TRƯỜNG ĐẠI HỌC CÔNG NGHỆ - ĐHQGHN
VIỆN TRÍ TUỆ NHÂN TẠO

KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU
BÁO CÁO

Phương pháp xác định vùng mua bán của tài sản số với tỷ suất lợi nhuận cao

Thành viên nhóm:	Phạm Nhật Quang	23020413
	Bùi Minh Quân	23020415
	Phan Quang Trường	23020443
	Hồ Lê Dương	22022641
Giảng viên hướng dẫn:	GS. Nguyễn Phương Thái	
	ThS. Ngô Minh Hương	

Hà Nội, 2025

Link mã nguồn: github.com/Bui-Minh-Quan/Data-Mining-Project
Link dữ liệu: kaggle.com/datasets/minhqunbi/data-mining-dataset

Mục lục

Danh mục từ viết tắt và thuật ngữ

1. Danh mục từ viết tắt

Viết tắt	Tiếng Anh	Tiếng Việt / Giải nghĩa
BTC	Bitcoin	Đồng tiền mã hóa có vốn hóa lớn nhất thị trường.
OHLCV	Open-High-Low-Close-Volume	Dữ liệu nến bao gồm giá mở cửa, cao nhất, thấp nhất, đóng cửa và khối lượng giao dịch.
EMA	Exponential Moving Average	Đường trung bình động lũy thừa, đặt trọng số cao hơn vào dữ liệu giá gần nhất.
RSI	Relative Strength Index	Chỉ số sức mạnh tương đối, đo lường tốc độ và sự thay đổi của biến động giá.
ATR	Average True Range	Khoảng dao động trung bình thực tế, thước đo độ biến động của thị trường.
SMC	Smart Money Concepts	Phương pháp phân tích dựa trên hành vi giao dịch của các tổ chức tài chính lớn (dòng tiền thông minh).
OB	Order Block	Khối lệnh - vùng giá có khối lượng giao dịch lớn, nơi các tổ chức để lại dấu vết.
FVG	Fair Value Gap	Khoảng trống giá trị - vùng mất cân bằng thanh khoản trên biểu đồ giá.
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên.
TF-IDF	Term Frequency-Inverse Document Frequency	Kỹ thuật thống kê đánh giá tầm quan trọng của một từ trong văn bản.
PCA	Principal Component Analysis	Phân tích thành phần chính (dùng để giảm chiều dữ liệu).

2. Giải thích thuật ngữ

Thuật ngữ	Giải thích
Bullish / Bull Trend	Xu hướng tăng giá hoặc tâm lý lạc quan của thị trường.
Bearish / Bear Trend	Xu hướng giảm giá hoặc tâm lý bi quan của thị trường.
Pump	(Trong phạm vi đề tài) Tín hiệu giá tăng mạnh và chạm mức chốt lời (Upper Barrier) trước khi chạm mức cắt lỗ.
Dump	(Trong phạm vi đề tài) Tín hiệu giá giảm mạnh và chạm mức cắt lỗ (Lower Barrier) trước.
Sideways	Trạng thái thị trường đi ngang, biên độ dao động hẹp, không rõ xu hướng tăng hay giảm.
Shock Volume	Khối lượng giao dịch đột biến (gấp nhiều lần trung bình), thường báo hiệu sự tham gia của dòng tiền lớn.
Look-ahead Bias	Lỗi "nhìn trước tương lai" trong kiểm thử mô hình, xảy ra khi sử dụng thông tin chưa xảy ra tại thời điểm dự báo để ra quyết định.
Overfitting	Hiện tượng quá khớp, khi mô hình học quá kỹ các nhiễu trong dữ liệu huấn luyện và hoạt động kém trên dữ liệu mới.
Backtest	Phương pháp kiểm thử chiến lược giao dịch trên dữ liệu lịch sử để đánh giá hiệu quả.
Drawdown	Mức sụt giảm tài khoản lớn nhất tính từ đỉnh vốn xuống đáy vốn trong một khoảng thời gian nhất định.

Chương 1

Giới thiệu

1.1 Bối cảnh và vấn đề nghiên cứu

Thị trường tiền mã hóa, với đại diện tiêu biểu là Bitcoin, đã và đang trở thành một trong những kênh đầu tư tài chính sôi động và thu hút dòng tiền lớn nhất toàn cầu trong thập kỷ qua. Khác với thị trường chứng khoán truyền thống, thị trường tiền mã hóa hoạt động 24/7 với biên độ dao động giá cực lớn, tạo ra cơ hội sinh lời hấp dẫn nhưng đồng thời cũng tiềm ẩn rủi ro thanh lý tài sản rất cao. Trong bối cảnh đó, sự biến động của giá không chỉ được định hình bởi quy luật cung cầu đơn thuần mà còn chịu tác động mạnh mẽ từ tâm lý đám đông, các sự kiện kinh tế vĩ mô và luồng tin tức truyền thông liên tục. Điều này đặt ra một bài toán thách thức cho các nhà đầu tư: Thách thức đặt ra là mô hình hóa các biến động phi tuyến tính và nhiễu của chuỗi thời gian tài chính để trích xuất các mẫu hình có ý nghĩa thống kê.

Vấn đề nghiên cứu của đề tài không nằm ở việc cố gắng dự đoán chính xác giá trị tương lai của Bitcoin, một nhiệm vụ được xem là bất khả thi trong lý thuyết bước đi ngẫu nhiên. Thay vào đó, trọng tâm của nghiên cứu là tìm kiếm các mẫu hình lặp lại trong lịch sử và lượng hóa chúng thành các luật giao dịch dựa trên xác suất thống kê. Bằng cách kết hợp dữ liệu lịch sử giá với dữ liệu phi cấu trúc từ tin tức, chúng em mong muốn xây dựng một hệ thống khai phá dữ liệu có khả năng nhận diện các điều kiện thị trường đặc thù. Khi các điều kiện này hội tụ, xác suất giá di chuyển theo một hướng nhất định sẽ cao hơn ngẫu nhiên, từ đó cung cấp cơ sở định lượng để ra quyết định mua hoặc bán một cách khoa học.

1.2 Lý do chọn đề tài

Lý do cốt lõi thúc đẩy nhóm lựa chọn đề tài này xuất phát từ chính niềm đam mê và trải nghiệm thực tế của các thành viên đối với thị trường tài chính. Với nền tảng kiến thức sẵn có về đầu tư chứng khoán và tiền mã hóa, nhóm nghiên cứu nhận thấy rằng phần lớn các quyết định giao dịch của nhà đầu tư cá nhân thường bị chi phối bởi cảm xúc hoặc những nhận định chủ quan thiếu cơ sở kiểm chứng. Trong khi đó, bản chất của giao dịch tài chính thành công là cuộc chơi của xác suất và quản trị rủi ro. Do đó, việc áp dụng các kỹ thuật Khai phá dữ liệu, cụ thể là các phương pháp truyền thống nhưng mạnh mẽ về mặt giải thích như phân cụm và Khai phá luật kết hợp, là bước đi cần thiết để chuyển hóa các kinh nghiệm đầu tư cảm tính thành các chiến lược có tính kỷ luật và có thể đo lường được.

Hơn nữa, việc lựa chọn phương pháp tiếp cận dựa trên luật kết hợp thay vì các mô hình học sâu phức tạp là một quyết định có chủ đích. Trong tài chính, khả năng giải thích của mô hình quan trọng không kém độ chính xác. Nhóm muốn tìm hiểu rõ **tại sao** một tín hiệu mua được đưa ra, liệu đó chỉ đơn thuần là do chỉ báo kỹ thuật, hay còn do tác động từ một chuỗi tin tức nào đó. Đề tài này, vì thế, không chỉ là một bài tập học thuật nhằm áp dụng các kỹ thuật xử lý dữ liệu lớn, mà còn là một nỗ lực nhằm xây dựng một công cụ hỗ trợ ra quyết định thực tế, phục vụ trực tiếp cho hoạt động đầu tư của chính các thành viên trong nhóm trong tương lai.

1.3 Đóng góp của đề tài

Đề tài tập trung giải quyết bài toán hỗ trợ ra quyết định giao dịch thông qua cách tiếp cận định lượng và khai phá dữ liệu. Các đóng góp chính của đề tài bao gồm:

- **Đề xuất quy trình tích hợp dữ liệu đa nguồn:** Xây dựng một đường ống xử lý dữ liệu đồng bộ hóa dữ liệu chuỗi thời gian-giá với dữ liệu phi cấu trúc như tin tức. Đặc biệt, nhóm áp dụng kỹ thuật chuẩn hóa động cho dữ liệu tin tức để giải quyết triệt để vấn đề nhìn trước tương lai thường gặp trong kiểm thử mô hình tài chính.
- **Định lượng hóa các khái niệm *smart money concepts (SMC)*:** Chuyển đổi các khái niệm phân tích kỹ thuật định tính phức tạp như khối lệnh và Khoảng trống giá trị thành các công thức toán học cụ thể, tạo tiền đề cho việc tự động hóa quá trình nhận diện tín hiệu.
- **Xây dựng hệ thống luật giao dịch có tính giải thích cao:** Thay vì sử dụng các mô hình hộp đen, nghiên cứu ứng dụng thuật toán Khai phá luật kết hợp để sinh ra các quy tắc giao dịch minh bạch. Điều này giúp nhà đầu tư hiểu rõ lý do đằng sau mỗi tín hiệu mua/bán, từ đó nâng cao niềm tin vào hệ thống.

1.4 Đối tượng và phạm vi nghiên cứu

1.4.1 Đối tượng nghiên cứu

- **Tài sản:** Bitcoin (BTC/USDT) trên thị trường giao ngay (Spot Market).
- **Dữ liệu:** Dữ liệu lịch sử giá nến phút và dữ liệu tiêu đề tin tức tiếng Anh liên quan đến thị trường Crypto.
- **Thuật toán:** Tập trung vào các kỹ thuật phân cụm (K-Means) và Khai phá luật kết hợp (FP-Growth).

1.4.2 Phạm vi nghiên cứu

- **Thời gian dữ liệu:** Dữ liệu được thu thập trong giai đoạn từ 01/01/2023 đến hết năm 2025.
- **Nguồn dữ liệu:** Dữ liệu giá được lấy từ sàn giao dịch Binance; dữ liệu tin tức được thu thập từ nền tảng tổng hợp CryptoPanic.
- **Giới hạn:** Đề tài chỉ tập trung vào việc phát hiện các điểm đảo chiều và xu hướng ngắn hạn, không bao gồm việc tối ưu hóa danh mục đầu tư hay giao dịch cao tần.

1.5 Cấu trúc báo cáo

Ngoài phần mở đầu và kết luận, báo cáo được tổ chức thành các chương như sau:

- **Chương 2:** Trình bày quy trình thu thập và tiền xử lý dữ liệu giá và tin tức.
- **Chương 3:** Mô tả chi tiết kỹ thuật tạo đặc trưng, bao gồm các chỉ báo kỹ thuật và SMC.
- **Chương 4:** Phân tích phương pháp phân cụm dữ liệu để xác định trạng thái thị trường và chủ đề tin tức.
- **Chương 5:** Trình bày quá trình khai phá luật kết hợp và xây dựng tập luật lai.
- **Chương 6:** Thực nghiệm, đánh giá hiệu quả mô hình qua backtest và phân tích kết quả.
- **Chương 7:** Tổng kết các kết quả đạt được và đề xuất hướng phát triển.

Chương 2

Chuẩn bị dữ liệu

2.1 Dữ liệu giá Bitcoin

2.1.1 Nguồn dữ liệu và thu thập

Để đảm bảo tính chính xác và đồng nhất của dữ liệu thị trường, báo cáo sẽ sử dụng nguồn dữ liệu từ Binance – sàn giao dịch tiền mã hóa có thanh khoản lớn nhất thế giới. Dữ liệu được thu thập thông qua thư viện Python `binance_historical_data`.

Báo cáo tập trung khai thác dữ liệu nến ở thị trường giao ngay trên khung thời gian 1 phút. Việc chọn khung thời gian nhỏ giúp mô hình nắm bắt được các biến động nhỏ nhất, mà dữ liệu theo ngày hoặc giờ có thể bỏ qua.

Trong từng khung thời gian mỗi nến sẽ có đặc trưng dữ liệu dạng OHLCV, bao gồm:

1. Giá mở đầu nến (Open);
2. Giá cao nhất nến (High);
3. Giá thấp nhất nến (Low);
4. Giá đóng kết nến (Close);
5. Khối lượng giao dịch (Volume).

2.1.2 Tiền xử lý

Dữ liệu tải về được chia thành các file nhỏ theo tháng và ngày. Quy trình tiền xử lý bao gồm 3 bước:

- Tổng hợp dữ liệu
- Chuẩn hóa thời gian
- Phân chia dữ liệu

Một vấn đề kỹ thuật phát sinh trong quá trình gộp dữ liệu là sự không đồng nhất về định dạng thời gian: một số bản ghi sử dụng đơn vị milliseconds (10^{13}), trong khi số khác sử dụng microseconds (10^{16}). Vậy nên, ta cần đưa tất cả về cùng một đơn vị chuẩn. Cuối cùng, dữ liệu được chia thành tập huấn luyện (Train) và kiểm thử (Test) dựa trên mốc thời gian (bảng 2.1).

Khác với dữ liệu dạng bảng thông thường, dữ liệu chuỗi thời gian tài chính có tính phụ thuộc thời gian. Việc áp dụng phương pháp chia ngẫu nhiên sẽ dẫn đến hiện tượng nhìn trước được dữ liệu, khi mô hình học được các mẫu hình từ tương lai để dự đoán quá khứ. Do đó, nghiên cứu áp dụng phương pháp chia tách theo trình tự thời gian để mô phỏng chính xác môi trường giao dịch thực tế.

Bảng 2.1: Thống kê phân bố dữ liệu

Tập dữ liệu	Thời điểm bắt đầu	Thời điểm kết thúc	Số mẫu	Tỷ lệ
Train	2023-01-01 00:00	2024-12-31 23:59	1,052,560	68.25%
Test	2025-01-01 00:00	2025-12-06 23:59	489,600	31.75%

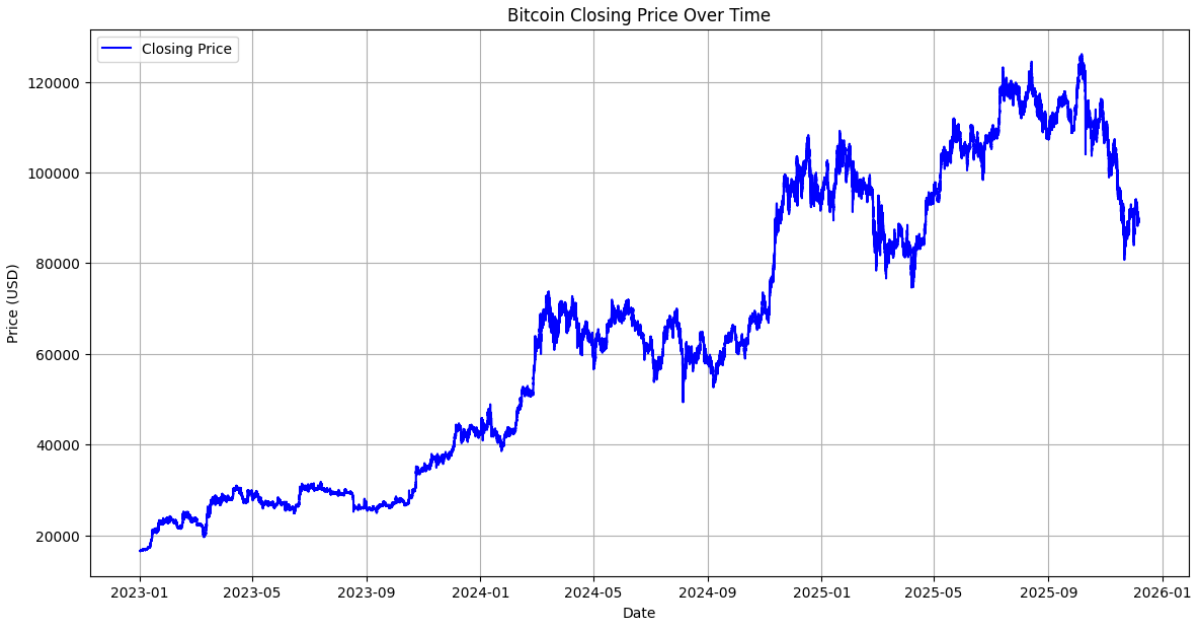
2.1.3 Đặc điểm dữ liệu

Bảng 2.2: Thống kê mô tả dữ liệu giá Bitcoin theo từng phút (2023-2025)

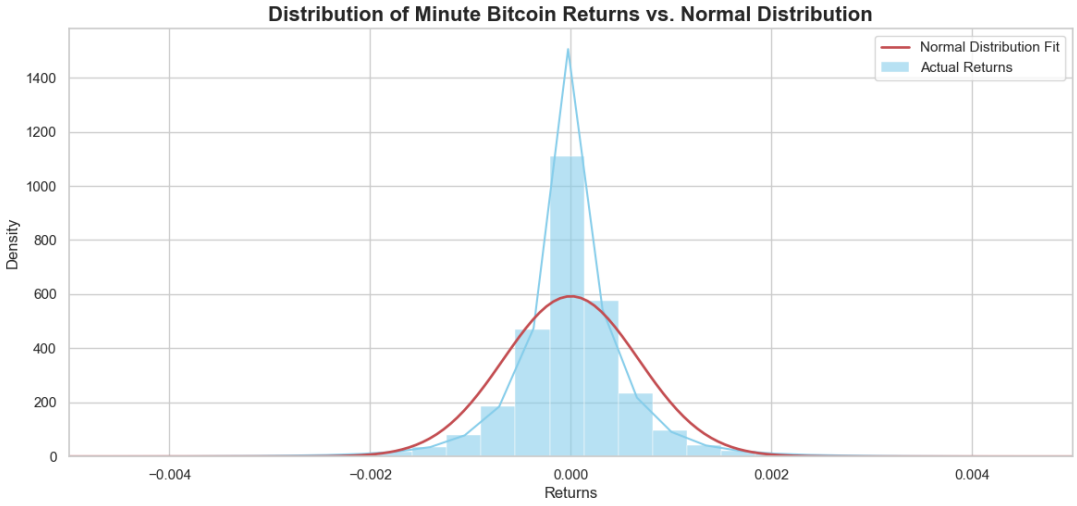
Thống kê	Giá đóng ($)	K.Lượng ($)	Lợi nhuận	Biến động nội
Trung bình	47,376.86	1,681,654	0.0002%	0.0654
Độ lệch chuẩn	21,659.57	2,969,644	0.0673%	0.0769
Min	16,505.87	1,121	-3.6800%	0.0000
Q1	27,728.17	346,496	-0.0253%	0.0211
Q2	42,696.65	771,385	0.0000%	0.0462
Q3	64,222.99	1,817,253	0.0257%	0.0849
Max	108,258.39	256,885,789	3.1600%	6.2983
Skewness	-	-	**-0.640**	**6.917**
Kurtosis	-	-	**82.958**	**171.868**

Sau khi thực hiện phân tích những đặc trưng cơ bản của dữ liệu (bảng 2.2), một số nhận xét có thể được rút ra như sau:

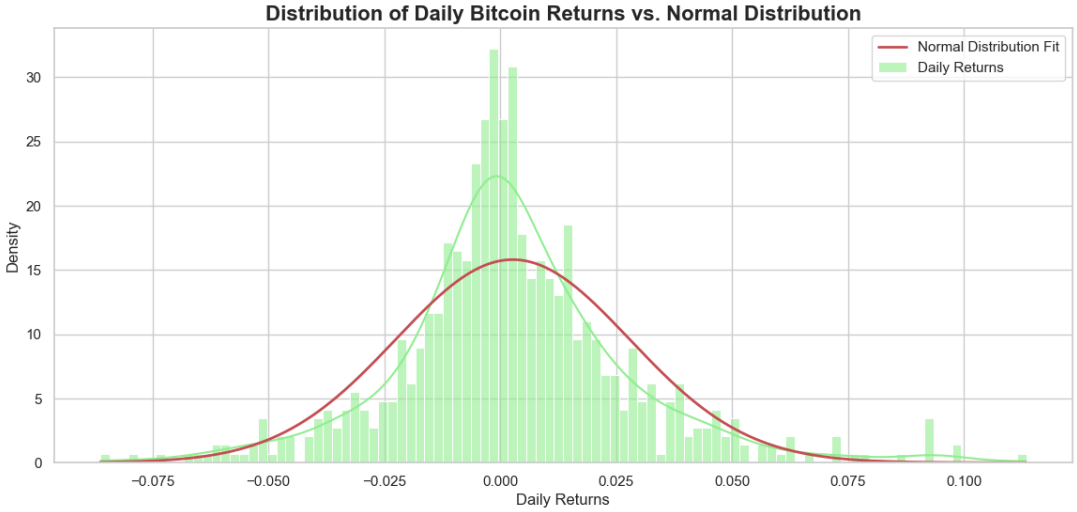
- Giá Bitcoin trải dài từ mức thấp nhất khoảng 16,500$ lên đến đỉnh điểm hơn 108,000$. Độ lệch chuẩn lớn (21,659$) cho thấy rủi ro biến động giá là rất cao.
- Chỉ số Kurtosis của lợi nhuận đạt mức cực đại (82.96), lớn hơn rất nhiều so với phân phối chuẩn. Điều này minh chứng cho khả năng xảy ra các sự kiện bất thường cao hơn nhiều so với phân phối chuẩn, tức là thị trường thường xuyên xuất hiện các cú sốc giá (tăng/giảm đột ngột) nằm ngoài dự đoán thông thường.
- Chỉ số Skewness âm (-0.64) cho thấy các phiên giảm giá thường có cường độ mạnh hơn so với các phiên tăng giá.



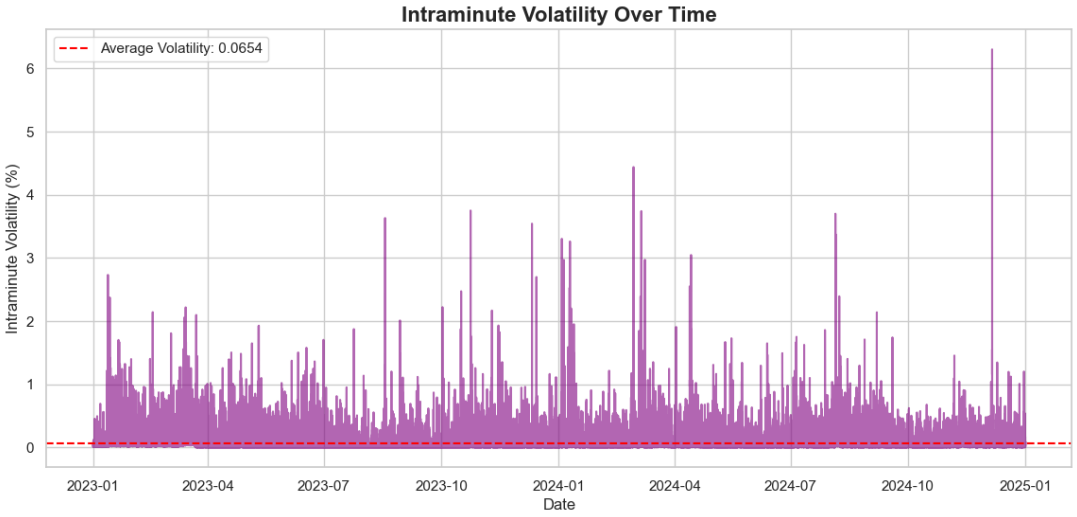
Hình 2.1: Giá đóng Bitcoin từ đầu năm 2023 đến nay



Hình 2.2: Biểu đồ Phân bố lợi nhuận theo phút của Bitcoin



Hình 2.3: Biểu đồ Phân bố lợi nhuận theo ngày của Bitcoin



Hình 2.4: Độ biến động nội tại theo phút của Bitcoin giai đoạn 2023-2025

Kiểm định tính dừng

Tính dừng là một giả định quan trọng trong nhiều mô hình thống kê và học máy. Chuỗi dữ liệu không dừng thường có các tham số thống kê (như trung bình, phương sai) thay đổi theo thời gian, gây khó khăn cho việc dự báo.

Nhóm đã sử dụng **Kiểm định Augmented dickey-fuller (ADF)** để đánh giá tính dừng của chuỗi giá đóng cửa (P_t) và chuỗi lợi nhuận (R_t). Kết quả kiểm định cho thấy chuỗi giá gốc (P_t) chấp nhận giả thuyết H_0 (chuỗi không dừng) với p-value > 0.05. Ngược lại, chuỗi lợi nhuận (R_t) bác bỏ H_0 với p-value ≈ 0, khẳng định tính dừng. Đây là cơ sở để nghiên cứu tập trung xây dựng đặc trưng dựa trên biến động tương đối (như RSI, ROC) thay vì sử dụng giá trị tuyệt đối trong các chương tiếp theo.

2.2 Dữ liệu tin tức (Dữ liệu văn bản)

2.2.1 Nguồn dữ liệu và thu thập

Đối với dữ liệu văn bản, báo cáo sử dụng bộ dữ liệu **Crypto News Dataset** [9]. Đây là tập hợp hơn 248,000 bản tin được thu thập từ *CryptoPanic* - nền tảng tổng hợp tin tức tiền mã hóa lớn nhất hiện nay.

Bộ dữ liệu bao gồm các thông tin liên quan đến 660 loại tiền mã hóa hàng đầu, với các trường dữ liệu:

- **Metadata:** `id`, `newsDatetime`, `sourceDomain`, `url`.
- **Nội dung:** `title` (tiêu đề), `description` (mô tả).
- **Phản ứng người dùng (User Reactions):** Các chỉ số cảm xúc được dán nhãn bởi cộng đồng như `positive`, `negative`, `important`, `toxic`, `liked`, `saved`, v.v.

Để phù hợp với phạm vi đề tài, ta tiến hành lọc và chỉ giữ lại các bản tin có trường `currencies` chứa mã "BTC" và nằm trong khung thời gian nghiên cứu (2023-2025).

2.2.2 Tiền xử lý văn bản

Dữ liệu văn bản thô từ web thường chứa nhiều nhiễu (HTML tags, URL, ký tự lạ) gây ảnh hưởng đến hiệu suất của thuật toán xử lý ngôn ngữ tự nhiên. Nhóm đã xây dựng hàm `clean_text` để chuẩn hóa dữ liệu tiêu đề.

Quy trình làm sạch bao gồm: Giải mã HTML, loại bỏ đường dẫn, xóa tên người dùng/hashtag và các ký tự phi ASCII.

2.2.3 Xử lý vấn đề dữ liệu thưa

Một thách thức lớn khi làm việc với dữ liệu tương tác người dùng là tính thưa (Sparsity). Thống kê sơ bộ cho thấy hầu hết các trường dữ liệu liên quan đến phản ứng của nhà đầu tư có giá trị bằng 0 chiếm tỷ trọng áp đảo (bảng 2.3).

Trường dữ liệu	Số lượng giá trị 0	Tỷ lệ (%)
"Toxic"	17,319	99.92%
"Disliked"	16,025	92.45%
"Negative"	15,865	91.53%
"Comments"	15,787	91.08%
"Lol"	15,281	88.16%
"Saved"	15,312	88.34%
"Important"	14,805	85.42%
"Liked"	14,247	82.20%
"Positive"	14,243	82.17%

Bảng 2.3: Thống kê tỷ lệ dữ liệu thưa trước khi xử lý

Trong bài toán Khai phá luật kết hợp, dữ liệu quá thưa sẽ dẫn đến việc thuật toán sinh ra hàng loạt luật tầm thường. Ví dụ: *Nếu Toxic = 0 thì Disliked = 0.* Những luật này, tuy có độ tin cậy cao nhưng không mang lại giá trị tri thức, làm lu mờ các luật quan trọng nhưng xuất hiện ít hơn.

Để khắc phục, ta áp dụng kỹ thuật **gộp đặc trưng**. Thay vì giữ nguyên các cột rời rạc, tổng hợp chúng thành các nhóm chỉ số có ý nghĩa bao quát hơn:

- `reaction_positive`: Tổng hợp các phản ứng tích cực (Positive + Liked + Important + Saved).
- `reaction_negative`: Tổng hợp các phản ứng tiêu cực (Negative + Disliked + Toxic).
- `total_reactions`: Tổng lượng tương tác.

Sau khi gộp, tỷ lệ dữ liệu thưa đã giảm đáng kể, giúp tập dữ liệu trở nên dày hơn, sẵn sàng cho việc xây dựng mô hình.

- `reaction_positive`: Tỷ lệ giá trị 0 giảm xuống còn **76.89%** (so với mức >82% của các cột thành phần).
- `reaction_negative`: Tỷ lệ giá trị 0 giảm xuống còn **85.94%** (cải thiện rõ rệt so với mức 99.92% của cột Toxic).
- `total_reactions`: Tỷ lệ giá trị 0 chỉ còn **72.53%**.

Chương 3

Kỹ thuật tạo đặc trưng dữ liệu

Mục tiêu cốt lõi của quá trình tạo đặc trưng trong nghiên cứu này là chuyển đổi dữ liệu thô từ nhiều nguồn khác nhau thành các tín hiệu định lượng giàu thông tin, giúp các thuật toán máy học có thể nhận diện được những mẫu hình tiềm ẩn. Đối với dữ liệu lịch sử giá, nếu chỉ sử dụng các giá trị OHLCV nguyên bản, thuật toán phân cụm sẽ có xu hướng gom nhóm dựa trên độ lớn của giá thay vì bản chất vận động của thị trường. Do đó, dữ liệu cần được chuyển đổi thành các đặc trưng đại diện cho xu hướng, động lượng và dấu vết của dòng tiền lớn để phục vụ việc xác định cấu trúc thị trường (ví dụ: thị trường đang tích lũy hay phân phối). Song song với đó, đối với dữ liệu tin tức, thách thức nằm ở việc lượng hóa ngôn ngữ tự nhiên thành các chỉ số cảm xúc và mức độ quan tâm của cộng đồng, từ đó tích hợp vào mô hình dự báo tổng thể.

Chương này trình bày chi tiết quy trình xây dựng đặc trưng cho hai nguồn dữ liệu trên, bắt đầu với các chỉ báo kỹ thuật và khái niệm tiền thông minh áp dụng cho giá Bitcoin, tiếp theo là kỹ thuật xử lý và tổng hợp đặc trưng từ dữ liệu văn bản.

3.1 Tạo đặc trưng với dữ liệu lịch sử giá Bitcoin

3.1.1 Các chỉ báo kỹ thuật

Nhóm các chỉ báo kỹ thuật được tính toán nhằm cung cấp cho mô hình góc nhìn định lượng về trạng thái hiện tại của thị trường. Nghiên cứu tập trung sử dụng bốn chỉ báo chính: Đường trung bình động lũy thừa, chỉ số sức mạnh tương đối, khoảng dao động trung bình thực tế và Khối lượng đột biến.

Đường trung bình động lũy thừa Việc lựa chọn EMA 50 và 200 dựa trên quy ước chung của các quỹ đầu tư để xác định xu hướng trung và dài hạn. Khác với đường trung bình động đơn giản (SMA), EMA đặt trọng số lớn hơn vào các dữ liệu giá gần nhất, giúp chỉ báo phản ứng nhạy bén hơn với các biến động mới. Công thức tính EMA tại thời điểm t được biểu diễn như sau:

$$EMA_t = \alpha \cdot P_t + (1 - \alpha) \cdot EMA_{t-1} \tag{3.1}$$

Trong đó, P_t là giá đóng cửa tại thời điểm hiện tại, và hệ số làm mượt $\alpha = \frac{2}{N+1}$ với N là chu kỳ của đường trung bình. Hai đường EMA chu kỳ 50 ($N = 50$) và 200 ($N = 200$) được thiết lập; vị trí tương đối của giá so với hai đường này sẽ cho biết thị trường đang trong xu hướng tăng hay giảm.

Chỉ số sức mạnh tương đối Chỉ báo RSI chu kỳ 14 được lựa chọn theo đề xuất nguyên bản của Wilder [**?**], đây là thiết lập tiêu chuẩn giúp cân bằng giữa độ nhạy và độ tin cậy của tín hiệu. RSI so sánh độ lớn của các đợt tăng giá so với giảm giá nhằm nhận diện các vùng quá mua hoặc quá bán. Chỉ số này được xác định dựa trên tỷ lệ sức mạnh tương đối (RS):

$$RSI = 100 - \frac{100}{1 + RS} \quad \text{với} \quad RS = \frac{\text{Mức tăng trung bình}}{\text{Mức giảm trung bình}} \tag{3.2}$$

Giá trị RS là tỷ số giữa mức tăng trung bình và mức giảm trung bình trong 14 phiên gần nhất. Khi $RSI > 70$, thị trường được xem là quá mua, và ngược lại khi $RSI < 30$ là vùng quá bán, nơi xác suất đảo chiều thường tăng cao.

Khoảng dao động trung bình thực tế Chủ yếu dùng để đo lường mức độ rủi ro và biến động. Trước hết, giá trị dao động thực của mỗi phiên được xác định là giá trị lớn nhất trong ba đại lượng: biên độ phiên hiện tại ($High - Low$), chênh lệch giữa giá cao nhất và giá đóng cửa phiên trước ($|High - Close_{prev}|$), và chênh lệch giữa giá thấp nhất và giá đóng cửa phiên trước ($|Low - Close_{prev}|$).

$$TR_t = \max(H_t - L_t, |H_t - C_{t-1}|, |L_t - C_{t-1}|) \tag{3.3}$$

Chỉ báo ATR sau đó được tính bằng cách lấy trung bình trượt của chuỗi TR trong 14 phiên. Một giá trị ATR cao cho thấy thị trường đang biến động mạnh, thường xuất hiện ở các giai đoạn bùng nổ hoặc sập giá, trong khi ATR thấp báo hiệu giai đoạn tích lũy nén giá.

Khối lượng đột biến Chỉ báo được sử dụng để phát hiện sự tham gia của các tổ chức tài chính lớn. Đây là một biến nhị phân, nhận giá trị 1 khi khối lượng giao dịch hiện tại (V_t) vượt quá hai lần mức trung bình của 20 phiên gần nhất ($MA_{20}(V)$), và nhận giá trị 0 trong các trường hợp còn lại:

$$ShockVol_t = \begin{cases} 1 & \text{nếu } V_t > 2 \times \text{MA}_{20}(V) \\ 0 & \text{nếu } V_t \leq 2 \times \text{MA}_{20}(V) \end{cases} \tag{3.4}$$

Sự đột biến về khối lượng này đóng vai trò như một tín hiệu xác nhận, gia tăng độ tin cậy cho các xu hướng giá đang diễn ra.

3.1.2 Khái niệm *Smart Money Concepts*

Bên cạnh các chỉ báo dao động truyền thống, báo cáo đi sâu vào việc trích xuất các đặc trưng hành vi dựa trên phương pháp *Smart Money Concepts*. Mục đích của phần này là tìm kiếm dấu vết của các nhà tạo lập thông qua việc phân tích cấu trúc nến và các vùng mất cân bằng cung cầu.

Quy trình bắt đầu bằng việc phân tích cấu trúc nến, để định lượng hành vi giá trong từng phiên giao dịch. Các thông số thành phần bao gồm kích thước thân nến ($Body$), độ dài râu nến trên ($Wick_{upper}$) và râu nến dưới ($Wick_{lower}$) được tính toán như sau:

$$\begin{aligned} Body &= |Close - Open| \\ Wick_{upper} &= High - \max(Open, Close) \\ Wick_{lower} &= \min(Open, Close) - Low \end{aligned}$$

Việc tách biệt này là tiền đề quan trọng, bởi lẽ râu nến dài thường thể hiện sự từ chối quyết liệt tại các vùng thanh khoản quan trọng. Dựa trên các thông số đó, ta suy ra được một số khái niệm.

Khối lệnh OB là vùng giá mà tại đó phe mua hoặc phe bán đã tham gia với khối lượng lớn, thường dẫn đến sự đảo chiều xu hướng. Một Khối lệnh tăng được xác định khi hội tụ đủ ba yếu tố: xuất hiện nến với râu dưới chiếm ưu thế, râu dưới dài hơn râu trên, và đi kèm với tín hiệu khối lượng đột biến ($ShockVol = 1$). Logic này được biểu diễn qua điều kiện:

$$\text{IsBullishOB} = (Wick_{lower} > 2 \cdot Body) \wedge (Wick_{lower} > Wick_{upper}) \wedge (ShockVol == 1) \tag{3.5}$$

Ngược lại, khối lệnh giảm được xác định khi nến có râu trên chiếm ưu thế, báo hiệu áp lực bán tháo mạnh:

$$\text{IsBearishOB} = (Wick_{upper} > 2 \cdot Body) \wedge (Wick_{upper} > Wick_{lower}) \wedge (ShockVol == 1) \tag{3.6}$$

Khoảng trống giá trị hợp lý FVG xuất hiện khi giá di chuyển quá nhanh về một hướng, tạo ra khoảng hở giữa râu của cây nến thứ nhất ($t-2$) và cây nến thứ ba (t). Về mặt thuật toán, một Bullish FVG được xác nhận khi giá thấp nhất của nến hiện tại cao hơn giá cao nhất của nến cách đó 2 phiên, với độ lớn khoảng trống vượt qua một ngưỡng nhiễu (δ):

$$\text{IsBullishFVG}_t = (L_t > H_{t-2}) \wedge ((L_t - H_{t-2}) > \delta) \wedge (C_t > O_t) \tag{3.7}$$

Trong đó L, H, C, O lần lượt là giá Low, High, Close, Open. Tương tự, Bearish FVG xuất hiện khi $H_t < L_{t-2}$. Những vùng FVG này đóng vai trò như các thanh khoản hút giá quay trở lại để giá có thể

tái cân bằng trong tương lai.

Việc kết hợp các chỉ báo kỹ thuật với các cấu trúc SMC giúp bộ dữ liệu không chỉ phản ánh được hành động giá mà còn cung cấp một phương pháp luận về cách di chuyển của giá. Đây là tiền đề quan trọng để thuật toán phân cụm ở chương sau có thể gom nhóm các trạng thái thị trường dựa trên bản chất hành vi thay vì chỉ dựa trên biên độ giá đơn thuần.

3.2 Tạo đặc trưng với dữ liệu tin tức

Dù dữ liệu tin tức thường được gắn liền với xử lý ngôn ngữ tự nhiên (sẽ được trình bày chi tiết trong chương 4), nhưng các siêu dữ liệu đi kèm như số lượng yêu thích, bình luận hay cảm xúc lại mang tính định lượng rất cao. Đây là thước đo định lượng trực tiếp cho tâm lý đám đông tại thời điểm tin tức được công bố.

Tuy nhiên, việc sử dụng trực tiếp sử dụng số lượng tương tác chứa đựng nhiều rủi ro. Ví dụ: 100 lượt yêu thích trong giai đoạn thị trường sôi động là con số bình thường, nhưng 100 lượt yêu thích trong giai đoạn thị trường ảm đạm lại là một sự đột biến lớn. Nếu chỉ sử dụng giá trị tuyệt đối hoặc chuẩn hóa theo toàn bộ tập dữ liệu, mô hình sẽ mắc lỗi do sử dụng thông tin của tương lai để tính toán cho quá khứ.

Để giải quyết vấn đề này, ta áp dụng kỹ thuật **Chuẩn hóa động**, tính toán số liệu dựa trên khung thời gian nhất định trước thời điểm cần tính. Cụ thể, tại mỗi thời điểm t, mức độ bất thường của phản ứng người dùng được tính toán dựa trên độ lệch chuẩn so với trung bình của 24 giờ trước đó. Báo cáo sử dụng Z-Score để giúp dữ liệu đầu vào của mô hình luôn phản ánh đúng bối cảnh thị trường tại thời điểm đó mà không vi phạm nguyên tắc nhân quả.

$$Z_t = \frac{X_t - \mu_{t-24h}}{\sigma_{t-24h}} \tag{3.8}$$

Sau khi tính toán được các chỉ số Z-Score liên tục, để phục vụ cho bài toán Khai phá luật kết hợp (vốn yêu cầu dữ liệu theo từng lớp cụ thể), ta tiến hành rời rạc hóa các giá trị này thành các nhãn.

Đối với chỉ số cảm xúc (`Sentiment_Z_Score`), dữ liệu được phân thành 3 nhóm: tiêu cực (Negative), trung tính (Neutral) và tích cực (Positive). Phân lớn các bản tin đều có phản ứng cảm xúc trung tính. Tương tự, đối với mức độ quan tâm (`Volume_Z_Score`), dữ liệu được chia thành mức thấp, trung bình và cao. Những điểm dữ liệu quan trọng mà mô hình cần tập trung khai thác là những sự kiện tạo được sự quan tâm đông đảo của cộng đồng.

Việc chuyển đổi từ dữ liệu thô sang các nhãn định tính dựa trên thống kê động này giúp loại bỏ nhiễu và làm nổi bật các tín hiệu thị trường thực sự có ý nghĩa.

Bảng 3.1: Thống kê thực hiện rời rạc hóa dữ liệu tin tức

Chỉ số	Phân lớp	Ngưỡng	Số lượng mẫu	Tỷ lệ (%)
`Sentiment_Z_Score`	Neutral	$[-0.5; 0.5]$	14,514	83,74%
	Negative	< -0.5	1,242	7,16%
	Positive	> 0.5	1,577	9,10%
`Volume_Z_Score`	Medium	$[-0.5; 2.0]$	14,514	83,74%
	Low	< -0.5	1,866	10,76%
	High	> 2.0	953	5,50%

Chương 4

Phân cụm dữ liệu

Sau khi đã trích xuất được các đặc trưng từ dữ liệu thô, bước tiếp theo là áp dụng thuật toán học máy không giám sát để phân chia dữ liệu thành các nhóm có tính chất tương đồng. Chương này trình bày quy trình phân cụm dữ liệu nhằm hai mục tiêu chính: xác định các pha của thị trường dựa trên hành vi giá, và gom nhóm các chủ đề tin tức dựa trên nội dung văn bản.

4.1 Phân cụm với dữ liệu giá Bitcoin

Mục tiêu của phần này là tự động gán nhãn cho từng phiên giao dịch vào một trong bốn pha của chu kỳ thị trường theo lý thuyết Wyckoff [12]: tích lũy, tăng trưởng, phân phối và suy thoái. Tuy nhiên, trong quá trình phân cụm, nhóm đã phát hiện sự tồn tại của pha cao trào thay cho pha phân phối, và đã chỉ ra lý do pha cao trào phù hợp với bài toán hơn pha phân phối.

4.1.1 Chuẩn bị dữ liệu

Để thuật toán phân cụm hoạt động hiệu quả, dữ liệu đầu vào không thể là giá trị OHLCV đơn thuần mà phải là các chỉ số đại diện cho trạng thái hiện tại của thị trường. Nhóm đã xây dựng bộ 5 đặc trưng thường dùng của phương pháp Wyckoff để mô tả xu hướng, hiệu suất nến và cấu trúc sóng.

Vị thế chiến thuật Đo lường độ lệch của giá hiện tại so với đường trung bình EMA 50. Chỉ số này cho biết giá đang quá biến động hay đang ở vùng cân bằng.

$$\text{Trend}_{\text{tactical}} = \frac{Close - \text{EMA}_{50}}{\text{EMA}_{50}} \times 100 \tag{4.1}$$

Độ trưởng thành của xu hướng Chỉ số này đo lường khoảng cách giữa xu hướng trung hạn (EMA 50) và dài hạn (EMA 200). Khoảng cách càng lớn chứng tỏ xu hướng đã diễn ra trong thời gian dài và có tính bền vững.

$$\text{Trend}_{\text{maturity}} = \frac{\text{EMA}_{50} - \text{EMA}_{200}}{Close} \times 100 \tag{4.2}$$

Hiệu suất nến Trong lý thuyết Wyckoff, nỗ lực (biên độ nến) phải tương xứng với kết quả (thân nến). Một cây nến có thân lớn và râu ngắn thể hiện sự dứt khoát (hiệu suất cao - thường thấy ở pha tăng trưởng/suy thoái), ngược lại nến thân nhỏ râu dài thể hiện sự lưỡng lự (hiệu suất thấp - thường thấy ở pha tích lũy/phân phối).

$$EffRegime = \frac{|Close - Open|}{High - Low + \epsilon} \text{ với } \epsilon \text{ vô cùng nhỏ, } \epsilon > 0 \tag{4.3}$$

Tỷ lệ khối lượng Tính bằng tỷ lệ giữa khối lượng hiện tại và trung bình 20 phiên. Tỷ lệ khối lượng tăng đột biến thường báo hiệu các điểm đảo chiều.

$$VolRegime = \frac{Vol}{\text{MA}_{20}} \tag{4.4}$$

Điểm cấu trúc Đây là đặc trưng quan trọng nhất vì nó có tính "ghi nhớ", khi đặc trưng này được tính cộng dồn theo thời gian với hệ số suy giảm (decay factor $\gamma = 0.98$). Điểm số sẽ tăng khi xuất hiện các khối lệnh tăng hoặc FVG, và giảm khi xuất hiện các tín hiệu suy yếu.

$$Score_t = Score_{t-1} \times \gamma + \text{Signal}_t \tag{4.5}$$

Trong đó Signal_t nhận giá trị $+1$ nếu là OB tăng, -1 nếu là OB giảm. Đặc trưng này giúp mô hình nhận diện được phe nào đang kiểm soát thị trường về mặt cấu trúc.

4.1.2 Diễn giải kết quả phân cụm

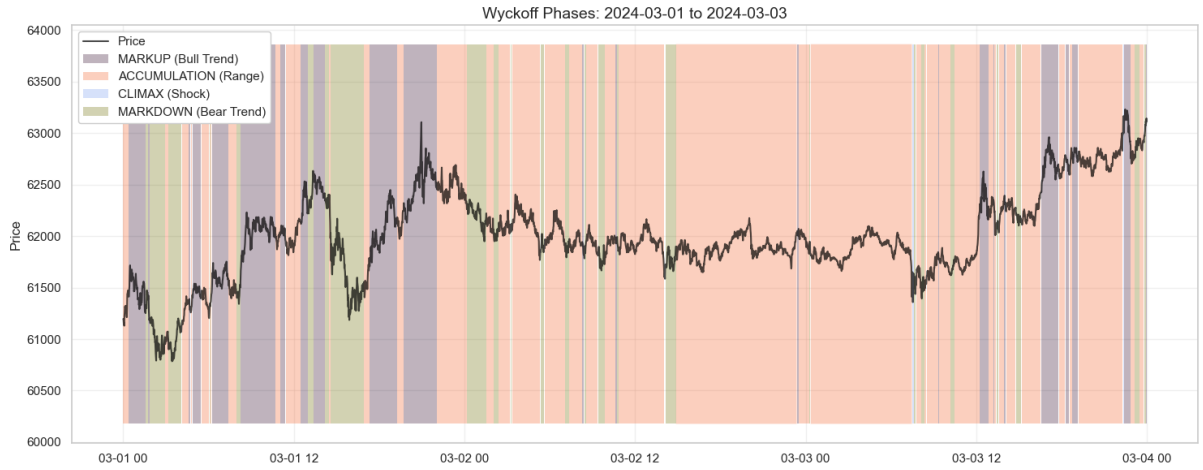
Với giả thuyết thị trường vận động theo 4 pha chính, phương án sử dụng thuật toán **K-Means** với số cụm $k = 4$ được tính đến. Dữ liệu trước khi đưa vào mô hình được chuẩn hóa để đưa về cùng một phân phối chuẩn. Sau khi thực hiện phân cụm, ta tiến hành phân tích giá trị trung bình của các đặc trưng trong từng cụm để gán nhãn ý nghĩa tài chính. Ta có bảng 4.1 tóm tắt các đặc điểm chính.

Bảng 4.1: Đặc điểm trung bình của các cụm

Cụm	$\text{Trend}_{\text{tactical}}$	$\text{Trend}_{\text{maturity}}$	$EffRegime$	$VolRegime$	$Score$	Số lượng mẫu
0	0.195	0.284	0.567	0.920	**0.781**	213,019
1	0.002	0.008	0.750	0.833	0.016	493,383
2	-0.011	0.004	0.699	**3.568**	0.049	59,253
3	-0.131	-0.188	0.529	0.896	**-0.295**	286,802

Dựa trên số liệu bảng 4.1, nhóm thực hiện biện luận và gán nhãn như sau:

- **Cụm 0 (Pha Tăng trưởng):** Đặc trưng bởi điểm cấu trúc rất cao (0.781) và độ chín xu hướng dương (0.284). Đây là giai đoạn phe Mua kiểm soát hoàn toàn, giá tăng trưởng ổn định.
- **Cụm 1 (Pha Tích lũy):** Các chỉ số xu hướng và cấu trúc đều tiệm cận 0, nhưng số lượng mẫu lại chiếm đa số (gần 500k mẫu). Đây là trạng thái thị trường đi ngang, biên độ hẹp, tương ứng với giai đoạn tích lũy hoặc tái tích lũy.
- **Cụm 2 (Pha Cao trào):** Điểm nổi bật nhất là khối lượng giao dịch cực lớn (gấp 3.5 lần trung bình). Đây là các điểm cao trào, nơi diễn ra sự trao tay ồ ạt giữa dòng tiền thông minh và đám đông, thường báo hiệu sự đảo chiều sắp xảy ra.
- **Cụm 3 (Pha Suy thoái):** Đặc trưng bởi điểm cấu trúc âm và xu hướng âm. Đây là giai đoạn giá giảm, phe Bán kiểm soát thị trường.



Hình 4.1: Kết quả phân cụm trạng thái thị trường theo lý thuyết Wyckoff trên biểu đồ giá Bitcoin (Giai đoạn mẫu: 01/03/2024 - 03/03/2024).

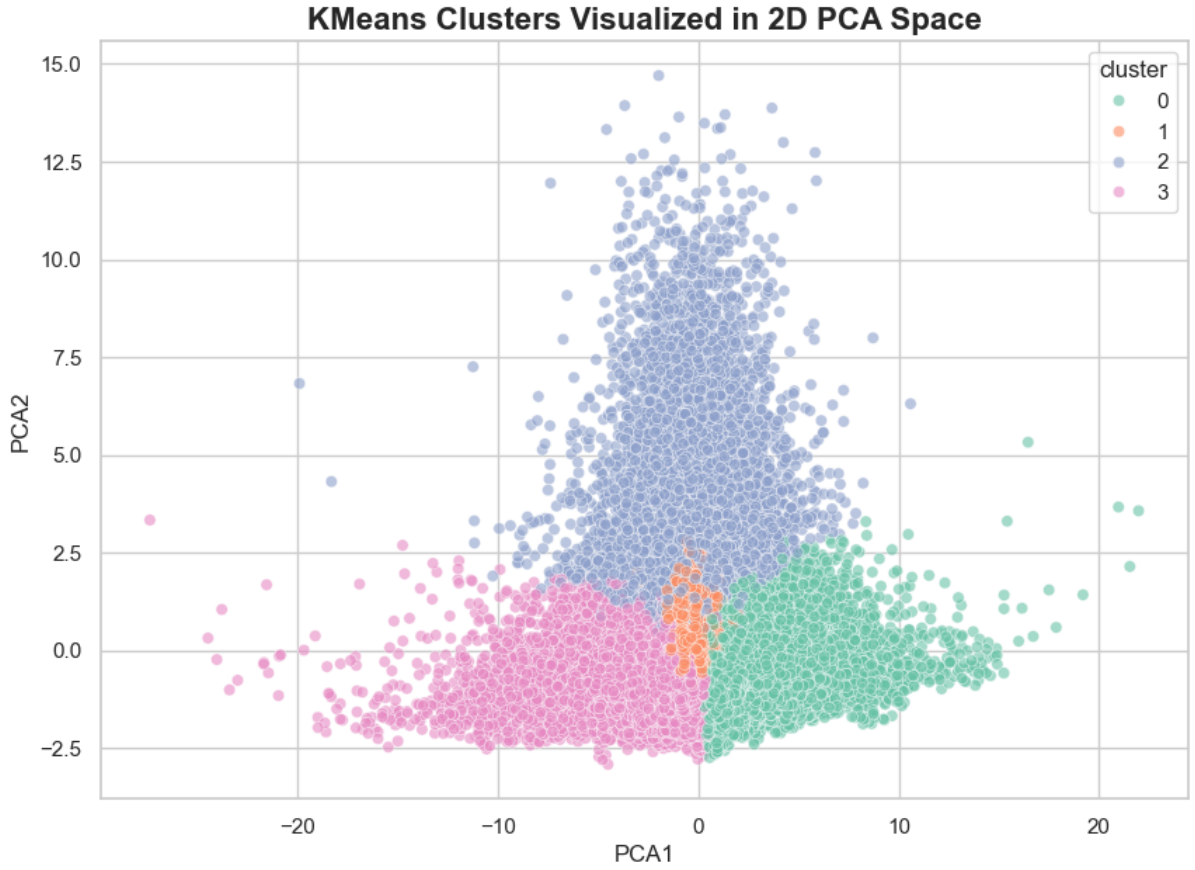
4.1.3 Đánh giá kết quả phân cụm

Để kiểm chứng chất lượng của các cụm được sinh ra, ta sử dụng hệ thống 4 chỉ số đánh giá nội tại. Kết quả thu được như sau:

- **Inertia (3336295.15):** Là tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm của nó. Giá trị này càng nhỏ càng tốt, tuy nhiên với kích thước dữ liệu lớn (hơn 1 triệu điểm), giá trị này ở mức chấp nhận được.
- **Silhouette Score (0.2191):** Đo lường mức độ tách biệt giữa các cụm, có giá trị từ -1 đến 1. Công thức:

  $$S = \frac{b - a}{\max(a, b)}$$

  Trong đó a là khoảng cách trung bình nội bộ cụm, b là khoảng cách trung bình đến cụm lân cận gần nhất. Giá trị 0.22 cho thấy các cụm có sự phân tách nhưng vẫn tồn tại sự chồng lấn. Điều này hoàn toàn phù hợp với dữ liệu tài chính liên tục, nơi ranh giới giữa tích lũy và tăng trưởng thường không sắc nét mà có sự chuyển biến dần dần.
- **Davies-Bouldin Index (1.4146):** Tỷ lệ giữa độ phân tán nội bộ và khoảng cách giữa các cụm. Giá trị càng thấp càng tốt. Mức 1.41 là một kết quả khá tốt đối với bài toán phân cụm dữ liệu chuỗi thời gian nhiều nhiễu.
- **Calinski-Harabasz Index (202523.28):** Tỷ số giữa phương sai giữa các cụm và phương sai trong cụm. Giá trị càng cao cho thấy các cụm càng đông đúc và tách biệt rõ ràng. Con số này khẳng định cấu trúc phân cụm có ý nghĩa thống kê cao.



Hình 4.2: Trực quan hóa các cụm dữ liệu Wyckoff trên không gian 2 chiều sử dụng PCA.

Tổng hợp lại các phân tích trên, mô hình K-Means đã thành công trong việc tách dòng dữ liệu giá hỗn loạn thành 4 trạng thái thị trường riêng biệt.

Một điểm đáng chú ý là kết quả thực nghiệm có sự khác biệt nhỏ so với lý thuyết Wyckoff ban đầu. Thay vì tách bạch thành 4 pha theo lý thuyết (tích lũy - tăng trưởng - phân phối - suy thoái), thuật toán phân cụm lại nhận diện được pha mang ý nghĩa biểu thị sự cao trào hơn là phân phối. Tuy vậy, điều đó lại giúp cho thuật toán phân tích tốt hơn, vì hai lý do:

- Thứ nhất, về mặt hành vi, pha Phân phối thường có đặc điểm đi ngang khá tương đồng với pha tích lũy về mặt biến động giá, khiến thuật toán khó phân biệt rạch ròi.
- Thứ hai, về mặt tín hiệu, pha cao trào lại sở hữu các đặc trưng cực đoan về khối lượng và biên độ. Việc mô hình tách được cụm này có ý nghĩa thực tiễn hơn nhiều trong giao dịch, bởi đây là những điểm đảo chiều xác suất cao mang lại cơ hội lợi nhuận cao nhất.

Vì vậy, dù không rập khuôn theo lý thuyết, bộ nhãn cụm [**tăng trưởng, tích lũy, cao trào, suy thoái**] vẫn đảm bảo tính bao quát, và thậm chí còn tốt hơn trong việc phát hiện các điểm bất thường của thị trường. Đây sẽ là nguồn dữ liệu chất lượng cho thuật toán khai phá luật kết hợp ở chương 5.

4.2 Phân cụm với dữ liệu tin tức

Một bài toán khác là phân nhóm dữ liệu văn bản, với mục tiêu là tự động phân loại hơn 200,000 tiêu đề tin tức thành các chủ đề lớn để phục vụ cho việc khai phá luật kết hợp.

4.2.1 Quy trình phân cụm

Thách thức Đối với dữ liệu tin tức, nhóm nghiên cứu đối mặt với rào cản lớn về tính phi cấu trúc của ngôn ngữ tự nhiên, xảy ra hiện tượng khó phân cụm, khó phát hiện mẫu do dữ liệu quá đa chiều. Các thuật toán phân cụm dựa trên khoảng cách như K-Means thường hoạt động kém hiệu quả trong không gian vector quá lớn, nơi khoảng cách Euclid giữa các điểm dữ liệu trở nên xấp xỉ nhau và mất đi ý nghĩa phân loại. Do đó, việc chuyển đổi trực tiếp văn bản thô sang các vector đặc trưng khổng lồ sẽ khiến mô hình bị nhiễu và không thể hội tụ chính xác.

Giải pháp Để giải quyết vấn đề này, nhóm đề xuất một quy trình xử lý tuần tự gồm ba bước:

1. Chuyển đổi không gian ngữ nghĩa thành các vector ngữ nghĩa. Ta sẽ biến đổi mỗi tiêu đề tin tức thành vector đa chiều sử dụng mô hình `all-MiniLM-L6-v2` [11]. Khi đó, mô hình có thể nhận thức được nghĩa của văn bản, từ đó thực hiện các phép tính số học nhằm phân tích văn bản.
2. Sử dụng thuật toán PCA để giảm chiều dữ liệu. Mục tiêu của bước này là nén các vector đa chiều xuống miền không gian thấp hơn, giúp giảm chiều dữ liệu hiệu quả, giảm nhiễu, mà vẫn giữ được các thành phần quan trọng nhất.
3. Sử dụng thuật toán K-Means để tách biệt các cụm dữ liệu. Bước giảm chiều dữ liệu trước đó đã phần nào khắc phục nhược điểm của K-Means khi phân cụm trên dữ liệu nhiều chiều.

Tham số tối ưu Để thực hiện giải pháp trên, trước hết ta cần tìm ra bộ tham số thích hợp nhất cho dữ liệu và bài toán. Nhóm sử dụng kỹ thuật tìm kiếm lưới để thử nhiều tổ hợp tham số khác nhau. Ta đánh giá hiệu suất của bộ tham số dựa trên **Silhouette Score**, với ý nghĩa phản ánh độ tách biệt và cô đặc của các cụm dữ liệu.

Nhóm đã tiến hành thử nghiệm toàn diện với các tổ hợp tham số khác nhau: số chiều PCA dao động trong tập $\{3, 4, 5, 10, 15\}$ và số lượng cụm K từ 4 đến 12. Kết quả chi tiết của quá trình thử nghiệm được trình bày tại bảng 4.2.

Quan sát bảng số liệu cho thấy một xu hướng rõ rệt: hiệu suất phân cụm giảm dần khi số chiều PCA tăng lên. Cụ thể, khi giữ nguyên số chiều ở mức cao (10 hoặc 15), chỉ số Silhouette chỉ đạt mức rất thấp do nhiễu. Ngược lại, việc giảm xuống không gian 3 chiều mang lại kết quả vượt trội nhất. Tại cấu hình **PCA = 3** và **Số cụm K = 5**, mô hình đạt chỉ số Silhouette cao nhất là **0.2883**. Do đó, nhóm quyết định lựa chọn bộ tham số này cho việc phân loại chủ đề tin tức.

Bảng 4.2: Kết quả grid Search (Silhouette Score) theo số chiều PCA và số cụm K

Số cụm (K)	**Silhouette Score theo số chiều PCA**				
	PCA = 3	**PCA = 4**	**PCA = 5**	**PCA = 10**	**PCA = 15**
4	0.2841	0.2281	0.1941	0.1225	0.0957
5	**0.2883**	0.2532	0.2164	0.1349	0.1054
6	0.2811	0.2581	0.2227	0.1335	0.1016
7	0.2607	0.2453	0.2207	0.1368	0.1112
8	0.2608	0.2351	0.2114	0.1315	0.1060
9	0.2520	0.2237	0.1970	0.1355	0.1101
10	0.2484	0.2125	0.1938	0.1357	0.1159
11	0.2475	0.2113	0.1923	0.1294	0.1191
12	0.2492	0.2105	0.1864	0.1280	0.1176

4.2.2 Diễn giải kết quả phân cụm

Sau khi thuật toán K-Means phân chia tập dữ liệu thành 5 cụm riêng biệt, thách thức tiếp theo là giải thích được ý nghĩa ngữ nghĩa của từng nhóm để gán các nhãn chủ đề phù hợp. Bằng cách sử dụng kỹ thuật TF-IDF để trích xuất các từ khóa trọng tâm có tần suất xuất hiện cao, kết hợp với việc kiểm tra ngẫu nhiên nội dung các tiêu đề, nhóm đã định danh được 5 nhóm tin tức chính.

Cụm 0 - Nhóm hệ sinh thái và hạ tầng Nhóm phản ánh sự vận động nội tại của công nghệ và hệ sinh thái blockchain. Đây là nhóm tin tức mang tính nền tảng, thường xuất hiện các báo cáo phân tích cơ bản về dự án hoặc xu hướng dòng tiền dịch chuyển giữa các hệ sinh thái. Ví dụ tiêu biểu là các thảo luận về việc liệu Altcoins có thay thế được Bitcoin, hay tương lai của ngành tiền điện tử.

```
- Bitcoin's Supply On Exchange Tightens: Could a New Bull Run Be Just Weeks Away?
- Bitcoin Exchange Depositing Transactions At 4-Year Low, Bottom Signal?
- Bitcoin Continues To Exit Exchanges As Supply Drops To New 2024 Low
- While Bitcoin Supply in Stock Exchanges is Falling, Tether (USDT) Supply is Rising!
- Major Exchange Experienced \$116 Million in Outflow in Ethereum, Bitcoin and USDT
```

Cụm 1 - Nhóm phân tích dòng tiền Cụm chủ đề này tập trung hoàn toàn vào số liệu của dòng tiền. Sự xuất hiện dày đặc của các từ khóa *etf, outflows, inflows, reserves, volume* chỉ ra rằng đây là nơi tập hợp các thông tin về dòng tiền thực tế. Đặc biệt, các tin tức về dòng vốn vào/ra khỏi các quỹ ETF, hay sự thay đổi dự trữ Bitcoin trên các sàn giao dịch được gom vào nhóm này. Đây là những tín hiệu khách quan nhất phản ánh sức khỏe cung cầu của thị trường, ví dụ như việc ghi nhận hàng trăm triệu USD rút ròng khỏi quỹ ETF trong một tuần sẽ có ảnh hưởng rất lớn.

```
- Nonprofit regulation group Better Markets urges SEC to reject Bitcoin ETF applications
- Justin Sun To Buy German Government's BTC Stash To Minimise Market Impact
- Millions in BTC, XMR possibly stolen after reports of darknet market 'exit scam'
```

Cụm 2 - Nhóm memecoins Nhóm này nhắm tới những đồng coin có vốn hóa từ bé tới rất bé, và giá có sự biến động rất lớn. Với các từ khóa như *doge, pepe, meme, pump, frenzy*, nhóm này bao gồm các tin tức liên quan đến Memecoins và các đợt tăng giá, giảm giá đột biến dựa trên hiệu ứng mạng xã hội. Những đồng coin này chỉ có dòng tiền từ các nhà đầu tư nhỏ lẻ, tuy nhiên lại tác động lớn đến biến động giá ngắn hạn và chỉ số cảm xúc toàn thị trường.

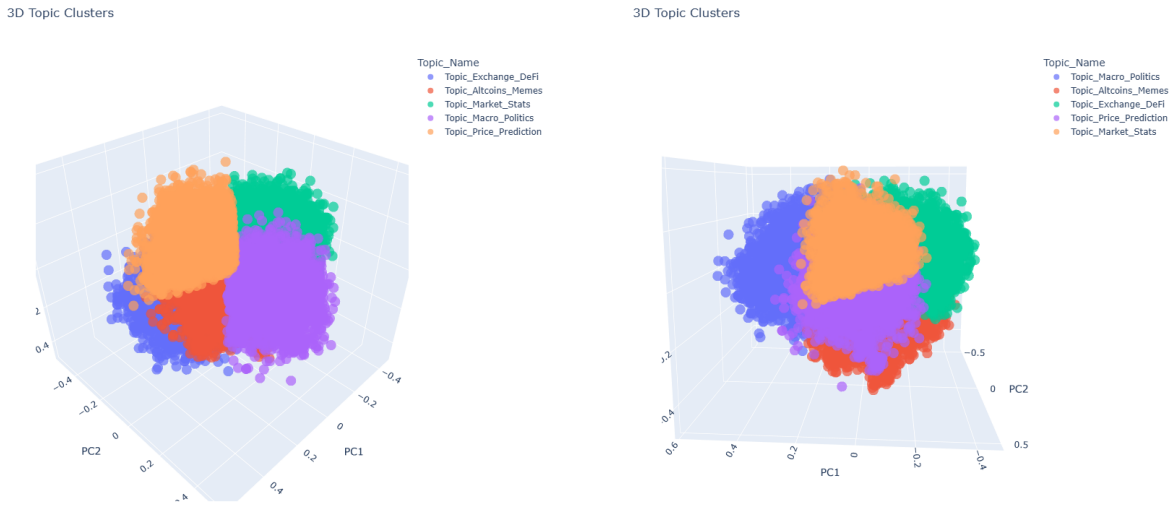
```
- Top Altcoins to Watch for a 100\% Rally Next Week: Curve DAO, Hedera, and JasmyCoin
- Crypto News Today: BTC Plunges, Altcoins Stumble in Sell-Off
- Top 5 Bitcoin-Like Altcoins That Could Make You A Millionaire In The Crypto Bull Run
- 3 Bullish Altcoins Predicted to Outshine Bitcoin (BTC) in July 2024
```

Cụm 3 - Nhóm kinh tế Vĩ mô và chính trị Nhóm chủ đề này bao gồm các yếu tố ngoại sinh, tuy nằm ngoài biểu đồ kỹ thuật nhưng có sức ảnh hưởng mang tính định hình xu hướng. Các từ khóa *trump, election, sec, regulation, fed* cho thấy sự nhạy cảm của thị trường tiền mã hóa đối với bối cảnh chính trị và chính sách pháp lý của các nước trên thế giới, đặc biệt là tại Mỹ. Các sự kiện như bầu cử Tổng thống, quyết định lãi suất của FED hay các vụ kiện tụng của SEC thường được phân loại vào nhóm này. Tác động của nhóm tin này thường mang tính cấu trúc và dài hạn hơn.

```
- Cathie Wood Of Ark Invest Believes Ethereum ETF Approval Linked To US Politics,
Sees Likelihood Of A Solana ETF As Well
- Trump's Crypto Portfolio Takes A Hit, Robert Kiyosaki Bats For Bitcoin
- What's next for the crypto industry following Donald Trump's election victory?
- Ethereum ETF Approved, Elon Musk Mourns Kabosu's Death, Trump Embraces Crypto Donations
```

Cụm 4 - Nhóm Dự báo và Kỳ vọng tương lai Nhóm tin tức này mang tính chất chủ quan, là tầm nhìn tương lai của nhà đầu tư với thị trường. Với các từ khóa *outlook, predict, target, bull run, crash*, đây là tập hợp các nhận định, đồn đoán và phân tích kỹ thuật từ các chuyên gia hoặc KOLs về hướng đi tương lai của giá. Việc phân tách này rất quan trọng, giúp mô hình phân biệt được đâu là sự kiện đã xảy ra, và đâu là kỳ vọng, từ đó đánh giá mức độ rủi ro của thị trường một cách chính xác hơn.

```
- Fantom (FTM) Price Prediction: Can This Bullish Pattern Prevent a 31\% Fall?
- Polkadot Price Sets Sights on New Highs: DOT Bullish Momentum Building
- Bitcoin Taps $60K, Ethereum, Dogecoin End The Week On A High: 'Next Few Weeks
Could Be Last Chance To Grab BTC At Cheap Prices'
```



(a) Topic visualization 1

(b) Topic visualization 2

Hình 4.3: Trực quan kết quả phân cụm ở không gian 3 chiều

Kết luận: Việc phân cụm thành công 5 chủ đề trên đóng vai trò then chốt cho giai đoạn khai phá luật kết hợp tiếp theo. Nó cho phép hệ thống không chỉ đánh giá tin tức dựa trên cảm xúc đang tiêu cực hay tích cực mà còn đặt nó vào ngữ cảnh cụ thể. Chẳng hạn, một tin tức tích cực thuộc nhóm *vĩ mô* có thể mang trọng số tín hiệu hoàn toàn khác biệt so với một tin tích cực thuộc nhóm *memecoins*, từ đó nâng cao độ chính xác của các quy tắc giao dịch được sinh ra.

Chương 5

Khai phá luật kết hợp

Sau khi đã hoàn tất công đoạn chuẩn bị dữ liệu và phân cụm, chương này tập trung vào việc khai phá luật kết hợp để tìm kiếm các mẫu hình giao dịch tiềm năng. Mục tiêu là tìm ra các quy tắc có dạng *"Nếu điều kiện A, B, C xảy ra thì xác suất giá tăng/giảm là bao nhiêu"*.

Quy trình thực hiện được chia thành ba giai đoạn: Định nghĩa biến mục tiêu và rời rạc hóa dữ liệu giá, Khai phá luật trên dữ liệu giá thuần túy, và Tích hợp dữ liệu tin tức để xây dựng hệ thống luật lai.

5.1 Chuẩn bị dữ liệu và biến mục tiêu

5.1.1 Định nghĩa biến mục tiêu

Trong các bài toán dự báo tài chính truyền thống, mô hình thường được huấn luyện để dự đoán giá đóng cửa của phiên tiếp theo sẽ tăng hay giảm. Tuy nhiên, trong giao dịch thực tế, cách tiếp cận này bộc lộ hạn chế lớn vì nó bỏ qua yếu tố quản trị rủi ro và thời gian nắm giữ vị thế. Một nhà giao dịch không chỉ quan tâm giá có tăng hay không, mà quan trọng hơn là liệu giá có tăng đủ mạnh để chạm mức chốt lời trước khi chạm mức cắt lỗ hay không.

Nhóm cần định nghĩa các biến mục tiêu, chính là 3 quyết định đầu tư: **mua, bán hay giữ**. Để gán nhãn dữ liệu, nhóm áp dụng phương pháp **Triple-Barrier Method**. Đây là kỹ thuật gán nhãn dữ liệu bằng cách thiết lập ba giới hạn cho mỗi điểm dữ liệu dựa trên một khung thời gian cố định. Đặc biệt, thay vì sử dụng các ngưỡng cố định (ví dụ: $\pm 1\%$), nhóm sử dụng chỉ báo **ATR** để thiết lập các rào chắn động. Điều này đảm bảo rằng mục tiêu lợi nhuận sẽ tự động nới rộng trong các giai đoạn thị trường biến động mạnh và thu hẹp lại khi thị trường đi ngang, phản ánh đúng bản chất biến động của rủi ro.

Ba rào chắn được thiết lập như sau:

1. **Rào chắn trên:** Đóng vai trò là mức chốt lời kỳ vọng. Được tính bằng $Close_t + 3.0 \times ATR_{14}$.
2. **Rào chắn dưới:** Đóng vai trò là mức cắt lỗ. Được tính bằng $Close_t - 3.0 \times ATR_{14}$.
3. **Rào chắn thời gian:** Giới hạn thời gian nắm giữ vị thế, được thiết lập là 60 phiên (tương ứng 60 phút) kể từ thời điểm vào lệnh.

Cơ chế gán nhãn hoạt động dựa trên nguyên tắc chạm: Nếu giá chạm rào chắn trên trước tiên trong vòng 60 phút, nhãn được gán là **Pump** (Tín hiệu mua). Ngược lại, nếu giá chạm rào chắn dưới trước, nhãn là **Dump** (Tín hiệu bán). Trường hợp giá không chạm rào chắn nào sau 60 phút, hoặc biến động quá mạnh chạm cả hai rào chắn (hiện tượng giá quét thanh khoản của bên mua và bên bán rồi mới đi theo hướng chính), nhãn sẽ được gán là **Sideways**.

5.1.2 Rời rạc hóa dữ liệu

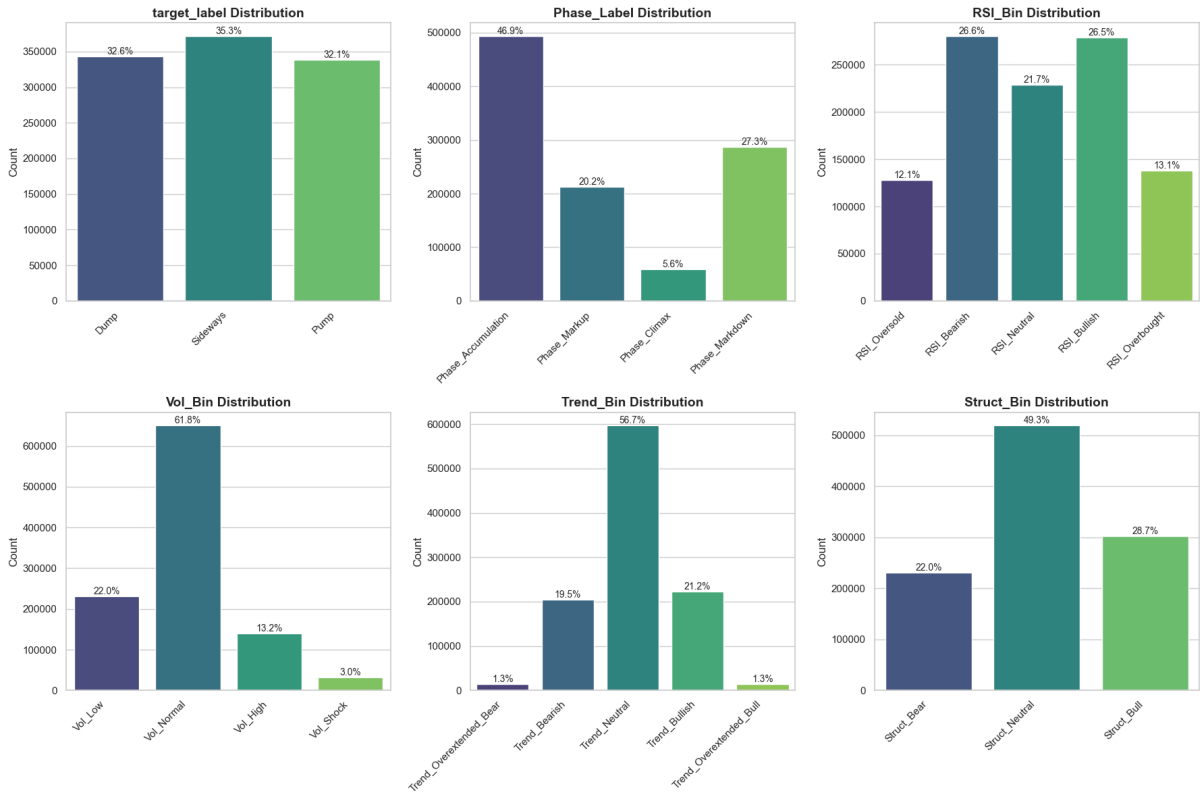
Thuật toán khai phá luật kết hợp (như FP-Growth hay Apriori) được thiết kế để làm việc trên dữ liệu dạng giao dịch (transactional data), nơi mỗi bản ghi là tập hợp các "mặt hàng" rời rạc. Trong khi đó, dữ liệu tài chính thu được từ các chương trước phần lớn là các biến số thực liên tục. Do đó, để chuyển đổi dữ liệu này thành định dạng phù hợp cho mô hình, cần phải thực hiện quá trình rời rạc hóa bằng

cách phân chia dữ liệu vào các khoảng giá trị, dựa trên các ngưỡng kỹ thuật tiêu chuẩn và đặc điểm phân phối thống kê.

Quy trình rời rạc hóa được áp dụng chi tiết cho từng nhóm đặc trưng như sau:

- **Chỉ số sức mạnh tương đối:** Thay vì chỉ sử dụng ngưỡng quá mua/quá bán kinh điển (70/30), nhóm chia nhỏ không gian giá trị thành 5 trạng thái để nắm bắt kỹ hơn các biến động trung gian. Cụ thể, vùng từ 45 đến 55 được tách riêng thành neutraal để lọc nhiễu khi thị trường không có xu hướng rõ ràng. Các giá trị trên 70 được gán nhãn **overbought** (quá mua) và dưới 30 là **oversold** (quá bán).
- **Tỷ lệ khối lượng:** Mức **normal** nằm trong khoảng 0.5 đến 1.5 lần trung bình. Nhãn **shock** khi giá trị gấp 3 lần trung bình, đại diện cho các sự kiện đột biến thanh khoản báo hiệu sự đảo chiều.
- **Vị thế chiến thuật (Trend Tactical):** Các trạng thái **overextended** được định nghĩa khi độ lệch vượt quá $\pm 0.6\%$, đây là vùng giá thường có xu hướng hồi quy về trung bình.
- **Điểm cấu trúc (Structure Score):** Phân thành 3 trạng thái: Bull (Tích cực), bear (Tiêu cực) và neutral (Trung tính).
- **Pha thị trường:** Sử dụng kết quả từ chương 4, ta biết được pha hiện tại của thị trường (markup, accumulation, climax, markdown).

Các quy tắc rời rạc hóa được tổng hợp tại bảng 5.1. Phân bố tần suất của từng nhãn được biểu diễn trong biểu đồ hình 5.1.



Hình 5.1: Phân bố tần suất các đặc trưng và nhãn mục tiêu trong tập dữ liệu giao dịch sau khi rời rạc hóa.

Đặc trưng	Ngưỡng phân chia	Nhãn
RSI (14)	$< 30, 30\text{-}45, 45\text{-}55, 55\text{-}70, > 70$	Oversold, Bearish, Neutral, Bullish, Overbought
Volume Regime	$< 0.5, 0.5\text{-}1.5, 1.5\text{-}3.0, > 3.0$	Low, Normal, High, Shock
Trend Tactical	$< -0.6, -0.6\text{-} -0.1, \ldots, > 0.6$	Overextended_Bear, ..., Overextended_Bull
Structure Score	$< -0.5, -0.5\text{-}0.5, > 0.5$	Struct_Bear, Struct_Neutral, Struct_Bull
Phase Cluster	Ánh xạ trực tiếp từ K-Means	Markup, Accumulation, Climax, Markdown

Bảng 5.1: Tổng hợp quy tắc rời rạc hóa và gán nhãn đặc trưng

5.2 Khai phá luật trên dữ liệu giá

Trước khi tích hợp các yếu tố ngoại sinh từ tin tức, nhóm tiến hành khai phá trên tập dữ liệu thuần kỹ thuật để thiết lập một mô hình cơ sở. Mục tiêu của bước này là trả lời câu hỏi: *"Liệu các chỉ báo kỹ thuật và cấu trúc thị trường đơn thuần có thể dự báo được các đợt biến động giá mạnh hay không?"*.

5.2.1 Thiết lập thuật toán và Tham số

Để xử lý lượng dữ liệu giao dịch lớn, nhóm lựa chọn thuật toán **FP-Growth**. So với **Apriori**, **FP-Growth** hiệu quả hơn đáng kể về mặt bộ nhớ và tốc độ do nó nén dữ liệu vào cấu trúc cây FP-Tree và không cần sinh các tập ứng viên lặp đi lặp lại.

Quy trình thực hiện bắt đầu bằng việc chuyển đổi dữ liệu dạng danh sách sang ma trận nhị phân, sử dụng `TransactionEncoder`. Sau đó, thuật toán được cấu hình với bộ tham số chiến lược như sau:

- **Độ hỗ trợ tối thiểu (min Support = 0.001):** Nhóm quyết định hạ ngưỡng này xuống mức rất thấp (0.1%). Lý do là các cơ hội giao dịch lợi nhuận cao (giá tăng mạnh hoặc giảm mạnh) thường là các sự kiện hiếm. Việc đặt ngưỡng quá cao, sẽ vô tình loại bỏ các mẫu hình giá trị này.
- **Độ tin cậy tối thiểu (min Confidence = 0.1):** Ngưỡng khởi điểm thấp để bao quát không gian tìm kiếm, sau đó sẽ lọc lại ở bước hậu xử lý.
- **Chiến lược lọc vùng mua/bán:** Để trích xuất các tín hiệu có giá trị thực chiến, nhóm áp dụng bộ lọc khắt khe hơn trên kết quả đầu ra:
  - **Vùng mua (Pump):** Yêu cầu $Lift > 1.1$ và $Confidence > 0.35$.
  - **Vùng bán (Dump):** Yêu cầu $Lift > 1.1$ và $Confidence > 0.40$ (Do tâm lý hoảng loạn thường dễ đoán hơn hưng phấn, nên ngưỡng tin cậy cho lệnh bán được đặt cao hơn).

5.2.2 Kết quả thực nghiệm và đánh giá

Bảng 5.2: Thống kê kết quả khai phá luật trên dữ liệu giá

Chỉ số	Giá trị
Tổng số luật tìm thấy	16,035
$Lift$ trung bình	1.52
Số lượng luật mua (Pump)	**1,358**
Max $Lift$ (Pump)	17.15
Max $Confidence$ (Pump)	37.51%
Số lượng luật bán (Dump)	**1,397**
Max $Lift$ (Dump)	17.26
Max $Confidence$ (Dump)	38.17%

Đánh giá kết quả:

- Với hơn 16,000 luật được tìm thấy, dữ liệu Bitcoin cho thấy các mẫu hình kỹ thuật không xuất hiện ngẫu nhiên mà có tính chu kỳ và lặp lại cao.
- Số lượng luật bán (**1,397**) và mua (**1,358**) khá tương đồng, cho thấy mô hình có khả năng nhận diện cơ hội ở cả hai kịch bản tăng và giảm giá.
- Giá trị **Max Lift** $>$ **17** khẳng định rằng khi các điều kiện đặc biệt của luật xuất hiện, xác suất xảy ra biến động giá mạnh cao gấp 17 lần so với thông thường. Đây là các tín hiệu tốt để giao dịch.
- Mức $Confidence$ xấp xỉ **38%** là một con số khả quan trong thị trường Bitcoin đầy biến động. Tuy nhiên, nó cũng nhắc nhở về việc cần kết hợp thêm các chiến lược cắt lỗ (stop-loss) để bù đắp cho tỷ lệ tín hiệu nhiễu còn lại.

5.3 Khai phá luật hỗn hợp

Sau khi đã thiết lập được mô hình cơ sở từ dữ liệu giá, bước tiến quan trọng tiếp theo là tích hợp luồng thông tin từ tin tức vào mô hình. Mục tiêu của phần này không chỉ là tìm ra các quy luật mới, mà là tìm ra các *ngữ cảnh thông tin* giúp giải thích tại sao giá lại biến động, từ đó nâng cao độ tin cậy của các quyết định giao dịch.

5.3.1 Chuẩn bị dữ liệu đa nguồn

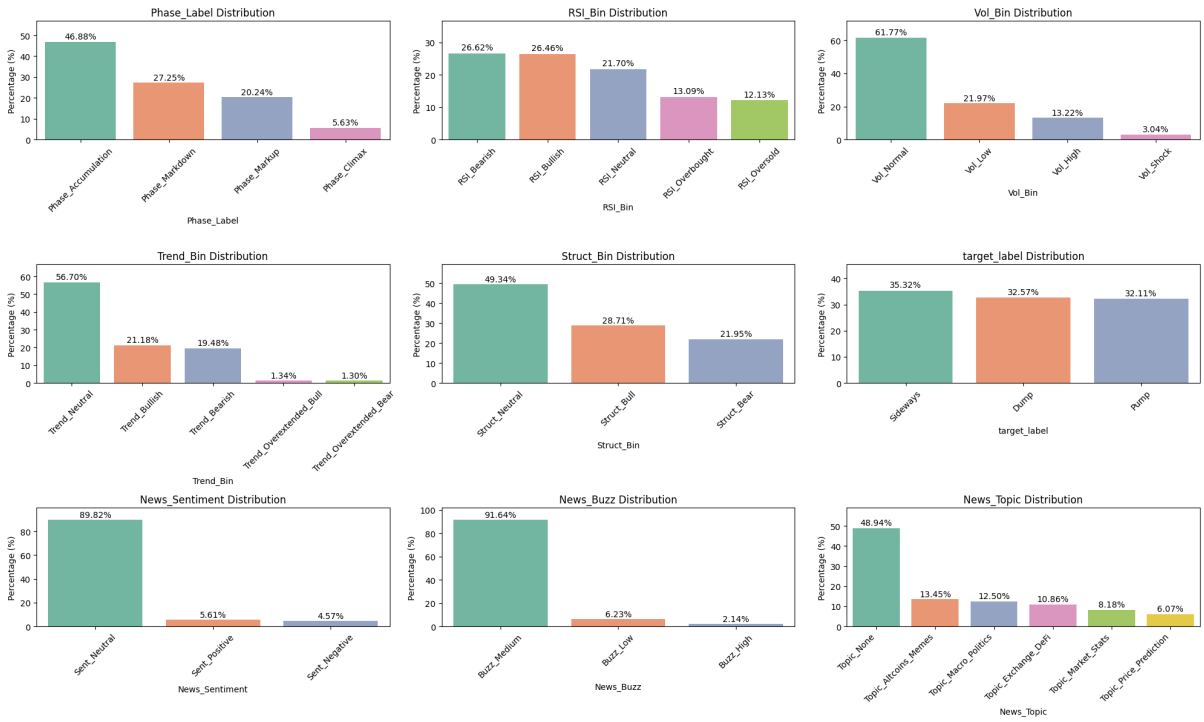
Thách thức lớn nhất trong việc kết hợp dữ liệu giá và tin tức nằm ở sự bất đồng bộ về thời gian. Dữ liệu giá có tính liên tục, ổn định và tần suất cao, trong khi tin tức xuất hiện ngẫu nhiên và rời rạc. Để giải quyết vấn đề này, nhóm xây dựng quy trình tích hợp gồm ba bước: Tổng hợp, đồng bộ hóa và hợp nhất.

Bước 1: Tổng hợp và rời rạc hóa tin tức theo giờ. Thay vì xử lý từng bản tin riêng lẻ, nhóm tiến hành tổng hợp các chỉ số Z-Score của Cảm xúc thị trường và Mức độ quan tâm theo khung thời gian 1 giờ. Sau đó, các giá trị này được chuyển đổi sang dạng nhãn. Ngoài ra, nhãn chủ đề tin tức phổ biến nhất trong giờ đó cũng được trích xuất.

Bước 2: Đồng bộ, xử lý độ trễ thông tin. Dữ liệu tin tức được dịch chuyển chậm hơn một chút so với thời gian giá thực. Điều này mô phỏng đúng thực tế, rằng thông tin cần có thời gian để tới được các nhà đầu tư, và thị trường cần có thời gian để phản ánh tin tức đó.

Bước 3: Hợp nhất dữ liệu. Nhóm sử dụng kỹ thuật `merge_asof backward`. Kỹ thuật này cho phép mỗi bản ghi giá tại thời điểm t tìm kiếm và kết hợp với bản ghi tin tức gần nhất trong quá khứ (trong phạm vi dung sai 2 giờ). Nếu không có tin tức nào trong khoảng thời gian này, các trạng thái mặc định sẽ được gán để lấp đầy dữ liệu.

Kết quả của quá trình này là một bộ cơ sở dữ liệu giao dịch thống nhất, nơi mỗi biến động giá đều được đặt trong bối cảnh tin tức tương ứng.



Hình 5.2: Phân bố tần suất các đặc trưng trong tập dữ liệu tích hợp

5.3.2 Khai phá mẫu kết hợp lai

Trên tập dữ liệu đã tích hợp, nhóm tiếp tục áp dụng thuật toán FP-Growth. Tuy nhiên, để tối ưu hóa hiệu suất và tập trung vào các tín hiệu có giá trị, nhóm áp dụng chiến lược **lọc nhiễu chủ động**:

- **Loại bỏ dữ liệu ít ý nghĩa:** Các giá trị mặc định như *Sent_Neutral, Buzz_Medium, Topic_None* chiếm tỷ trọng lớn nhưng mang ít thông tin dự báo. Việc loại bỏ chúng khỏi tập giao dịch giúp thuật toán tập trung tìm kiếm các mối liên kết khi có sự kiện tin tức thực sự (tích cực/tiêu cực hoặc tin nóng).
- **Giới hạn độ dài luật (Max Length = 6):** Để tránh sinh ra các luật quá phức tạp và khó giải thích, cũng như giảm bớt hiện tượng quá khớp, độ dài tối đa của tập phổ biến được giới hạn ở mức 6 phần tử.

Kết quả thực nghiệm: Sau quá trình lọc nhiễu và khai phá, ta thu được tổng cộng **24,188** luật kết hợp. Trong đó:

- **Số lượng luật mua:** 1,821 luật. Max Lift đạt **17.15** và Max Confidence đạt **37.88%**.
- **Số lượng luật bán:** 1,882 luật. Max Lift đạt **17.26** và Max Confidence đạt **39.69%**.

Mặc dù chỉ số lift cực đại thấp hơn so với mô hình thuần giá do điều kiện kết hợp khắt khe hơn, nhưng độ tin cậy trung bình lại có sự cải thiện nhẹ. Quan trọng hơn, các luật sinh ra đều mang tính giải thích cao, liên kết trực tiếp hành động giá với các sự kiện thực tế.

5.3.3 Đánh giá hiệu quả

Để kiểm chứng giả thuyết rằng dữ liệu tin tức giúp cải thiện chất lượng các quy tắc giao dịch, nhóm đã tiến hành so sánh trực tiếp các chỉ số thống kê giữa tập luật sinh ra từ mô hình hỗn hợp và mô hình thuần kỹ thuật. Kết quả thực nghiệm được tổng hợp trong Bảng 5.3.

Chỉ số	Chiến lược Kỹ thuật	Chiến lược Hỗn hợp	Mức cải thiện (%)
Tổng số luật tìm thấy	16035	24188	+50.84%
Độ tin cậy trung bình	29.27%	29.95%	+2.31%
Max dump confidence	38.17%	39.68%	+3.97%
Số lượng luật mua	1358	1821	+34.09%
Số lượng luật bán	1397	1882	+34.71%
Max pump lift	17.15	17.15	+0.0%

Bảng 5.3: So sánh hiệu quả giữa mô hình Hỗn hợp và mô hình Kỹ thuật

Đánh giá kết quả:

- Mô hình Hỗn hợp cho thấy sự gia tăng về số lượng luật được khai phá. Điều này cho thấy việc tích hợp thêm các đặc trưng tin tức đã giúp thuật toán FP-Growth phát hiện được nhiều mẫu hình thị trường hơn so với khi chỉ sử dụng dữ liệu giá thuần túy.
- Độ tin cậy trung bình của các luật tăng nhẹ. Mặc dù mức tăng không quá lớn, kết quả này cho thấy việc bổ sung thông tin tin tức giúp các luật giao dịch trở nên nhất quán hơn, thay vì chỉ dựa vào các dao động ngắn hạn của giá.
- Số lượng luật mua tăng 34.09%, số lượng luật bán tăng 34.71%. Mô hình có khả năng nhận diện được nhiều kịch bản mua/bán hơn và từ đó nâng cao tính linh hoạt trong chiến lược giao dịch.
- Chỉ số Max dump confidence tăng từ 38.17% lên 39.68%, cho thấy các tín hiệu bán mạnh trở nên đáng tin cậy hơn khi kết hợp tin tức. Trong khi đó, chỉ số Max pump lift giữ nguyên, chứng tỏ việc có thêm luật mua và luật bán không làm suy giảm sức mạnh của các luật có hiệu quả cao nhất.

Chương 6

Thử nghiệm và đánh giá

Mục tiêu của chương này là mô phỏng lại quá trình giao dịch thực tế trên tập dữ liệu kiểm thử từ 01/01/2025 đến 06/12/2025. Để đảm bảo tính khách quan, nhân quả và tránh hiện tượng quá khớp, toàn bộ quy trình từ xử lý dữ liệu đến ra quyết định đều phải tuân thủ nghiêm ngặt nguyên tắc: "Tại thời điểm T, hệ thống chỉ được sử dụng thông tin có sẵn từ T trở về trước".

6.1 Chuẩn bị dữ liệu giá

6.1.1 Trích xuất đặc trưng dữ liệu

Quy trình tạo đặc trưng trên tập kiểm thử (Test) được thực hiện hoàn toàn tương tự như trên tập huấn luyện (Train) để đảm bảo sự nhất quán cho mô hình. Các nhóm chỉ báo đã đề cập, bao gồm: Chỉ báo kỹ thuật cơ bản (RSI, EMA), các đặc trưng Smart Money (OB, FVG) và các đặc trưng cấu trúc Wyckoff.

Tuy nhiên, một điểm khác biệt cốt lõi nằm ở bước *chuẩn hóa dữ liệu* trước khi đưa vào mô hình phân cụm K-Means. Thay vì tính toán trung bình (μ) và độ lệch chuẩn (σ) trực tiếp trên tập test, nhóm sử dụng lại các tham số thống kê đã học được từ tập train.

$$X_{test\_scaled} = \frac{X_{test} - \mu_{train}}{\sigma_{train}} \tag{6.1}$$

Nếu chúng ta chuẩn hóa lại tập Test bằng μ_{test}, sẽ đồng nghĩa với việc ta để dữ liệu tương lai ảnh hưởng đến dữ liệu hiện tại (biết trước giá trung bình của năm 2025). Việc tái sử dụng μ_{train} đảm bảo tính công bằng, mô phỏng đúng bối cảnh thực tế khi hệ thống vận hành. Mặc dù các đặc trưng thị trường có thể thay đổi theo thời gian, nhưng do các biến đầu vào của chúng ta (như độ lệch EMA, RSI, $Efficiency$) đều là các đại lượng tương đối, nên việc sử dụng tham số từ quá khứ vẫn đảm bảo độ chính xác cao.

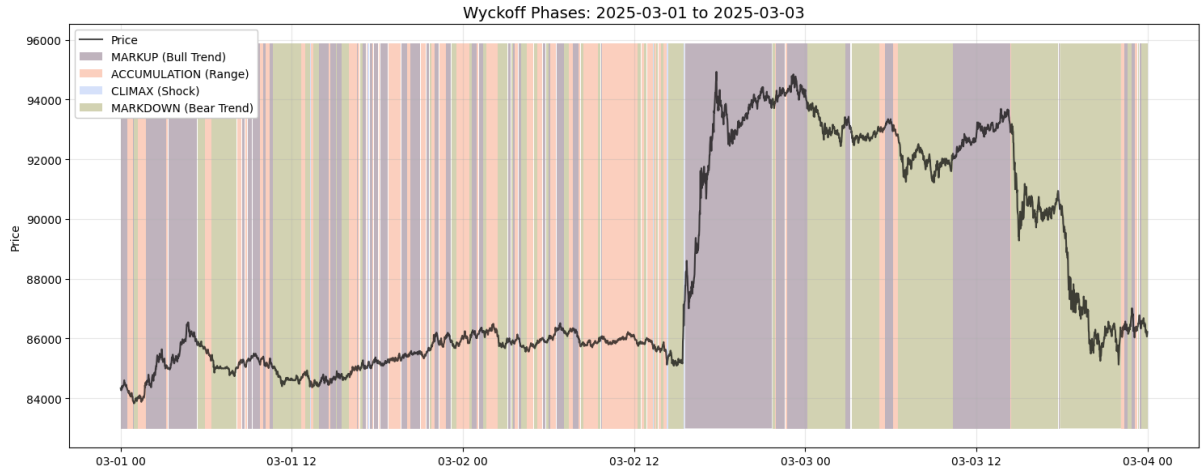
6.1.2 Dự báo trạng thái Wyckoff và tạo tập giao dịch

Sau khi đã có đầy đủ các đặc trưng thô, bước tiếp theo là chuyển đổi chúng thành định dạng giao dịch để có thể khớp lệnh với các luật kết hợp đã tìm được. Quy trình này bao gồm 6 bước tuần tự:

1. **Tải dữ liệu:** Đọc dữ liệu Train (để lấy tham số chuẩn hóa) và dữ liệu Test.
2. **Tạo đặc trưng thô:** Áp dụng các hàm tính toán chỉ báo cho cả hai tập dữ liệu.
3. **Đồng bộ hóa dữ liệu:** Loại bỏ các hàng có giá trị `NaN` để đảm bảo hai tập dữ liệu tương thích về cấu trúc.
4. **Chuẩn hóa dữ liệu:** Sử dụng `StandardScaler` trên tập Train. Điều này giúp bộ chuẩn hóa học được dữ liệu phân phối chuẩn quá khứ.
5. **Dự báo pha thị trường:** Sử dụng mô hình K-Means (đã lưu từ giai đoạn huấn luyện) để dự báo cụm pha thị trường cho tập Test.

6. **Rời rạc hóa:** Chuyển đổi các giá trị số liên tục và nhãn cụm thành các nhãn văn bản (ví dụ: *RSI_ Oversold, Phase_ Markup*) để sẵn sàng cho việc tra cứu luật.

Kết quả trả về là một bộ dữ liệu chứa các giao dịch của năm 2025, với đầy đủ các nhãn trạng thái (Phase, RSI, Volume...) đã được đồng bộ hóa với hệ tri thức của mô hình.



Hình 6.1: Trực quan hoá kết quả của mô hình phân cụm (K-Means) trên một phần tập Test.

6.2 Chuẩn bị dữ liệu tin tức

Song song với việc tái tạo dữ liệu giá, dòng dữ liệu tin tức cho năm 2025 cũng cần được xử lý và chuẩn hóa để trích xuất các tín hiệu đầu vào cho mô hình. Quy trình này bao gồm hai công đoạn chính: xử lý các chỉ số phản ứng định lượng và phân loại chủ đề dựa trên nội dung văn bản.

6.2.1 Xử lý văn bản và trích xuất đặc trưng phản ứng

Quy trình xử lý bắt đầu bằng việc làm sạch văn bản và lọc dữ liệu để chỉ giữ lại các tin tức liên quan trực tiếp đến Bitcoin trong giai đoạn kiểm thử (từ 01/01/2025 trở đi). Sau đó, các chỉ số tương tác thô (like, share, comment...) được tổng hợp và chuẩn hóa bằng kỹ thuật Chuẩn hóa động trong khung thời gian 24 tiếng. Điều quan trọng cần nhấn mạnh là việc sử dụng phương pháp tính động thay vì lấy trung bình trên toàn bộ dữ liệu sẽ giúp mô hình thích nghi với sự thay đổi trong hành vi người dùng theo thời gian. Một tin tức có 100 like trong giai đoạn ảm đạm sẽ có Z-Score cao hơn nhiều so với tin tức tương tự trong giai đoạn sôi động, qua đó phản ánh đúng mức độ dị thường của sự kiện.

Cuối cùng, các giá trị Z-Score liên tục sẽ được rời rạc hóa thành các nhãn (*Sent_ Positive, Buzz_ High...*) để tương thích với bộ luật kết hợp.

6.2.2 Phân loại chủ đề dựa trên Từ khóa

Đối với nhiệm vụ phân loại chủ đề cho tập kiểm thử, nhóm áp dụng một phương pháp tiếp cận khác biệt và hiệu quả hơn so với giai đoạn huấn luyện. Thay vì chạy lại mô hình SBERT và K-Means (vốn tốn kém tài nguyên và khó kiểm soát nhãn cụm mới sinh ra), nhóm sử dụng phương pháp **Bag-of-Words** dựa trên bộ từ điển từ khóa đã được trích xuất từ tập train.

Quy trình cụ thể như sau:

1. **Xây dựng từ điển chủ đề:** Từ kết quả phân cụm ở Chương 4, nhóm đã lưu lại danh sách 1000 từ khóa phổ biến nhất đại diện cho mỗi chủ đề (ví dụ: *"sec", "etf"* cho Topic_Macro_Politics).

2. **So khớp từ khóa:** Với mỗi tiêu đề tin tức mới trong năm 2025, hệ thống sẽ tách từ và đếm số lượng từ khóa trùng khớp với từng bộ từ điển chủ đề.

3. **Gán nhãn:** Tiêu đề sẽ được gán cho chủ đề có số lượng từ khóa trùng khớp nhiều nhất. Trong trường hợp có sự tranh chấp (số lượng trùng khớp bằng nhau), hệ thống ưu tiên các chủ đề nhạy cảm như *Topic_Price_Prediction* hoặc *Topic_Altcoins_Memes* để bắt tín hiệu tốt hơn.

Phương pháp này đảm bảo tính nhất quán của hệ thống nhãn chủ đề giữa quá khứ và tương lai, đồng thời đem lại hiệu năng cao cho hệ thống.

6.3 Xây dựng chiến lược hỗn hợp

Sau khi đã có tập dữ liệu giao dịch và kho tri thức là các luật kết hợp, bước tiếp theo là xây dựng một cơ chế định lượng để chuyển đổi các mẫu hình rời rạc thành một con số tín hiệu duy nhất. Hệ thống này đóng vai trò như bộ tư duy trung tâm, quyết định cường độ của tín hiệu **mua** hoặc **bán** tại từng thời điểm.

6.3.1 Bộ lọc và trọng số hóa quy tắc

Trong hàng nghìn luật được sinh ra từ thuật toán FP-Growth, không phải luật nào cũng có giá trị dự báo như nhau. Một số luật có độ tin cậy thấp hoặc độ nâng không đáng kể cần phải được loại bỏ để tránh gây nhiễu. Mục tiêu là đảm bảo một quy tắc vừa có độ chính xác cao, vừa có tính liên kết mạnh sẽ có trọng số lớn hơn trong quyết định cuối cùng. Nhóm thực hiện lọc và gán trọng số theo quy trình sau:

1. **Lọc luật hành động:** Chỉ giữ lại các luật mà vế phải là kết quả giao dịch mong muốn (pump hoặc dump).
2. **Áp dụng ngưỡng chất lượng:** Các luật phải thỏa mãn ngưỡng $confidence \geq 0.3$ và $lift > 1.0$.
3. **Tính toán trọng số:** Đây là bước quan trọng nhất. Thay vì coi các luật là bình đẳng, nhóm gán cho mỗi luật một trọng số dựa trên chất lượng của nó:

$$Power(R) = Confidence(R) \times Lift(R) \tag{6.2}$$

6.3.2 Thuật toán tổng hợp tín hiệu

Trong thực tế, tại một thời điểm t, trạng thái thị trường (bao gồm hàng chục đặc trưng như RSI, Volume, News...) có thể kích hoạt đồng thời nhiều luật khác nhau, thậm chí là các luật đối nghịch.

- Ví dụ: Tập đặc trưng $\{A, B\}$ kích hoạt luật mua.
- Nhưng đồng thời, tập đặc trưng $\{C, D\}$ trong cùng thời điểm đó lại kích hoạt luật bán.

Nếu chỉ chọn một luật duy nhất, hệ thống sẽ bị phiến diện. Do đó, nhóm sử dụng cơ chế *cộng hưởng tín hiệu*. Tại mỗi cây nến, thuật toán sẽ tính tổng sức mạnh của tất cả các luật mua khớp lệnh ($Score_{Bull}$) và tất cả các luật bán khớp lệnh ($Score_{Bear}$). Tín hiệu ròng là hiệu số giữa hai lực lượng này:

$$Net\_Signal_t = \sum_{r \in Rules_{Buy}} Power(r) - \sum_{r \in Rules_{Sell}} Power(r) \tag{6.3}$$

Nếu $Net\_Signal > 0$, phe mua đang áp đảo và ngược lại. Độ lớn của giá trị này thể hiện mức độ đồng thuận của hệ thống.

6.3.3 Thiết lập ngưỡng hành động

Biến tín hiệu đầu ra từ mô hình chấm điểm (Net_Signal) bản chất là một giá trị liên tục, đại diện cho cán cân cung cầu tại từng thời điểm. Để chuyển đổi đại lượng này thành các quyết định giao dịch nhị phân dứt khoát, nhóm nghiên cứu thiết lập một **ngưỡng trong khoảng** ± 5.0.

Cơ chế ra quyết định hoạt động dựa trên nguyên tắc **đồng thuận cao**. Cụ thể, một lệnh Mua chỉ được kích hoạt khi Net_Signal vượt quá ngưỡng $+5.0$, đồng nghĩa với việc tổng sức mạnh của các luật Mua phải vượt trội hơn các luật Bán ít nhất 5 đơn vị. Ngược lại, lệnh Bán sẽ được thực thi khi áp lực bán áp đảo, đẩy tín hiệu xuống dưới mức -5.0. Khoảng giá trị nằm giữa biên độ này (từ -5.0 đến 5.0) được định nghĩa là vùng Nắm giữ. Đây là vùng đệm chiến lược đóng vai trò như một bộ lọc nhiễu, ngăn hệ thống thực hiện các giao dịch sai lầm khi xu hướng thị trường chưa thực sự rõ ràng.

6.4 Đánh giá các chiến lược so sánh

Giai đoạn cuối cùng và quan trọng nhất của nghiên cứu là đưa các mô hình vào môi trường giả lập để đánh giá hiệu quả đầu tư thực tế. Trong phần này, nhóm trình bày chi tiết về thiết kế kịch bản thử nghiệm, cơ chế vận hành của bộ máy mô phỏng và các tham số chiến lược được áp dụng.

6.4.1 Các chiến lược so sánh

Để kiểm chứng giả thuyết nghiên cứu về vai trò của thông tin trong việc dự báo biến động giá, nhóm áp dụng phương pháp kiểm thử đối chứng trên tập dữ liệu năm 2025. Mục tiêu cốt lõi của thiết kế này là cô lập tác động của biến số tin tức để đo lường chính xác đóng góp của nó vào hiệu quả giao dịch cuối cùng. Thí nghiệm được cấu trúc xoay quanh ba chiến lược hoạt động song song, đại diện cho ba tư duy đầu tư khác nhau trên thị trường.

Chiến lược Mua và nắm giữ thụ động *Buy & Hold*. Theo chiến lược này, hệ thống giả định thực hiện một lệnh mua Bitcoin tại thời điểm mở cửa của giai đoạn kiểm thử (01/01/2025) và duy trì vị thế bất chấp mọi biến động cho đến khi kết thúc kỳ. Đây là thước đo tiêu chuẩn vàng trong mọi bài toán tài chính nhằm đánh giá sự biến động của thị trường. Việc so sánh với chiến lược này là bắt buộc để xác định xem liệu các nỗ lực giao dịch chủ động, với độ phức tạp về thuật toán và chi phí giao dịch phát sinh, có thực sự mang lại hiệu quả so với mức tăng trưởng tự nhiên của tài sản hay không.

Chiến lược Kỹ thuật *Technical Only*. Chiến lược đại diện cho trường phái giao dịch định lượng truyền thống dựa trên chỉ báo kỹ thuật. Hệ thống ra quyết định hoàn toàn dựa trên các tập luật kết hợp được khai phá từ dữ liệu giá và khối lượng, bao gồm các *chỉ báo động lượng (RSI), cấu trúc pha Wyckoff và các mô hình nến SMC*. Chiến lược này hoạt động dựa trên giả định của phân tích kỹ thuật rằng giá sẽ nói lên câu chuyện và các mẫu hình hành vi trong quá khứ sẽ có xu hướng lặp lại. Tuy nhiên, điểm yếu cố hữu của phương pháp này là thiếu vắng sự nhạy bén với các sự kiện vĩ mô hoặc các tin tức thiên nga đen, dẫn đến nguy cơ sinh ra nhiều tín hiệu nhiễu trong các giai đoạn thị trường biến động phi kỹ thuật.

Chiến lược Hỗn hợp AI *AI Hybrid*. Đây là chiến lược đề xuất chính của nghiên cứu. Khác với hai hướng tiếp cận trên, chiến lược này tích hợp thêm chiều dữ liệu phi cấu trúc từ tin tức (cảm xúc, mức độ quan tâm, chủ đề) vào bộ quy tắc ra quyết định. Tín hiệu giao dịch chỉ được kích hoạt khi và chỉ khi có sự đồng thuận giữa tín hiệu kỹ thuật và luồng thông tin. Ví dụ, một tín hiệu mua kỹ thuật sẽ chỉ được thực thi nếu đi kèm với tin tức tích cực hoặc sự bùng nổ về mức độ quan tâm của cộng đồng. Kỳ vọng đặt ra cho mô hình lai ghép này là khả năng lọc nhiễu vượt trội, giúp hệ thống hạn chế các giao dịch sai lầm và tối ưu hóa tỷ lệ lợi nhuận trên rủi ro.

6.4.2 Mô hình mô phỏng giao dịch

Để đảm bảo tính thực tế, nhóm đã xây dựng một mô hình kiểm thử dựa trên dữ liệu của từng cây nến phút. Mô hình này mô phỏng các ràng buộc thực tế như phí giao dịch và quy tắc quản lý vốn.

1. **Vốn khởi điểm:** 10,000 USD.
2. **Phí giao dịch:** 0.1% trên tổng giá trị lệnh (mức phí tiêu chuẩn trên nhiều sàn giao dịch).
3. **Trạng thái vị thế:** Tại mỗi thời điểm, tài khoản chỉ có thể ở một trong hai trạng thái:
    - **Full cash:** Nắm giữ 100% tiền mặt. Khi có tín hiệu mua, hệ thống sẽ dùng toàn bộ tiền để mua Bitcoin (sau khi trừ phí).
    - **Full crypto:** Nắm giữ 100% Bitcoin. Khi có tín hiệu bán, hệ thống sẽ bán toàn bộ lượng coin để chuyển về tiền mặt.
4. **Định giá tài sản:** Tại mỗi bước thời gian, tổng giá trị danh mục được cập nhật theo giá thị trường hiện hành để theo dõi biến động tài sản theo thời gian thực.

6.5 Kết quả thực nghiệm

Trong toàn bộ gần 490,000 phút giao dịch, cả chiến lược Kỹ thuật và chiến lược Hỗn hợp chỉ thực hiện số lượng lệnh rất hạn chế (lần lượt là khoảng 1,439 và 1,464 lệnh tổng cộng). Đáng chú ý, tỷ lệ thời gian hệ thống ở trạng thái chờ (HOLD) chiếm tới hơn **99.7%**. Điều này khẳng định mô hình không hoạt động theo kiểu giao dịch nhiều, mô hình kiên nhẫn chờ đợi và chỉ hành động khi cơ hội có xác suất thắng cao xuất hiện.

Ngoài ra, trong giao dịch thuật toán, chi phí giao dịch đóng vai trò là biến số ngoại sinh then chốt, quyết định trực tiếp đến khả năng sinh lời thực tế của mô hình. Để kiểm chứng độ bền vững của các chiến lược đề xuất, nghiên cứu tiến hành phân tích độ nhạy trên 4 kịch bản phí giao dịch khác nhau: 1.0%, 0.5%, 0.1% (mức chuẩn) và 0.0% (lý thuyết). Kết quả được phân tích trên hai khía cạnh chính là hiệu suất lợi nhuận và mức độ rủi ro.

6.5.1 Phân tích hiệu suất lợi nhuận ròng

Bảng 6.1 trình bày chi tiết lợi nhuận ròng của ba chiến lược qua các kịch bản phí. Tại mức phí tiêu chuẩn 0.1%, chiến lược AI Hybrid thể hiện ưu thế vượt trội khi đạt mức lợi nhuận cao nhất +10.92%, đánh bại hoàn toàn chiến lược Kỹ thuật (+8.39%) và thị trường chung đang suy giảm (Buy & Hold lỗ -4.87%). Kết quả này khẳng định rằng việc tích hợp dữ liệu tin tức giúp hệ thống nắm bắt được các cơ hội giao dịch chất lượng cao mà phân tích kỹ thuật đơn thuần thường bỏ lỡ.

Đáng chú ý, khi xét đến các kịch bản phí cao (0.5% - 1.0%), sự khác biệt về tính hiệu quả giữa hai mô hình thuật toán được bộc lộ rõ rệt. Chiến lược Kỹ thuật sụp đổ hoàn toàn với mức lỗ lên tới -21.69% do tần suất giao dịch quá dày đặc (361 lệnh), khiến chi phí bào mòn toàn bộ lợi nhuận. Ngược lại, chiến lược AI Hybrid, nhờ bộ lọc tin tức khắt khe, chỉ thực hiện 182 lệnh và duy trì mức lỗ thấp hơn đáng kể (-5.85%), gần tương đương với mức sụt giảm tự nhiên của thị trường. Điều này chứng minh khả năng kháng nhiễu và bảo toàn vốn tốt hơn của mô hình lai trong môi trường giao dịch khắc nghiệt.

Bảng 6.1: Tổng hợp Lợi nhuận ròng (%) của các chiến lược theo mức phí

Chiến lược	**Mức phí giao dịch**			
	1.0%	**0.5%**	**0.1% (Chuẩn)**	**0.0% (Lý thuyết)**
Buy & Hold	**-4.96%**	**-4.91%**	-4.87%	-4.86%
Technical Only	-21.69%	-21.69%	+8.39%	+12.37%
AI Hybrid	-5.85%	-5.85%	**+10.92%**	**+12.96%**

6.5.2 Phân tích rủi ro

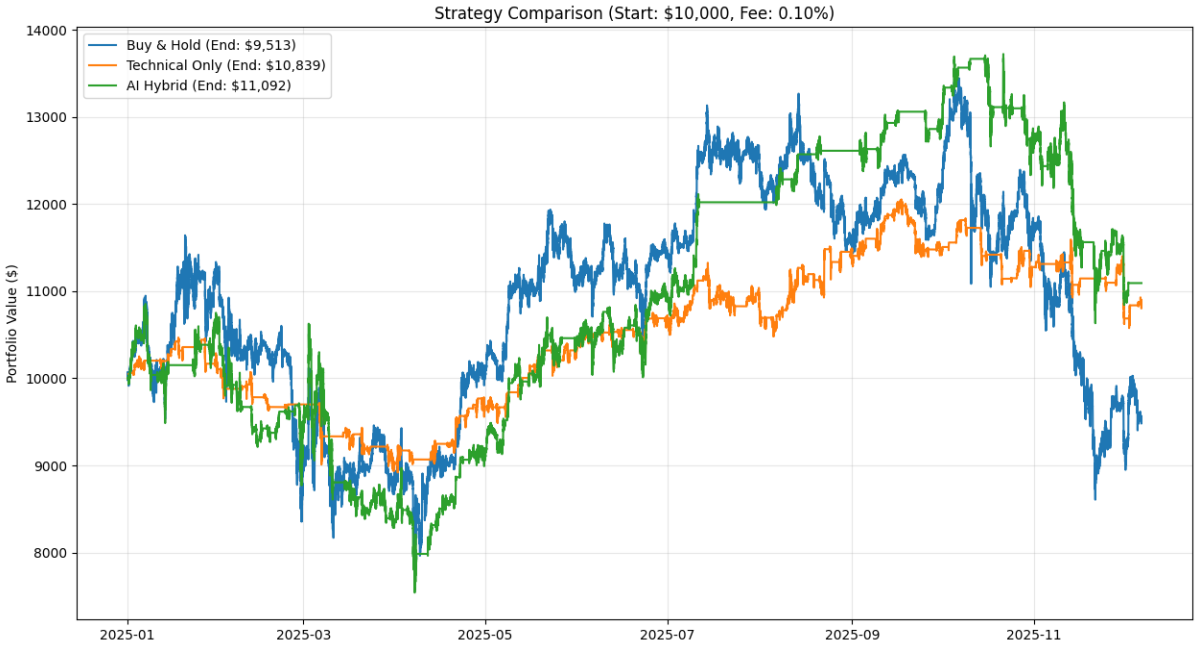
Bên cạnh lợi nhuận, các chỉ số về độ sụt giảm lớn nhất của tài khoản và giá trị tài sản đỉnh được tổng hợp tại Bảng 6.2 để đánh giá rủi ro và lợi nhuận của chiến lược. Một quan sát quan trọng tại kịch bản phí 0.1% là độ sụt giảm của chiến lược AI Hybrid (-30.54%) lại cao hơn so với chiến lược Kỹ thuật (-14.80%). Hiện tượng này phản ánh sự đánh đổi cốt lõi trong triết lý giao dịch của hai mô hình. Chiến lược Kỹ thuật hoạt động dựa trên các dao động ngắn hạn, thường chốt lời hoặc cắt lỗ rất nhanh, giúp đường cong vốn mượt mà hơn nhưng lại hạn chế tiềm năng tăng trưởng, thể hiện qua mức đỉnh tài sản chỉ đạt 12,052 USD.

Trái lại, chiến lược AI Hybrid mang đặc tính bám theo xu hướng. Khi có sự xác nhận của tin tức, mô hình chấp nhận giữ vị thế lâu hơn để tận dụng tối đa đà tăng, đẩy giá trị tài sản lên mức đỉnh cao nhất là 13,721 USD. Hệ quả tất yếu là mô hình phải chấp nhận các đợt rung lắc mạnh hơn của thị trường trước khi tín hiệu đảo chiều được xác nhận, dẫn đến mức sụt giảm tạm thời lớn hơn. Tuy nhiên, xét trên hiệu quả tổng thể, chiến lược AI Hybrid vẫn đảm bảo mức sụt giảm thấp hơn so với việc nắm giữ thụ động (Buy & Hold: -35.99%), đồng thời mang lại lợi nhuận cuối cùng cao nhất.

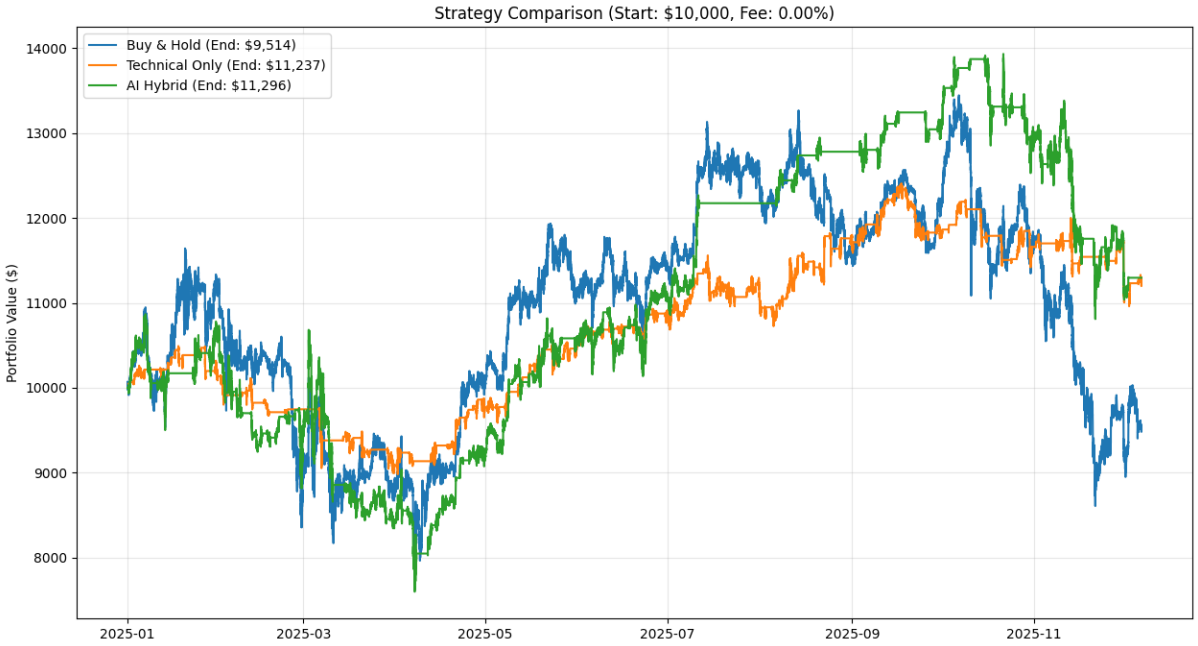
Tổng kết lại, kết quả thực nghiệm chứng minh rằng việc tích hợp dữ liệu tin tức vào mô hình giao dịch mang lại hiệu quả vượt trội về mặt lợi nhuận ròng trong điều kiện thị trường thực tế. Mặc dù phải chịu mức độ biến động tài sản lớn hơn trong ngắn hạn so với chiến lược thuần kỹ thuật, nhưng chiến lược AI Hybrid tạo ra mức đỉnh tài sản cao hơn và tránh được rủi ro bào mòn vốn do chi phí giao dịch, đáp ứng tốt mục tiêu tối ưu hóa tăng trưởng tài sản dài hạn.

Bảng 6.2: Tổng hợp chỉ số rủi ro và giá trị tài sản đỉnh theo mức phí

Chiến lược	**Độ sụt giảm lớn nhất (%)**				**Giá trị tài sản cao nhất ($)**			
	1.0%	**0.5%**	**0.1%**	**0.0%**	**1.0%**	**0.5%**	**0.1%**	**0.0%**
Buy & Hold	-35.99	-35.99	-35.99	-35.99	**13,432**	**13,438**	13,444	13,445
Technical Only	-25.21	-25.21	-14.80	-14.41	10,253	10,253	12,052	12,405
AI Hybrid	-34.43	-34.43	-30.54	-30.09	11,995	11,995	**13,721**	**13,933**



Hình 6.2: So sánh giá trị tài sản của các chiến lược theo thời gian (mức phí giao dịch 0.1%)



Hình 6.3: So sánh giá trị tài sản của các chiến lược theo thời gian (mức phí giao dịch 0%)

Chương 7

Tổng kết và phương hướng phát triển

7.1 Tổng kết

Báo cáo đã hoàn thành mục tiêu xây dựng một hệ thống hỗ trợ ra quyết định giao dịch Bitcoin tự động dựa trên sự kết hợp giữa kỹ thuật khai phá dữ liệu truyền thống và xử lý ngôn ngữ tự nhiên. Kết quả thực nghiệm cho thấy mô hình Hỗn hợp (AI Hybrid) không chỉ vượt trội hơn so với chiến lược mua và nắm giữ thụ động, mà còn khắc phục được nhược điểm giao dịch của các phương pháp phân tích kỹ thuật đơn thuần. Điểm mạnh cốt lõi của phương pháp đề xuất nằm ở khả năng tích hợp ngữ cảnh thông tin vào dữ liệu giá. Bằng cách sử dụng tin tức như một bộ lọc nhiễu, hệ thống đã chuyển đổi tư duy giao dịch từ việc phản ứng với mọi biến động giá sang một chiến thuật thông minh hơn, chỉ tham gia khi có sự đồng thuận cao giữa tín hiệu kỹ thuật và tâm lý đám đông. Bên cạnh đó, việc sử dụng các luật kết hợp cho phép người sử dụng hiểu rõ lý do đằng sau mỗi quyết định mua bán, khác biệt hoàn toàn với tính chất không thể lý giải của các mô hình học sâu hiện đại.

Tuy nhiên, nghiên cứu vẫn tồn tại những hạn chế nhất định cần nhìn nhận khách quan. Điểm yếu lớn nhất của mô hình hiện tại là độ trễ trong việc phản ứng với các tin tức thời gian thực do cơ chế tổng hợp dữ liệu theo khung giờ. Điều này khiến hệ thống có thể bỏ lỡ các cơ hội giao dịch chớp nhoáng hoặc phản ứng chậm trước các cú sập giá bất ngờ. Ngoài ra, chiến lược bám theo xu hướng dựa trên sự xác nhận của tin tức cũng buộc mô hình phải chấp nhận mức độ biến động tài sản cao hơn trong ngắn hạn để đi hết con sóng lớn, đòi hỏi nhà đầu tư phải có tâm lý vững vàng và khả năng chịu đựng rủi ro tốt. Hơn nữa, độ phức tạp trong việc vận hành và bảo trì luồng dữ liệu đa nguồn cũng là một rào cản không nhỏ khi triển khai hệ thống vào thực tế.

7.2 Phương hướng phát triển

Để khắc phục các hạn chế nêu trên và nâng cao hiệu suất của hệ thống trong tương lai, nhóm đề xuất một số hướng cải thiện tiềm năng. Trước hết, về mặt thuật toán, việc thay thế hoặc bổ sung các luật kết hợp tĩnh bằng các mô hình học sâu chuỗi thời gian như LSTM (Long Short-Term Memory) hay thậm chí là Transformer, đều là những hướng đi hứa hẹn. Các mô hình này có khả năng nắm bắt tốt hơn sự phụ thuộc thời gian dài hạn và các mối quan hệ phi tuyến tính phức tạp mà thuật toán FP-Growth có thể bỏ sót. Đồng thời, cơ chế Học tăng cường có thể được áp dụng để mô hình có thể thích ứng tốt với các giai đoạn của thị trường, thay vì sử dụng các giá trị cố định.

Về mặt dữ liệu, không gian đặc trưng cần được mở rộng thêm các nguồn thông tin định lượng khác bên cạnh giá và tin tức. Cụ thể, việc tích hợp dữ liệu On-chain (như dòng tiền vào-ra của những nhà đầu tư vốn lớn, tỷ lệ đòn bẩy, v.v.) sẽ cung cấp cái nhìn sâu sắc hơn về hành vi của các nhà tạo lập thị trường. Song song với đó, hệ thống xử lý ngôn ngữ tự nhiên cần được nâng cấp lên cơ chế thời gian thực (real-time) thay vì xử lý theo lô, nhằm đánh giá tác động của tin tức ngay khi chúng vừa được công bố. Cuối cùng, việc bổ sung thêm các quy tắc quản trị vốn nâng cao, như tự động điều chỉnh khối lượng lệnh dựa trên độ biến động hiện tại của thị trường, sẽ giúp giảm thiểu rủi ro lỗ vốn, và làm mượt đường cong tăng trưởng vốn trong dài hạn.

Tài liệu tham khảo

[1] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[2] Elie Bouri, Rangan Gupta, and Seyed Mehdi Hosseini. Herding behaviour in cryptocurrencies. *Finance Research Letters*, 29:216–221, 2019.

[3] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12, 2000.

[4] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3(1):3415, 2013.

[5] Marcos Lopez de Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[6] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[7] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[8] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[9] Soheil Rahsaz. Github: Crypto news dataset, 2025. https://github.com/soheilrahsaz/cryptoNewsDataset.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[11] Team Sentence-Transformers. `all-MiniLM-L6-v2` embedding model, 2022. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[12] Richard D Wyckoff. *The Richard D. Wyckoff Method of Trading and Investing in Stocks*. Wyckoff Associates, 1931.