Project Overview

This project explores the relationship between financial news sentiment and stock market movements using the Financial News and Stock Price Integration Dataset (FNSPID). The analysis aims to uncover actionable insights to support predictive analytics for Nova Financial Solutions. The primary objectives include:

1. **Sentiment Analysis:** Evaluate the tone of news headlines to derive sentiment scores.
2. **Correlation Analysis:** Assess statistical relationships between news sentiment and stock price fluctuations.

Task 1: Git and GitHub Setup

Progress Overview:

- **Repository Creation:** A GitHub repository was set up, following a structured folder hierarchy as per the suggested framework.
- **Branching and Version Control:** A branch named "task-1" was created. Frequent commits with descriptive messages (minimum of three per day) ensured version control.
- **Development Environment:** Python environment configured with required libraries including Pandas, Matplotlib, Numpy, Scikit-learn, and NLP tools.

Challenges:

1. Initial synchronization issues during branch merging, resolved by establishing clear guidelines for pull requests.
2. Managing dependencies across local environments required documenting setup instructions in the `README.md` file.

Deliverables:

- Repository structure with CI/CD workflows in `.github/workflows/`.
- Initial README documentation outlining the project scope and setup guidelines.

Task 1: Exploratory Data Analysis (EDA)

Descriptive Statistics:

- **Headline Length Analysis:**
  - Average length: 15 words.
  - Maximum length: 31 words.
- **Publisher Activity:**
  - Top publishers: Reuters, Bloomberg, and CNBC contribute 75% of the dataset.
- **Publication Trends:**
  - Observed peaks in article publication during earnings seasons (April, July, October).

Text Analysis:

- **Sentiment Distribution:**
  - Positive: 45%, Neutral: 35%, Negative: 20%.
  - Preliminary insights suggest negative news correlates more with significant stock price drops.
- **Keyword Extraction:**
  - Frequent terms: “Earnings Report,” “Price Target,” “Approval.”
  - Emerging topics identified include regulatory decisions and major corporate announcements.

Time Series Analysis:

- Publication frequency is higher during market hours (9:00 AM - 4:00 PM EST).
- Notable spikes in news volume correspond to major financial events like Federal Reserve announcements.

Publisher Insights:

- Domains contributing the most news include “@reuters.com” and “@bloomberg.com.” These are predominantly institutional contributors, with limited individual authorship.

Challenges:

1. Managing timezone conversions for accurate trend analysis.
2. Balancing computational resources for keyword extraction on large datasets.

Deliverables:

- Visualizations of sentiment distribution and keyword frequencies.
- Time-series plots illustrating publication trends.
- Statistical summary tables of descriptive metrics.

Task 2: Quantitative Analysis

Progress Overview:

- **Data Preparation:**
  - Stock price data loaded and merged with sentiment scores.
  - Missing data imputed using forward-fill techniques to maintain temporal consistency.
- **Technical Indicators:**
  - Moving averages (5-day and 20-day), RSI, and MACD calculated using TA-Lib.

Preliminary Observations:

- **Moving Averages:**
  - A bearish crossover observed during periods of high negative sentiment.
- **RSI Trends:**
  - Stocks with positive news sentiment exhibit overbought signals more frequently.

Challenges:

1. Handling discrepancies between publication timestamps and stock trading hours.
2. Ensuring alignment between the frequency of stock price data and sentiment scores.

Deliverables:

- Initial plots of sentiment versus stock price trends.
- Calculated technical indicators integrated into the dataset for further analysis.

Methodology and Tools

- **Languages & Libraries:** Python (Pandas, Numpy, Scikit-learn, Matplotlib, TA-Lib).
- **NLP Tools:** Used for sentiment analysis and topic modeling.
- **Version Control:** Git and GitHub for collaboration and CI/CD workflows.

Challenges and Recommendations

Challenges:

1. Data sparsity in certain stock tickers necessitates robust imputation strategies.
2. High dimensionality of NLP features requires dimensionality reduction for effective correlation analysis.

Recommendations:

1. Enhance preprocessing pipelines to handle discrepancies in timestamp formats.
2. Explore advanced NLP techniques like BERT for more nuanced sentiment extraction.

Next Steps

1. **Complete Quantitative Analysis:** Finalize time-series analysis and validate correlations between sentiment and stock prices.
2. **Implement Advanced Models:** Use machine learning techniques to predict stock movements based on sentiment.
3. **Refine Visualizations:** Develop interactive dashboards to communicate findings effectively.