SCHOOL OF
COMPUTING &
DATA SCIENCE
The University of Hong Kong

Smarter Investment using Big Data, Data Science and Algorithmic Trading

CHAN Chun Hei (King)

UID: 3035684908

Supervisor: Prof SM Yiu

Outline

1. Introduction

2. Methodology

3. Results and Key Findings

4. Future Plan

5. Conclusion

What are
Algorithmic Trading Strategies ?

Introduction – Problem Statement
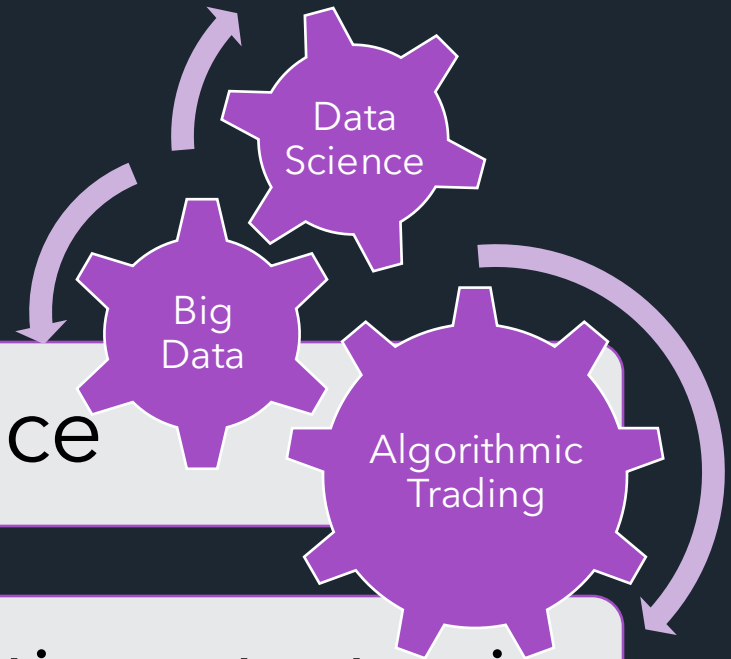
Hard to be consistent

Risk of overfitting

New forms of data

Introduction – Objectives

Evaluation
- Compare strategies performance

Enhancement
- Propose improvements on existing strategies

Exploration
- Explore textual data and web scraping

Visualisation
- Visualise results and integrate functions

Data Science

Big Data

Algorithmic Trading

Introduction – Project Overview

Algorithmic Trading	• Experiments (Baseline + Enhancements)
Trend Analysis	• Experiments (Baseline + Enhancements
News Analysis	• Proof-of-Concept • Analysis of Potential Challenges

Methodology

Data Collection

Algorithmic Trading Models

Back Testing using Evaluation Metrics

Historical Data:

Data cleaning, preprocessing, and aggregation:

Visualisation:

Methodology

Data Collection

↓

Algorithmic Trading Models

↓

Back Testing using Evaluation Metrics

Preliminary Approach:

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

Strategies:

Confidence Interval	Crossover	Local Maximum and Minimum

Methodology

Data Collection

↓

Algorithmic Trading Models
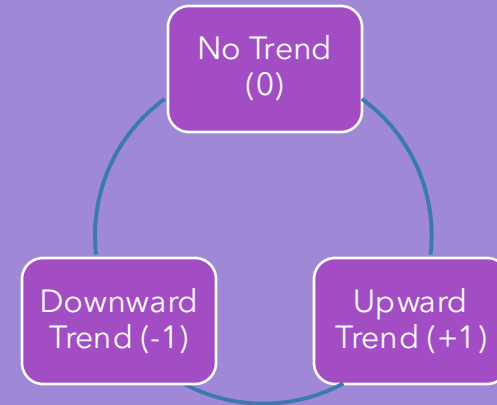
↓

Back Testing using Evaluation Metrics

Back Testing:

Evaluation Metrics:

1. Annualised Rate of Return (ARR)
2. Sharpe Ratio
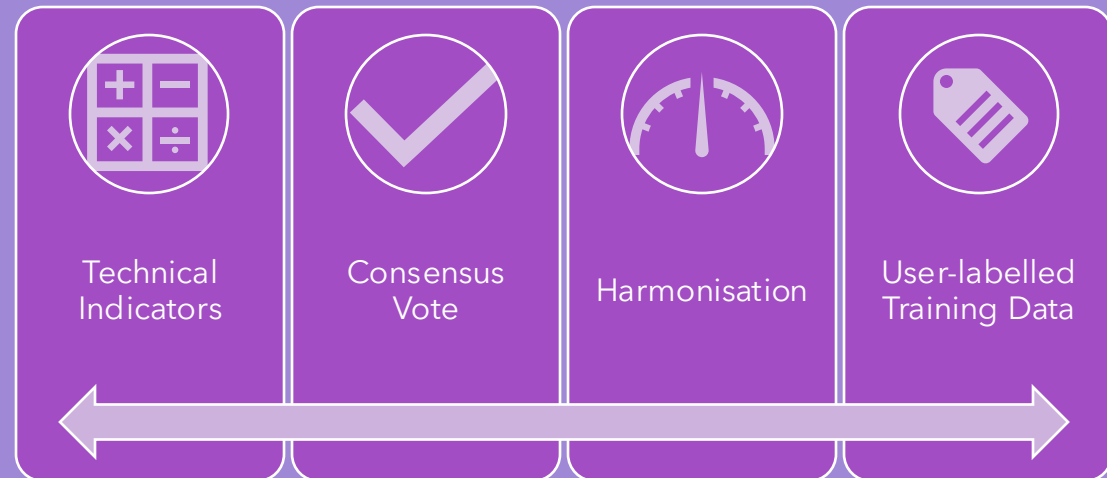3. Win Rate
4. Maximum Drawdown
5. Profit Factor
6. Alpha

Methodology

Trend Analysis

↓

News Analysis

↓

Dashboard

Idea:

No Trend (0)

Downward Trend (-1)

Upward Trend (+1)

Strategies:

Technical Indicators

Consensus Vote

Harmonisation

User-labelled Training Data

Methodology

Trend Analysis

↓

News Analysis

↓

Dashboard

Pipeline:

Raw Text → Parsing → Semantic Analysis

System Prompt:

You are an advanced financial analysis assistant specializing in market impact assessment. Your role is to analyse provided articles, news, or reports and generate insightful, data-driven assessments on how they may affect market conditions, sector performance, and specific stocks.

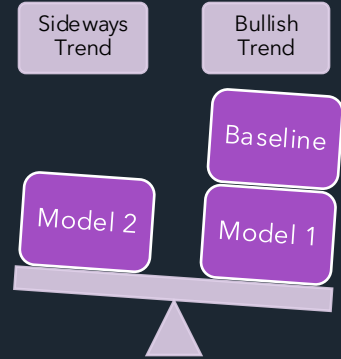Key Responsibilities:
…

Output Guidelines:
…

Methodology

Trend Analysis

↓

News Analysis

↓

Dashboard

Interactive visualisation:

plotly | Dash

Key Finding 1 - Dependence on Market Trend

Model	Bullish (Upward) Market	Sideways (No trend) Market
Baseline	✅ Profit (+18.6% return)	❌ Cannot profit (-9.4% return)
Model 1	✅ Profit (+13.4% return)	⭕ Outperform, but cannot profit (+6.5% alpha, -0.3% return)
Model 2	❌ Underperform (-6.8% alpha)	✅ Outperform (+8.9% alpha, +3.6% return)

- Key Insights:
  - ⚠️ Trade-off between trend profitability and sideways-market resilience → Adaptive trend detection is essential

Baseline – Algorithmic Trading

- Simple moving average crossover
- S&P 500 (SPY) data from 1 Jan 2022 to 30 Jun 2024
  - Sidewalk trend (Jan '22 – Feb '23)
  - Bullish trend (Mar '23 – Apr '24)

Baseline – Algorithmic Trading

- Annualised Rate of Return: 3.6%
  → Profitability

- Dependent on the trend:
  - Jan '22 – Feb '23 (Sidewalk): -9.4%
  - Mar '23 – Apr '24 (Bullish): 18.6%

Baseline			
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

Baseline – Algorithmic Trading

- Strategy cannot profit consistently

- Two major problems:
  - Unable to profit because of frequent unwanted signals and inability to capture peaks and troughs during sidewalk trend
  - Unable to effectively capture profit due to lagging property during bullish trend

Baseline			
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

16

Two Major Problems

Frequent unwanted signals during sidewalk

- Focus on reducing frequent unwanted signals when price moves around moving average → Model 1
- Focus on capturing peaks and troughs → Model 2

Lagging property of moving average during bullish trend

Results – Algorithmic Trading

Model 1 - Moving average confidence interval

For *each trading day*

    if *price > moving average - 1 × σ when crossover* **then**

        Buy()

    else if *price < moving average + 2 × σ when crossover* **then**

        Sell()

END For

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()

END For

Results – Algorithmic Trading

Model 1 – Moving average confidence interval (Trend Following)

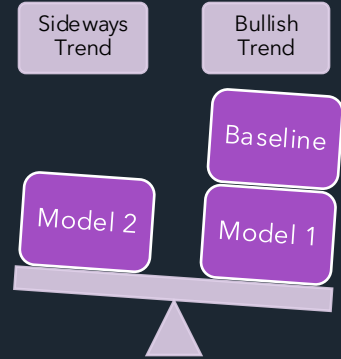**Model 2 – Relative strength index local maximum and minimum (Mean Reversion)**

Results – Algorithmic Trading

- Model 1
  - Less unwanted trades
  - Problem: Still cannot profit from sidewalk
  - Annualised Rate of Return: 6.3%
    - Sidewalk (-0.3%): Reduced lost and outperforms market ($\alpha = 0.065$)
    - Bullish (13.4%): Less profit

- Model 2
  - Successfully profit during sidewalk
  - Less unwanted signals
  - Sacrificed profit from bullish trend
  - Annualised Rate of Return: 5.1%
    - Sidewalk (3.6%): Profit and outperforms market ($\alpha = 0.089$)
    - Bullish (2.1%): Sacrificed more 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			
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

Key Finding 1 - Dependence on Market Trend

Model	Bullish (Upward) Market	Sideways (No trend) Market
Baseline	✅ Profit (+18.6% return)	❌ Cannot profit (-9.4% return)
Model 1	✅ Profit (+13.4% return)	⭕ Outperform, but cannot profit (+6.5% alpha, -0.3% return)
Model 2	❌ Underperform (-6.8% alpha)	✅ Outperform (+8.9% alpha, +3.6% return)

- Key Insights:
  - ⚠️ Trade-off between trend profitability and sideways-market resilience → Adaptive trend detection is essential
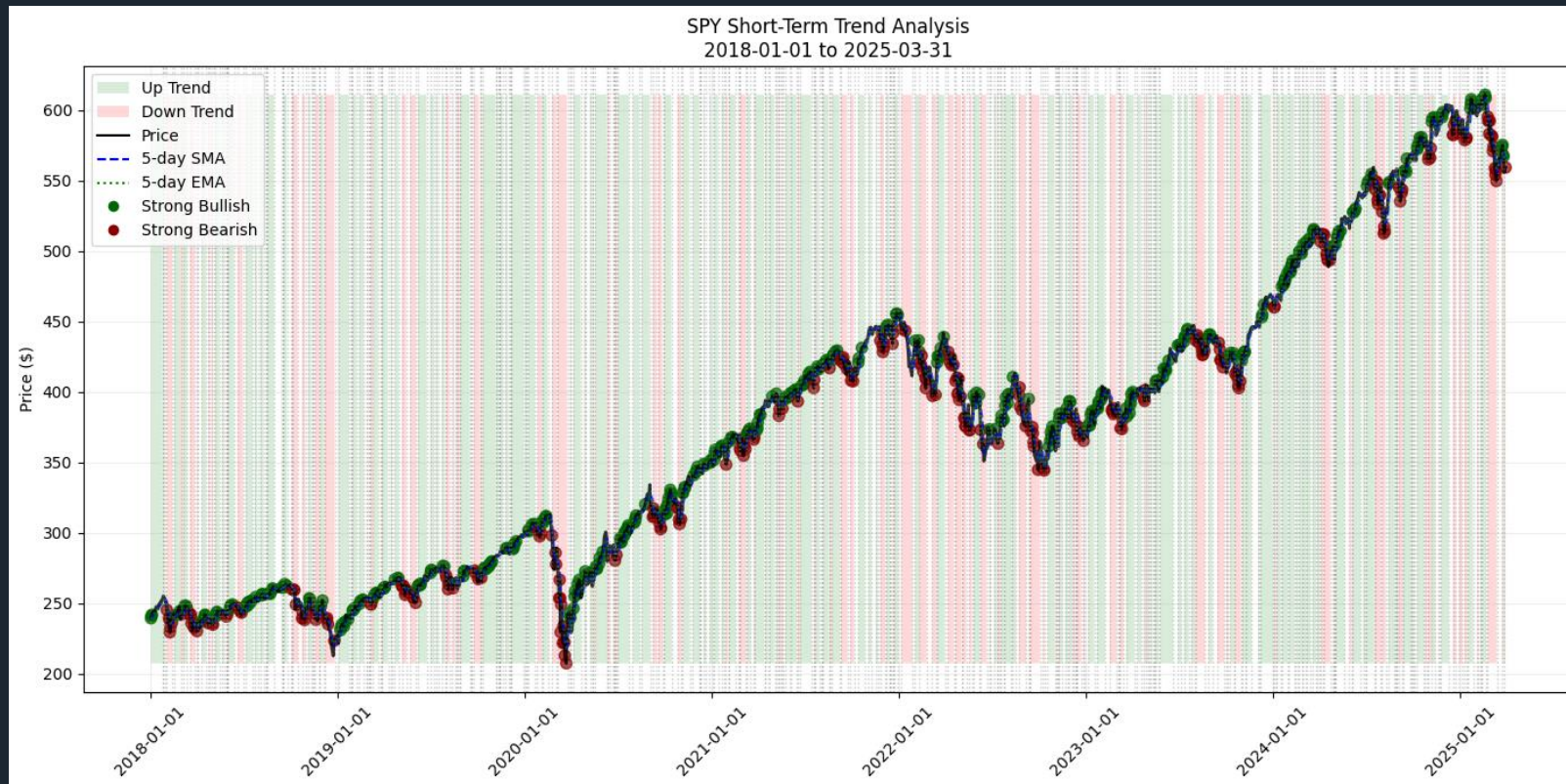
Baseline – Trend Analysis

- Simple parallel ensemble model
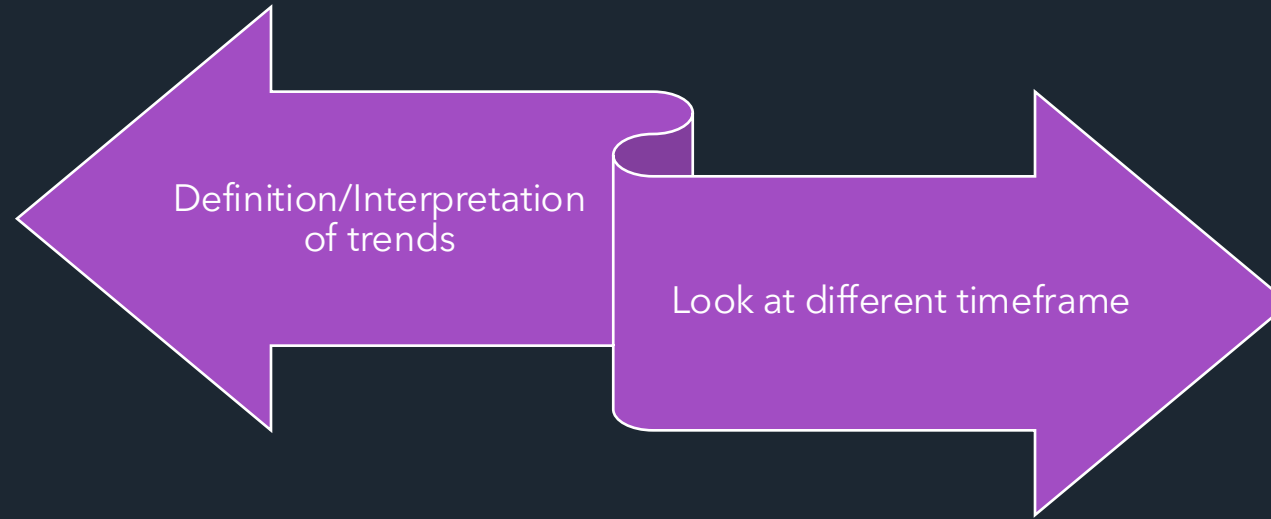- S&P 500 (SPY) data from 1 Jan 2018 to 31 Mar 2025
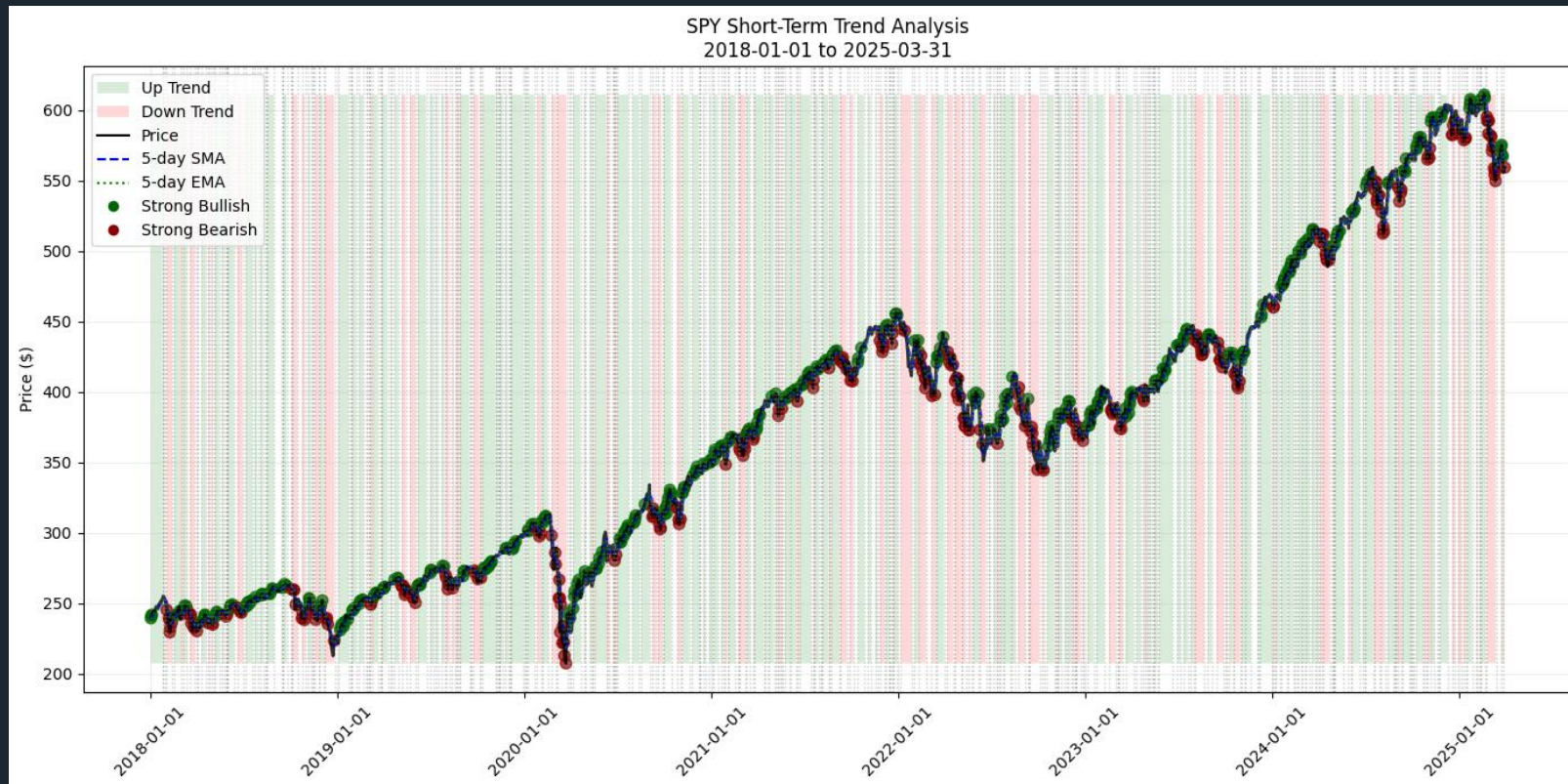
Baseline – Trend Analysis

- Two major problems:
  - Fragmented classifications, model treats each day independently
  - Low flexibility

Key Finding 2 - The Subjectivity Problem

Definition/Interpretation of trends

Look at different timeframe

- 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 → Model 3
  - User-labelled training for personalised alerts → Model 4

Results – Trend Analysis

Model 3 - Dynamic Parallel Ensemble Model

Results – Trend Analysis

- Dynamic Parallel Ensemble Model
  - Harmonisation reduces fragmentation
  - More flexibility
    - Indicator window
    - Harmonisation period
   - Classification threshold

Harmonisation:
- Averages trend values across a predefined window (current day ± n surrounding days)
- If n=1, project will apply a 3-day rolling average centering in the middle
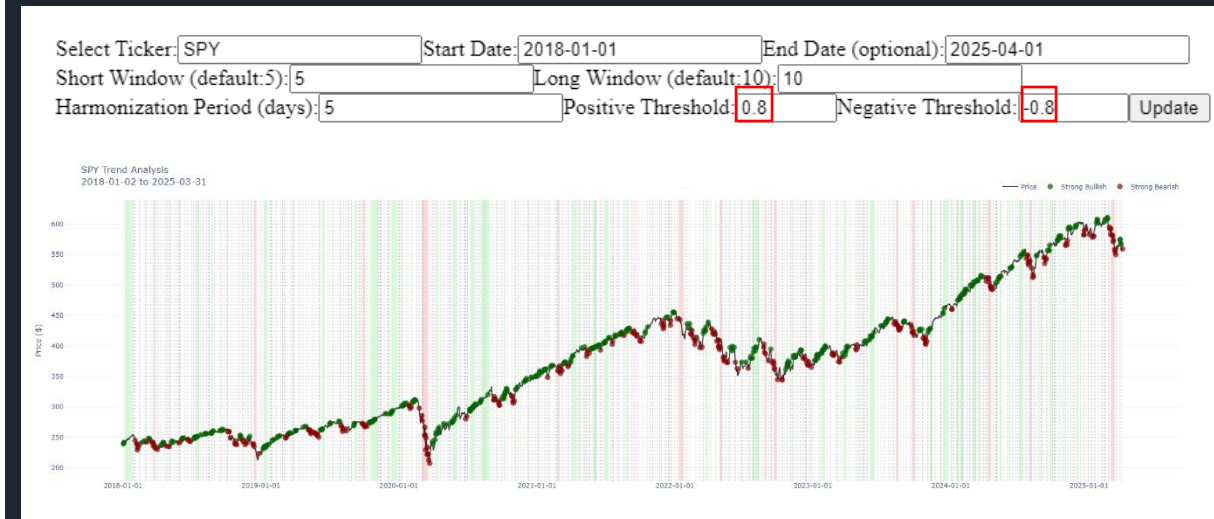
Results – Trend Analysis

- Experiments on the effects of parameter
    - Controlled Setup (Model 3): Windows = (5, 10), Harmonisation Period = 5, Threshold = ±0.5
    - Setup A: Windows = (50, 100), Harmonisation Period = 5, Threshold = ±0.5
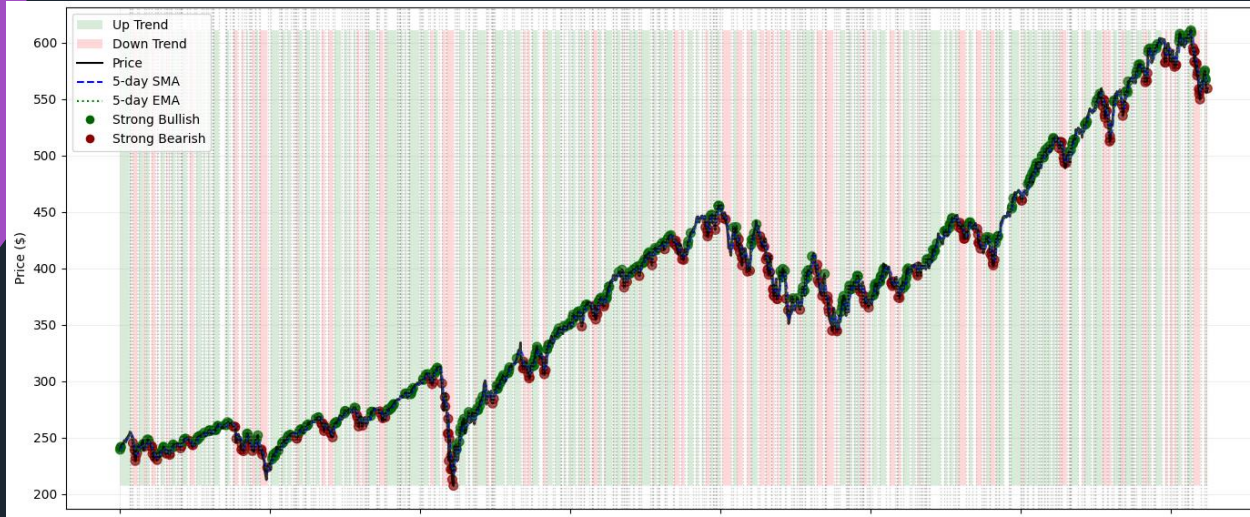   - Setup B: Windows = (5, 10), Harmonisation Period = 5, Threshold = ±0.8
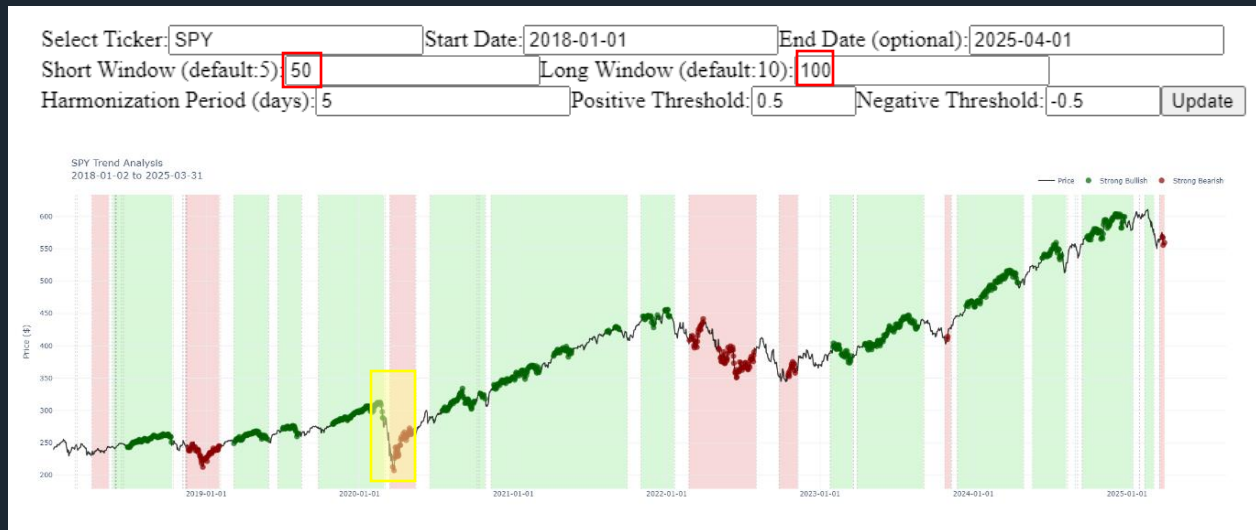
Setup A

Setup B

Results – Trend Analysis
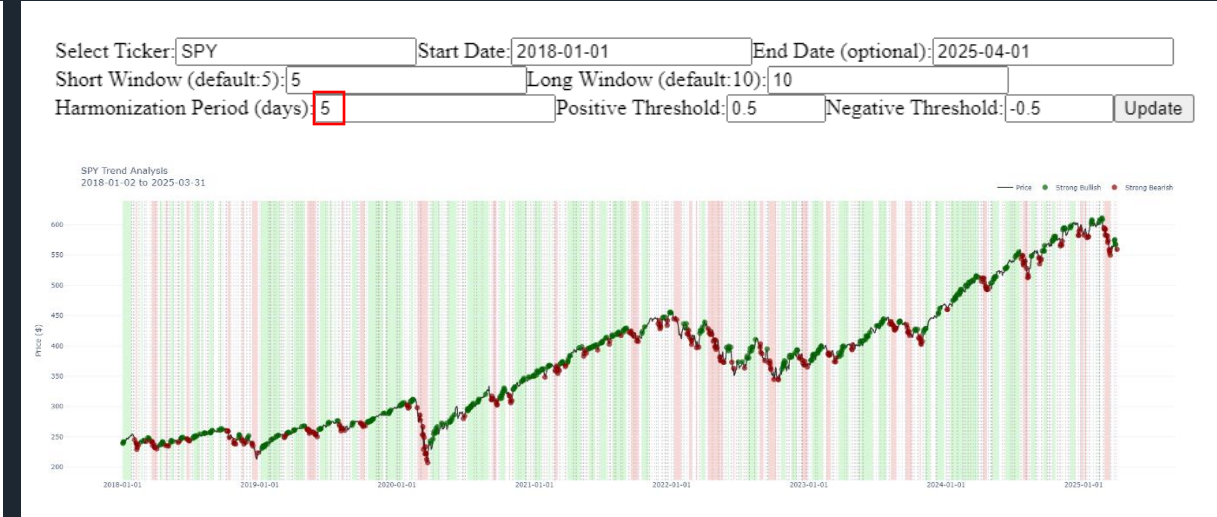

Baseline

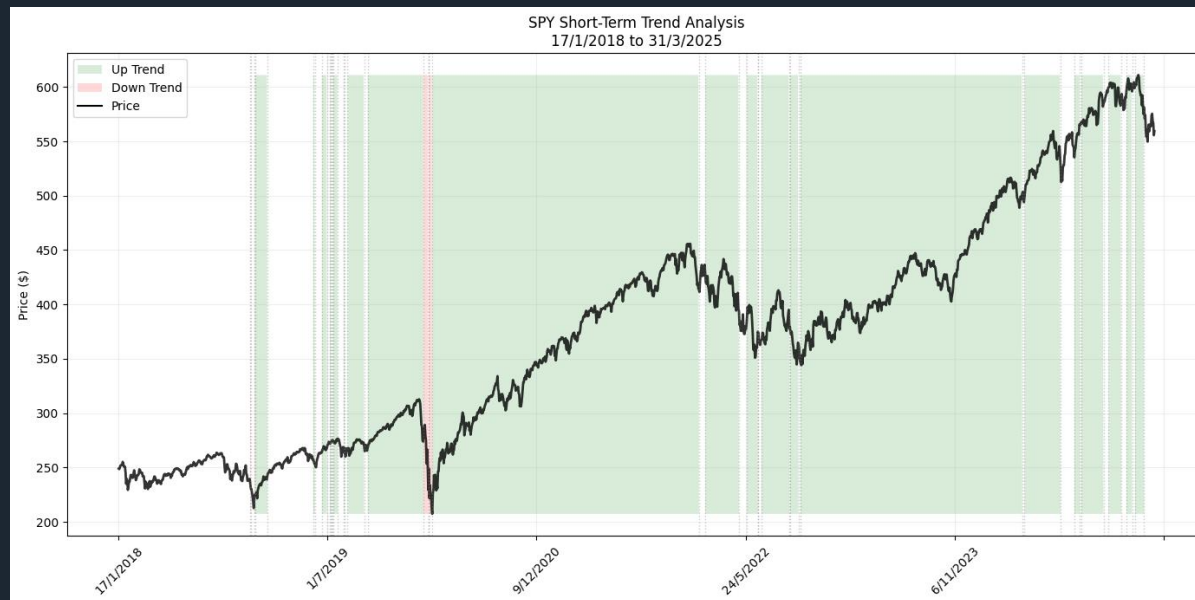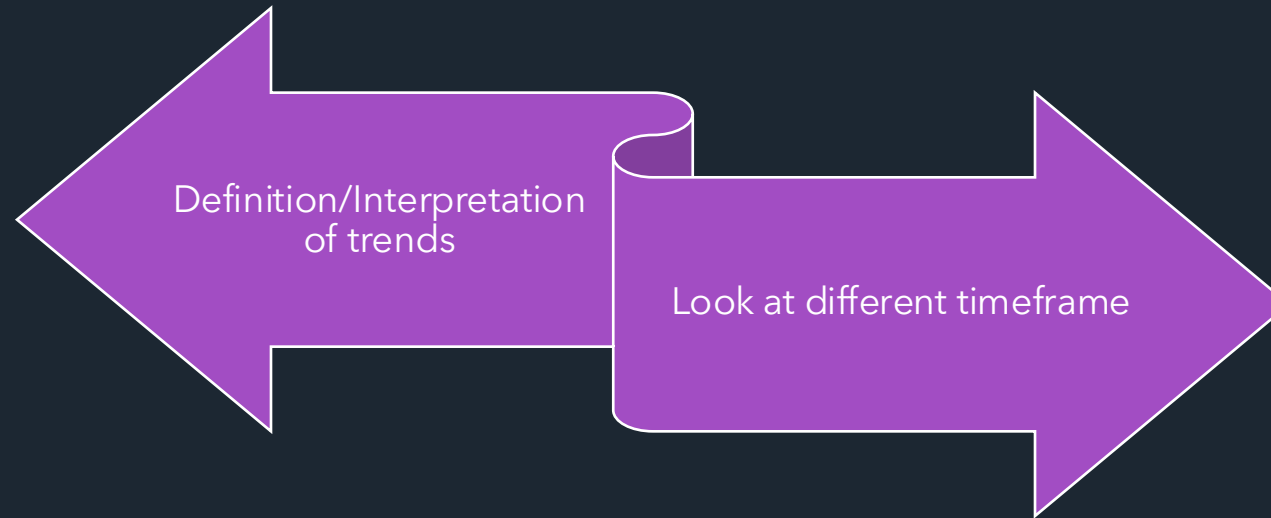Controlled Setup (Model 3)

Setup A

Setup B

Results – Trend Analysis

- Model 4: LSTM Neural Network
  - Ability to learn from sequential data
  - Human-labelled training data reflects user preferences on trends
  - Adapt to individual interpretation styles and trading philosophies intuitively and directly

SPY Short-Term Trend Analysis
17/1/2018 to 31/3/2025

Key Finding 2 - The Subjectivity Problem

Definition/Interpretation of trends

Look at different timeframe

- 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:

Approach	Strengths	Weaknesses
Model 3 – Ensemble Voting (Post-customisation)	• Interpretable rules • Customisable predictions	• Misses sequential patterns
Model 4 – LSTM Hybrid (Pre-customisation)	• Customisable training data • Captures multi-day momentum	• Black box nature of AI • More computationally expensive

Key Finding 3 - 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 → Contextual understanding is useful

Results – News Analysis

- LLM powered semantic analysis shows technical feasibility, but more refinement is needed

Extracted News Articles about SPY

[1] Title: Market Outlook: Next Week's Most Important Events (NYSEARCA:SPY) (31 分鐘前)
By: Seeking Alpha
Website: https://seekingalpha.com/article/4776438-market-outlook-next-week-most-important-events
[2] Title: The Smartest S&P 500 ETF to Buy With $500 Right Now (18 小時前)
By: The Motley Fool
Website: https://www.fool.com/investing/2025/04/20/the-smartest-sp-500-etf-to-buy-with-500-right-now/
[3] Title: Fed Policy Shift and Global Supply Chain Concerns: Impact on $SPY Trading (3 天前)
By: Blockchain News
Website: https://blockchain.news/flashnews/fed-policy-shift-and-global-supply-chain-concerns-impact-on-spy-trading
[4] Title: SPY ETF News, 4/18/2025 (2 天前)
By: The Globe and Mail
Website: https://www.theglobeandmail.com/investing/markets/stocks/HSY/pressreleases/31947199/spy-etf-news-4182025/
[5] Title: SPY, QQQ Call Volumes Spiked Minutes Before Tariff Pause Announcement: Alexandria Ocasio-Cortez Demands Disclosure From Congress Members (1 週前)
By: Benzinga
Website: https://www.benzinga.com/government/regulations/25/04/44728276/spy-qqq-call-volumes-spiked-minutes-before-tariff-pause-announcement-alexandria-ocasio-cortez-demands-disclosure-from-congress-members
[6] Title: Only 1 Of These 3 Unusually Active SPY Put Options Makes For A Good Bull Put Spread (3 天前)
By: Barchart.com
Website: https://www.barchart.com/story/news/31938797/only-1-of-these-3-unusually-active-spy-put-options-makes-for-a-good-bull-put-spread

Generated Analysis

Analysis:

Summary: The article discusses the benefits of investing in the Vanguard S&P 500 ETF amid market volatility caused by tariffs and economic concerns.

Key Impacts:
- Historic average return rate of 10% annually for the S&P 500.
- Vanguard's S&P 500 ETF offers exposure to 500 large U.S. publicly traded companies with a low expense ratio of 0.03%.
- The S&P 500 index has shown consistent long-term growth despite short-term fluctuations.
- Investing in a diversified ETF like VOO can provide stability during market turmoil caused by tariffs.
- Market volatility is expected due to uncertain tariff policies, potentially affecting short-term gains in the S&P 500.

Top Affected Stocks/Sectors:
- Vanguard S&P 500 ETF (VOO) - The highlighted ETF likely to benefit from investors seeking stability amid market uncertainties.
- Large-cap U.S. publicly traded companies across sectors like technology, financials, healthcare, consumer goods, and energy.

Suggested Actions:
- Monitor Market Volatility: Keep track of tariff-related developments and their impact on market stability.
- Consider Investing in Diversified ETFs: Look into ETFs like VOO for exposure to a broad range of companies during uncertain market conditions.
- Focus on Long-Term Growth: Emphasize long-term investment strategies given the historical growth trends of the S&P 500.
- Stay Informed: Keep abreast of changing economic conditions and tariff policies to make informed investment decisions.

AI generated content. Not financial advice. Investors should cross-validate with the latest data and risk tolerance.

Key Finding 4 – Difficulty in NLP and Sentiment Analysis

- Challenges:

Linguistic Complexity	Temporal Relevance Gap	Nonlinear Relationship	Source Credibility Blind Spot	Actionability Deficit	Human Behavioural Complexity
Ambiguity and domain-specific semantics	Value between a 30-min and 1-week old news	Sentiment ≠ Price Action	Credibility between established publishers and less regulated sources	Recommendations remain too generic	Markets often react more strongly to negative sentiment than positive
		Macro trends often overshadow sentiment → isolation is hard			
		Hard to prove the correlation between event and price movement			

- Ethical concerns:

Limitations

- Risk of hallucination
- Black box problem
- Accountability Issues

Suggestions

- Users must remain aware of the limitations, including potential biases stemming from skewed training data
- AI-generated content not verified by human analysts should carry clear disclaimers identifying it as machine-produced

Future Work

Advanced Model Architecture

- Reinforcement learning for dynamic hyperparameter optimization
- Large language model for more advanced interpretation of complex market data

Adaptive Ensemble Framework

- Development of ensemble classifiers that combine LSTM, technical indicators, and NLP outputs

Generative Adversarial Network for Trend Analysis

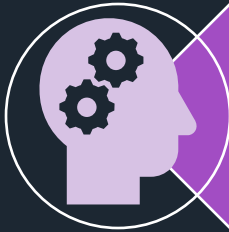- Exploration of GAN for more advanced trend analysis

Research on Textual Data

- Advancement of natural language capabilities for financial text analysis

Conclusion – Achievement of Objectives

Raw Market Data

Patterns Analysis

Contextual Understanding

Trend Analysis

News Analysis

Algorithmic Trading

Evaluated different algorithmic trading strategies across diverse market regimes

Enhanced baseline algorithmic trading models through statistical methods

Enhanced trend analysis algorithm through technical indicator refinement, harmonisation, and LSTM integration

Demonstrated the feasibility of textual data integration while analysing its challenges

Developed an interactive dashboard prototype for strategy visualization

Conclusion – Significance of Work

Bridges classical technical analysis with contemporary AI techniques

Provided empirical evidence of market-regime dependencies that inform both academic research and practical strategy development

Established a framework for personalised algorithmic trading through configurable interfaces and adaptive model architectures

Q&A