B5W1: Predicting Price Moves with News Sentiment

Dive into real-world financial news and discover how headlines shape stock swings, learn to analyze sentiment scores, run technical indicators, and link them to daily returns.

CHALLENGE OVERVIEW

This project focuses on the detailed analysis of a large corpus of financial news data to discover correlations between news sentiment and stock market movements. This challenge is designed to refine your skills in Data Engineering (DE), Financial Analytics (FA), and Machine Learning Engineering (MLE).

This challenge will enhance your ability to analyze complex data sets, demonstrate adaptability, and employ innovative thinking skills that are crucial for the demanding environment at Nova Financial Insights. This project will not only deepen your understanding of essential financial analytic techniques.

Engage with as many tasks as possible. The volume and complexity of the tasks are designed to simulate the pressures and deadlines typical in the financial analytics field.

BUSINESS OBJECTIVE

Nova Financial Solutions aims to enhance its predictive analytics capabilities to significantly boost its financial forecasting accuracy and operational efficiency through

advanced data analysis. As a Data Analyst at Nova Financial Solutions, your primary task is to conduct a rigorous analysis of the financial news dataset. The focus of your analysis should be two-fold:

- **Sentiment Analysis**: Perform sentiment analysis on the 'headline' text to quantify the tone and sentiment expressed in financial news. This will involve using natural language processing (NLP) techniques to derive sentiment scores, which can be associated with the respective 'Stock Symbol' to understand the emotional context surrounding stock-related news.
- **Correlation Analysis**: Establish statistical correlations between the sentiment derived from news articles and the corresponding stock price movements. This involves tracking stock price changes around the date the article was published and analyzing the impact of news sentiment on stock performance. This analysis should consider the publication date and potentially the time the article was published if such data can be inferred or is available.

Your recommendations should leverage insights from this sentiment analysis to suggest investment strategies. These strategies should utilize the relationship between news sentiment and stock price fluctuations to predict future movements. The final report should provide clear, actionable insights based on your analysis, offering innovative strategies to use news sentiment as a predictive tool for stock market trends.

DATASET OVERVIEW

Financial News and Stock Price Integration Dataset

FNSPID (Financial News and Stock Price Integration Dataset), is a comprehensive financial dataset designed to enhance stock market predictions by combining quantitative and qualitative data.

The structure of the data is as follows

- **headline**: Article release headline, the title of the news article, which often includes key financial actions like stocks hitting highs, price target changes, or company earnings.
- **url**: The direct link to the full news article.
- **publisher**: Author/creator of article.

- **date**: The publication date and time, including timezone information(UTC-4 timezone).
- **stock**: Stock ticker symbol (unique series of letters assigned to a publicly traded company). For example (AAPL: Apple)

TEAM

Facilitator:

- Mahlet
- Kerod
- Rediet
- Rehmet

KEY DATES

- **Challenge Introduction** - 8:00 AM UTC time on Wednesday 28 May 2025.
- **Interim Submission - 8:00 PM UTC time on Friday 30 May 2025.**
- **Final Submission** - 8:00 PM UTC time on Tuesday 03 June 2025.

COMMUNICATION & SUPPORT

- **Slack channel:** #all-week1
- **Office hours:** Mon–Fri, 08:00–15:00 UTC on Zoom

LEARNING OBJECTIVES

By the end of this week, you will be able to:

1. Configure a reproducible Python data-science environment with GitHub integration.
2. Perform Exploratory Data Analysis (EDA) on text and time series data.
3. Compute technical indicators (MA, RSI, MACD) using TA-Lib and PyNance.
4. Run sentiment analysis on news headlines with NLP tools.
5. Measure correlation between news sentiment and daily stock returns.
6. Document findings and write a concise, publication-style report.

PROJECT PLANNING - EDA & STATS

- Tasks:
  - Data Understanding
  - Exploratory Data Analysis (EDA)
  - Statistical thinking
- KPIs:
  - Proactivity to self-learn - sharing references.
  - EDA techniques to understand data and discover insights,
  - Demonstrating Stats understanding by using suitable statistical distributions and plots to provide evidence for actionable insights gained from EDA.

DELIVERABLES AND TASKS TO BE DONE

Task 1: Git and GitHub

- Tasks:
  - Setting up Python environment
  - Git version control
  - CI/CD
- Key Performance Indicators (KPIs):
  - Dev Environment Setup.
  - Relevant skill in the area demonstrated.
- Suggested folder structure:

```
├── .vscode/
│   └── settings.json
├── .github/
│   └── workflows
│       ├── unittests.yml
├── .gitignore
├── requirements.txt
```

```
├── README.md
├── src/
│   ├── __init__.py
├── notebooks/
│   ├── __init__.py
│   └── README.md
├── tests/
│   ├── __init__.py
└── scripts/
    ├── __init__.py
    └── README.md
```

Minimum Essential To Do

- Create a github repository that you will be using to host all the code for this week.
- Create at least one new branch called ”task-1” for your analysis
- Commit your work at least three times a day with a descriptive commit message
- Perform Exploratory Data Analysis (EDA) analysis on the following:
    - **Descriptive Statistics**:
        - Obtain basic statistics for textual lengths (like headline length).
        - Count the number of articles per publisher to identify which publishers are most active.
        - Analyze the publication dates to see trends over time, such as increased news frequency on particular days or during specific events.
    - **Text Analysis(Topic Modeling)**:

        - Use natural language processing to identify common keywords or phrases, potentially extracting topics or significant events (like "FDA approval", "price target", etc.).
    - **Time Series Analysis**:
        - How does the publication frequency vary over time? Are there spikes in article publications related to specific market events?
        - Analysis of publishing times might reveal if there's a specific time when most news is released, which could be crucial for traders and automated trading systems.
    - **Publisher Analysis**:
        - Which publishers contribute most to the news feed? Is there a difference in the type of news they report?
        - If email addresses are used as publisher names, identify unique domains to see if certain organizations contribute more frequently.

Task 2: Quantitative analysis using pynance and TaLib

- Tasks:
    - Use additional finance data
    - Load and prepare the data.
        - Load your stock price data into a pandas DataFrame. Ensure your data includes columns like Open, High, Low, Close, and Volume.
    - Apply Analysis Indicators with TA-Lib
        - You can use TA-Lib to calculate various technical indicators such as moving averages, RSI (Relative Strength Index), and MACD (Moving Average Convergence Divergence)
    - Use PyNance for Financial Metrics
    - Visualize the Data
        - Create visualizations to better understand the data and the impact of different indicators on the stock price.
- KPIs
    - Proactivity to self-learn - sharing references.
    - Accuracy of indicators
    - Completeness of Data Analysis

Minimum Essential To Do:

- Merge the necessary branches from task-1 into the main branch using a Pull Request (PR)
- Create at least one new branch called "task-2" for the ongoing development of the dashboard.
- Commit your work with a descriptive commit message.
- Prepare Your Data
- Calculate Basic Technical Indicators
- Visualize Data

Task 3: Correlation between news and stock movement

- Tasks:
  - **Date Alignment**: Ensure that both datasets (news and stock prices) are aligned by dates. This might involve normalizing timestamps.
  - **Sentiment Analysis**: Conduct sentiment analysis on news headlines to quantify the tone of each article (positive, negative, neutral).Tools: Use Python libraries like nltk, TextBlob for sentiment analysis.
  - **Analysis:**
    - **Calculate Daily Stock Returns**: Compute the percentage change in daily closing prices to represent stock movements.
    - **Correlation Analysis**: Use statistical methods to test the correlation between daily news sentiment scores and stock returns.
- KPIs
  - Proactivity to self-learn - sharing references.
  - Sentiment Analysis
  - Correlation Strength

Minimum Essential To Do:

- Merge the necessary branches from task-2 into the main branch using a Pull Request (PR)
- Create at least one new branch called "task-3" for the ongoing development of the dashboard.
- Commit your work with a descriptive commit message.

- Data preparation
  - **Normalize Dates**: Align dates in both news and stock datasets to ensure each news item matches the corresponding stock trading day.
  - **Perform Sentiment Analysis**: Use a simple and effective sentiment analysis tool to assign sentiment scores to headlines.
  -
- Calculate Stock Movements
  - **Compute Daily Returns**: Calculate daily percentage changes in stock prices to represent movements.
- Correlation Analysis
- **Aggregate Sentiments**: Compute average daily sentiment scores if multiple articles appear on the same day.
- **Calculate Correlation**: Determine the Pearson correlation coefficient between average daily sentiment scores and stock daily returns.

DUE DATE (SUBMISSION)

Friday (30 May, 2025): 8:00 PM (UTC)

- GitHub Link to your main branch
- Interim report - Covering task-1 partial progress task-2
  - **Length**: Maximum 3 pages.
  - **Focus**: Summarize initial findings, methodology, and any challenges encountered. Keep the report concise and informative.

Tuesday (03 Jun, 2025): 8:00 PM (UTC)

- GitHub Link to your main branch
- FInal report : Covers all **Week-1 work** in detail.
  - **Length**: Up to 10 pages, including a maximum of 10 plots.
  - **Format**: Written in a style suitable for publication as a **Medium Blog**.

Feedback

You may not receive detailed comments on your interim submission but will receive a grade.

OTHER CONSIDERATIONS

- **Documentation:** Encourage detailed documentation in code and report writing.
- **Collaboration:** Emphasise collaboration through Github issues and projects.
- **Communication**: Regular check-ins, Q&A sessions, and a supportive community atmosphere.
- **Flexibility:** Acknowledge potential challenges and encourage proactive communication.
- **Professionalism:** Emphasise work ethics and professional behavior.
- **Time Management:** Stress the importance of punctuality and managing time effectively.

TUTORIALS SCHEDULE

In the following, the **Bold** indicates morning sessions, and *ITALIC* indicates afternoon sessions.

- Day 1: `
  - **Introduction to the Challenge (Mahlet)**
  - *Stock market data and analysis (Kerod)*
- Day 2:
  - **Introduction to YFinance, pynance (Rehmet)**
  - *Modular Programming with Python Scripts and Jupyter (Rediet)*
- Day 3:
  - **Introduction to quantitative and time series analysis (Rediet)**
  - *Data visualization and interpretation (Rehmet)*
- Day 4:
  - **Correlation analysis(Kerod)**
  - *Q&A (Mahlet & Kerod)*
- Day 5:
  - **Q&A (Rehmet)**

Feedback

You will receive comments/feedback in addition to a grade.

REFERENCES

- **Stock Market**

  - https://www.investopedia.com/terms/s/stockmarket.asp
  - https://www.investopedia.com/terms/s/stock-analysis.asp
- **Python Testing**
  - https://machinelearningmastery.com/a-gentle-introduction-to-unit-testing-in-python/
  - https://docs.python-guide.org/writing/tests/
  - https://realpython.com/python-testing/
- **Python Packages:**
  - https://textblob.readthedocs.io/en/dev/
  - https://github.com/mqandil/pynance
  - https://github.com/ta-lib/ta-lib-python
  -
- **Data Engineering**
  - What is Data Engineer: Role Description, Skills, and Background | AltexSoft
- **Version control – Git**
  - What is version control | Atlassian
  - Learn Git branching -- interactive way to learn Git
  - Git with large files
  - Which files to not track and how to not track them? | Atlassian
  - .gitignore docs
  - Conventional commits -- lightweight convention on top of commit messages.
- **CI/CD**
  - What is Continuous Integration | Atlassian
  - DevOps Pipeline | Atlassian
  - 7 Popular Open Source CI/CD Tools - DevOps.com
  - Setting up a CI/CD pipeline on Github