FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Min Wang	Germany	professional@minwang-ai.com	
Louckmane Diawara	Burkina Faso	loukmanediawara597@gmail.com	
Ashish Kumar	India	ashishexam89580@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	**Min Wang**
Team member 2	**DIAWARA Louckmane**
Team member 3	**Ashish Kumar**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.
N/A

1. Part 1: Assessing Models with Alternative Data

Q1. Data Understanding

The paper (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024) integrates market and technical data to predict stock market movements in emerging markets using neural networks. The authors employ **daily historical data** from Yahoo Finance for three exchange-traded funds (ETFs): ECH (Chile), EWZ (Brazil), and IVV (S&P 500). The raw attributes include *Open, High, Low, Close, Adjusted Close, and Volume* prices for each trading day between December 2009 and January 2020. From these primary variables, over **210 technical indicators** were derived using the Pandas Technical Analysis (Pandas TA) library, expanding the feature space to 216 daily features (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024).

Technical indicators such as moving averages (SMA, EMA), momentum oscillators (RSI, Stochastic RSI), and trend-based metrics (Bollinger Bands, Balance of Power, and Z-score) are mathematically constructed transformations of price and volume data. They are crucial for machine learning models as they summarize price behavior and market psychology, thereby capturing non-linear dependencies in time series. Using these engineered features reduces noise and enables the neural network to identify turning points, volatility regimes, and cyclic patterns essential for forecasting stock price trends in emerging markets.

Figure 1: IVV Price History (all). Source: Yahoo Finance.

Q2. Security Understanding (IVV)

The **iShares Core S&P 500 ETF (IVV)** is one of the largest and most liquid exchange-traded funds (ETFs) globally, managed by BlackRock. It tracks the **S&P 500 Index**, providing exposure to 500 of the largest publicly traded companies in the United States across multiple sectors. IVV represents a benchmark for developed market performance, emphasizing diversification, stability, and low cost.

As of October 2025, IVV holds approximately **US$699.9 billion** in assets under management (AUM) with an **expense ratio of 0.03%**. Its sector allocation is led by information technology (27.7%), healthcare (13.4%), and consumer discretionary (12.0%) Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024; iShares 2025. The fund has exhibited strong growth since 2009, reflecting the resilience of the U.S. equity market. Historical price data from Yahoo Finance show IVV's price rising from around **US$100 in 2009** to over **US$660 in October 2025**. The **52-week range** is approximately **US$484−677**, with a **P/E ratio of 29.64** and a **3-year standard deviation of 13.37%**.

The authors of the reference paper treat the problem as a **classification task** rather than regression, since the goal is to predict the *direction* of price movement—whether the opening price will rise or fall. The dependent variable Γ_t is defined as:

$$\Gamma_t = \begin{cases} 1, & \text{if } Open_t - Open_{t-1} > 0 \\ \\ 0, & \text{otherwise.} \end{cases}$$

Two alternative ways to define the classification variable include:

1. Using the sign of daily returns based on closing prices instead of opening prices, i.e., $\Gamma_t = 1$ if $(Close_t - Close_{t-1})/Close_{t-1} > 0$.

2. Defining multiple categories such as "strong increase," "neutral," and "decline" using percentage thresholds, e.g., $> +0.5\%$, between -0.5% and $+0.5\%$, and $< -0.5\%$.

This binary classification approach simplifies the learning process and aligns the model with directional trading strategies commonly used by investors.

Q3. Methodology Understanding

Following the paper's structure, Section 2 (*Materials and Methods*) can be reorganized into two main sections:
Section 2. Data – with subcategories:

- 2.1 Stocks Analyzed

- 2.2 Methodological Approach Based on Data Mining (CRISP-DM)

- 2.3 Technical Indicators

- 2.4 Class Assignment

- 2.5 Data Normalization

- 2.6 Data Cleaning

  Section 3. Methodology – with subcategories:

- 3.1 MLP for Predictive Analysis

- 3.2 Statistical Measures for Feature Selection

- 3.3 LASSO Regularization

- 3.4 Tree-based and Correlation-based Feature Selection

- 3.5 Principal Feature Analysis, MAD, and Dispersion Ratio

- 3.6 Cross-validation and Experimental Design

Descriptive statistics (e.g., Pearson correlation) summarize linear relationships among variables, while model-based methods (e.g., LASSO and MLP) focus on predictive learning. The optimization of technical indicators was achieved through iterative feature selection and cross-validation. The authors improved predictive power by retaining only the top 5% of salient features, reducing computational cost by 84.7% and improving accuracy by roughly 2%. Optimizing indicators ensures the neural network receives only high-information features, enhancing generalization and avoiding overfitting (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024).

Q4. Feature Understanding

In the context of the paper, a **feature** refers to any variable derived from raw financial data that serves as input to the model, such as technical indicators like RSI, MACD, or Bollinger Band Percent (BBP). A feature differs from a **method**, which is an analytical procedure (e.g., LASSO or Chi-Squared selection), and from a **model**, which learns to map features to outputs (e.g., MLP neural network). The study categorizes features into momentum, volatility, trend, volume, statistics, and cycle indicators.

The optimization of technical indicators involves statistical filtering to identify the most informative subset using measures such as variance, correlation, and dispersion ratio. By selecting a minimal yet representative set (Selected(5)), the authors achieved near-optimal accuracy (78–80%) using only 5% of total features. This demonstrates that feature optimization enhances interpretability, reduces noise, and aligns with dimensionality reduction principles.

Q5. Optimization Understanding

Cross-validation is a model validation method used to evaluate how the results of a model will generalize to unseen data by repeatedly training and testing on different data splits. **k-fold cross-validation** divides the dataset into k subsets, training on $k-1$ folds and testing on the remaining one, cycling through all folds and averaging the performance to mitigate overfitting (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024).

Jaccard distance measures the dissimilarity between two sets and is defined as $J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$. In this paper, it quantifies the similarity between feature subsets selected for different ETFs (e.g., $J(ECH, EWZ) = 0.33$, $J(ECH, IVV) = 0.64$). A lower Jaccard distance implies greater overlap in selected features.

Compared to Euclidean and Manhattan distances, which measure geometric distances between numerical points, the Jaccard distance is categorical and set-based, focusing on overlap rather than magnitude. The authors define the **optimal solution** as the feature subset that maximizes predictive accuracy while minimizing computational cost and redundancy—specifically, the Selected(5) set that achieved the best cross-validated accuracy.

2. Following Steps

Step 1. Financial Problem

The primary financial problem addressed by Sagaceta-Mejía et al. (2024) is the difficulty of accurately predicting **stock market trend direction in emerging markets**, where volatility, nonlinear dependencies, and data scarcity make traditional models unreliable. Emerging markets such as Brazil and Chile experience rapid structural and macroeconomic changes, high liquidity variation, and sectoral imbalances, which cause nonstationarity in price dynamics. Accurately anticipating market movements is essential for investors seeking to optimize timing and risk exposure in exchange-traded funds (ETFs) linked to these economies.

The authors propose a data-driven methodology that integrates **optimized technical indicators** and a **multilayer perceptron (MLP)** classifier to determine the next-day movement of ETF opening prices. By filtering and ranking over 200 candidate indicators from Yahoo Finance data using statistical selection techniques, their approach identifies the most salient features that improve prediction accuracy while reducing computational cost. The model's goal is to provide **actionable predictive insight**—helping investors decide when to enter or exit positions—rather than forecasting continuous price levels.

Predicting stock movements in emerging markets differs markedly from developed markets like the United States (represented by IVV) because of distinct **volatility regimes, liquidity conditions, and information efficiency**. Emerging markets often show stronger cyclical dependencies and react more sensitively to macroeconomic shocks. As shown in Table 1 of the reference paper, the iShares MSCI Chile ETF (ECH) and iShares MSCI Brazil ETF (EWZ) have concentrated exposure to materials, energy, and financial sectors, whereas IVV is dominated by technology and healthcare. These structural differences imply that feature sets useful for developed markets—such as volume-based or sentiment-driven indicators—may not generalize to emerging markets. The paper therefore tailors its optimization process to improve robustness under higher volatility and lower information transparency conditions (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024).

Step 2. Application and Main Results

The application phase focuses on evaluating how optimized indicator selection improves both accuracy and efficiency in ETF trend prediction. Using 10-fold cross-validation, the authors test multiple subsets of features (denoted as *Selected(n)*). The optimal subset, *Selected(5)*, contains only about 5 % of all indicators yet achieves comparable or superior accuracy to the full 216-feature set.

- For emerging-market ETFs (ECH and EWZ), classification accuracy increased to ≈ **80 %** while reducing training time by ≈ **85 %**.

- For the developed-market ETF (IVV), the model achieved **78.5 %** accuracy with only nine key features.

- The most salient indicators include Balance of Power (BOP), Bollinger Band Percent (BBP), Correlation Trend Indicator (CTI), Williams %R (WILLR), and On-Balance Volume (AOBV) for emerging markets, and Price–Volume Rank (PVR) plus TTM Trend (TTM_-TRND) for IVV.

The results demonstrate that feature optimization substantially enhances predictive performance by filtering redundant or low-variance variables. Figure 5 in the reference paper shows a clear improvement trend up to $n = 5$, after which excessive dimensionality reduction leads to information loss. The study further reports an average accuracy improvement of **13.6 %** and an average training-time reduction of **84.7 %** when applying early stopping with optimized features.

Practical implications: For investors in emerging markets, the model offers a quantitative decision-support system for *market timing and risk management*. By interpreting optimized indicators, investors can identify favorable entry or exit points, mitigate exposure during high volatility, and complement qualitative macroeconomic analysis. The approach illustrates how alternative data—technical indicators derived from raw price and volume series—can be leveraged with machine learning to improve investment decisions.

Key takeaway: Efficient feature selection not only enhances accuracy but also ensures interpretability and scalability. The methodology bridges the gap between data mining

and financial modeling by delivering a computationally efficient neural-network framework adaptable to both emerging and developed markets (Sagaceta-Mejía, Sánchez-Gutiérrez, and Fresán-Figueroa 2024).

Step 3. Replication

See the Jupyter notebook file `ivv_replication_paper_aligned.ipynb`.

PART 2: A USER GUIDE FOR FINANCIAL ANALYSIS WITH GOOGLE TRENDS INDEX

2.1 Sources of Data

Alternative data refers to information collected from nontraditional channels, outside of standard financial statements or regulatory filings. Among the ten subcategories of alternative data identified by Sun et al. (2024), *Internet search and web traffic data*—particularly the Google Trends Index—has gained significant relevance in financial analytics. This user guide focuses on the use of **Google Search Volume Index (SVI)** as a proxy for investor attention and sentiment.

The main data source is **Google Trends**, a public platform offering normalized measures of search frequency for specific keywords over time. In the paper *"The impact of Google searches, Put-Call ratio, and Trading Volume on stock performance using Wavelet Coherence analysis"* (Vasileiou and Tzanakis 2024), Google Trends data was used to quantify investor attention toward AMC Theatres, a leading "meme stock." The study confirmed that spikes in Google search activity closely mirrored trading anomalies and price surges.

In practice, SVI data can be collected directly from `https://trends.google.com` or programmatically via the `pytrends` Python library—an unofficial API that allows automated queries, keyword comparisons, and regional filtering. Complementary data, such as stock prices and volumes, may be retrieved through `yfinance` or Bloomberg terminals for cross-variable analysis.

2.2 Types of Data

Google Trends Index offers a range of structured and unstructured data suitable for a financial analysis as below:

- **Search Volume Index (SVI) Time Series :** a metric between 0 and 100, normalized, which corresponds to relative popularity of a specified keyword within a specified period. One example is, a score of 100 corresponds to the highest observed search volume.

- **Interest by Region :** a geographic breakdown of search interest which tells you where search interest is concentrated. This is useful for market segmentation or for identifying local retail investor sentiment.

- **Related Queries and Topics :** these are contextually associated searches which help understand the investor psyche. For example, "AMC short squeeze" vs "AMC stock forecast" are different but related queries.

- **Trending or Rising Searches :** these are the searches that are gaining momentum and are, therefore, leading indicators of a shift in market attention.

- **Temporal Frequency :**this varies by the time period and region you select. It can be daily or weekly. For financial studies, weekly data is synchronous, aligning with the asset returns, is preferable.

2.3 Quality of Data

Google Trends data is readily available and highly dynamic, but the quality and reproducibility of the data should be thoughtfully considered.

- **Normalization and sampling bias :** the index is normalized to relative values within the time period and region chosen, meaning absolute search volumes are inaccessible. In addition, Google uses random sampling, which implies that historical values may be somewhat.

- **Temporal Resolution :** weekly granularity limits detection of intra-week sentiment changes,which may be crucial for high-frequency trading.

- **Reproducibility :** results can vary with different time windows or overlapping keywords, making methodological transparency essential.

- **Data Cleaning :** analysts must ensure removal of NaN values, synchronization of timestamps, and consistent time zones when merging SVI with financial data (Vasileiou & Tzanakis, 2022).

2.4 Ethical Issues

Using search data introduces several ethical and methodological considerations:

- **Privacy and Anonymity :** google Trends data is aggregated and anonymized; no individual-level identifiers are accessible. Nonetheless, ethical research mandates acknowledgment of this aggregation to prevent overinterpretation of behavioral signals.

- **Market Manipulation Risks :** awareness of popular search terms can lead to herding behavior or speculative bubbles, as witnessed during the meme-stock phenomenon.

- **Reproducibility and Transparency :** given the non-deterministic sampling method, replicable research requires storing and timestamping downloaded datasets.

- **Interpretation Risks:** search data reflects attention, not necessarily informed sentiment.A spike in searches could indicate curiosity, panic, or rumor propagation, not a deliberate investment decision.

2.5 Python Code to Import and Structure Data

Below is a demonstration of how to import and structure Google Trends data for financial analysis:

```python
1  # libraries
2  from pytrends.request import TrendReq
3  import pandas as pd
4
5  # Initialize the Pytrends connection
6  pytrends = TrendReq(hl='en-US', tz=360)
7
8  # timeframe
9  kw_list = ["AAPL"]                          # Apple stock ticker
10 timeframe = "2024-10-01 2025-10-01"
11
12 #payload
13 pytrends.build_payload(kw_list, timeframe=timeframe)
14
15 # Retrieve Search Volume Index (SVI) over time
16 search_interest_df = pytrends.interest_over_time()
17
18 # Display first rows
19 print("SVI over time:")
20 print(search_interest_df.head())
21
22 # Research Interest by Region
23 regional_interest = pytrends.interest_by_region(resolution='COUNTRY',
      inc_low_vol=True)
24
25 # Display first rows
26 print("\nInterest by Region:")
27 print(regional_interest.head())
```

In this project, we gather data on the stock ticker **AAPL**, representing **Apple Inc.**, using the **Pytrends Python library**. After importing the necessary libraries (**pytrends** and **pandas**), we establish a connection to **Google Trends** while defaulting to the **English language**. Also, we set the **time zone to 6 hours ahead of Coordinated Universal Time (UTC+6)**. This step ensures the **Google Trends response is properly timestamped**. We define our pri-

mary keyword of interest, **"AAPL"**, and specify our desired **time range of one year**, from **October 1, 2024, to October 1, 2025**. The **build_payload()** function effectively customizes the request before it is sent, capturing all necessary details such as **keywords, time frame, and geographical location**.

Using **interest_over_time()** function, we obtain the **Search Volume Index (SVI)** and save it as a **pandas DataFrame**. For each week of the period we specified, this table gives a relative measure of the **amount of attention Apple received from the public**. When considering actionable **Apple Inc. stock data**, **high SVI values** reflect periods of pronounced interest during key events, such as **product launches, earnings announcements, or high-profile news coverage**. The **isPartial column** indicates whether the **most recent observed period of data is complete or partial**.

Pytrends also offers other useful features to enrich the analysis. **Interest by region** (**interest_by_region()**) allows us to geographically map where the searches originate, which can shed light on the **concentration of investors or customers** across different areas. **Related queries** (**related_queries()**) reveal the **terms searched in connection with AAPL**, helping to distinguish whether users are seeking **product information or investment insights**. The **keyword suggestions function** (**suggestions()**) helps identify **additional relevant terms to track**, while **real-time trending searches** (**realtime_trending_searches()**) can capture **sudden spikes in interest**, signaling events with **potential market impact**.

2.6 Exploratory Data Analysis (EDA)

To demonstrate the usefulness of this kind of data, we build a simple trading algorithm in Python. The algorithm is not implemented on livestock market feeds but is instead backtested on historical price data retrieved via yfinance. The following basic steps are undertaken for strategy development and testing:

1. The SVI dataframe is concatenated with stock price data. All dates are converted to `datetime` objects and used as indexes. Stock prices are resampled to weekly values to ensure alignment.

2. Stock returns are calculated using a simple `pct.change()` function. This makes calculating strategy returns easier.

3. The combined dataframe is structured into a strategy-ready format (with trading signals, positions, and cumulative returns).
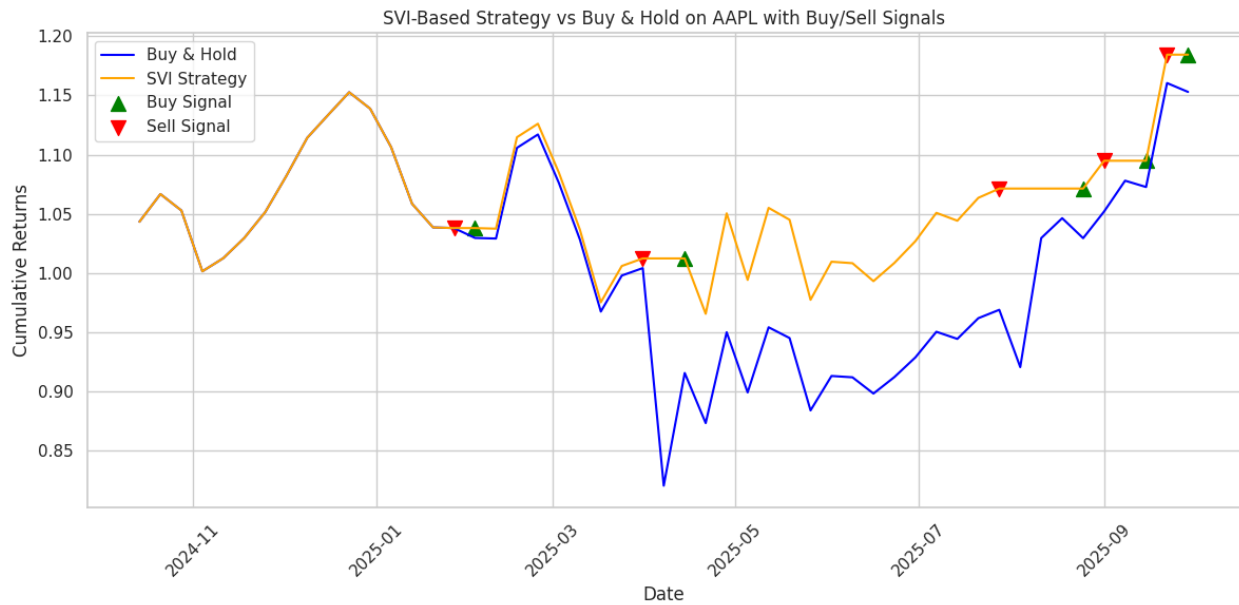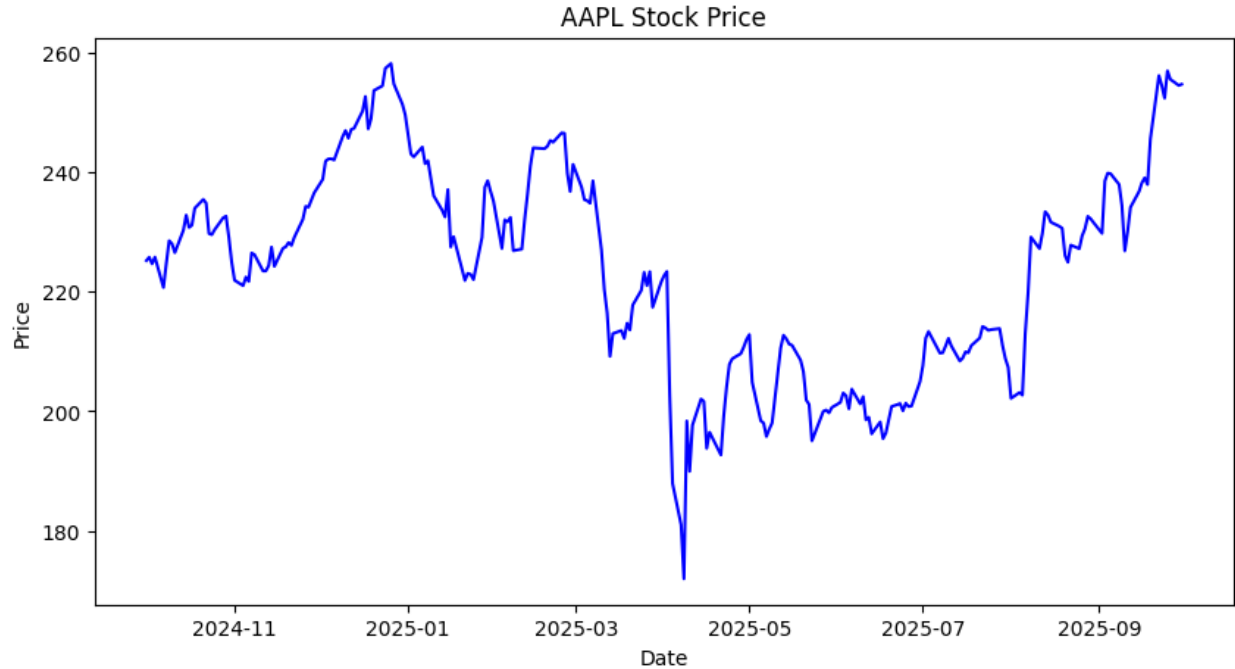
Figure 2: SVI For AAPL
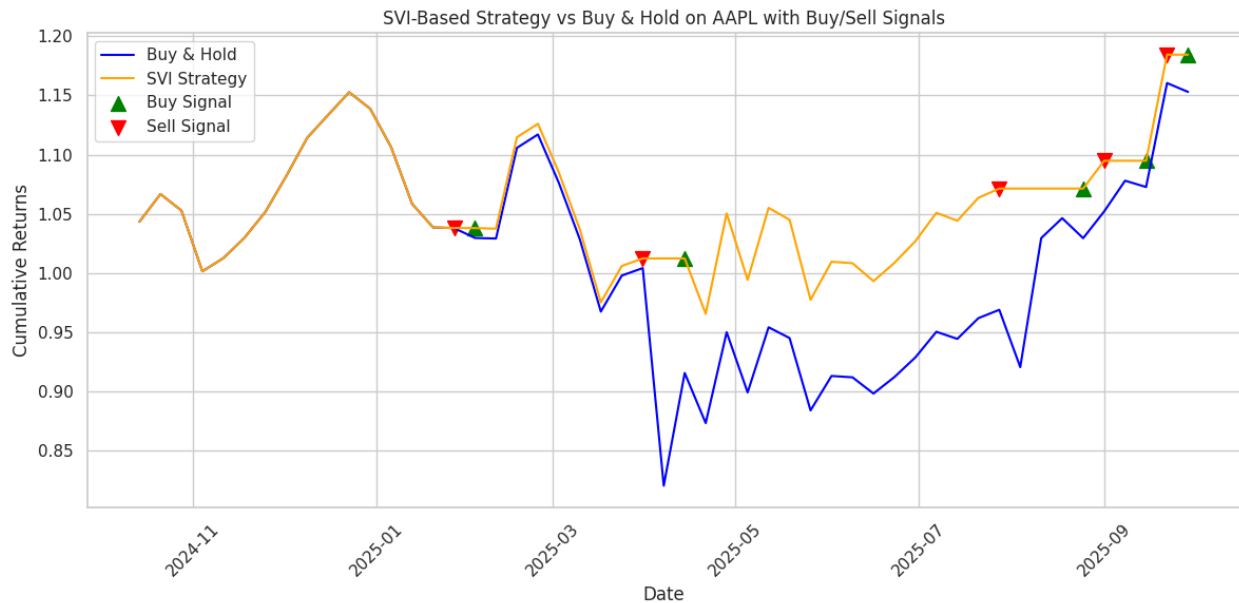

Figure 3: AAPL Stock Price

Figure 4: SVI vs Buy and Hold onAAPL

$$\text{Correlation Matrix: SVI vs AAPL Stock Price} = \begin{bmatrix} 1.000 & -0.129 \\ -0.129 & 1.000 \end{bmatrix}$$

Table 1: Correlation Matrix: SVI vs AAPL Stock Price

	SVI	Adj_Close
SVI	1.000	-0.129
Adj_Close	-0.129	1.000

2.7 Short Literature Review

Our analysis shows a **negative correlation** between **Search Volume Index (SVI)** and **AAPL stock price** (**Correlation = -0.129**), which is consistent with some **behavioral finance literature**. Existing literature explains how **spikes in market participant attention** are accompanied by **uncertainty and panic in the market**, causing prices to decline.

Da, Engelberg & Gao (2011): They demonstrated how **increased Google searches** on specific financial topics **preceded and predicted market volatility**, as well as larger than expected price movements, thus supporting the idea that **attention of market participants is indicative of the short-term price movement potential** of the underlying asset.

Preis, Moat & Stanley (2013): Showed how **SVI for stock price related financial query searches** can **predict significant market variations**, including price drops.

According to **Shannon's Information Theory**, informationally "**worried**" investors look for "**risky**" assets on the internet, spiking **SVI**. Though **attention signals** are external and can help **speculative strategy formation**, they **lack direct profitability** due to **noise and ambiguous intent**.

Evidence from our **correlation matrix** confirms this: as **interest increases (SVI increases)**, the **price of the underlying stock declines slightly**. This is consistent with the **negative correlation observed** in these studies. This correlation is **weak** as expected in our single stock example; however, the literature suggests...

Conclusion : The empirical evidence reinforces that SVI is a viable behavioral proxy for investor sentiment. Its high frequency and accessibility make it useful for lightweight trading strategies and market analysis, although careful handling is required to avoid lookahead bias, overfitting, and misinterpretation of noisy search data.

References

iShares, BlackRock (2025). *iShares Core S&P 500 ETF (IVV) Overview and Holdings*. Online report. Accessed October 2025. URL: `https://www.ishares.com/us/products/239726/ishares-core-sp-500-etf`.

Sagaceta-Mejía, Alma Rocío, Máximo Eduardo Sánchez-Gutiérrez, and Julián Alberto Fresán-Figueroa (2024). In: *Economics* 18.1, p. 20220073. DOI: `doi:10.1515/econ-2022-0073`. URL: `https://doi.org/10.1515/econ-2022-0073`.

Vasileiou, Evangelos and Polydoros Tzanakis (2024). "The impact of Google searches, Put-Call ratio, and Trading Volume on stock performance using Wavelet Coherence analysis: The AMC case". In: *Journal of Behavioral Finance* 25.1, pp. 111−119.