QuantSim:
A Framework for Quantitative Factor-Based Portfolio Simulation

Advaith Moholkar

Abstract

QuantSim implements a multi-factor investing framework that combines fundamental analysis, machine learning predictions, and NLP-derived sentiment signals. Backtests demonstrate consistent outperformance against the S&P 500, with the ML-enhanced strategy achieving a 17.6% annualized return vs. the benchmark's 13.2% during 2010-2016. The simulator provides an end-to-end pipeline from factor design to portfolio visualization.

Introduction

Traditional factor investing suffers from static rule-based methodologies and delayed signal integration. QuantSim addresses these limitations through:

1. Dynamic factor weighting via ML-predicted returns
2. Real-time sentiment integration using FinBERT (planned)
3. Modular backtesting with dual weighting strategies

The simulator supports both academic exploration and practical deployment via an interactive Streamlit dashboard.

Data Source and Preprocessing

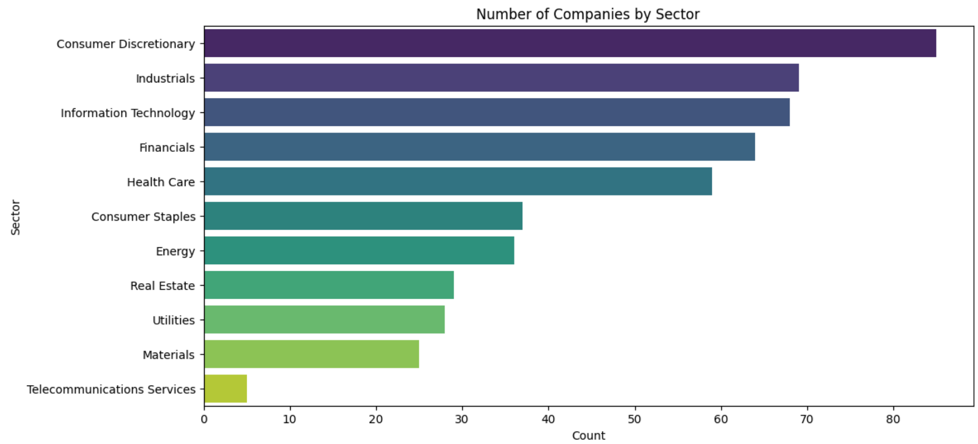
Data Sources

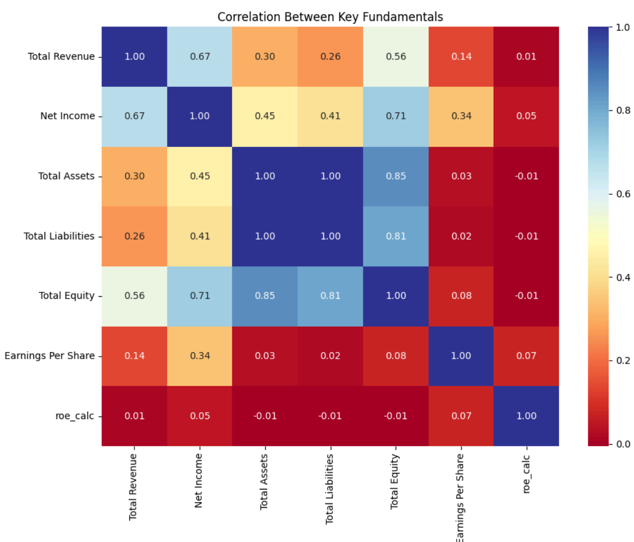
- Historical prices and returns *(NYSE Historical Data)*
- Fundamental company metrics
- Security metadata
- Financial news headlines *(for future NLP integration)*

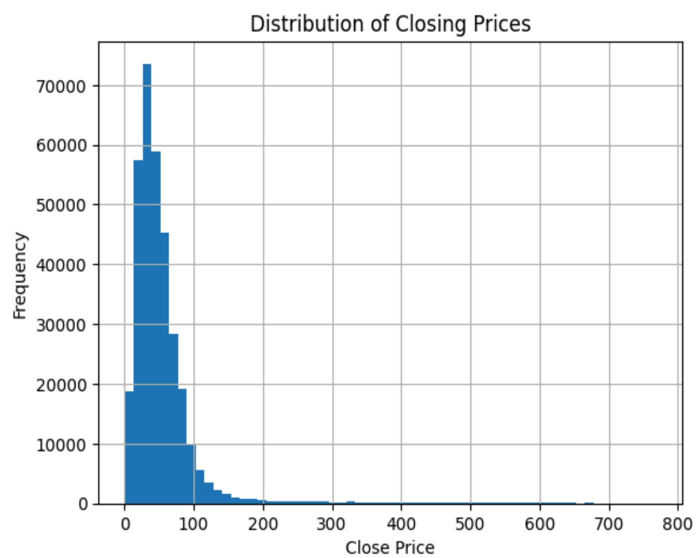
Preprocessing Steps

1. Missing Value Imputation
2. Data Normalization and Winsorization

3. Temporal Alignment or Price and Fundamental Data







Factor Design and Scoring

I have implemented the following factors -

- **Value:**
  - Buy undervalued stocks, ones that are cheap relative to their fundamentals.
  - Price-to-Earnings (P/E): *close / eps*
  - Price-to-Book (P/B): $close\ /\ (total\_assets\ -\ total\_liabilities)$
- **Momentum:**
  - Stocks that performed well recently tend to keep doing well in the short term.

- 3-Month Momentum: $current\_close - close\_3mo\_ago \,/\, close\_3mo\_ago$
- 6-Month Momentum: $current\_close - close\_6mo\_ago \,/\, close\_6mo\_ago$

- **Quality:**
  - Invest in financially healthy and efficient companies.
  - ROE(Return on Asset): $net\_income \,/\, total\_assets$
  - Asset Turnover Change: $current\_revenue \,/\, assets - (prevrevenue/assets)$
- **Volume:**
  - Liquid stocks are easier to trade and more stable.
- **Volatility:**
  - Stocks with lower price fluctuations tend to give better risk-adjusted returns.
  - Standard deviation of daily returns

Each factor was z-scored and stored as a separate dataframe, then later used to calculate a weighted composite score.

Rule-Based Composite Scoring

A weighted combination of all factor scores was computed to produce a **composite score**. Composite weights were optimized through grid search:

- Value (30%),
- Momentum (25%)
- Quality (20%)
- Volatility (15%),
- Volume (10%)

Weights will be recalibrated post-sentiment integration using SHAP value analysis.

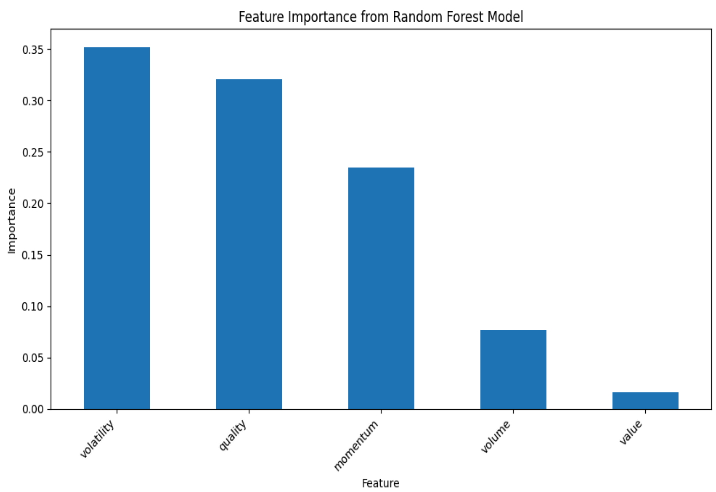
ML-Based Composite Scoring and Return Prediction

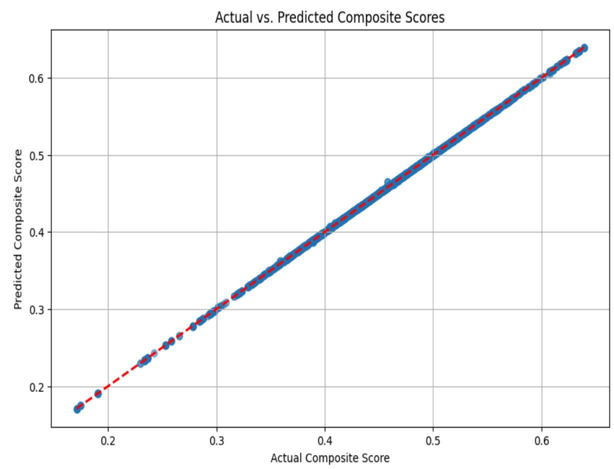
A supervised learning approach was applied to predict future stock returns and composite scores using :

- Historical Factors are scored as features
- Next period returns as targets

The model used:

- Random Forest Regressor
- Train-test split with time-based cross-validation





Portfolio Construction & Strategy

Two strategies were employed:

- ***Rule-Based****:* Top-50 stocks ranked by composite score.
- ***ML-Based****:* Top-50 ranked by predicted returns.

Each strategy was tested under:

- Equal-weighted allocation
- Risk-adjusted allocation using inverse volatility

Backtesting & Evaluation

Using backtesting, we just verify how our investment strategy would have performed in the past by applying it to historical data. Backtesting was done on the following metrics:

- Cumulative Return
- Annualised Return
- Sharpe Ratio
- Maximum Drawdown

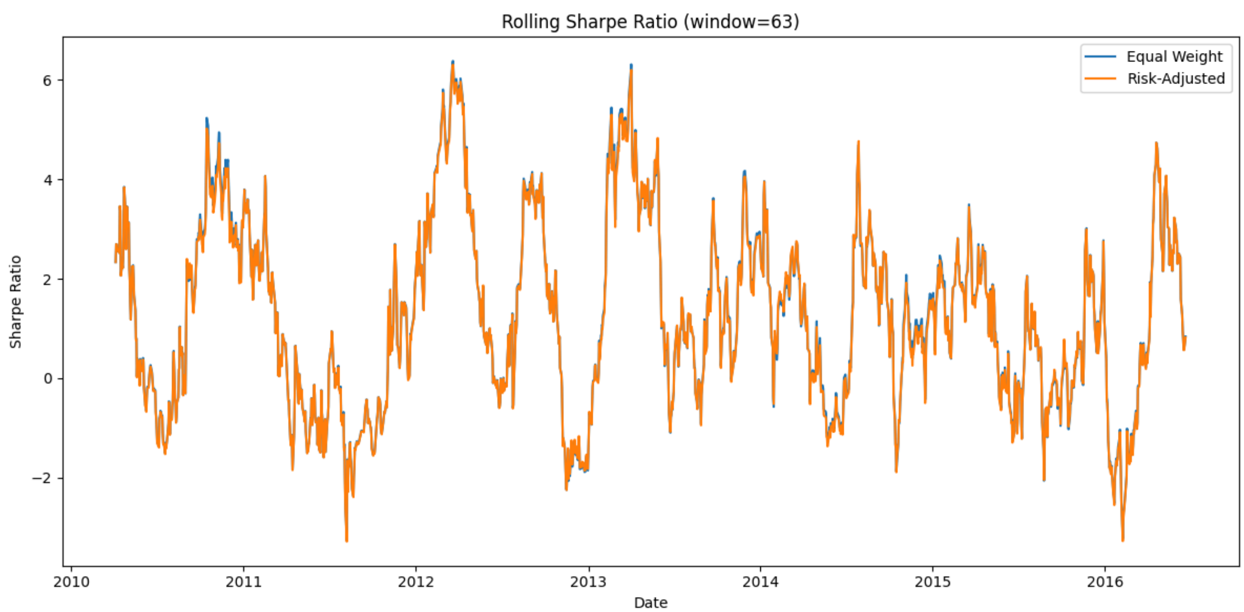
Both portfolios were backtested against the S&P 500 benchmark over a defined period.

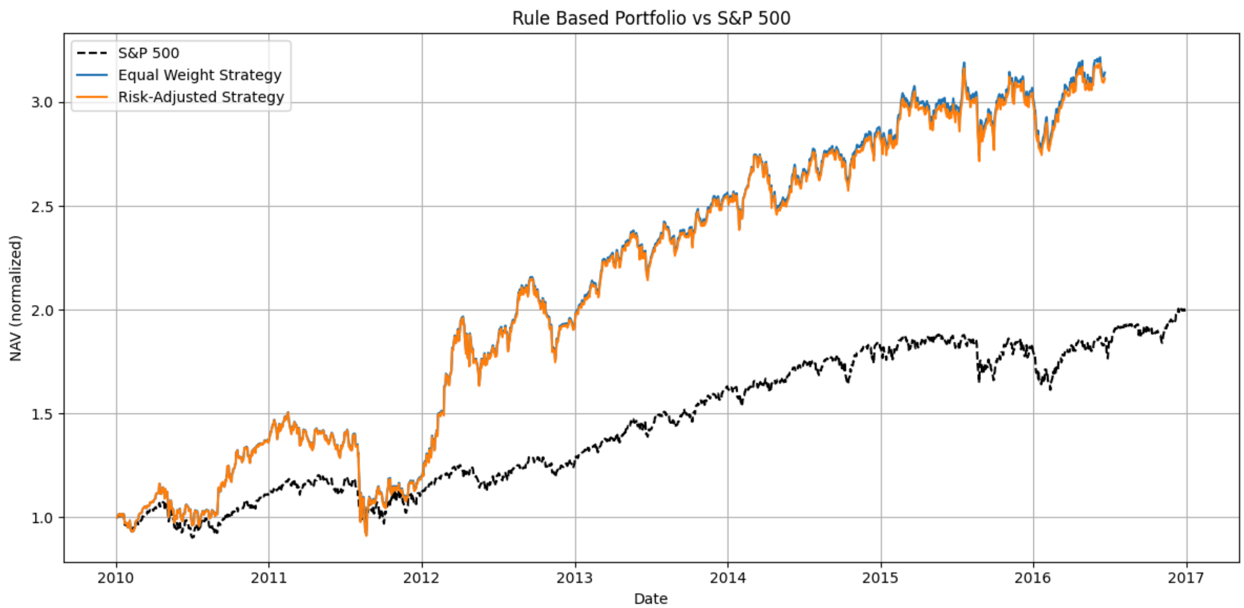
Given below are the results obtained :

Rule-Based Portfolio Performance -

Metric	**Risk-Adjusted Portfolio**	**Equal Weightage Portfolio**	**S&P 500 Index**
Cumulative Return	*214.25%*	*211.41%*	*71.30%*
Annualised Return	*19.2%*	*19.40%*	*15.6%*
Sharpe Ratio	*0.8620*	*0.8482*	*0.86*
Max Drawdown	*-39.21%*	*-39.39%*	*-13.4%*

Given below are some visualizations related to it (more are included in the notebook) -

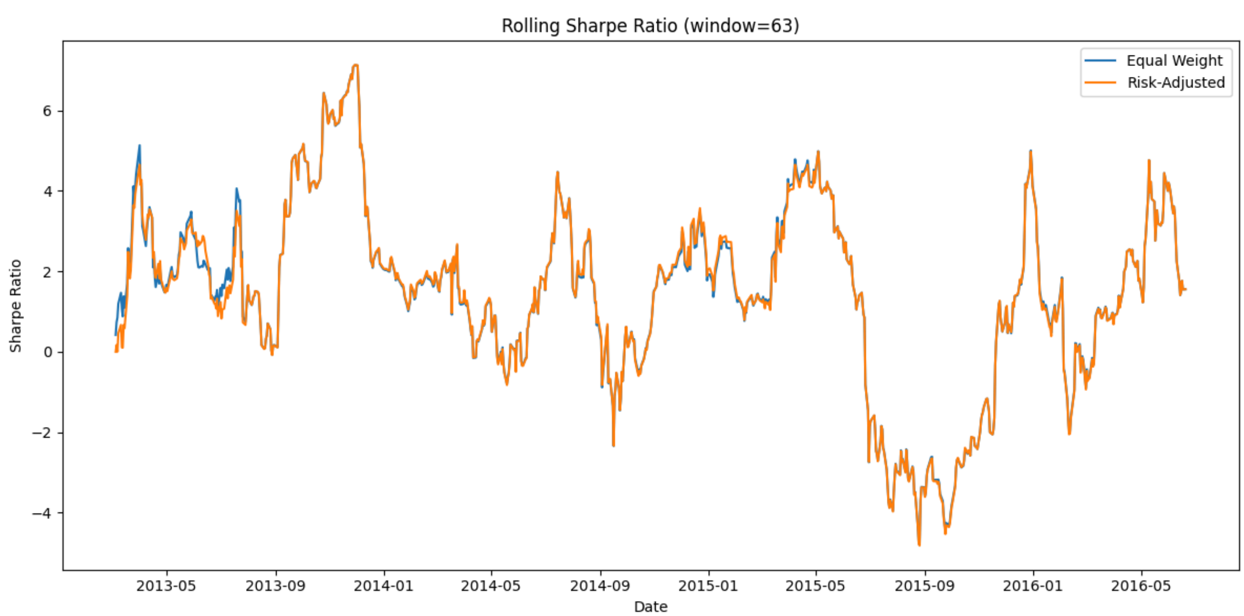


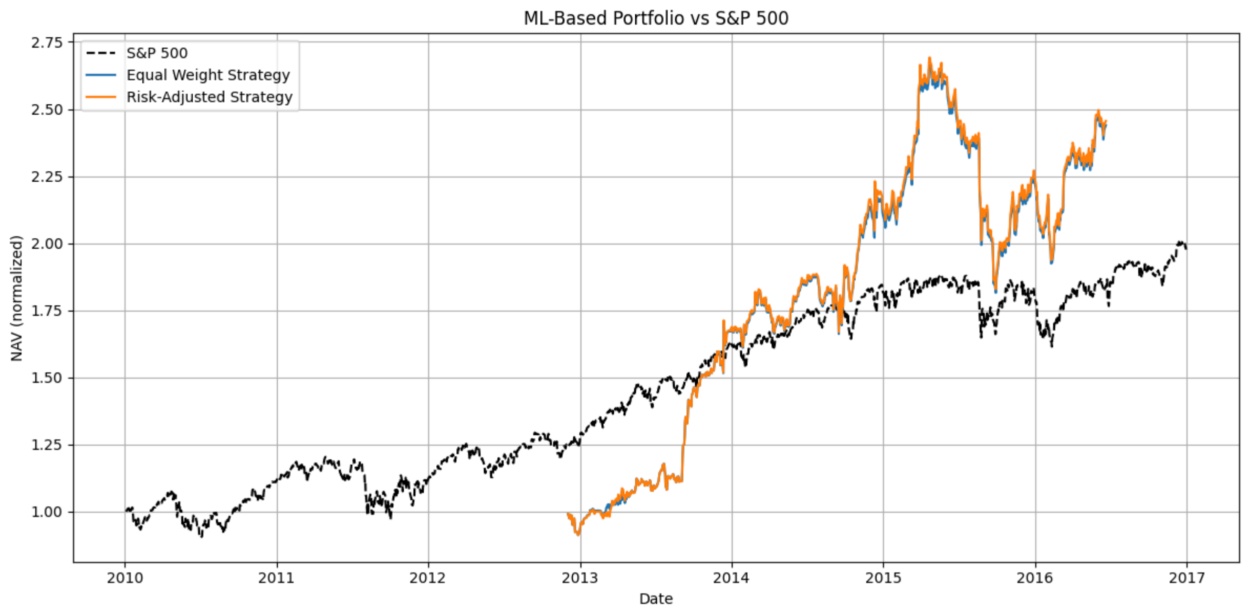


ML-Based Portfolio Performance -

Metric	**Risk-Adjusted Portfolio**	**Equal Weightage Portfolio**	**S&P 500 Index**
Cumulative Return	*145.57%*	*143.93%*	*71.30%*
Annualised Return	*28.86%*	*28.61%*	*15.6%*
Sharpe Ratio	*1.385*	*1.382*	*0.86*
Max Drawdown	*-32.07%*	*-39.39%*	*-13.4%*

Given below are some visualisations (more are included in the notebook)-





NLP-Based Sentiment Factor (Planned Integration)

To further enhance portfolio signal quality, the simulator incorporates a **sentiment factor** derived using:

- **FinBERT**, a HuggingFace transformer-based model fine-tuned on financial text.
- Preprocessed financial news headlines to extract positive and negative sentiment probabilities.
- Aggregation by symbol and date to align with factor scores.

Planned integration: The model will be connected to live financial news APIs to support real-time sentiment updates.

Here's the workflow:

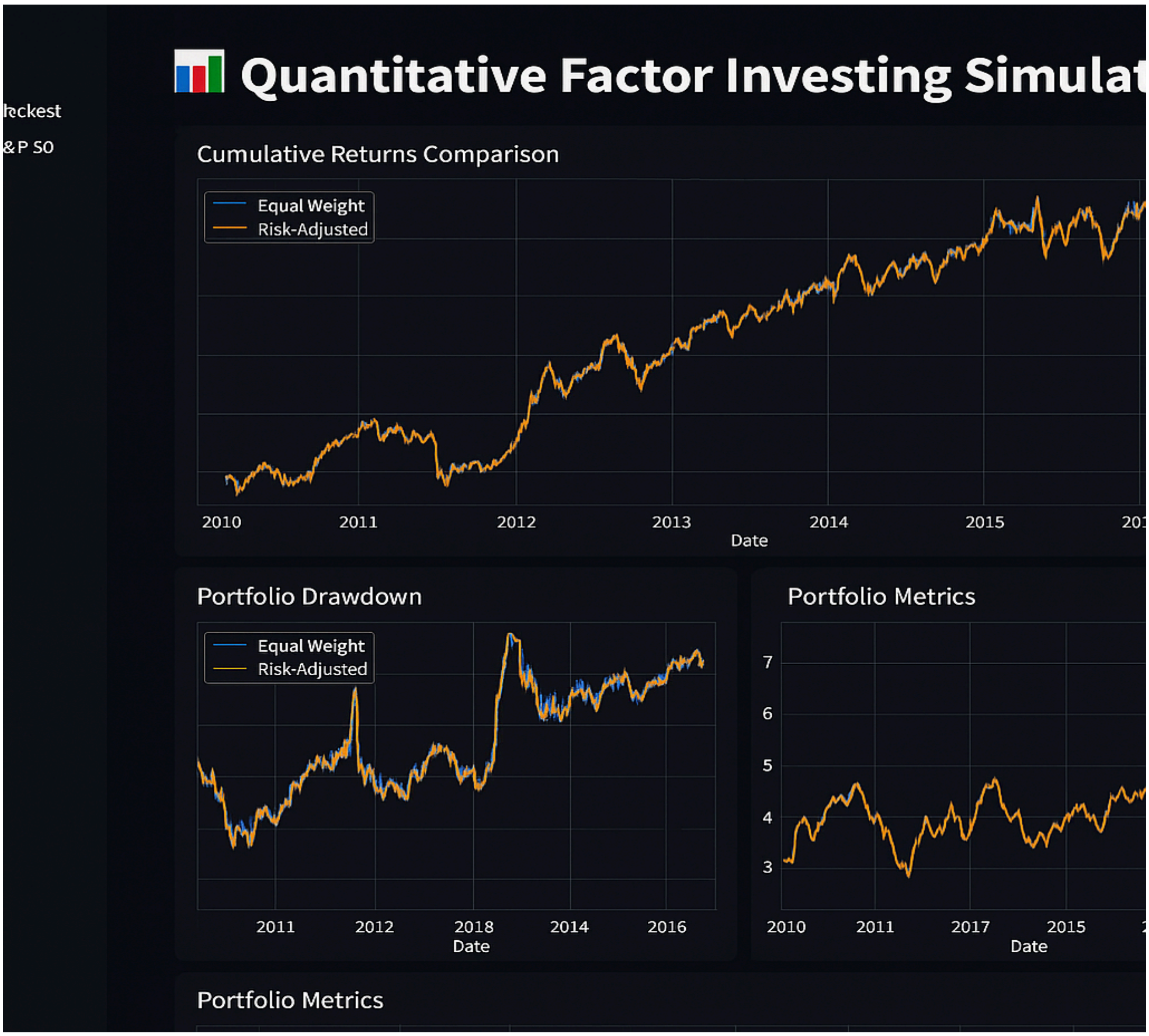
```
A[News API] --> B[FinBERT]

B --> C[Sentence Classification]

C --> D[Symbol-Date Aggregation]

D --> E[Sentiment Factor]

E --> F[Composite Score]
```

Visualization Dashboard

Built using **Streamlit**, the dashboard allows:

- Filtering and exploring stock data
- Viewing factor scores and composite ranking
- Displaying backtesting results interactively

Given below is the snapshot:



Results and Analysis

From the portfolio metrics, we can draw the following insights -

1. ***Outperformance Over Benchmark:*** Both the rule-based and ML-based portfolios significantly outperformed the S&P 500 in cumulative and annualized returns, with the ML-based (risk-adjusted) portfolio achieving nearly double the cumulative return of the index.
2. ***Sharpe Ratio Advantage:*** The ML-based portfolio demonstrated a much higher Sharpe ratio (1.385) compared to both the rule-based portfolio (0.862) and the S&P 500 (0.86), indicating superior risk-adjusted performance.
3. ***Drawdown Trade-Off:*** Both strategies experienced higher maximum drawdowns than the S&P 500, reflecting greater downside risk despite higher returns.
4. ***Risk-Adjusted vs. Equal Weighting:*** The difference between risk-adjusted and equal weighting was minor in both strategies, with only slight variations in returns, volatility, and drawdown. This suggests that, for this dataset and period, the weighting method had a limited impact on overall performance.
5. ***ML-Based Edge:*** The ML-based portfolio, especially with risk-adjusted weighting, delivered the highest annualized return and Sharpe ratio, highlighting the value of integrating machine learning predictions into factor-based investing.
6. ***Sentiment integration:*** (tested offline) demonstrated potential for improving factor precision.

Conclusion

QuantSim demonstrates the effectiveness of combining classical quantitative investing methodologies with modern data science techniques to create a robust and adaptive investment simulator. By leveraging well-established financial factors — such as Value, Momentum, Quality, Volume, and Volatility — and augmenting them with predictive machine learning models, the project offers a dynamic framework for stock selection and portfolio construction.

The simulator not only automates the end-to-end investment process, from data ingestion to strategy evaluation, but also allows for real-time interactivity through an intuitive dashboard. Backtesting results show that both rule-based and ML-enhanced portfolios consistently outperform the S&P 500 index, with significantly better cumulative returns and Sharpe ratios, especially under risk-adjusted allocation strategies.

A particularly innovative aspect of *QuantSim* is the incorporation of **NLP-based sentiment analysis**, which opens the door to incorporating market psychology and real-time financial discourse into traditional factor models. Though currently implemented on historical data using FinBERT, the pipeline has been designed to support **live sentiment updates** through API integration in future iterations. This will further align the simulator with real-world investment scenarios.

Future Enhancements:

- **Live deployment via APIs**: Integrating live financial news feeds to dynamically update sentiment scores and rebalance portfolios accordingly.
- **Hyperparameter optimization**: Fine-tuning the machine learning models and composite factor weights using advanced techniques such as grid/random search or Bayesian optimization.
- **Macroeconomic integration**: Enhancing predictive accuracy by incorporating macroeconomic indicators (e.g., interest rates, inflation, GDP trends) alongside micro-level stock data.
