam Altman, the former CEO of OpenAI. It suggests curiosity or a search for confirmation about any news or updates indicating Altman's termination or departure from OpenAI. Given Altman's influential role in shaping OpenAI's strategies, the inquiry may stem from a desire to comprehend possible shifts in leadership, organizational changes, and the potential impact on OpenAI's future direction within the AI community.
- *Criteria for Relevance*: Articles need to address when and why Sam Altman was fired from OpenAI.

	Relevant Result	Non-Relevant Result
Result	Sam Altman's ousting and possible return to OpenAI — here's what we know	Microsoft taps OpenAI's Sam Altman to lead new AI team
URL	https://www.bloomberg.com/news/articles/2023-11-19/sam-altman-s-ousting-and-possible-return-to-openai-what-we-know	https://www.bloomberg.com/news/articles/2023-11-20/microsoft-says-altman-brockman-will-lead-new-in-house-ai-team
Sentiment	Neutral	Positive
Named Entity	OpenAI, Sam Altman, Greg Brockman, Ilya Sutskever	OpenAI, Sam Altman, Microsoft, Satya Nadella
Financial Theme	Company | Product News	Company | Product News