Predicting Stock Movements Based on News Sentiment Using NLP

Problem Identification

Problem Statement:

Can we predict short-term stock price movements using the sentiment of financial news headlines, rather than relying on historical stock price trends? This project aims to explore whether natural language in financial news article headlines can be quantified and used to accurately predict whether a stock's price will rise or fall.

Context:

Stock prices are traditionally analyzed using time-series data and technical indicators. However, markets often react to qualitative news content–mergers, leadership changes, economic events, and more–before these changes are reflected in historical prices. The rise of NLP presents a promising avenue to exploit this information. This project diverges from conventional models by using only the sentiment extracted from news headlines to predict market behavior.

Criteria for Success:

- A working model that classifies news headlines as indicators of stock price increase or decrease
- Achieving classification accuracy significantly above random chance (e.g., >60%)
- A well-documented analysis pipeline with reproducible code
- A validation strategy showing the model's generalizability to unseen data

Scope of the Solution Space:

This project will:

- Use NLP to analyze the sentiment of financial news headlines
- Link news article headlines to the appropriate company ticker(s)
- Use stock price changes after the headline publication as the target variable
- Explore different machine learning classifiers
- Evaluate the performance using standard metrics

It will NOT:

- Use time-series data or historical price trends as features
- Perform long-term predictions (focusing on the short-term: 1-day - 3-day window after publication)

Constraints:

- "Sentiment" is inherently subjective, models may misclassify nuanced headlines
- Headlines may not always cause stock price movement: correlation is not causation

- Linking headlines to appropriate entities may not always work (may be ambiguously linked, or no market data available for that time/entity)
- Limited availability or time alignment between headline publication time and market hours

Stakeholders:
- Investors who want tools to interpret new impact
- Financial analysts who want to integrate NLP models into decision-making tools
- Traders/portfolio managers who might benefit from timely sentiment-based alerts
- Researchers in behavioral finance/data science

Data Sources:
- Primary source: Kaggle Dataset -> Financial News Headlines + Tickers (Contains labeled financial headlines and associated tickers)
- Stock price data: Using Yahoo Finance (via finance) to get stock closing prices over time, for time periods immediately after headline dates

Approach

1. Data cleaning & preprocessing:
    a. Associate each headline with its ticker(s) and timestamp, and connect to closing stock price immediately after headline date
    b. Remove missing values, outliers, etc.
2. Sentiment analysis:
    a. Use NLP models to assign sentiment scores to each headline (pretrained?)
    b. Convert sentiment into categorical classes: positive, negative, neutral
3. Label generation:
    a. Calculate short-term stock price change after each headline, labeling as increase (1) or decrease (0) based on some threshold
4. Modeling:
    a. Train classification models (X=sentiment score, Y=stock movement, 1 or 0) -> sentiment analysis != prediction (classification learns to relate sentiment to stock movement)
    b. Evaluate performance on held-out test data
5. Evaluation:
    a. Use classification metrics to determine accuracy/precision
    b. Analyze and extract insights of which headlines have strongest predictive powers