Intelligent Factor-Based Stock Selection System

MSML651

Jingjing Peng

Motivations

-- Financial markets are increasingly complex and influenced by sector dynamics, macro trends, corporate operations, and investor sentiment.

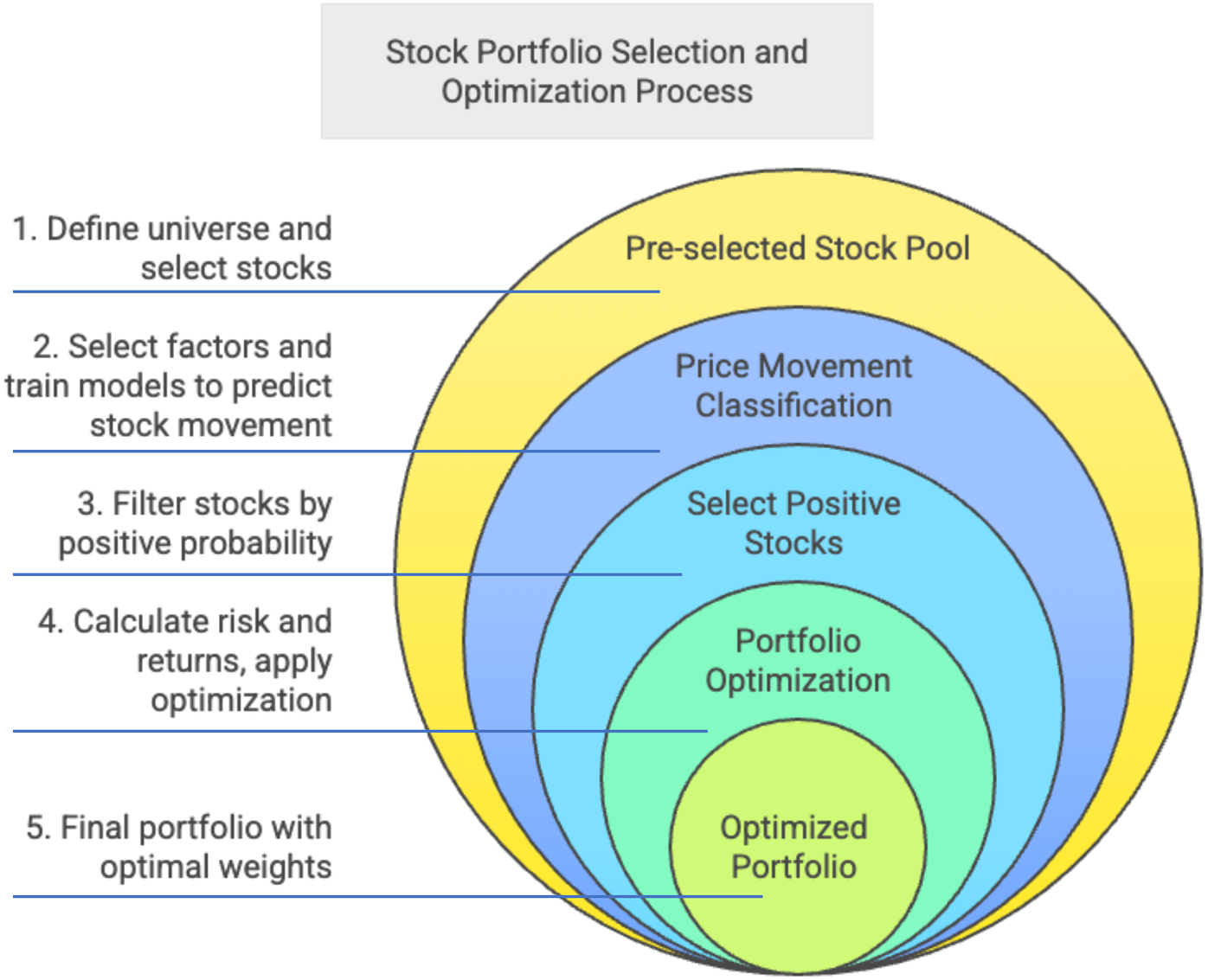
-- Large institutions rely on sophisticated models and vast resources to drive their investment strategies, making precise, data-driven decisions. I am interested in knowing how it works.

-- The project aims to develop an quantitative, data-driven stock selection and portfolio optimization system.

-- I break this problem into several main steps: data collection, modeling, stock selection, prediction, and portfolio optimization.

Project Overview

The system is designed to outperform traditional equal-weight or passive investment strategies by using quantitative data analytics, machine learning, and optimization methods.



Step 1: Pre-selected Stock Pool

Data Sources

5 sectors × 5 stocks = 25 stocks total

Historical Data

Daily prices, volume, returns, volatility

Macro Factors

Interest rates, GDP growth, unemployment, PMI

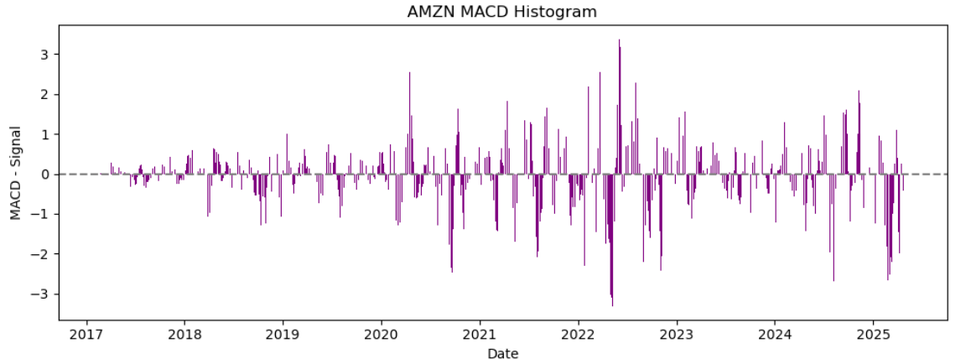
Text Analysis

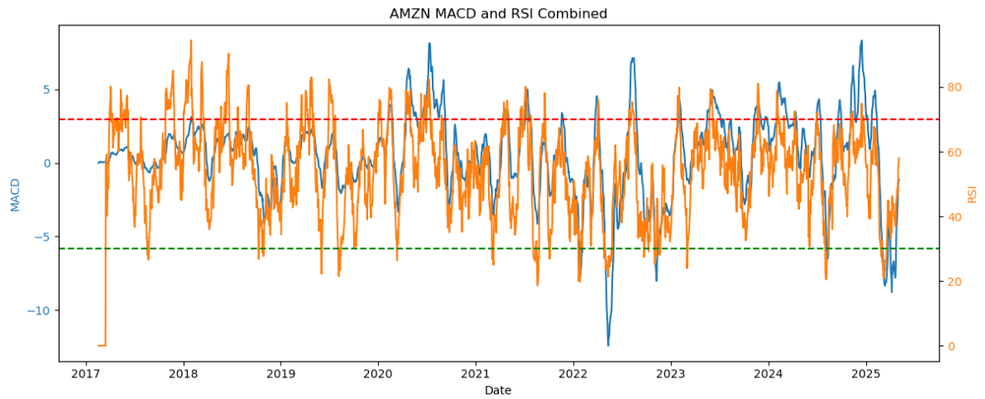
10-K, 10-Q reports with sentiment scores

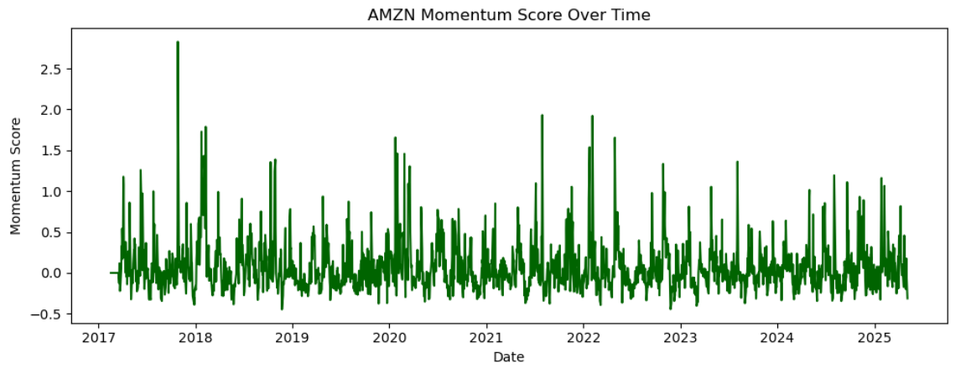
Here I chose **5 key sectors** to ensure coverage across diverse industries, capturing sector-specific risks and opportunities.
Within each sector, the selected stocks are well-known, actively traded, and have available historical and financial data.
The inclusion of **text-based sentiment (10-K, 10-Q)** and **macro factors** provides both quantitative and qualitative signals, aiming to improve predictive power across sectors.

```
# Stock industry groups
industry_groups =
'AI Healthcare': ["GH", "EXAS", "ILMN", "TDOC", "MDT"],
'Fintech': ["PYPL", "COIN", "AFRM", "SOFI", "UPST"],
'Clean Energy': ["TSLA", "ENPH", "FSLR", "PLUG", "NEE"],
'Cloud and Big Data': ["AMZN", "MSFT", "GOOGL", "SNOW", "CRM"],
'Semiconductor': ["NVDA", "AMD", "INTC", "ASML", "TSM"]
```

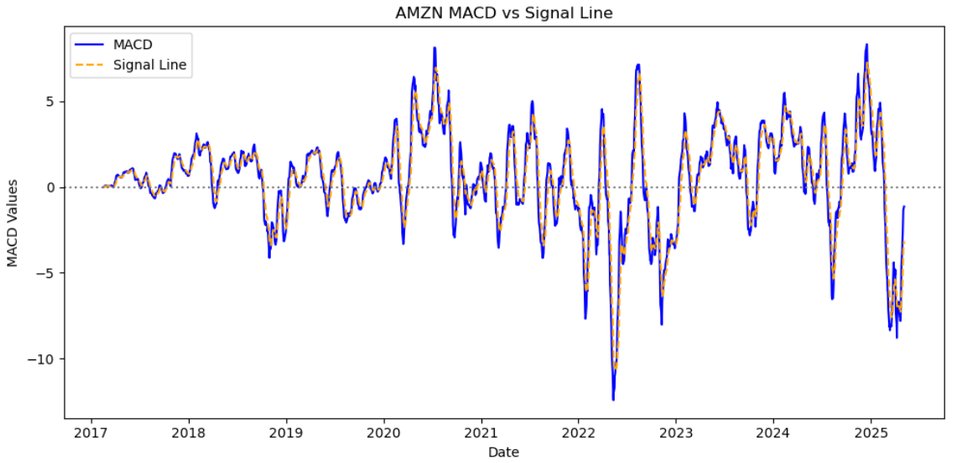
Data preview - Historical Stock Data

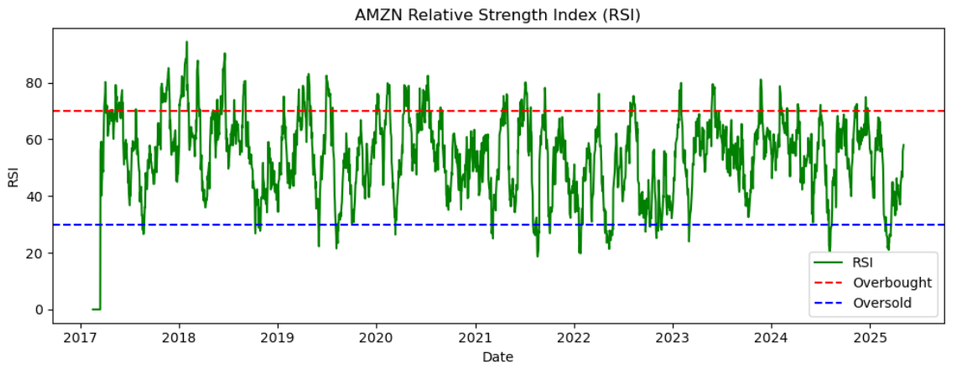




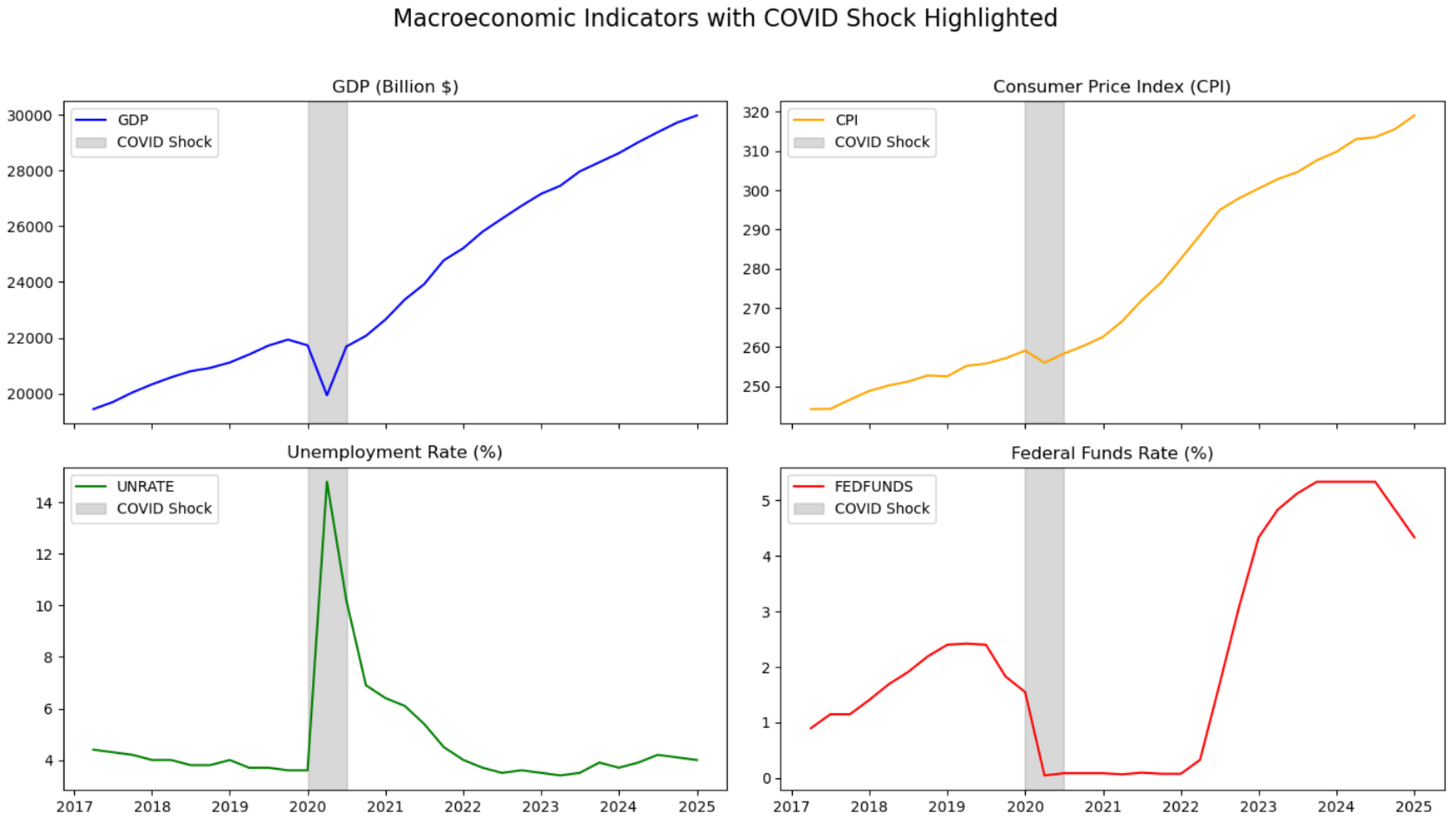


This is a **preview of the factors from historical stock data** used in the analysis, using AMZN (Amazon) stock as an example. These are factors from daily trading information.





Data preview – Macro Factors



These are some key macroeconomic indicators used as input features:

- **GDP (Gross Domestic Product)** reflects overall economic growth.
- **CPI (Consumer Price Index)** tracks inflation levels.
- **Unemployment Rate (UNRATE)** signals labor market conditions.
- **Federal Funds Rate (FEDFUNDS)** represents central bank policy and interest rate environment.

The highlighted **COVID-19 shock period** shows the major macro disruptions that influenced market dynamics, helping explain data patterns and why incorporating macro factors is essential for robust stock modeling.

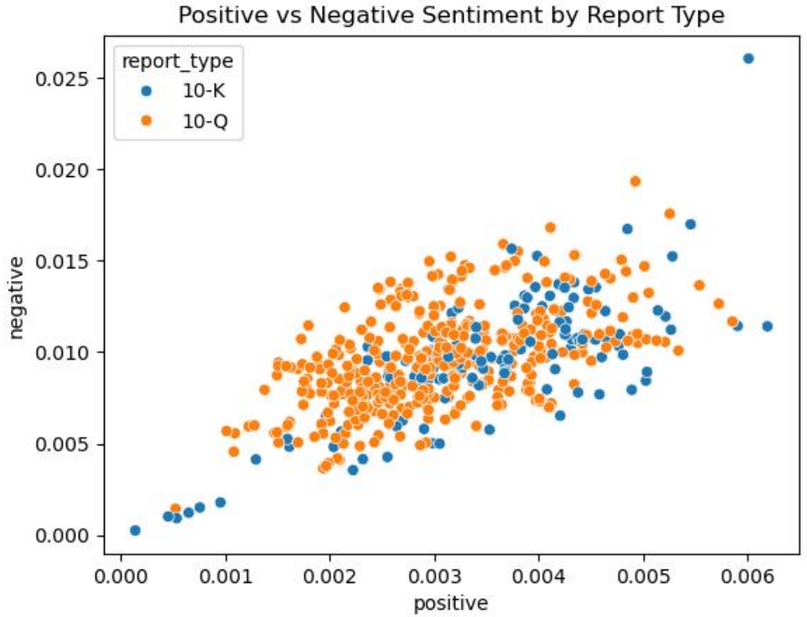
Data preview - Reports

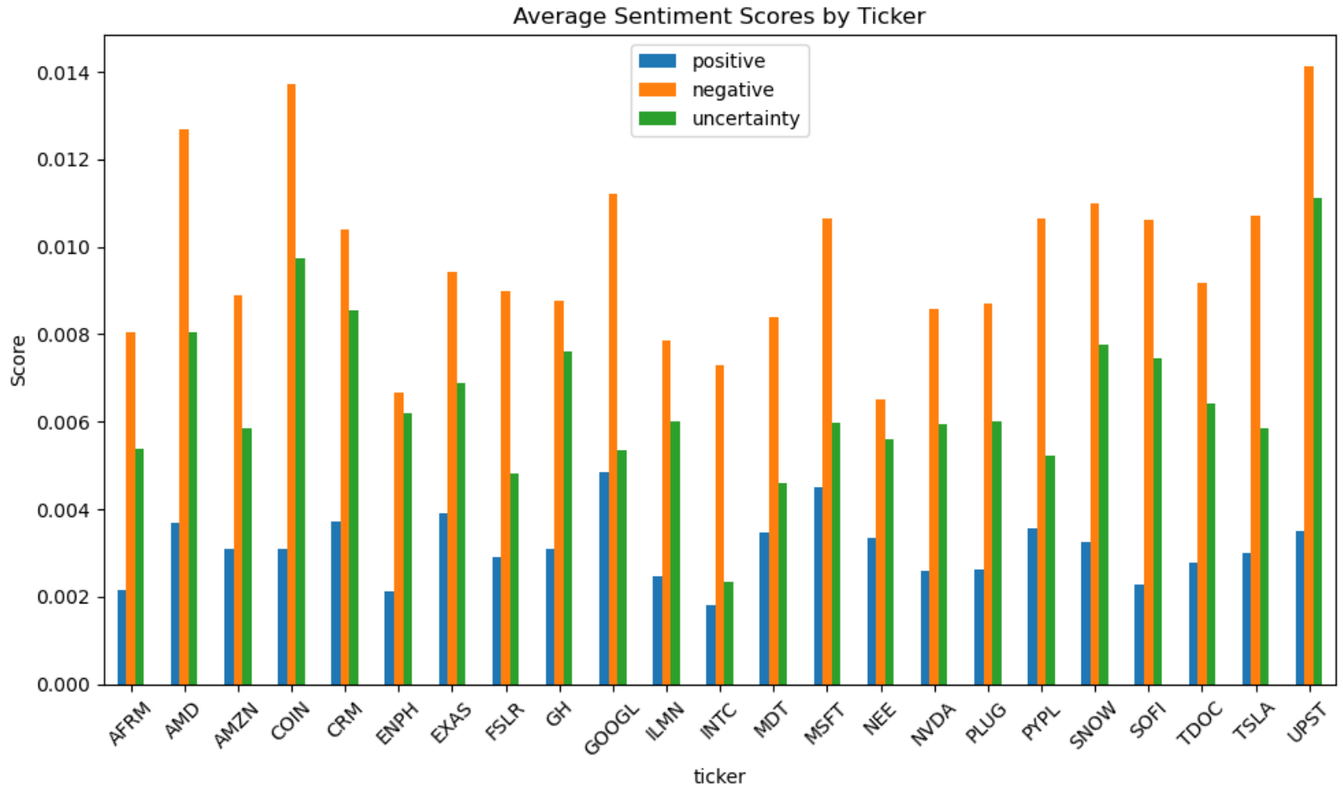
The original data from the 10-K and 10-Q reports provides detailed financial and operational information about each publicly traded company.

The data size of 25 companies is estimated to total 7.5GB, and the cleaned textual data is approximately 190 million characters, which is about 190 megabytes of pure text.

Large, unstructured text datasets (SEC filings), data cleaning and transformation.

- Downloaded SEC EDGAR 10-K and 10-Q filings (2019–2024) for 25 stocks.
- Cleaned text: removed headers, HTML tags, special characters, meaningless sections.
- Applied Loughran-McDonald dictionary using a custom sentiment classifier.
- Extracted sentiment scores: positive, negative, uncertainty.
- Aggregated scores by ticker and report type, saved to CSV.
- Spark pipelines could further optimize the workflow for scale





Step 2: Price Movement Classification

Goal: Build a predictive model to classify stock movements (up / stable / down) one month ahead.

Input Features

Features:

- Technical indicators (RSI, MACD, Bollinger Bands, etc.)
- Macroeconomic factors (GDP, CPI, unemployment rate, FED funds)
- Sentiment scores

Label Definition

- 1 → Up (> threshold)
- 0 → Stable (within threshold)
- -1 → Down (< -threshold)
- Shifted for XGBoost: -1 → 0, 0 → 1, 1 → 2

Methods

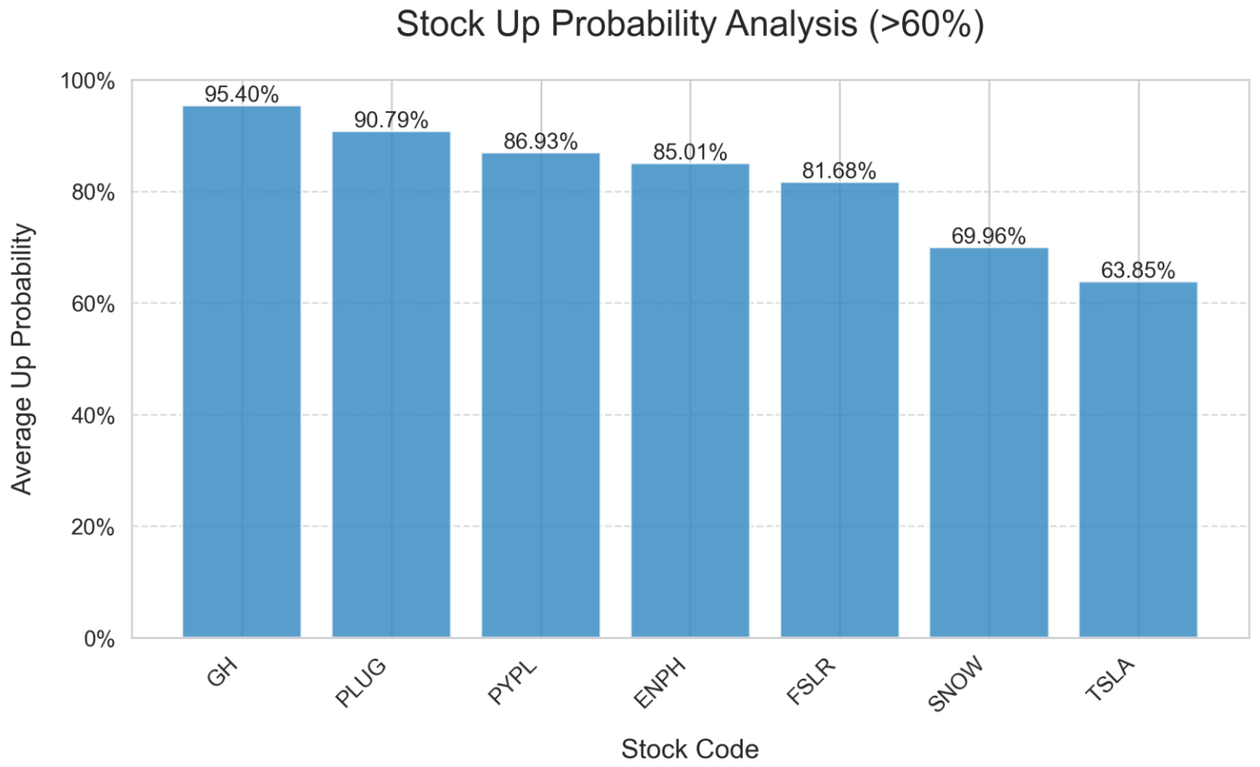
- XGBoost
- Logistic Regression (Pending)

The different factors are linked using stock name and dates.

Model result

Train and test period:

```yaml
train:
  start_time: "2019-01-01"
  end_time: "2024-12-31"
  fit_start_time: "2019-01-01"
  fit_end_time: "2022-12-31"
  valid_start_time: "2023-01-01"
  valid_end_time: "2024-12-31"
  batch_size: 64
  epochs: 100
  early_stop: 10
```

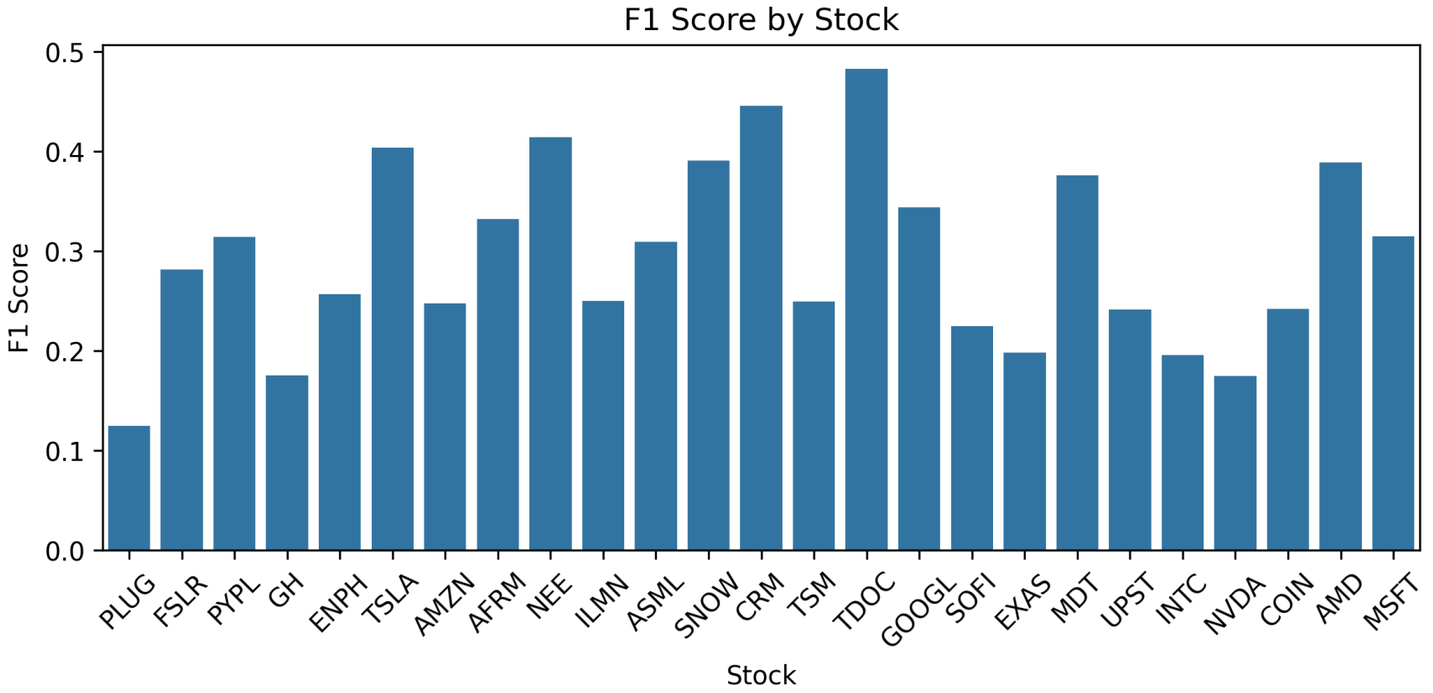


Model evaluation

Top-performing predictions (F1 ≈ 0.4–0.48):
- TDOC, CRM, NEE, TSLA, SNOW
- These models perform well, balancing correct predictions and low false positives/negatives.

Medium-performing predictions (F1 ≈ 0.3–0.4):
- FSLR, PYPL, AFRM, GOOGL, MSFT, AMD, ENPH
- These have reasonable prediction quality but still leave room for improvement.

Low-performing predictions (F1 < 0.2):
- PLUG, GH, EXAS, NVDA
- Models here struggle, possibly due to limited data, low signal in features, or difficulties capturing market trends.



Overall observation:
There is clear variability across stocks; high-scoring stocks could be prioritized for investment or deeper analysis, while low-scoring stocks may need data quality checks, feature reengineering, or model adjustments.
Improving multi-class predictions (rise/fall/stable) remains a key challenge and opportunity for further development.

Step 3: Selecting Positive Stocks

Threshold Setting

Select stocks with probability > 0.6

Refined Pool

Final selection of bullish stocks for optimization

Selected Investment Candidates

(Stocks with **average predicted up probability > 60%**)

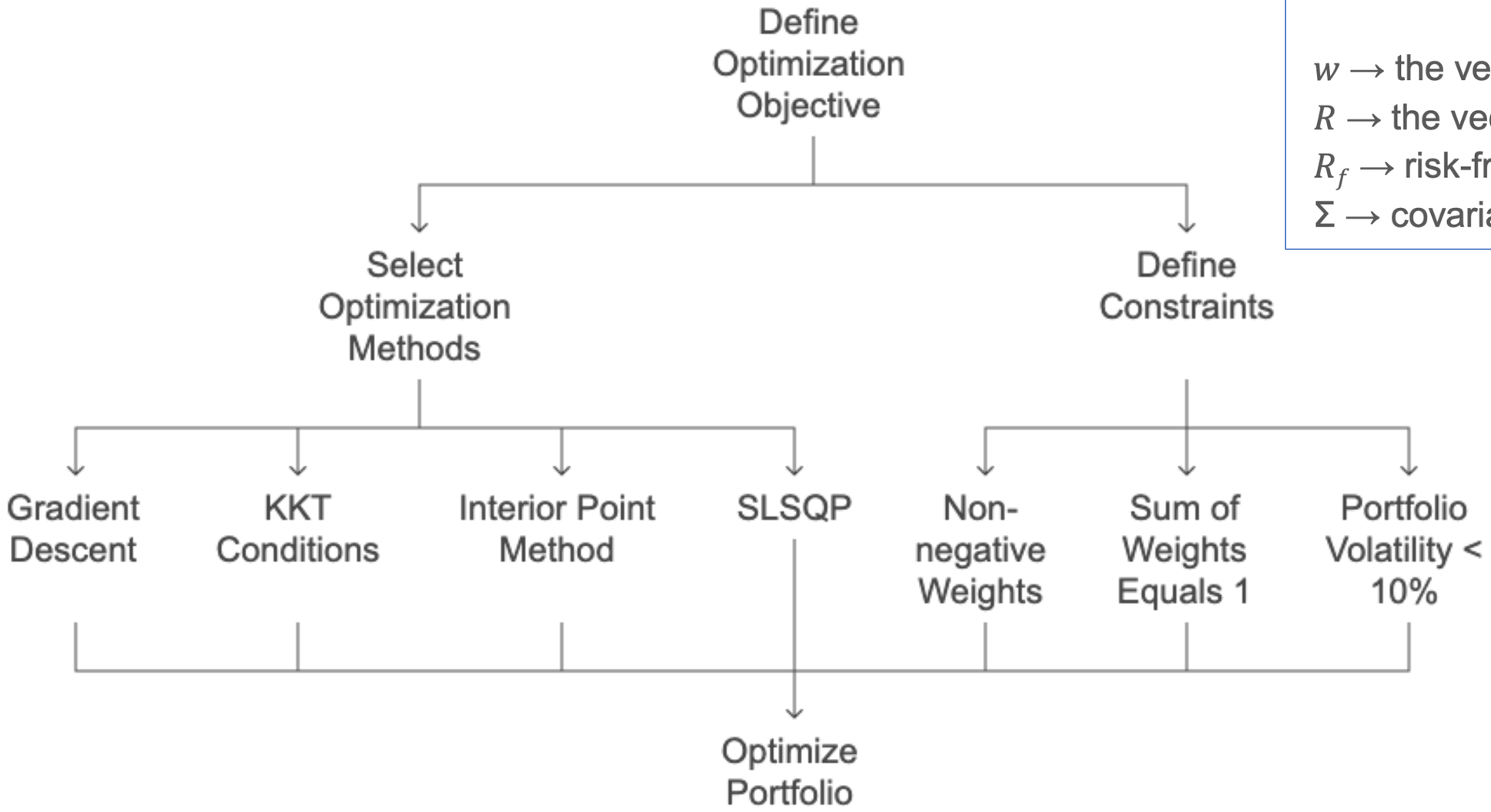
- ENPH (Enphase Energy)
- PYPL (PayPal)
- GH (Guardant Health)
- PLUG (Plug Power)
- FSLR (First Solar)
- GH (Guardant Health)

Step 4: Portfolio Optimization

Objective Function:

$$\max_{w} \frac{w^\top (R - R_f)}{\sqrt{w^\top \Sigma w}}$$

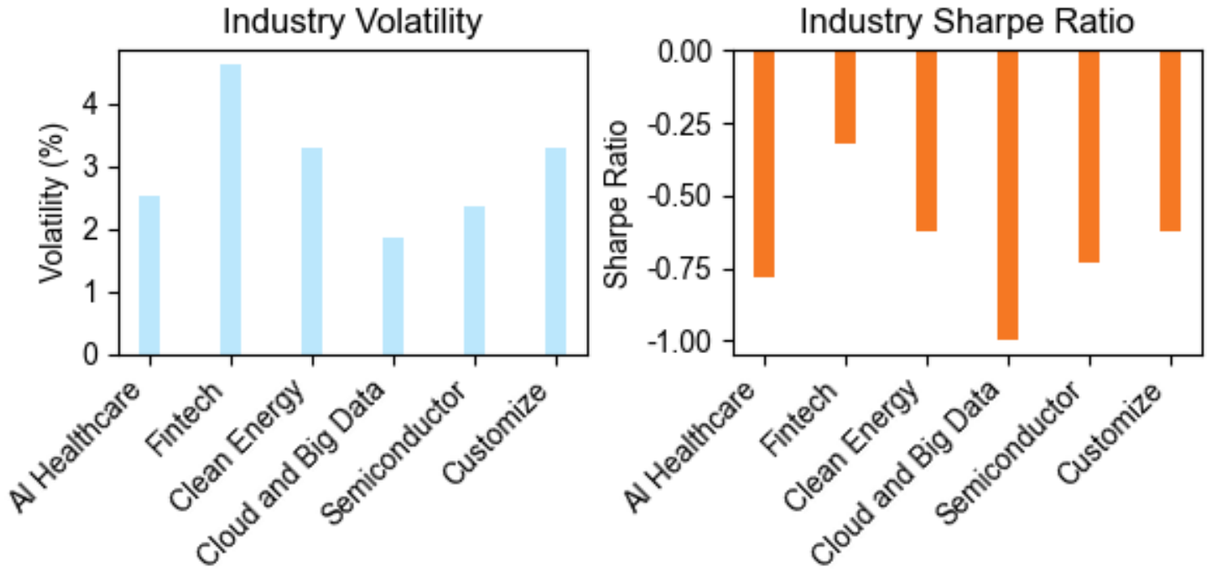
$w \rightarrow$ the vector of portfolio weights
$R \rightarrow$ the vector of expected returns
$R_f \rightarrow$ risk-free rate (scalar)
$\Sigma \rightarrow$ covariance matrix of returns



Step 5: Portfolio Optimization Results

Optimizer Results Comparison – Comparing Across Sectors

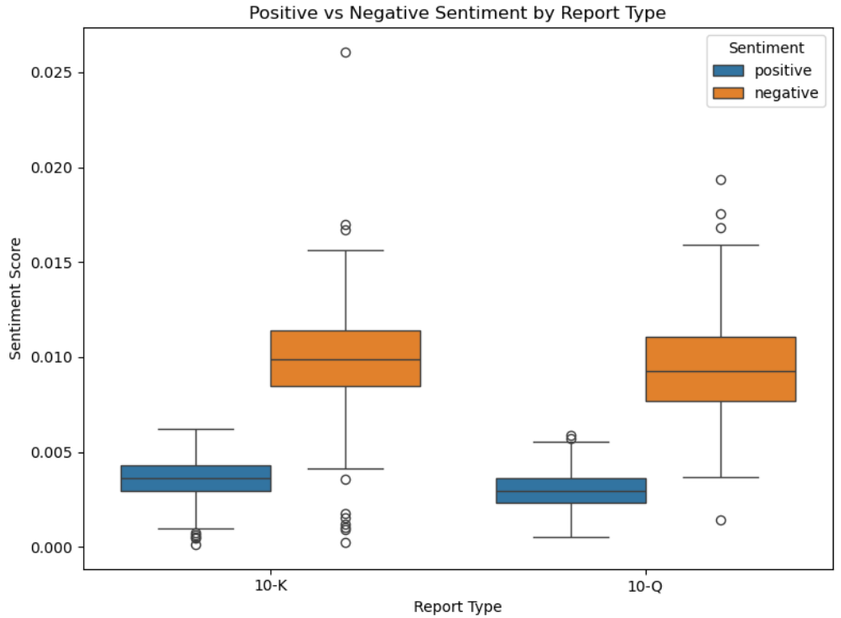
- Unfortunately, the current evaluation results are not ideal and do not indicate that the stock selection approach is successful. It fails to outperform single-sector combinations. I am currently investigating the underlying issues.
- All industries show negative Sharpe ratios, suggesting underperformance relative to the risk-free rate or poor reward for risk taken.
- One issue I realized is that the criteria for selecting stocks need to be adjusted — not only focusing on whether they are in an uptrend, but also considering the magnitude of their returns.
- Perhaps the quantity I selected is still far from enough, insufficient to break through the advantage comparison between industries.



Technical Highlights

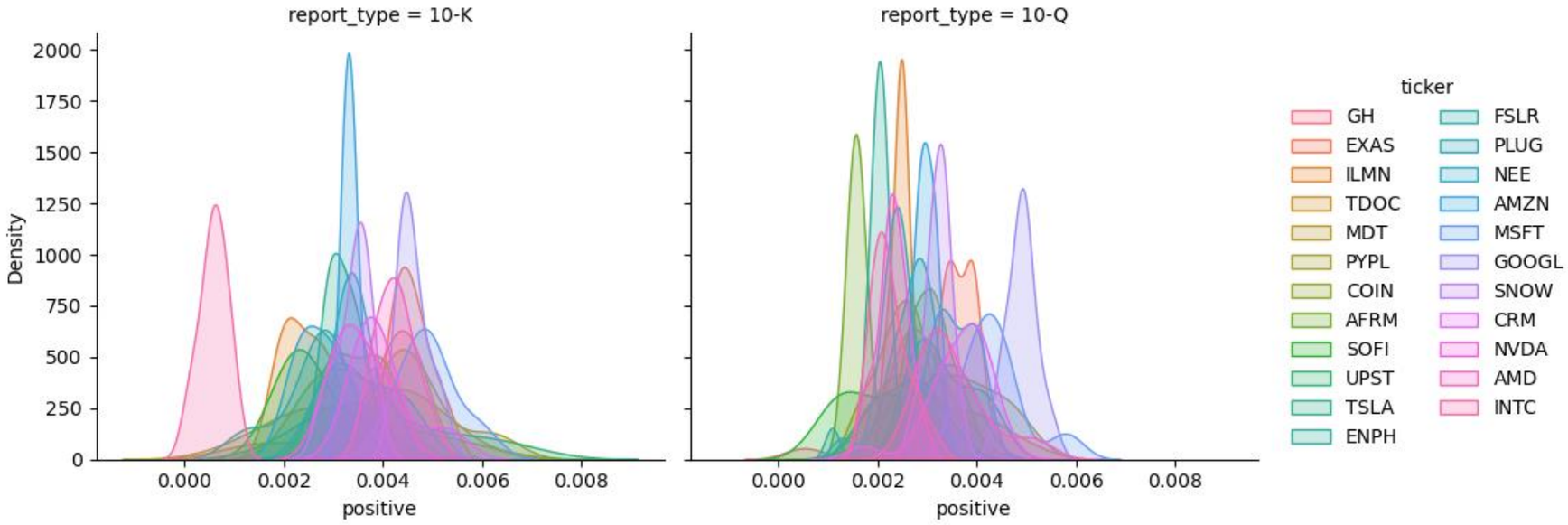
Stock Factors Processing

Annual report cleaning and sentiment feature extraction



Model Integration

Combined classification and and optimization pipeline



Conclusions

Achievements

- Went through the whole process

Future Improvements

- Checking potential problems
- Incorporate news data, like news for each stocks and Tariffs, inflation, and other macroeconomic factors.
- Implement streaming capabilities

References

Yang, Xiao, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. "Qlib: An ai-oriented quantitative investment platform." *arXiv preprint arXiv:2009.11189* (2020).

Thank You !

Questions?