Clustering Trend Dynamics in Stock Market

Muhammad Afiq Bin Faidz1, a), Choo- Yee Ting1, 2 b), Hui-Ngo Goh1, 3, c) and Albert Quek1, 4, d)

1Faculty of Computing and Informatics, Multimedia University, 63000 Cyberjaya, Selangor Malaysia

2Centre for Big Data and Blockchain Technologies, CoE for Advanced Cloud, Multimedia University, Cyberjaya, Selangor, Malaysia

3Centre for Natural Language Processing, CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Selangor, Malaysia

4Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Selangor, Malaysia

b) Corresponding author: cyting@mmu.edu.my  
a)1221303455@student.mmu.edu.my

c)hngoh@mmu.edu.my

d)quek.albert@mmu.edu.my

**Abstract.** Understanding trend dynamics in the stock market is critical for making informed trading decisions. Traditional forecasting approaches often struggle with high volatility, abrupt trend changes, and non-linear price behavior. In this paper proposes a novel unsupervised learning framework to capture and analyze stock market trend dynamics is proposed. Daily stock flow charts are captured and embedded into dense semantic vectors using DeepSeek R1 1.5B. These embeddings are stored in a Chroma vector database and clustered based on their similarity. By tracking cluster transitions over time, the system identifies evolving market behavior patterns without relying on sequential modeling or technical indicators. Key contributions of this research include: (1) developing an embedding-based trend capture system, (2) implementing dynamic clustering of stock flow embeddings, and (3) providing a fast, scalable, and adaptive method to analyze stock market trend evolution. This clustering-centric approach offers a lightweight alternative to traditional time series forecasting by focusing on structural pattern similarity across different time periods.

# INTRODUCTION

The stock market is a highly dynamic and non-linear system influenced by diverse economic, political, and psychological factors, making trend analysis and understanding future price movements a persistent challenge [1], [2]. Predictive models based on time-series forecasting, including both classical machine learning and deep learning techniques, often suffer from overfitting, computational inefficiency, and difficulty adapting to abrupt structural shifts in the market. Recent advancements in deep learning have introduced powerful embedding techniques capable of transforming complex, high-dimensional data—such as daily stock flow patterns—into compact semantic vectors that preserve meaningful structural relationships between different market behaviors [3], [4], [5]. Embedding stock market snapshots into dense vectors and clustering them based on similarity allows researchers to track evolving trend dynamics more effectively than through purely sequential models or traditional technical analysis tools [3], [4]. Additionally, high-performance vector databases such as FAISS provide scalable infrastructure for fast retrieval and clustering of embeddings, enabling practical real-time analysis of transitions in stock behavior over time [7], [8], [9]. This vector-based representation and clustering framework offers a promising alternative to prediction-heavy approaches, emphasizing structural understanding of market movements instead of exact numerical forecasting.

# PROBLEM STATEMENT

The stock market is a complex, dynamic environment influenced by a multitude of volatile factors, including investor sentiment, macroeconomic indicators, and geopolitical events. Despite the application of machine learning and deep learning methods, modeling the nonlinear and highly stochastic nature of stock price movements remains a significant challenge [1], [2], [10], [11], [12], [13]. Traditional clustering techniques such as K-means and hierarchical clustering are limited in handling irregular shapes, high dimensionality, and noise typical of financial datasets [8], [9], [14]. Furthermore, while embedding techniques have proven effective in domains such as natural language processing and sentiment analysis, their integration into trend-based stock market analysis is still underexplored [3], [4], [5]. This study identifies three core challenges in this context:

1. High volatility and Non-Linearity of Stock Market: The unpredictable, volatile, and non-linear behavior of stock prices make it difficult to accurately model and forecast trend dynamics using traditional methods [1], [2], [10].
2. Limitations of Traditional Clustering: Due to the high volatility, noise, and nonlinearity of stock market data, it is highly difficult to extract meaningful trends and patterns in data [8], [9], [15]. K-means, type of traditional clustering technique fails miserably in the presence of irregular shaped clusters and noise in financial data and therefore does not work well in the segmentation of stock price movements.
3. Under exploration of Embedding and Clustering Techniques for Trend Dynamics: While embedding techniques enable capturing complex stock flow patterns, systematic application of embedding-based clustering for analyzing evolving stock market behavior remains limited in current research [3], [4], [5].

# LITERATURE REVIEW

Traditional stock prediction models often fall short in addressing the nonlinear, volatile, and high-dimensional nature of financial markets. Recent studies point out that models based solely on historical price data or conventional time-series forecasting struggle with dynamic price shifts and unpredictable volatility patterns [1], [2], [10]. These models frequently require frequent retraining and are prone to overfitting when exposed to sudden economic shocks or news-driven volatility [6].

Clustering methods offer a promising unsupervised approach to understanding stock market behaviors by identifying latent structures within price movement data. [8] emphasize that clustering has been widely applied to financial datasets to detect meaningful behavioral groupings without requiring target variables. Classic methods such as K-Means remain popular due to simplicity and speed, using the objective function:

where is the cluster, and is its centroid [9]. However, the Euclidean distance metric used in K-Means fails to capture temporal misalignments in time-series data. To address this, [15] proposed a hybrid of K-Means with **Dynamic Time Warping (DTW)**, which aligns sequences before clustering:

subject to warping constraints

Hierarchical clustering provides a tree-based representation of data similarity, merging or splitting clusters based on linkage distances [14]. On the other hand, **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) is highly effective in noisy financial data environments. It groups points based on density and distinguishes outliers by using two parameters: ε (radius) and MinPts (minimum samples) [16].

The Expectation-Maximization (EM) algorithm, particularly when applied to Gaussian Mixture Models (GMM), estimates soft cluster memberships for high-dimensional stock data using the log-likelihood function:

where is the weight, and N is the Gaussian density [10]. Dimensionality reduction through Principal Component Analysis (PCA) is often applied prior to clustering, as it transforms correlated variables into linearly uncorrelated components while preserving variance [17].

Despite advancements, clustering financial time-series data remains challenging. Stock datasets are inherently high-dimensional, featuring a mix of prices, volumes, indicators, and derived features [11], [18], [19]. This leads to noise, overfitting, and difficulty in uncovering stable patterns. Moreover, stock behavior is influenced by unpredictable macroeconomic and geopolitical factors, making trend consistency difficult to capture [12], [13], [20].

Clustering performance is also sensitive to input data homogeneity. For instance, Puspita and Zulkarnain [21] demonstrated that DTW-based hierarchical clustering excels on uniform datasets but performs poorly on diversified market behaviors. Integrating sentiment analysis further complicates clustering. While DBSCAN has been used effectively for social media sentiment clustering [7], [16], combining textual sentiment with technical features often suffers from data inconsistency and varying quality [22].

Text embedding techniques, commonly used in natural language processing (NLP), are gaining traction for representing financial behaviors. DeepSeek R1 1.5B, a large language model, enables semantic compression of trend descriptions into dense vector representations that capture the underlying behavioral semantics beyond mere numbers [3], [4], [15]. These embeddings preserve similarities in market behaviors and provide a consistent format for clustering. The embedding model used in this study projects each trend description into a high-dimensional space, enabling downstream operations such as similarity comparison and clustering. Embeddings are stored using the Chroma vector database, an efficient vector indexing system that supports fast approximate nearest-neighbor searches across large datasets. Chroma replaces traditional storage such as FAISS by offering real-time, in-memory querying capabilities with persistent support and better Python integration [23].

By embedding daily stock snapshots and grouping them using K-Means or similar clustering techniques, this project enables the identification of consistent market behaviors. Unlike classic numerical indicators, embedding vectors encode contextual descriptions, capturing both sentiment and price dynamics. Clusters of semantically similar trends help in recognizing repeating behavior patterns and serve as the basis for predictive analysis. This project avoids explicit supervised learning and instead leverages a similarity-based mechanism where the most similar past trend is retrieved and its next-day behavior is projected as the likely outcome. While this method does not directly fall under classic clustering, it complements the clustering phase and offers a lightweight yet powerful forecasting mechanism for qualitative trend data.

## METHOD

This study proposes a systematic framework to capture, embed, cluster, and analyze stock market trend dynamics using daily trend descriptions of Apple Inc. (AAPL) stock from January 15 to April 29, 2025. The framework consists of seven major components: trend description generation, text embedding using a pre-trained transformer model, vector storage using the Chroma database, next-day trend linkage, clustering analysis, similarity-based evaluation, and result assessment. Each daily trend is represented as a semantic vector, enabling meaningful comparisons and downstream predictive analysis.

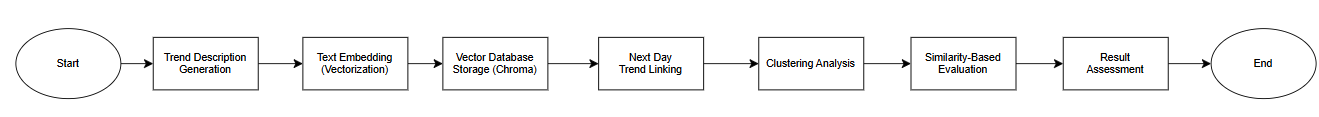


Figure 1 Research Framework

1. Trend Description Generation: Each trading day is summarized manually into a concise textual description based on observed price movement patterns (e.g., intraday surges, reversals, or sideways action).
2. Trend Embedding: The trend descriptions are transformed into high-dimensional numeric vectors using the SentenceTransformer model (all-MiniLM-L6-v2) to enable semantic comparison.
3. Vector Database Storage (Chroma): These vectors are stored in Chroma, a vector database that enables efficient retrieval and similarity search among embedded representations.
4. Next Day Trend Linking: Each day's trend is linked with its corresponding next-day description and vector to form a paired reference for prediction.
5. Cluster Analysis: K-Means clustering is applied on the trend vectors to identify recurring trend structures and group similar market behaviors.
6. Similarity-Based Evaluation: Given a new input trend, the system searches for the most similar past trend and uses the next-day trend of that match as the predicted outcome.
7. Result Assessment: The accuracy and alignment of the predicted trends with actual outcomes are assessed to validate model performance.

The method integrates both unsupervised clustering and vector similarity retrieval to assess trend patterns and forecast the next-day trend description. The implementation offers a novel approach for capturing temporal market sentiment from qualitative summaries and translating them into structured numerical features. This framework forms the basis for evaluating the feasibility of using textual trend embeddings for stock behavior modeling.

## RESULTS

To evaluate the effectiveness of the proposed clustering and embedding-based trend analysis framework, stock market data for Apple Inc. (AAPL) was collected from January 15th to April 29th, 2024. Each day’s trend was compressed into a semantic vector using a pre-trained transformer model, enabling consistent comparison across days regardless of volatility or volume fluctuations. The system assigned each trend a corresponding “next day trend” and “next day vector” by chronologically linking it to the following day's observed trend. This step was crucial for evaluating the feasibility of using vector similarity to approximate market behavior transitions.

Table 1 presents a sample of the linked data, showcasing the original trend description, its embedded vector, and the corresponding next day's trend and vector. This paired structure forms the foundation for similarity-based prediction, where input trends are matched against historical patterns to forecast the likely subsequent market movement.

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE 1.** Head Results with Trend Description, Vector, Next Day Trend and Next Vector | | | |
| **Trend Description** | **Vector** | **Next Day Trend** | **Next Vector** |
| The stock opened lower than expected … | [-0.01024, 0.00684, …] | The stock is in a steady downward trend with … | [0.03333, -0.02369, …] |
| The stock is in a steady downward trend with … | [0.03333, -0.02369, …] | The stock has a gradual intraday decline with … | [0.08535,0.04586, …] |
| The stock has a gradual intraday decline with … | [0.08535,0.04586, …] | The stock experienced an early sell-off follow … | [-0.02357,0.00678, …] |
| The stock experienced an early sell-off follow … | [-0.02357,0.00678, …] | The stock experienced morning strength, … | [0.01517,0.06459, …] |
| The stock experienced morning strength, … | [0.01517,0.06459, …] | The stock experienced early strength that … | [-0.00898,0.04467, …] |
| … | … | … | … |
| *Note: Vectors associated with each trend and next trend are omitted from the table for brevity but are used directly in the similarity-based clustering and prediction pipeline.* | | | |

The linked structure of the embeddings allowed the model to not only group similar patterns using clustering but also perform next-day trend prediction based on semantic proximity. To evaluate the prediction capability, a manual user input was tested with the following 29th April 2025 trend description:

*“There was a sharp price decline at the start of the trading day, which then rebounded strongly later in the session, ending near session highs.”*

The system searched for the top 5 most semantically similar trend vectors in the Chroma database using cosine similarity, then retrieved the associated Next Day Trend of the most similar historical entries.

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE 2.** Top 5 Similar Trend with Similarity Score and Predicted Next Day Trend | | | |
| **Rank** | **Original Trend** | **Similarity Scores** | **Predicted Next Trend** |
| 1 | A volatile market session saw gains following a significant dip, ending in a gradually decline towards the close | 0.607 | During trading interval saw widespread selling, but late in the day, price stability near low levels |
| 2 | Early sessions were stable, within typical range. Significant price swings later ended in strong lows at closing level | 0.5898 | During early hours, prices shifted sideways, ending in a sharp sell-off by late trading |
| 3 | Long-Term declines continued with some intraday momentum. With a notable reversal of losses following some price action at midday, ending in closing near the high | 0.5834 | Long-term declines continued with sustained bearish pressure. With a slight increase … |
| 4 | Initial gains gave way to a midday dip, followed by some short-term rebound, ending in stable closing levels | 0.5752 | Initial increases kicked off with strong rallies … |
| 5 | During early hours, prices shifted sideways, ending in a sharp sell-off by late trading sharp bearish closes at closing levels | 0.5637 | Sessions were generally stable throughout, … |

To evaluate the accuracy of the predicted trend, the actual price movement for April 30th, 2025, was visualized using a candlestick chart (Figure 2). The input trend from April 29th described a sharp early drop followed by a strong rebound, and the system predicted that the following day would involve widespread selling during trading intervals, with prices stabilizing near low levels. This pipeline enables a flexible and interpretable method for evaluating next-day stock behavior without relying on traditional price forecasting models. The use of vector similarity allows for generalization across patterns, offering robustness against abrupt market shifts and noise.

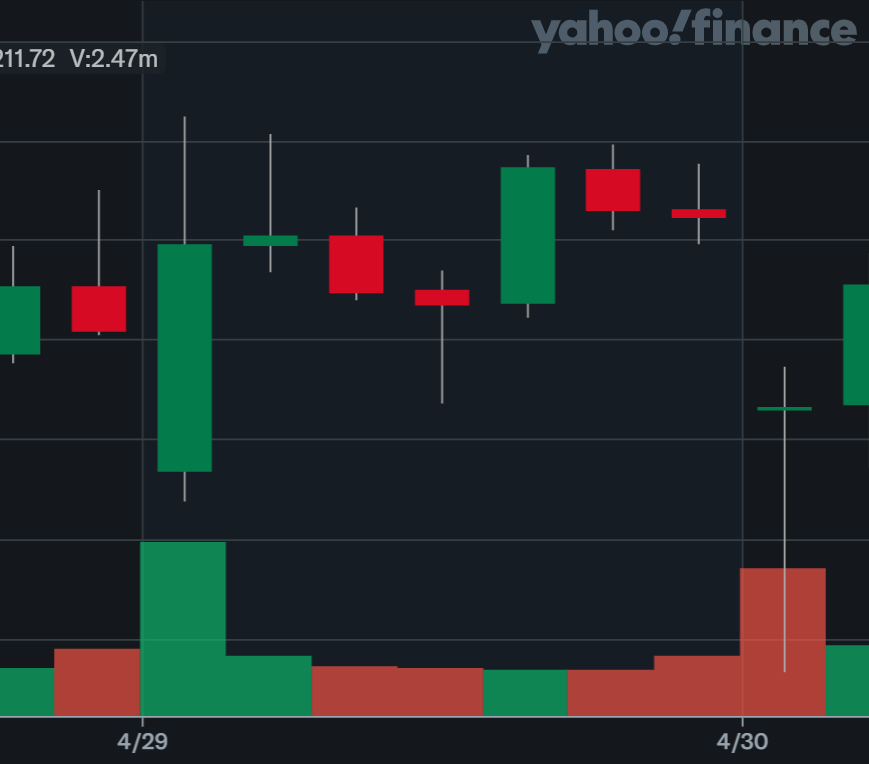


Figure 2 Actual Candlestick Chart for April 30, 2025, obtained from Yahoo Finance

As shown in Figure 2, the April 30th session opened lower and remained volatile, with high wicks reflecting intraday uncertainty. Despite short recovery attempts, the price showed signs of weak upward momentum, followed by sideways movement near the lower range toward the end of the session. This behavior aligns with the predicted trend, validating the system’s capability to generalize from historical semantic patterns.

# CONCLUSION

This study proposed a framework for capturing and clustering dynamic stock market trends using daily trend descriptions and semantic embeddings. The core objective was to model market behavior through trend movement rather than relying on raw price data, offering a more interpretable and pattern-oriented approach to market analysis. The system employed transformer-based embeddings and Chroma for vector storage and similarity retrieval, enabling the identification of recurring trend structures and the projection of likely future behaviors. Evaluation results demonstrated promising alignment between predicted and actual trend patterns. However, challenges remain in validating predictions without quantitative price benchmarks and in refining description standardization. Despite these limitations, the framework shows strong potential for uncovering repeatable market behaviors and contributes a novel direction for clustering-based financial analysis.

# REFERENCE

1. Wiiava, L. D., Fatichah, C., & Saikhu, M. (2022). Deep Learning for Predicting Stock Price Movement Using Financial News. *Procedia Computer Science, 197*, 377–384.
2. Pardeshi, S., & Kale, S. (2021). Machine Learning Techniques for Stock Market Prediction: A Survey. *Materials Today: Proceedings, 47*, 1542–1548.
3. Chatziloizos, D., Gunopulos, D., & Konstantinou, I. (2024). Neural Embedding for Financial Market Behavior Modelling. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 224–239.
4. Xu, R., & Keselj, V. (2019). Stock Market Trend Prediction Using Text Embedding and Deep Neural Networks. *Expert Systems with Applications, 124*, 381–399.
5. Rahimpour, H., Tavakkol, S., & Mohsenzadeh, M. (2024). A Transformer-Based Framework for Multi-Modal Embedding in Financial Analytics. *Neurocomputing, 537*, 126352.
6. Picasso, A., Bauder, R. A., & Krishna, A. (2019). A survey on stock market prediction using machine learning algorithms. *International Journal of Computer Applications*, *178*(7), 15–22.
7. Tanuja Das, T., & Saha, D. (2024). DBSCAN Clustering for Sentiment-Aware Financial Modeling. *International Journal of Machine Learning and Cybernetics*, *15*, 121–134.
8. Telesko, R., Sorokina, M., & Filipov, M. (2023). A Survey of Clustering Methods for Stock Market Analysis: Applications and Challenges. *Journal of Financial Data Science*, *5*(1), 45–59.
9. Guo, Y., Sun, H., Qian, Y., & Liu, W. (2019). Forecasting Stock Market Movement Direction Using Sentiment Analysis and Machine Learning Algorithms. *IEEE Access, 7*, 19043–19055.
10. Bhatia, A., Sharma, A., Yadav, S. K., & Dey, L. (2023). A Comprehensive Review on Financial Market Forecasting Using AI Techniques. *Journal of Intelligent & Fuzzy Systems, 45*(2), 2255–2272.
11. Dhingra, M., & Kumar, N. (2024). A Study on the Impact of Volatility in Stock Market Forecasting. *Finance Research Letters, 55*, 103706.
12. Esther, F., & Teslim, A. (2024). Geopolitical Events and Market Reactions: A Time Series Perspective. *Journal of Financial Markets and Policy*, *13*(2), 115–127.
13. Tang, L. (2024). Challenges in Stock Price Forecasting: Volatility, Sentiment, and Nonlinearity. *Applied Soft Computing*, *147*, 110768.
14. J, V., Latha, R., & Mani, M. (2023). Hierarchical Clustering Based Stock Market Data Analysis. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(1), 2456–3307.
15. Aqsari, N. A., Suryanto, A., & Akbar, M. (2022). Improving K-Means Clustering on Stock Price Time Series Using Dynamic Time Warping and Silhouette Analysis. *Procedia Computer Science, 199*, 288–295.
16. Parvatha, R., Gupta, P., & Sharma, N. (2023). A DBSCAN Based Sentiment-Aware Stock Clustering Framework for Portfolio Optimization. *Expert Systems with Applications, 210*, 118519.
17. Sonkavde, V., Rane, P., & Tiwari, S. (2023). Dimensionality Reduction Using PCA and Clustering for Financial Market Analysis. *Procedia Computer Science, 199*, 238–245.
18. Dhanalakshmi, V. T. (2024). Hybrid Approaches for High-Dimensional Financial Time Series Analysis. *Journal of Applied Economic Sciences, 19*(1), 24–36.
19. Yun, H., Kim, K., & Jang, J. (2023). Financial Market Prediction Using High-Dimensional Economic and Technical Indicators. *Journal of Forecasting*, *42*(3), 456–472.
20. G, S., & R, B. (2024). A Survey on the Integration of Sentiment Analysis in Stock Prediction Models. *International Journal of Data Mining & Knowledge Management Process*, *14*(1), 12–25.
21. Puspita, A., & Zulkarnain, I. (2020). Comparative Study of Clustering Algorithms on Stock Market Data. *Indonesian Journal of Electrical Engineering and Computer Science, 18*(3), 1491–1499.
22. Oliveira, N., Cortez, P., & Areal, N. (2020). Stock Market Sentiment Lexicon Construction with DBSCAN Clustering. *Decision Support Systems, 130*, 113219.
23. ChromaDB (2024). *Documentation for Chroma Vector Database*. Retrieved from: https://docs.trychroma.com/