Week 1

Interim Report for Predicting Price Moves with News Sentiment

Date: May 31, 2025

By: Sifrayenesh Wolde

sifrawolde@gmail.com

1. Introduction

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 conducting a rigorous financial news dataset analysis.

Current Progress

2.1 Project Structure & Setup

The repository has been organized to ensure modularity, reproducibility, and maintainability:

- ✓ Src/– Contains core Python modules:
    - o data_loader.py– Fetches news data from APIs (e.g., Alpha Vantage, NewsAPI).
    - o Eda.py– Includes functions for statistical analysis and visualization.
- ✓ notebook/– Jupyter notebooks for interactive EDA (exploratory.ipynb).
    - o test/– Unit tests (test_eda.py) to validate functionality.
- ✓ scripts/ – Utility scripts (preprocess.py) for data cleaning.
- ✓ CI/CD Pipeline – GitHub Actions (unittest.yml) for automated testing.
- ✓ Environment Management _ requirements.txt for dependencies.
- ❖ **Completed:**
- Basic API integration for news data.
- Initial EDA functions (summary stats, sentiment analysis).
- Unit test framework setup.

- ❖ **In Progress:**

- Enhancing data preprocessing for NLP tasks.
- Expanding test coverage.

2. Key Findings (Preliminary Analysis)

3.1 Data Insights

- **News Sentiment vs. Stock Movement:** Initial observations suggest a correlation between positive/negative news and short-term stock price changes.

- **Volume of News Matters:** Stocks with higher news coverage show more volatility.

- **Challenges in Data Cleaning:**

    - o Noise in headlines (e.g., "Apple launches product" vs. "Apple the fruit").

    - o Time zone mismatches between news timestamps and market data.

3.2 Technical Observations

- **API Limitations:** Some free APIs have rate limits (e.g., News API allows 100 requests/day).
- **Data Storage:** Need efficient storage for large news datasets (considering SQLite or Parquet).

3. Challenges Faced

Challenge	Impact	Mitigation Strategy
API Rate Limits	Slows down data collection	Implement caching, use multiple APIs
Noisy Text Data	Affects sentiment analysis	Improve NLP preprocessing (stopword removal, entity recognition)
Time Alignment	News vs. stock timestamps mismatch	Normalize timestamps to market hours
Test Data Generation	Mocking realistic news data is difficult	Use synthetic datasets for testing

4. Next Steps

5.1 Immediate Tasks

✓ Enhance Data Preprocessing

- Implement NLP techniques (TF-IDF, word embeddings).
- Handle missing data and outliers.

✓ Expand EDA

- Add visualizations (time-series trends, sentiment distribution).
- Compare multiple stocks' news impact.

✓ Improve Testing

- Increase unit test coverage.

- Add integration tests for API calls.

5.2 Medium-Term Goals

🔨 Build Predictive Models

- Experiment with:

- o **Sentiment-based models** (Logistic Regression, LSTM).
- o **Time-series forecasting** (ARIMA, Prophet).
- Evaluate accuracy using backtesting.

🔨 **Deployment**

- Dockerize the application for reproducibility.
- Set up a lightweight Flask/FastAPI demo.

7. Conclusion

The project is progressing as planned, with foundational components in place. The next critical steps involve **improving data quality, expanding analysis, and building predictive models**. Challenges like API limitations and noisy data are being addressed systematically.