Smarter Investment using Big Data, Data Science and Algorithmic Trading

GROUP
fyp24033

Chan, Chun Hei

Supervisor: Prof. Yiu Siu Ming

Introduction

Background

Algorithmic trading automates order execution using pre-programmed rules, a practice that began in the 1970s with the shift to electronic trading systems. By 2019, it accounted for approximately 92% of all equity trading volume (Kissell, 2020). The global algorithmic trading market was valued at USD 3.1 billion in 2023 and is projected to grow at a compound annual growth rate (CAGR) exceeding 13% from 2024 to 2032 (Global Market Insights, 2024).

Problem Statement

While algorithmic trading has revolutionized financial markets, several key challenges remain. First, the inherent complexity of financial markets makes it difficult for trading algorithms to generate consistent profits. Second, many models suffer from overfitting, performing well on historical data but poorly when applied to new market conditions. Third, most existing algorithms rely solely on structured numerical data, failing to utilize the wealth of publicly available unstructured data from news articles, social media posts, and other alternative sources.

Hard to be consistent

Risk of overfitting

New forms of data

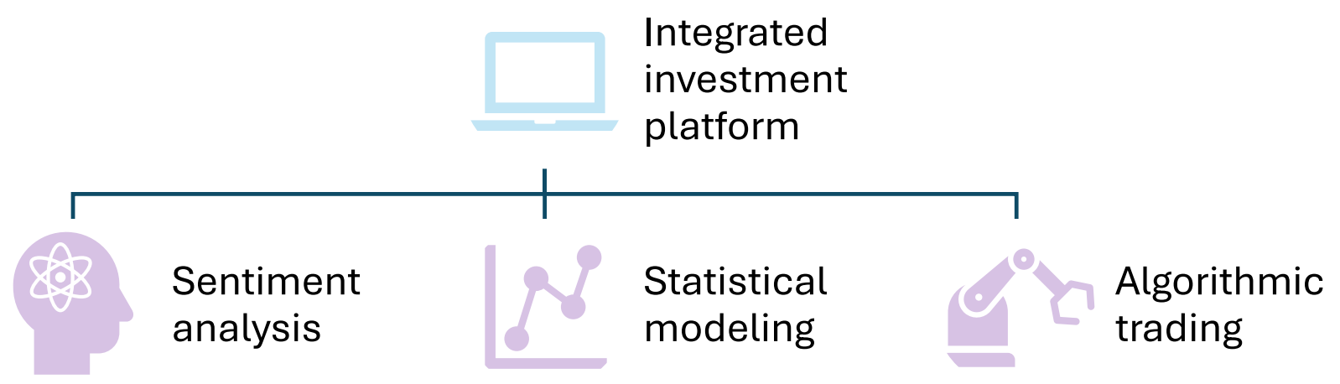
Objectives

This research aims to evaluate and compare algorithmic trading strategies using numerical data across diverse market conditions, while investigating the underlying factors driving their performance. A key focus involves enhancing these strategies through the systematic integration of textual data from news and social media sources, with rigorous assessment of their feasibility and added value. The study will develop interactive dashboards to effectively visualize trading performance and analytical insights. Ultimately, the project seeks to create an integrated AI-driven investment platform that combines statistical modeling, sentiment analysis, and algorithmic trading to provide data-driven investment recommendations while minimizing behavioral biases.

Evaluation	• Compare strategies performance
Enhancement	• Propose improvements on existing strategies
Exploration	• Explore textual data and web scraping
Visualisation	• Visualise results and integrate functions

Methodology

The project is divided into several key components. Together, they combine into an AI-driven investment platform.



Integrated investment platform

Sentiment analysis

Statistical modeling

Algorithmic trading

Data Collection

This study combines quantitative and qualitative data analysis for robust trading insights:

- Numerical data: Historical price/volume from Yahoo Finance API (high-frequency, reliable)
- Textual data: News/social media via web scraping (BeautifulSoup/Selenium) + native APIs

Python libraries such as Pandas and NumPy were employed for data cleaning, preprocessing, and aggregation because of their powerful data manipulation capabilities and efficiency in handling large datasets.

Key Advantages

✓ Multi-source approach ensures data completeness
✓ API + scraping combo maximizes coverage
✓ Industry-standard tools guarantee reproducibility

Algorithmic Trading Models

Baseline Model: Traditional moving average crossover strategy

Advanced Approaches:

- Moving averages confidence interval
- Relative strength index local maximum and minimum

Baseline: Traditional moving average crossover

```
For each trading day
    if price > moving average when crossover then
        Buy()
    else
        Sell()
ENDFor
```

Model 1: Moving averages confidence interval

```
For each trading day
    if price > moving average - n×σ when crossover then
        Buy()
    else if price < moving average + n×σ when crossover then
        Sell()
ENDFor
```

Model 2: Relative strength index local maximum and minimum

```
For each trading day
    if rsi < 30 and previous_rsi < rsi then
        Buy()
    else if rsi > 70 and previous_rsi > rsi then
        Sell()
ENDFor
```

Back Testing Framework:

- Platform: QuantConnect (open-source with built-in historical data)
- Key Metrics (Cuthbertson et al., 2010) (Sukma et al., 2024):
  - Performance: Annualized Return (ARR), Win Rate
  - Risk: Sharpe Ratio, Max Drawdown
  - Efficiency: Profit Factor, Alpha

Why This Matters

Our multi-strategy evaluation provides:

✓ Structured progression in evaluating the effectiveness of different methodologies
✓ Objective comparison across classical and AI-driven methods
✓ Quantifiable risk/reward assessment via financial metrics
✓ Reproducible results using open-source tools

Trend Analysis

Trend definition: upward (+1), down trend (-1) and no trend (0).

Parallel Ensemble Model

Multiple technical indicators:

- Simple & exponential moving averages
- MACD (Moving average convergence divergence)
- RSI (Relative strength index)
- Rolling slope

Consensus vote by summation:

- Total votes >=3: Upward trend (Strong if total votes >=4)
- Total votes <=-3: Downward trend (Strong if total votes <=-4)
- Otherwise: No trend

LSTM Neural Network

Human-labeled training

- Tailored to user-defined trend thresholds

Enhanced feature set:

- Bolinger Bands
- Volume change
- Volume simple moving average
- Momentum

Model architecture:

- LSTM(128, return_sequences=True, input_shape=input_shape),
- Dropout(0.4),
- LSTM(64, return_sequences=False),
- Dropout(0.3),
- Dense(32, activation='relu'),
- Dense(num_classes, activation='softmax’)

Key Advantages

✓ Hybrid Approach: Combines rule-based and AI-driven classification
✓ Adaptability: Customizable trend thresholds (user-defined labels)
✓ Robustness: Dropout layers prevent overfitting

News Analysis

Article Processing

- LLM summarization: GPT models extract key insights from news/articles
- Sentiment scoring: Fine-tuned models classify sentiment (Positive/Negative/Neutral)

Why It Matters

✓ Augments numeric data with qualitative insights
✓ Captures market-moving events pre-price reaction
✓ Adaptable to multiple languages/sources

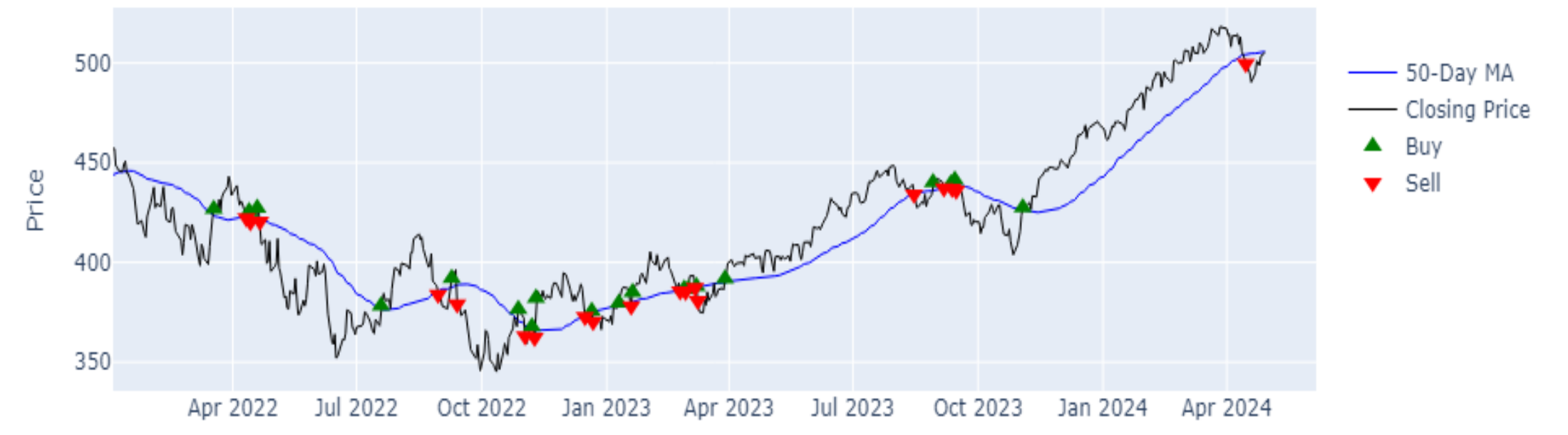
Results

This section outlines the results of the baseline model and two advanced model inspired by the two major problems identified from the baseline. It also shows the results of the trend detection component of this project.

Baseline – Simple Moving Average Crossover

Overview:

- Ticker: SPY (S&P 500)
- Period: Jan 2022 – Apr 2024



Key Insights:

- Trend-dependent performance:
  - Profitable in bullish markets (+18.56% return)
  - Loses in sideways markets (-9.42%) due to false signals
- Late entries/exits during trends (lagging MA)

Two Major Problems:

- Unable to profit because of frequent unwanted signals during sideways
  - Frequent signals when price moves around moving average
  - Cannot capture peaks and troughs
- Unable to effectively capture profit due to lagging property during bullish trend

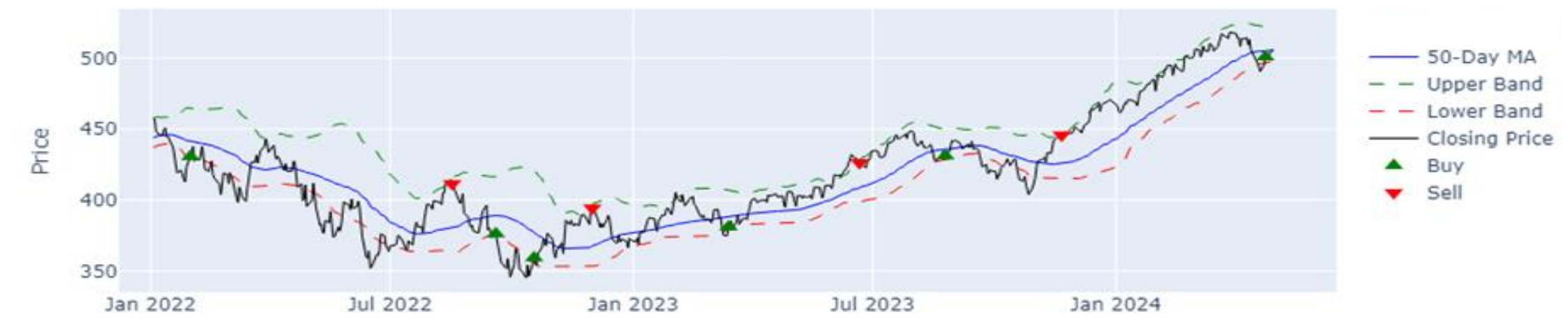
Baseline – Simple moving average crossover

Evaluation Metrics	U.S. Stock	Jan‘22 – Feb‘23	Mar’23 – Apr’24
Annualized Rate of Return	3.566%	-9.424%	18.557%
Sharpe Ratio	-0.1	-0.815	0.927
Win Rate	26%	27%	29%
Average Win	6.23%	1.51%	13.33%
Average Loss	-1.50%	-1.94%	-1.00%
Profit-Loss Ratio	4.16	0.78	13.33
Maximum Drawdown	17.500%	15.900%	6.100%
Alpha	-0.01	-0.054	0

Model 1 – Moving average confidence interval

Key Insights:

- Less unwanted trades
- Problem: Still cannot profit from sideways
- Annualised Rate of Return: 6.3%
  - Sideways (-0.3%): Reduced lost and outperforms market (α=0.065)
  - Bullish (13.4%): Less profit



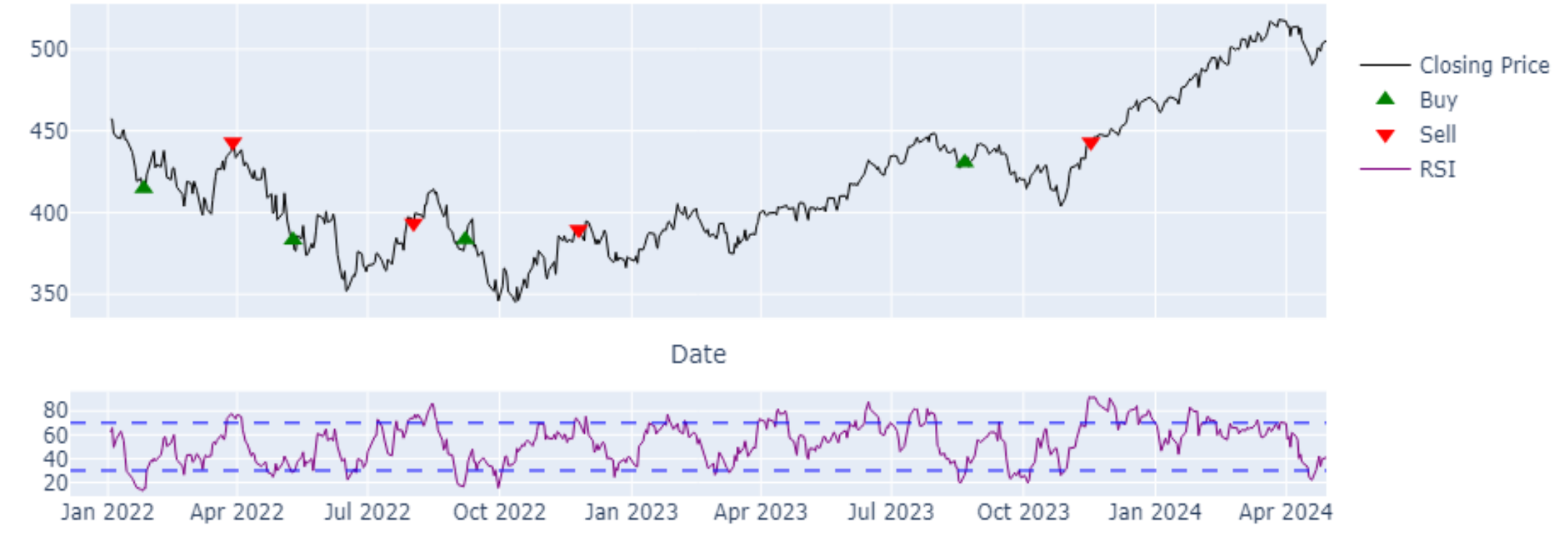
Model 1 – Moving average confidence interval

Evaluation Metrics	U.S. Stock	Jan‘22 – Feb‘23	Mar’23 – Apr’24
Annualized Rate of Return	6.333%	-0.294%	13.385%
Sharpe Ratio	0.108	-0.061	0.543
Win Rate	60%	50%	100%
Average Win	6.95%	4.44%	8.22%
Average Loss	-2.30%	-4.58%	0%
Profit-Loss Ratio	3.03	0.97	0
Maximum Drawdown	20.500%	20.500%	8.600%
Alpha	0.013	0.065	-0.021

Model 2 – Relative strength index local maximum and minimum

Key Insights:

- Successfully profit during sideways
- Less unwanted signals
- Sacrificed profit from bullish trend
- Annualised Rate of Return: 5.1%
  - Sideways (3.6%): Profit and outperforms market (α=0.089)
  - Bullish (2.1%): Sacrificed more profit



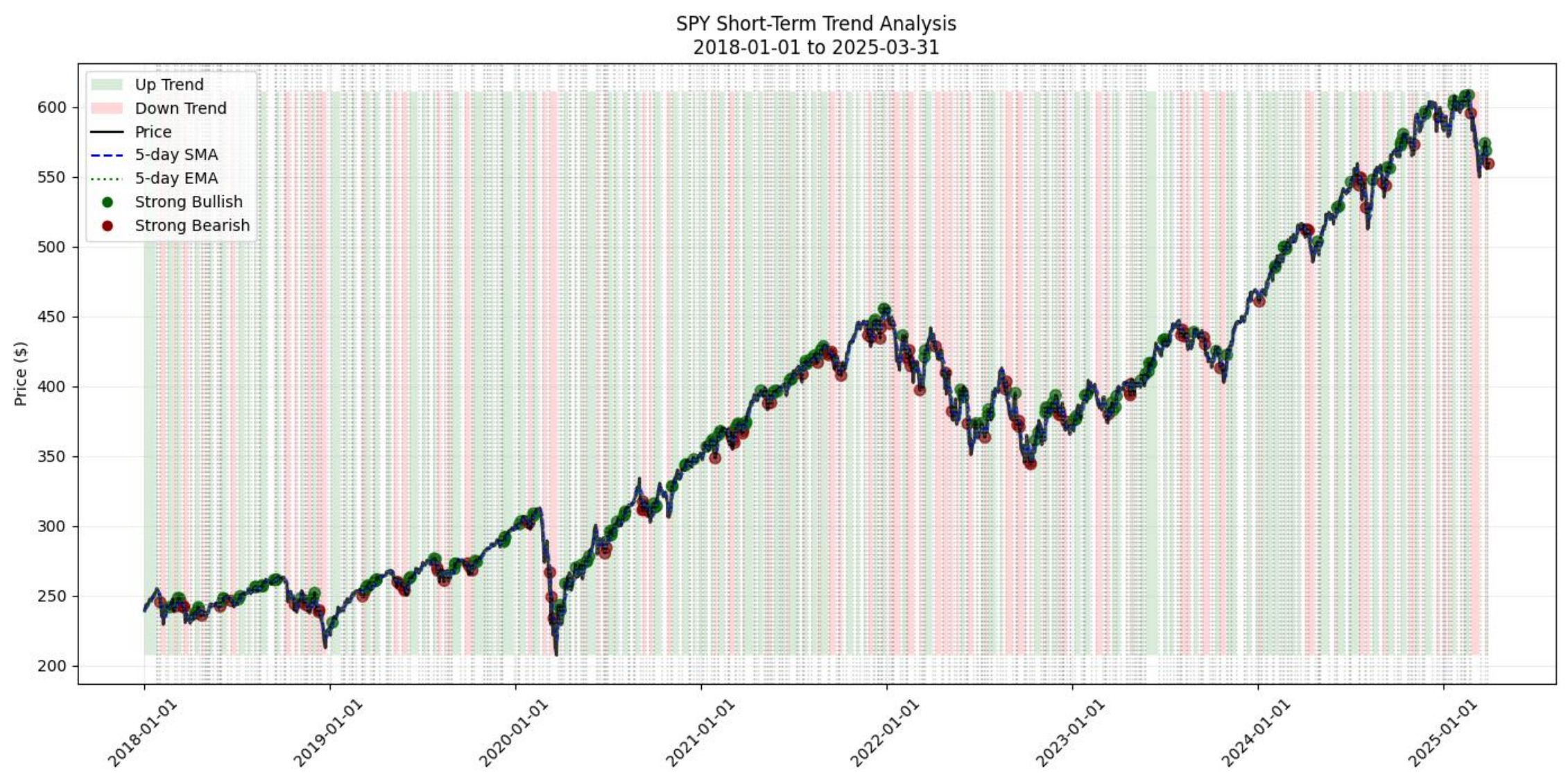
Model 2 – Relative strength index local maximum and minimum

Evaluation Metrics	U.S. Stock	Jan ‘22 – Feb ‘23	Mar ’23 – Apr’24
Annualized Rate of Return	5.105%	3.597%	2.148%
Sharpe Ratio	0.041	0.099	-0.718
Win Rate	67%	50%	100%
Average Win	3.59%	3.36%	2.51%
Average Loss	-1.22%	-1.22%	0%
Profit-Loss Ratio	2.96	2.76	0
Maximum Drawdown	14.600%	14.600%	8.600%
Alpha	0.004	0.089	-0.068

Parallel Ensemble Model

Overview:

- Ticker: SPY (S&P 500)
- Period: Jan 2018 – Mar 2025



Key Insights:

- Fragmented Analysis
  - Treats each trading day as an independent event
- Cannot show overall trend
  - Fails to recognize multi-day momentum patterns

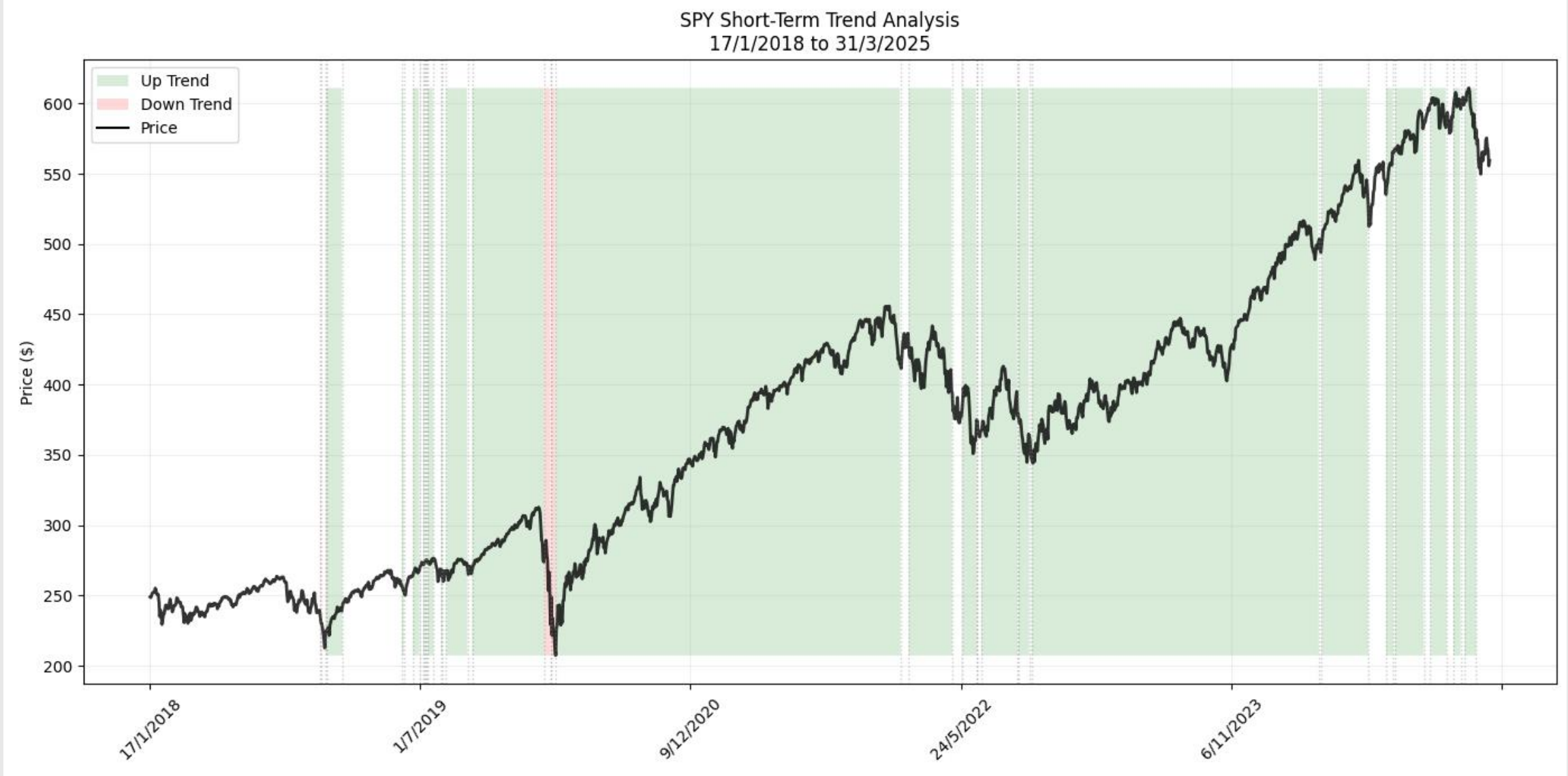
Major Problems:

- Each day is analysed individually and does not consider the sequence of days
- Less flexible and cannot reflect user’s own interpretation
  - Fixed threshold voting
  - Cannot adapt to different asset volatilities and user-defined risk preferences

LSTM Neural Network

Key Insights:

- Sequential Intelligence
  - Analyzes price movements as continuous time-series
  - Captures multi-day trends (bullish/bearish momentum)
- Personalised Classification
  - Trained on user-labeled data (adapts to personal interpretations)
  - Supports dynamic threshold adjustments



Key Findings

Dependence on Market Trend

- Baseline:
  - Excels in bullish markets (+18.6% returns)
  - Fails catastrophically in sideways markets (-9.4% returns, 27% win rate) due to frequent unwanted signals
- Model 1:
  - Excels in bullish markets (+13.4% returns)
  - Better performance than baseline but cannot profit from sideways markets (-0.3% returns, 50% win rate) due to less unwanted trades
- Model 2:
  - Outperforms in sideways/ranging markets (+8.9% alpha)
  - Underperforms during strong trends (-6.8% alpha)
- Key Insights:
  - ⚠ Trade-off between trend profitability and sideways-market resilience → Adaptive trend detection is essential

Importance of Time-series Data

- Hybrid (LSTM + Technical Indicators): Captures multi-day trends by looking at time-series patterns
- Ensemble (Technical Indicators Voting): Decision based on daily snapshots

The Subjectivity Problem

- Investor Polarisation:
  - Each investor interprets and defines bearish/bullish market differently
    - Conservative traders: Define "bullish" as ≥5 confirming indicators
    - Quantitative funds: Use statistical thresholds (e.g., 2σ moves)
  - Each investor looks at different time frame during analysis
    - Day traders (1hr) vs. funds (1mo)
- Key Insights:
  - ⚠ Hard to build a model that suit every investor’s preferences → Important to let investors customise their own model that align with their individual risk preferences
- Solution Framework:
  - Configurable thresholds in dashboards
  - User-labeled training for personalised alerts

Approach	Strengths	Weaknesses
LSTM Hybrid	• Captures multi-day momentum • Customisable classification	• More computationally expensive • Black box nature of AI
Ensemble Voting	• Interpretable rules	• Misses sequential patterns

Patterns Vary over Time

- Experiments show that trend patterns are different across different periods
- The difference in patterns make it hard to define and classify trend in a structured and predictable manner
- Case Studies:
  - Financial Crisis 2008: Protracted bearish trends (12+ months)
  - COVID-19: Rapid V-shaped recovery (3-month anomaly)
  - Trade-war 2025: Very volatile due to unpredictable policies
- Root Cause:
  - Macroeconomic shocks (e.g., Rapidly changing policies) and geopolitical events (e.g., wars) rewrite trend playbooks unpredictably

Difficulty in NLP and Sentiment Analysis

- Challenges:
  - Linguistic Complexity
    - Sarcasm/ambiguity in headlines (Haripriya & Patil, 2024)
    - Domain-specific semantics
    - Multilingual coverage needs
  - Temporal Alignment
    - Latency between: News release → NLP processing → Trading signal
- Breakthroughs:
  - Rise of LLM distillation
  - Real-time news summarization by LLM

Conclusion

Achievement of Objectives

This project successfully:

- Evaluated traditional and machine learning-based trading strategies across market regimes
- Enhanced baseline models through technical indicator refinement and LSTM integration
- Explored the feasibility of text data integration and anslysed its strengths and challenges
- Developed an interactive dashboard for visualization, laying the foundation for an integrated AI-driven investment platform

Summary of Key Findings

Market-Regime Dependency:

- Trend-following (moving average based) strategies excel in bullish markets but fail in sideways conditions
- Mean-reversion strategies show the inverse pattern, highlighting the need for adaptive approaches.

Temporal Modeling Superiority:

- LSTM hybrid model improved trend continuity over daily snapshot ensemble voting model

Investor-Centric Gaps:

- Traders demanded customizable thresholds, underscoring the limitations of one-for-all models

Significance of the Work

This research:

✓ Bridges technical analysis with modern ML/AI techniques
✓ Identifies the inherent difficulties, providing valuable insights for both academic research and practical applications in financial trading
✓ Demonstrates the value of hybrid models (Technical indicators + LSTM + sentiment analysis)
✓ Provides a framework for personalized strategy development via configurable dashboards

Future Directions

To address current limitations, future work will focus on:

- Reinforcement Learning: For dynamic hyperparameter optimization across market regimes
- Ensemble Trend Classification: Combining LSTM, technical indicators, and NLP outputs to switch strategies adaptively
- Real-Time Web Scraping: Reducing latency for live analysis
- LLM Research: Advancing unstructured text analysis capabilities
- Multi-Asset Validation: Testing generalizability to crypto/commodities markets
- Black Swan Analysis: Test the performance of the model during events

The ultimate goal is a self-adapting system that balances trend capture and sideways-market resilience — an all-rounded platform for traders in the dynamic real-world markets.