Xây dựng hệ thống xử lý dữ liệu lớn trong phân tích và dự đoán xu hướng thị trường chứng khoán Việt Nam

Viện Trí tuệ nhân tạo - Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội
Đào Tự Phát

Ngày 26 tháng 10 năm 2025

Tóm tắt nội dung

Thị trường chứng khoán, với hàng triệu giao dịch mỗi ngày, tạo ra một khối lượng dữ liệu khổng lồ (Big Data). Việc áp dụng các công nghệ xử lý dữ liệu lớn để phân tích nguồn dữ liệu này đã trở thành một công cụ hữu hiệu, giúp các nhà đầu tư đưa ra quyết định thông minh và hạn chế rủi ro. Bài tập này trình bày việc mô phỏng một hệ thống Big Data hoàn chỉnh để lưu trữ và xử lý dữ liệu chứng khoán. Dự án tập trung vào dữ liệu giao dịch lịch sử của thị trường Việt Nam. Kiến trúc hệ thống được xây dựng dựa trên hai công nghệ cốt lõi: **Hadoop (HDFS)** được sử dụng cho lưu trữ phân tán và **Apache Spark** được dùng cho xử lý dữ liệu song song. Dữ liệu được thu thập từ API công khai **vnstock**. Dự án minh họa một quy trình hoàn chỉnh từ khâu thu thập, lưu trữ phân tán đến xử lý dữ liệu chứng khoán quy mô lớn.

1 Giới thiệu

1.1 Bối cảnh và tính cấp thiết của đề tài

Thị trường chứng khoán là một môi trường phức tạp với hàng triệu lượt giao dịch diễn ra mỗi ngày, tạo ra một kịch bản không ngừng chuyển động. Khối lượng dữ liệu khổng lồ này (Big Data) vừa là một thách thức, vừa là một cơ hội lớn.

Việc áp dụng các công nghệ xử lý dữ liệu lớn vào phân tích môi trường chứng khoán đã trở thành một công cụ hữu hiệu cho các nhà đầu tư. Các phân tích sâu từ dữ liệu lớn, thay vì chỉ dựa vào trực giác hay các phân tích đơn lẻ, sẽ giúp các nhà đầu tư đưa ra những quyết định thông minh hơn, dự đoán xu hướng và hạn chế rủi ro trên một thị trường đầy biến động.

1.2 Mục tiêu dự án

Bài tập này hướng đến việc mô phỏng một hệ thống Big Data hoàn chỉnh sử dụng các công nghệ cốt lõi là Hadoop và Spark cho hai nhiệm vụ chính:

- **Lưu trữ:** Xây dựng một hệ thống tệp phân tán (HDFS) có khả năng lưu trữ dữ liệu chứng khoán lịch sử quy mô lớn.
- **Xử lý:** Xây dựng một cụm tính toán phân tán (Spark) để thực thi các tác vụ phân tích, thống kê và dự đoán trên tập dữ liệu đó.

2 Kiến trúc hệ thống

2.1 Công nghệ

Hệ thống tích hợp các công nghệ cốt lõi sau:

- **Hadoop (HDFS & YARN):** Được sử dụng làm nền tảng chính. HDFS đóng vai trò lưu trữ phân tán, và YARN chịu trách nhiệm quản lý tài nguyên cụm.
- **Spark:** Được sử dụng làm framework xử lý dữ liệu song song.

2.2 Kiến trúc hệ thống

2.2.1 Cụm lưu trữ HDFS

Đây là thành phần chịu trách nhiệm lưu trữ dữ liệu phân tán.

- **1** `namenode`**:** Quản lý metadata của hệ thống tệp.
- **4** `datanode`**:** Lưu trữ các block dữ liệu thực tế.

2.2.2 Cụm quản lý tài nguyên YARN

Hệ thống sử dụng YARN để quản lý và điều phối tài nguyên (CPU, RAM) cho toàn bộ cụm Hadoop.

- `resourcemanager`**:** Dịch vụ trung tâm, tiếp nhận yêu cầu từ các ứng dụng và cấp phát tài nguyên.
- `nodemanager`**:** Các dịch vụ chạy trên các node worker, chịu trách nhiệm khởi chạy và giám sát các tác vụ theo chỉ thị của ResourceManager.
- `historyserver`**:** Dịch vụ lưu trữ và cung cấp giao diện web để theo dõi lịch sử của các ứng dụng đã hoàn thành.

2.2.3 Cụm xử lý Spark

Một cụm Spark độc lập được thiết lập song song để thực thi các phân tích dữ liệu.

- `spark-master`**:** Dịch vụ quản lý của cụm Spark, nhận các tác vụ xử lý từ Jupyter Notebook và phân phối chúng đến các Spark Worker.
- `spark-worker`**:** Dịch vụ thực thi, trực tiếp đọc dữ liệu từ HDFS và thực hiện các phép tính toán, biến đổi dữ liệu.

3 Thu thập dữ liệu

Toàn bộ dữ liệu của thị trường chứng khoán Việt Nam được thu thập từ API của **VnStock**. Dự án này đã thu thập thông tin của 1722 công ty, đối với mỗi công ty, API được gọi để lấy dữ liệu cổ phiếu lịch sử từ ngày giao dịch đầu tiên của công ty đến ngày thu thập dữ liệu (ngày 26/10/2025).

Sau khi thu thập và loại bỏ dữ liệu bị lỗi, còn lại 1687 công ty. Dữ liệu của mỗi công ty được ghi vào một tệp CSV và bao gồm 6 trường: `Date` (ngày giao dịch), `Open` (giá mở cửa), `High` (giá cao nhất), `Low` (giá thấp nhất), `Close` (giá đóng cửa), `Volume` (khối lượng giao dịch). Các tệp CSV này tuân thủ nghiêm ngặt định dạng dữ liệu thị trường (OHLCV).

Sau khi dữ liệu được thu thập, chúng được gửi đến hệ thống tệp **HDFS** trước khi được xử lý bởi **Spark**.

4 Phân tích dữ liệu

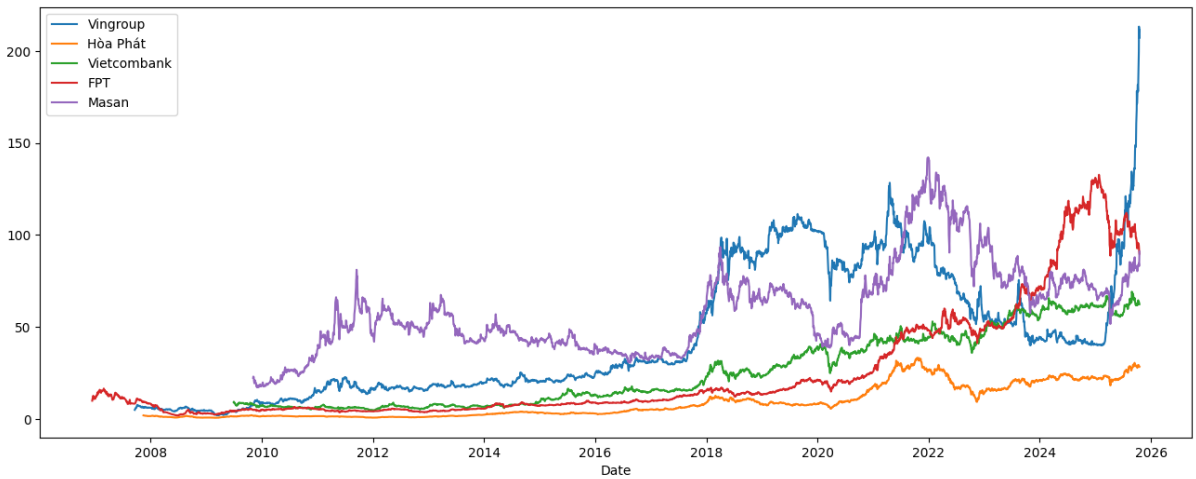
Để hiểu rõ hơn về hành vi của thị trường chứng khoán Việt Nam, dự án này đã tiến hành phân tích dữ liệu lịch sử từ 5 cổ phiếu lớn trên thị trường chứng khoán Việt Nam từ các lĩnh vực khác nhau, bao gồm: **Tập đoàn Vingroup (VIC)**, **Tập đoàn Hòa Phát (HPG)**, **Ngân hàng Vietcombank (VCB)**, **Tập đoàn FPT (FPT)**, và **Tập đoàn Masan (MSN)**.

Phân tích này, được thực hiện bằng PySpark, tập trung vào các chỉ số tài chính chính từ dữ liệu giao dịch hàng ngày (giá mở, cao, thấp, đóng cửa và khối lượng giao dịch).

4.1 Phân tích xu hướng (Trend Analysis)

4.1.1 Giá trung bình hàng ngày (Stock Daily Mean Price)

Để phân tích xu hướng dài hạn, dự án này tính toán giá trung bình hàng ngày ($Mean = \frac{High+Low}{2}$) cho mỗi cổ phiếu để làm mượt các biến động trong ngày.



Hình 1: Biểu đồ đường so sánh giá trung bình (Mean Price) theo ngày.

Biểu đồ đường (Hình 1) thể hiện sự biến động giá của các mã này từ khoảng năm 2008 đến cuối năm 2025.

- Xu hướng chung: Cả 5 mã đều cho thấy xu hướng tăng trưởng dài hạn, tuy nhiên với mức độ biến động và quỹ đạo tăng trưởng rất khác nhau.
- Vingroup (VIC - Xanh dương): Thể hiện sự tăng trưởng tương đối ổn định cho đến khoảng năm 2017, sau đó bước vào một chu kỳ tăng trưởng mạnh mẽ, đạt đỉnh vào khoảng năm 2021. Sau giai đoạn điều chỉnh, cổ phiếu này có một cú tăng vọt đột biến vào cuối năm 2025, trở thành mã có giá trị cao nhất trên biểu đồ.
- Hòa Phát (HPG - Cam): Gần như đi ngang ở mức giá rất thấp trong phần lớn thời gian, trước khi có một đợt tăng trưởng đột biến vào năm 2021. Đây là mã có thị giá thấp nhất trong nhóm được phân tích.
- Vietcombank (VCB - Xanh lá) & FPT (Đỏ): Cả hai mã này đều có chung đặc điểm là tăng trưởng chậm, ổn định trong thập kỷ đầu tiên (khoảng 2008-2018). Từ 2019 trở đi, cả hai bắt đầu một xu hướng tăng tốc mạnh mẽ và bền bỉ hơn.
- Masan (MSN - Tím): Mã này thể hiện mức độ biến động cao, thường xuyên dẫn đầu thị trường về giá trong giai đoạn 2018-2025, với các đỉnh lớn vào năm 2022 và 2025, nhưng cũng xen kẽ các đợt sụt giảm sâu.

4.1.2 Trung bình động (Moving Averages)

Phần này sử dụng hai chỉ báo Trung bình động đơn giản (SMA - simple moving averages) để phân tích xu hướng giá của các cổ phiếu trên thị trường Việt Nam:

- SMA 50 ngày (Đường nét đứt màu cam): Đại diện cho xu hướng ngắn hạn đến trung hạn.
- SMA 200 ngày (Đường nét đứt màu đỏ): Đại diện cho xu hướng dài hạn.

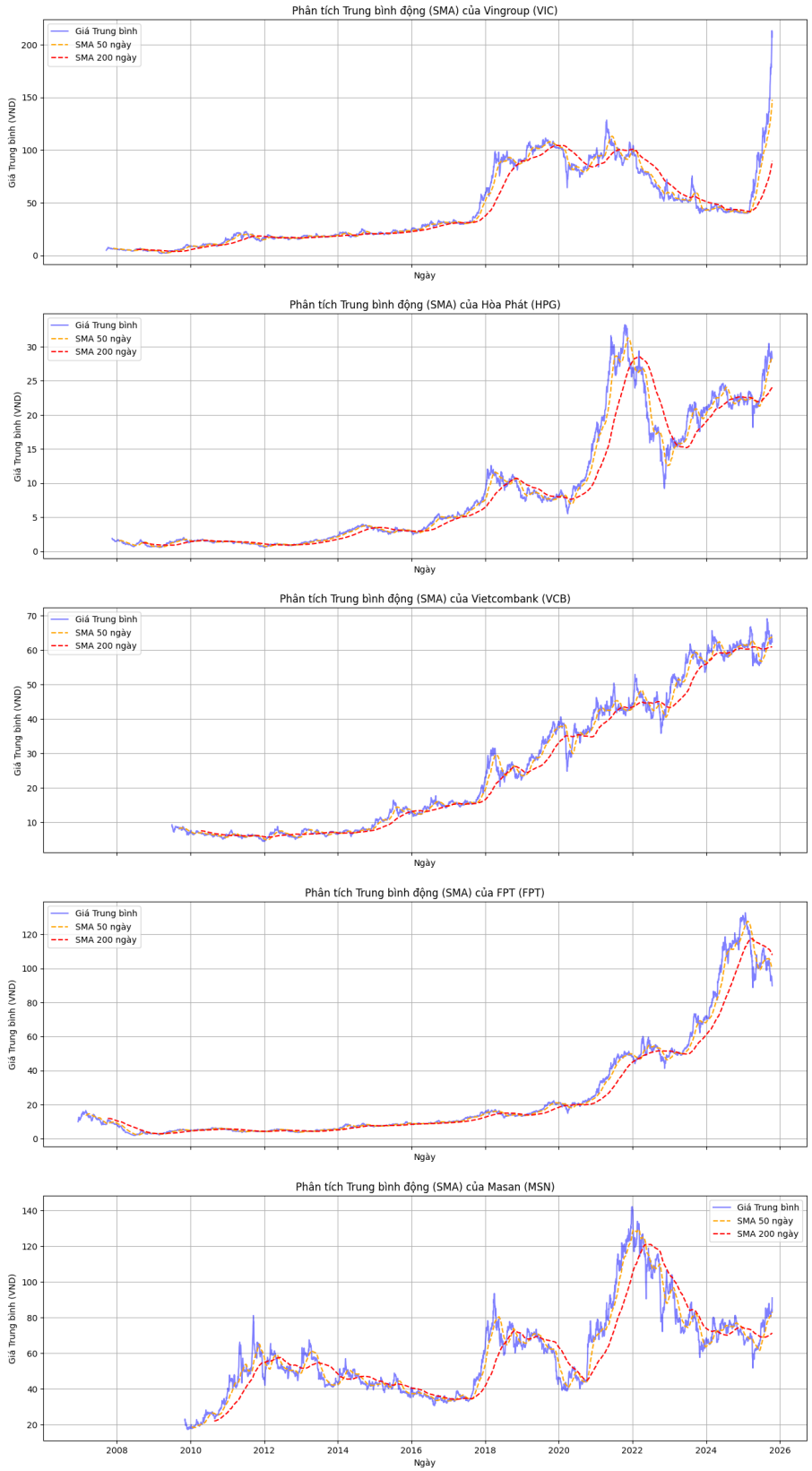
Sự giao cắt giữa hai đường này cung cấp các tín hiệu giao dịch quan trọng:

- Giao cắt vàng (Golden Cross): Khi đường SMA 50 cắt lên trên đường SMA 200, báo hiệu một xu hướng tăng giá mạnh mẽ.
- Giao cắt tử thần (Death Cross): Khi đường SMA 50 cắt xuống dưới đường SMA 200, báo hiệu một xu hướng giảm giá.

Biểu đồ của Vingroup (VIC) cho thấy tính chu kỳ rõ rệt.

- Một tín hiệu *Golden Cross* xuất hiện vào khoảng giữa năm 2017, bắt đầu một chu kỳ tăng giá mạnh mẽ kéo dài đến 2018 và duy trì mức giá cao cho đến đầu năm 2022.
- Tín hiệu *Death Cross* xảy ra vào khoảng quý 2 năm 2022, xác nhận một xu hướng giảm giá dài hạn.
- Đến cuối giai đoạn phân tích (2025), giá cổ phiếu đang vận động bên dưới cả hai đường SMA, tuy nhiên đường SMA 50 đang có dấu hiệu đi ngang và tiệm cận SMA 200, cho thấy khả năng có thể sớm kết thúc xu hướng giảm.

Hòa Phát (HPG) thể hiện một trong những chu kỳ tăng trưởng mạnh mẽ nhất.



Hình 2: Phân tích Trung bình động (SMA) của 5 mã cổ phiếu lớn.

- Giai đoạn 2016-2018 chứng kiến một xu hướng tăng giá khi SMA 50 liên tục nằm trên SMA 200.
- Đáng chú ý nhất là tín hiệu *Golden Cross* vào giữa năm 2020, khởi đầu cho một đợt siêu tăng trưởng kéo dài đến cuối năm 2021.
- Tín hiệu *Death Cross* rõ ràng vào đầu năm 2022 đã báo trước một đợt điều chỉnh sâu và kéo dài.
- Một tín hiệu *Golden Cross* mới đã xuất hiện vào giữa năm 2023, và kể từ đó, SMA 50 duy trì vị thế trên SMA 200, xác nhận HPG đang ở trong một chu kỳ tăng giá mới, bất chấp các đợt điều chỉnh ngắn hạn.

Vietcombank (VCB) là cổ phiếu có xu hướng tăng trưởng dài hạn ổn định và bền vững nhất trong nhóm.

- Kể từ năm 2012, đường SMA 50 gần như luôn nằm trên đường SMA 200, cho thấy một xu hướng tăng dài hạn rất mạnh mẽ.
- Ngay cả trong các đợt sụt giảm mạnh của thị trường (như đầu năm 2020), giá có thể tạm thời giảm xuống dưới SMA 200, nhưng không hề xảy ra tín hiệu *Death Cross*. Điều này cho thấy sức mạnh nội tại và khả năng phục hồi nhanh của cổ phiếu.
- Tại thời điểm cuối biểu đồ, VCB tiếp tục duy trì xu hướng tăng giá một cách rõ ràng.

Tương tự như VCB, FPT cũng cho thấy một xu hướng tăng trưởng dài hạn ấn tượng.

- Một tín hiệu *Golden Cross* vào giữa năm 2020 đã xác nhận một chu kỳ tăng giá rất mạnh và bền bỉ.
- Cổ phiếu này cũng cho thấy khả năng phục hồi tốt trong các đợt điều chỉnh. Trong suốt giai đoạn 2022-2023, dù thị trường chung khó khăn, FPT vẫn duy trì được SMA 50 trên SMA 200.
- Hiện tại, FPT đang ở trong một xu hướng tăng giá rõ ràng, khi giá liên tục bám sát và nằm trên đường SMA 50.

Masan (MSN) là cổ phiếu có tính chu kỳ và biến động mạnh, tương tự như VIC.

- Biểu đồ cho thấy các đỉnh giá lớn vào các năm 2011, 2018 và 2021.
- Tín hiệu *Death Cross* vào đầu năm 2022 rất rõ ràng, dẫn đến một đợt giảm giá mạnh.
- Kể từ giữa năm 2023 cho đến nay, đường SMA 50 liên tục nằm dưới đường SMA 200. Điều này xác nhận rằng MSN vẫn đang nằm trong một xu hướng giảm giá dài hạn, mặc dù đã có những nỗ lực phục hồi giá trong ngắn hạn.

Tổng kết và so sánh

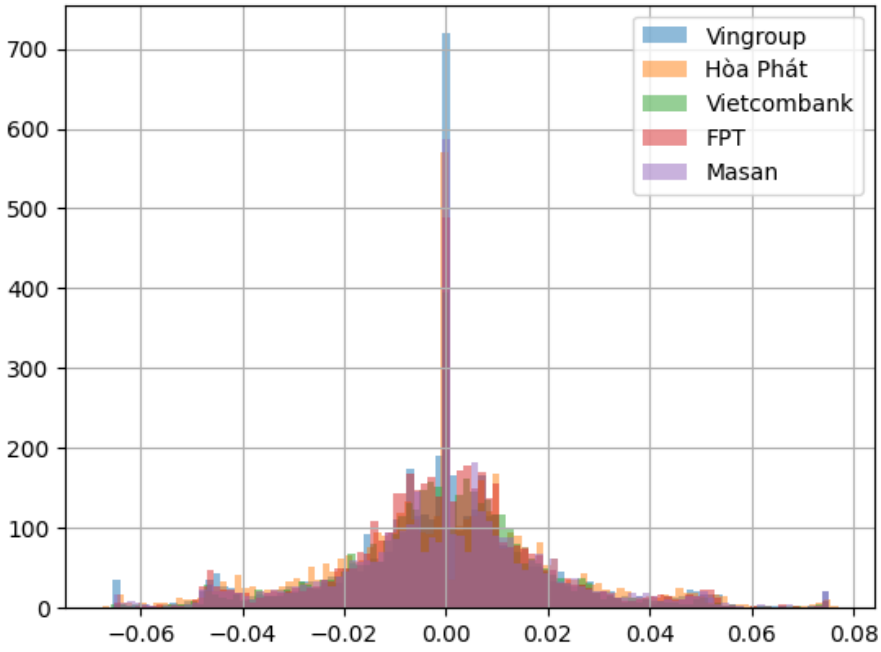
- Xu hướng tăng bền vững: VCB và FPT là hai cổ phiếu thể hiện xu hướng tăng trưởng dài hạn mạnh mẽ và ổn định nhất, ít bị ảnh hưởng bởi các chu kỳ giảm giá lớn (không xuất hiện *Death Cross* trong những năm gần đây).

- Phục hồi và tăng trưởng mới: HPG đã trải qua một đợt giảm giá sâu nhưng đã xác nhận xu hướng tăng trở lại với tín hiệu *Golden Cross* vào năm 2023.
- Tính chu kỳ và điều chỉnh: VIC và MSN cho thấy tính chu kỳ rõ rệt. Cả hai đều đã trải qua *Death Cross* vào năm 2022 và (tính đến cuối biểu đồ 2025) vẫn đang trong quá trình tìm kiếm sự ổn định hoặc vẫn nằm trong xu hướng giảm dài hạn (đặc biệt là MSN).

4.2 Phân tích biến động (Volatility Analysis)

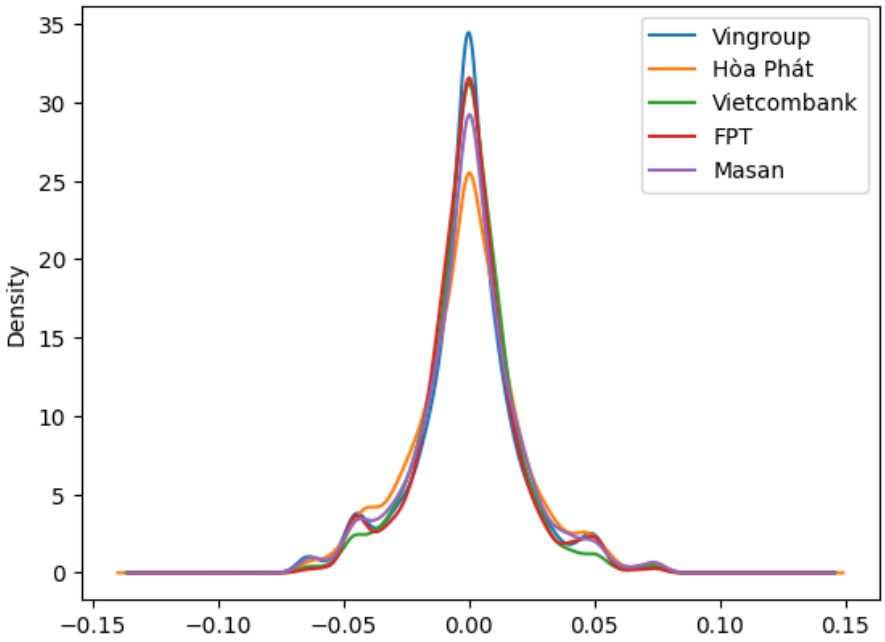
4.2.1 Tỷ suất sinh lợi hàng ngày (Daily Percentage Return)

Để đo lường rủi ro và mức độ biến động, dự án đã tính toán tỷ suất sinh lợi hàng ngày để đo lường hiệu suất và mức độ biến động của từng cổ phiếu. Độ lệch chuẩn của các tỷ suất sinh lợi này được sử dụng làm thước đo định lượng cho sự biến động.



Hình 3: Biểu đồ phân phối (Histogram) tỷ suất sinh lợi hàng ngày.

- Phân phối tập trung: Cả hai biểu đồ (Histogram tại Hình 3 và Biểu đồ Mật độ tại Hình 4) đều xác nhận rằng phần lớn tỷ suất sinh lợi hàng ngày của cả 5 mã đều tập trung rất cao xung quanh mốc 0. Điều này cho thấy các thay đổi lớn (lãi/lỗ) xảy ra với tần suất thấp, và thị trường phần lớn là ổn định trong ngày.
- So sánh độ biến động: Khi so sánh 5 mã với nhau, **Hòa Phát (HPG - màu cam)** thể hiện rõ sự khác biệt.
  - Trên biểu đồ Histogram (Hình 3), cột trung tâm (quanh mốc 0) của HPG thấp hơn đáng kể so với các mã còn lại, cho thấy HPG có ít ngày giao dịch gần như không đổi.
  - Biểu đồ Mật độ (Hình 4) làm rõ điều này: đường cong của HPG có đỉnh (peak) thấp nhất và phần "vai"rộng nhất, minh chứng cho việc HPG có nhiều ngày



Hình 4: Biểu đồ mật độ (KDE) của tỷ suất sinh lợi hàng ngày.

ghi nhận biến động ở mức vừa phải hơn.

- Dựa trên phân tích phân phối, **HPG** được xác định là cổ phiếu có độ biến động cao nhất trong nhóm được quan sát. Ngược lại, **Vingroup** (VIC - màu xanh dương) có đỉnh cao và nhọn nhất, cho thấy tính ổn định cao nhất (biến động thấp nhất).

4.2.2 Biến động theo thời gian (Volatility Over Time)

Phần này phân tích rủi ro của các cổ phiếu bằng cách sử dụng **Độ biến động trôi 30 ngày (30-Day Rolling Volatility)**. Chỉ báo này được tính bằng độ lệch chuẩn của phần trăm thay đổi giá hàng ngày trong một cửa sổ 30 ngày.

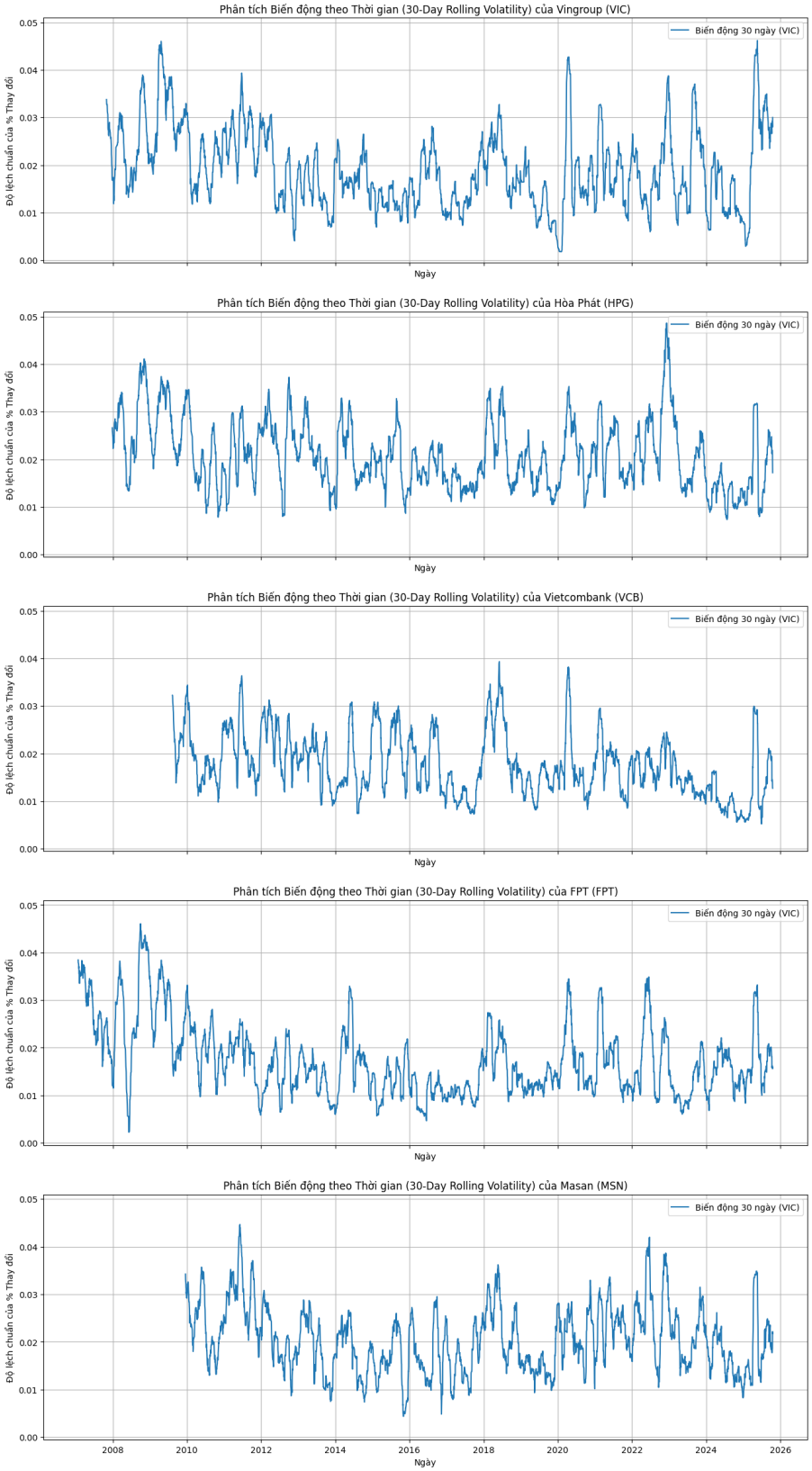
Một giá trị cao trên biểu đồ cho thấy một giai đoạn bất ổn, rủi ro cao và giá dao động mạnh (cả tăng lẫn giảm). Ngược lại, một giá trị thấp cho thấy một giai đoạn ổn định, rủi ro thấp và giá ít biến động.

Biểu đồ của Vingroup (VIC) cho thấy mức độ biến động cao và diễn ra thường xuyên trong suốt lịch sử giao dịch.

- Các đỉnh biến động rất rõ rệt, đặc biệt là trong các giai đoạn khủng hoảng (như 2008-2009) hoặc khi giá cổ phiếu tạo đỉnh và đáy lớn (như 2018, 2020 và 2022-2023).
- Điều này cho thấy VIC là một cổ phiếu có tính đầu cơ cao và nhạy cảm mạnh với các tin tức thị trường cũng như chu kỳ kinh doanh của chính tập đoàn. Mức độ rủi ro khi nắm giữ cổ phiếu này tương đối cao.

Độ biến động của Hòa Phát (HPG) có tính chu kỳ rất rõ ràng, gắn liền với các chu kỳ tăng trưởng và suy thoái của ngành thép.

- Giai đoạn 2020-2022 chứng kiến một đợt bùng nổ về biến động. Điều này trùng khớp hoàn hảo với giai đoạn siêu tăng trưởng và sau đó là đợt sụt giảm sâu mà



Hình 5: Phân tích Biến động trôi 30 ngày của 5 mã cổ phiếu.

chúng ta đã thấy trong phân tích SMA.

- Giai đoạn 2012-2016 và 2023-2024 có độ biến động tương đối thấp, cho thấy các giai đoạn "tích lũy"hoặc tăng trưởng ổn định.
- Biến động của HPG bùng nổ tại các điểm uốn quan trọng của xu hướng giá.

Vietcombank (VCB) là một trong những cổ phiếu có độ biến động thấp và ổn định nhất trong nhóm phân tích.

- Ngoại trừ các cú sốc toàn thị trường như khủng hoảng 2008 và đại dịch COVID-19 (đầu 2020), VCB duy trì một mức biến động nền rất thấp, đặc biệt là trong giai đoạn 2012-2019.
- Điều này củng cố cho phân tích SMA, cho thấy VCB là một cổ phiếu điển hình, có tính ổn định cao, rủi ro thấp hơn và phù hợp cho đầu tư dài hạn.

Biểu đồ của FPT cho thấy một câu chuyện thú vị về sự "trưởng thành"của một cổ phiếu.

- Trong giai đoạn đầu (2008-2011), FPT có độ biến động rất cao, tương tự như VIC.
- Tuy nhiên, kể từ khoảng năm 2012 trở đi, có một xu hướng giảm rõ rệt và bền vững của độ biến động. Cổ phiếu này ngày càng trở nên ổn định hơn khi công ty phát triển và mô hình kinh doanh trở nên vững chắc.
- Ngay cả trong cú sốc COVID-19, đỉnh biến động của FPT cũng thấp hơn đáng kể so với VCB hay HPG. Đây là minh chứng cho một cổ phiếu tăng trưởng ổn định với rủi ro ngày càng giảm dần.

Tương tự như VIC, Masan (MSN) cũng là cổ phiếu có độ biến động cao, mang tính chu kỳ và sự kiện.

- Các đỉnh biến động lớn (2011, 2018, 2021-2022) tương ứng chính xác với các chu kỳ giá lớn (tạo đỉnh và tạo đáy) đã được xác định trong phân tích SMA.
- Mức biến động nền của MSN luôn ở mức cao, cho thấy đây là một cổ phiếu có rủi ro lớn, với các dao động giá mạnh và khó lường, phù hợp hơn với các nhà đầu tư chấp nhận rủi ro cao.

Tổng kết và so sánh

- Nhóm ổn định: VCB và FPT. VCB ổn định một cách nhất quán, trong khi FPT đã "trưởng thành"từ một cổ phiếu biến động cao thành một cổ phiếu ổn định.
- Nhóm biến động cao: VIC và MSN. Cả hai đều cho thấy rủi ro cao, biến động mạnh và thường xuyên, phản ánh tính chất kinh doanh và mức độ đầu cơ cao.
- Nhóm biến động theo chu kỳ: HPG. Rủi ro của HPG tăng vọt trong các giai đoạn chuyển giao chu kỳ giá lớn, nhưng lại tương đối "hiền hòa"trong các giai đoạn tích lũy.

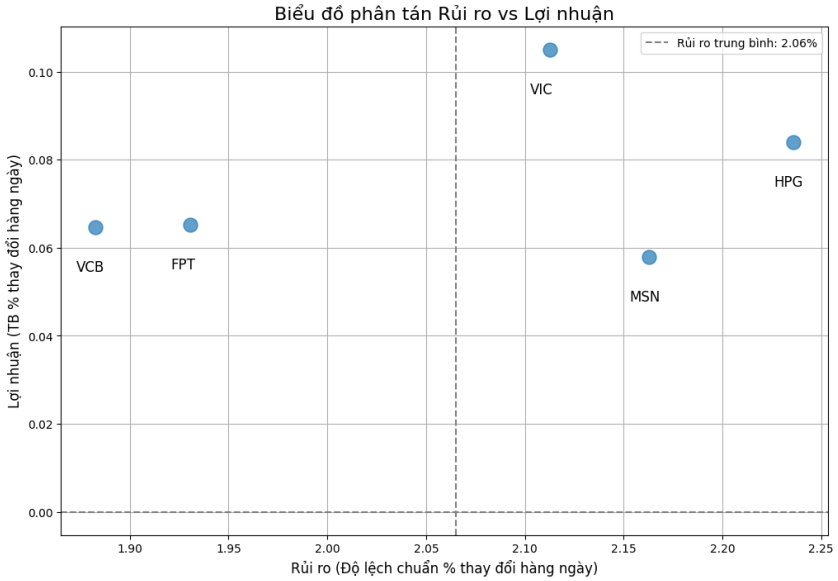
4.3 Phân tích tương quan (Correlation Analysis)

4.3.1 Rủi ro và lợi nhuận (Risk and Return)

Biểu đồ phân tán (Hình 6) cung cấp một cái nhìn tổng quan về hiệu suất đầu tư của 5 mã cổ phiếu bằng cách trực quan hóa hai yếu tố then chốt:

- Trục X (Rủi ro): Được đo bằng độ lệch chuẩn của phần trăm thay đổi giá hàng ngày. Giá trị càng cao, cổ phiếu càng rủi ro hoặc biến động.
- Trục Y (Lợi nhuận): Được đo bằng trung bình của phần trăm thay đổi giá hàng ngày. Giá trị càng cao, lợi nhuận trung bình hàng ngày càng lớn.

Đường đứt nét dọc thể hiện mức rủi ro trung bình (2.06%) của nhóm.



Hình 6: Biểu đồ phân tán Rủi ro vs Lợi nhuận.

Từ biểu đồ, chúng ta có thể chia các cổ phiếu thành ba nhóm riêng biệt:

Nhóm 1: Ổn định, hiệu suất tốt

- Cổ phiếu: FPT và VCB.
- Hai cổ phiếu này nằm ở góc phần tư phía trên bên trái, đây là vị trí "lý tưởng"nhất. Chúng mang lại mức lợi nhuận hàng ngày trung bình cao (khoảng 0.065% - 0.066%) trong khi mức độ rủi ro lại thấp hơn đáng kể so với mức trung bình (khoảng 1.90% - 1.94% so với 2.06%).
- Đây là những cổ phiếu phù hợp cho các nhà đầu tư tìm kiếm sự tăng trưởng ổn định với mức độ biến động thấp. Điều này hoàn toàn nhất quán với phân tích biến động ở phần trước, nơi VCB và FPT được xác định là hai cổ phiếu ổn định nhất.

Nhóm 2: Tăng trưởng cao, rủi ro cao

- Cổ phiếu: HPG và VIC.
- Nhóm này chấp nhận rủi ro cao hơn mức trung bình để đổi lấy lợi nhuận vượt trội.
- HPG: Có mức rủi ro cao nhất trong nhóm (khoảng 2.24%), nhưng cũng mang lại mức lợi nhuận rất cao, đứng thứ hai (khoảng 0.084%). Điều này phản ánh đúng bản chất chu kỳ và biến động mạnh của ngành thép.

- VIC: Điểm dữ liệu tại (x ≈ 2.08, y ≈ 0.105)cho thấy mức rủi ro cao hơn trung bình một chút, nhưng lại mang về mức *lợi nhuận trung bình hàng ngày cao nhất* trong 5 cổ phiếu. Đây là đặc điểm của một cổ phiếu tăng trưởng có rủi ro đi kèm.

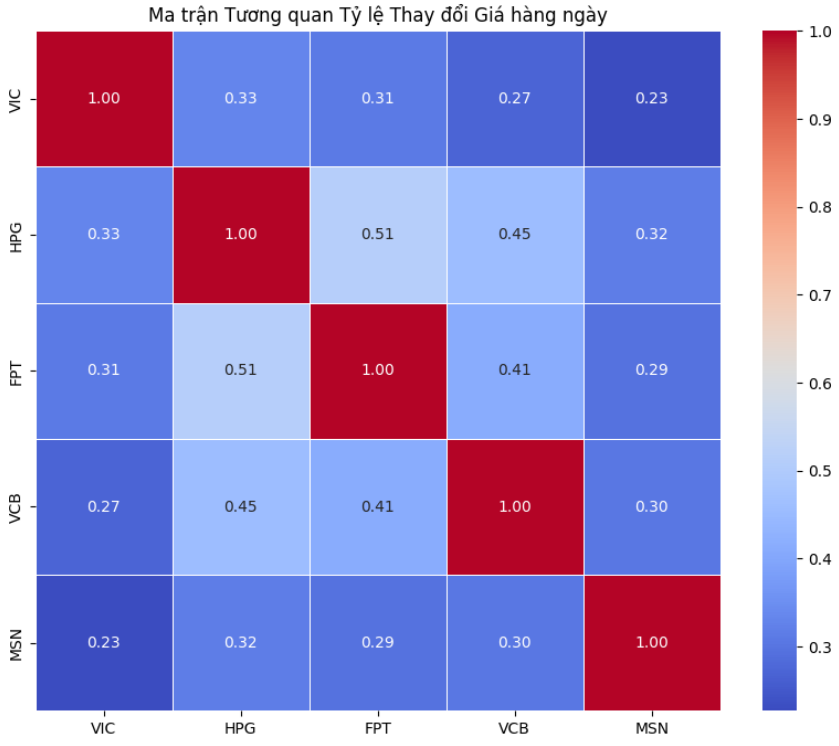
Nhóm 3: Rủi ro cao, hiệu suất thấp

- Cổ phiếu: MSN.
- MSN nằm ở vị trí kém mong muốn nhất. Cổ phiếu này có mức rủi ro cao (khoảng 2.17%), cao hơn đáng kể so với mức trung bình, nhưng lại mang về mức lợi nhuận trung bình hàng ngày (khoảng 0.058%) *thấp nhất* trong cả 5 cổ phiếu.
- Điều này cho thấy trong giai đoạn phân tích, nhà đầu tư phải gánh chịu rủi ro cao khi nắm giữ MSN nhưng không được đền bù bằng một mức lợi nhuận tương xứng so với các cổ phiếu khác.

4.3.2 Ma trận tương quan (Correlation Heatmap)

Phần này kiểm tra mức độ mà các cổ phiếu di chuyển cùng chiều với nhau. Ma trận tương quan (Hình 7) đo lường mối quan hệ tuyến tính của *tỷ lệ thay đổi giá hàng ngày* giữa các cặp cổ phiếu.

- Giá trị 1.00 (Màu đỏ thẫm): Tương quan dương hoàn hảo. Hai cổ phiếu luôn di chuyển cùng hướng.
- Giá trị gần 0 (Màu xanh nhạt): Không có tương quan. Chuyển động của hai cổ phiếu gần như độc lập với nhau.
- Giá trị dương (Màu xanh đến cam): Tương quan dương. Hai cổ phiếu có xu hướng di chuyển cùng hướng. Giá trị càng cao, mức độ đồng điệu càng lớn.



Hình 7: Ma trận Tương quan Tỷ lệ Thay đổi Giá hàng ngày.

Một điểm đáng chú ý là tất cả các hệ số tương quan đều là số dương (từ 0.23 đến 0.51). Điều này là hoàn toàn bình thường và có thể dự đoán được, vì tất cả đều là các cổ phiếu hàng đầu trong cùng một thị trường Việt Nam. Chúng đều chịu ảnh hưởng chung từ các yếu tố vĩ mô, tâm lý thị trường và dòng tiền lớn.

Nhóm Tương quan Cao (High Correlation)

Đây là các cổ phiếu có xu hướng di chuyển "đồng pha"mạnh mẽ nhất:

- HPG và FPT (0.51): Đây là cặp có tương quan mạnh nhất trong nhóm. Mặc dù thuộc hai ngành hoàn toàn khác nhau (Thép và Công nghệ), sự biến động giá hàng ngày của chúng lại rất đồng điệu. Điều này cho thấy cả hai có thể rất nhạy cảm với tâm lý chung của thị trường hoặc là hai cổ phiếu được các quỹ đầu tư lớn ưa chuộng.
- HPG và VCB (0.45): Mối tương quan mạnh, phản ánh vai trò "trụ cột"của cả hai trên thị trường.
- FPT và VCB (0.41): Tương tự, cặp FPT và VCB cũng cho thấy mối liên hệ chặt chẽ.

HPG, FPT và VCB tạo thành một "cụm"các cổ phiếu có xu hướng vận động cùng nhau.

Nhóm Tương quan Thấp (Low Correlation)

Đây là các cổ phiếu có chuyển động giá hàng ngày ít liên quan nhất đến nhau:

- VIC và MSN (0.23): Đây là cặp có hệ số tương quan thấp nhất trong ma trận. Điều này chỉ ra rằng chuyển động giá hàng ngày của Vingroup và Masan gần như độc lập với nhau. Các yếu tố thúc đẩy giá của chúng có thể mang tính đặc thù của doanh nghiệp (tin tức, kết quả kinh doanh riêng) hơn là tâm lý thị trường chung.
- VIC và VCB (0.27): Cũng là một mức tương quan tương đối thấp.

Phân tích tương quan là nền tảng cho việc đa dạng hóa danh mục để giảm thiểu rủi ro.

- Một danh mục đầu tư kết hợp các cổ phiếu có tương quan thấp VIC và MSN sẽ mang lại lợi ích giảm thiểu rủi ro tốt hơn nhiều so với một danh mục kết hợp các cổ phiếu có tương quan cao như HPG và FPT.
- Nắm giữ đồng thời cả HPG, FPT và VCB có thể không mang lại hiệu quả đa dạng hóa cao, vì khi một cổ phiếu giảm do yếu tố thị trường, hai cổ phiếu còn lại cũng có khả năng cao sẽ giảm theo.
- VIC và MSN, với mức tương quan thấp với các cổ phiếu còn lại, đóng vai trò là những nhân tố "đa dạng hóa"tốt nhất trong nhóm này.

4.4 Phân tích Động lượng (Momentum Analysis)

Phân tích khối lượng là một công cụ thiết yếu để xác nhận sức mạnh của một xu hướng giá. Các nguyên tắc cơ bản là:

- Xu hướng được xác nhận: Giá tăng đi kèm với khối lượng tăng (thể hiện sự quan tâm mua mạnh mẽ) hoặc Giá giảm đi kèm với khối lượng tăng (thể hiện áp lực bán tháo mạnh mẽ).

- Xu hướng yếu đi: Giá tăng nhưng khối lượng giảm (cho thấy sự cạn kiệt của phe mua) hoặc Giá giảm nhưng khối lượng giảm (cho thấy sự cạn kiệt của phe bán).
- Đỉnh/Đáy: Các đột biến về khối lượng cực lớn thường xuất hiện tại các điểm đảo chiều quan trọng, được gọi là "Đỉnh mua"hoặc "Đáy bán".

Vingroup (VIC) cho thấy mối quan hệ rất rõ ràng giữa các biến động giá lớn và khối lượng.

- Chu kỳ tăng giá 2017-2018 được hỗ trợ bởi nhiều đợt tăng đột biến về khối lượng, xác nhận sức mạnh của xu hướng.
- Giai đoạn giảm giá 2022-2023 cũng đi kèm với khối lượng bán tăng cao, cho thấy áp lực bán tháo.
- Điểm đáng chú ý nhất là vào cuối năm 2025: một đợt tăng giá gần như thẳng đứng được đi kèm với một đợt bùng nổ khối lượng ở mức cao lịch sử. Đây là một ví dụ điển hình của "Đỉnh mua", cho thấy sự hưng phấn cực độ và thường báo hiệu một đỉnh cao chính sắp được thiết lập.

Hòa Phát (HPG) là một ví dụ kinh điển về việc khối lượng xác nhận xu hướng chu kỳ.

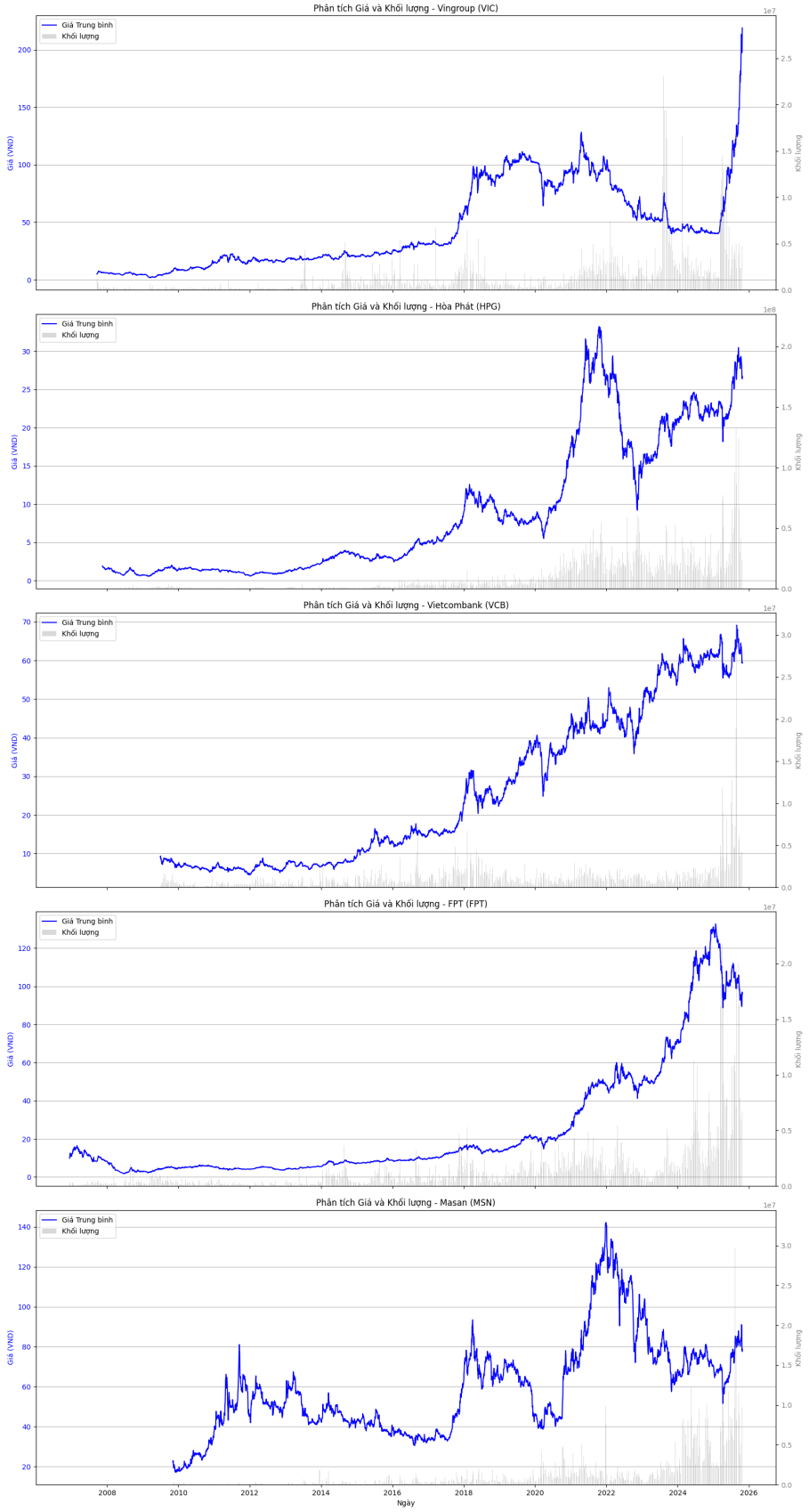
- Giai đoạn trước 2020, khối lượng giao dịch tương đối thấp và ổn định.
- Giai đoạn "siêu chu kỳ"2020-2021 chứng kiến giá tăng mạnh mẽ, được hỗ trợ hoàn hảo bởi mức tăng khối lượng giao dịch trung bình lên một tầm cao mới.
- Đợt sụt giảm năm 2022 cũng có khối lượng giao dịch rất cao, xác nhận một đợt bán tháo mạnh. Giai đoạn phục hồi 2023-2025 tiếp tục duy trì mặt bằng khối lượng cao, cho thấy HPG đã trở thành một cổ phiếu "quốc dân"với sự quan tâm lớn của thị trường.

So với các cổ phiếu khác, khối lượng của Vietcombank (VCB) tương đối ổn định, củng cố hình ảnh một cổ phiếu an toàn.

- Khối lượng giao dịch của VCB duy trì ở mức thấp và ổn định trong suốt xu hướng tăng dài hạn từ 2012-2019.
- Các đột biến về khối lượng thường chỉ xuất hiện trong các đợt hoảng loạn ngắn hạn (ví dụ: cú sụp đổ COVID-19 đầu 2020), cho thấy đây là các đợt "rũ bỏ"nhà đầu tư yếu thế.
- Kể từ năm 2024, khối lượng đã có sự gia tăng đáng kể, đi kèm với việc giá phá vỡ đỉnh lịch sử. Điều này xác nhận sức mạnh của đợt tăng giá hiện tại.

Biểu đồ của FPT cho thấy sự "trưởng thành"của một cổ phiếu.

- Giai đoạn 2012-2020: Giá tăng trưởng đều đặn nhưng khối lượng giao dịch lại rất thấp và đi ngang. Điều này cho thấy một giai đoạn "tích lũy"dài hạn, chủ yếu là do các nhà đầu tư tổ chức nắm giữ.
- Kể từ 2020: Tương tự HPG, FPT bước vào một chu kỳ tăng trưởng mới với sự bùng nổ về khối lượng. Giá tăng mạnh đi kèm với thanh khoản cao, xác nhận FPT đã trở thành một cổ phiếu thu hút sự quan tâm mạnh mẽ của dòng tiền đại chúng.



Hình 8: Phân tích Giá (đường màu xanh) và Khối lượng (thanh màu xám).

Masan (MSN) là cổ phiếu mà ở đó các đỉnh và đáy của giá gần như luôn được báo hiệu bởi các đột biến khối lượng cực lớn.

- Các đỉnh giá lớn vào năm 2011, 2018 và 2021 đều trùng với các "Đỉnh mua"với khối lượng giao dịch khổng lồ.
- Tương tự, các đáy lớn (như 2016, 2020, 2023-2024) cũng thường được hình thành sau các đợt bán tháo với khối lượng lớn.
- Điều này cho thấy khối lượng là một chỉ báo cực kỳ quan trọng để xác định các điểm đảo chiều lớn đối với một cổ phiếu mang tính chu kỳ cao như MSN.

5 Dự đoán xu hướng giá cổ phiếu

Để kiểm tra khả năng dự đoán, em đã xây dựng và đánh giá một số mô hình học máy và học sâu. Một trong những kiến trúc cốt lõi được lựa chọn để thử nghiệm là **LSTM (Long Short-Term Memory)**. Đây là một dạng Mạng Nơ-ron Hồi quy (RNN) chuyên dụng, có khả năng học hỏi và nhận diện các xu hướng phụ thuộc dài hạn (long-term dependencies) trong dữ liệu chuỗi thời gian, rất phù hợp với bài toán giá cổ phiếu.

Các mô hình được xây dựng dựa trên dữ liệu lịch sử của cổ phiếu **SAM - Công ty Cổ phần SAM Holdings**, một mã có nguồn dữ liệu dồi dào từ ngày 28/7/2000. Dữ liệu này được chia thành: tập huấn luyện (trước năm 2025) và tập thử nghiệm (năm 2025).

5.1 Mô hình LSTM đơn biến

Mô hình đầu tiên được thử nghiệm là LSTM đơn biến, chỉ sử dụng dữ liệu giá trung bình lịch sử làm đầu vào để dự đoán giá trung bình của ngày tiếp theo.

Hình 9: So sánh giá trị thực tế (Real SAM Stock Price - màu đỏ) và giá trị dự đoán (Predicted SAM Stock Price - màu xanh) của mô hình LSTM đơn biến.

Kết quả dự đoán của mô hình LSTM đơn biến trên tập dữ liệu thử nghiệm được trực quan hóa tại Hình 9.

Từ biểu đồ, có thể rút ra các nhận xét sau:

- Bám sát xu hướng: Đường dự đoán (màu xanh) đã bám sát rất tốt theo xu hướng chung của đường giá trị thực tế (màu đỏ). Mô hình đã học thành công các mô hình tăng và giảm dài hạn của cổ phiếu SAM, bao gồm cả cú sụt giảm mạnh (khoảng thời gian 50-60) và giai đoạn tăng trưởng chính (từ 125 đến 155).
- Độ trễ (Lag): Mô hình thể hiện một độ trễ nhỏ, đây là đặc điểm thường thấy của các mô hình chuỗi thời gian đơn biến. Điều này rõ ràng nhất tại các đỉnh và đáy nhọn. Ví dụ, tại đỉnh cao nhất (khoảng thời gian 155), đường dự đoán màu xanh đạt đỉnh hơi chậm và thấp hơn so với giá thực tế.
- Độ chính xác về hướng: Mặc dù có độ trễ, mô hình dự đoán rất chính xác về hướng di chuyển của giá (ví dụ: mô hình đã dự đoán chính xác xu hướng giảm ngay sau đỉnh cao 155). Nó cũng xử lý tốt các giai đoạn biến động phức tạp (ví dụ, từ 160-200), cho thấy khả năng của LSTM trong việc mô hình hóa dữ liệu chuỗi thời gian.

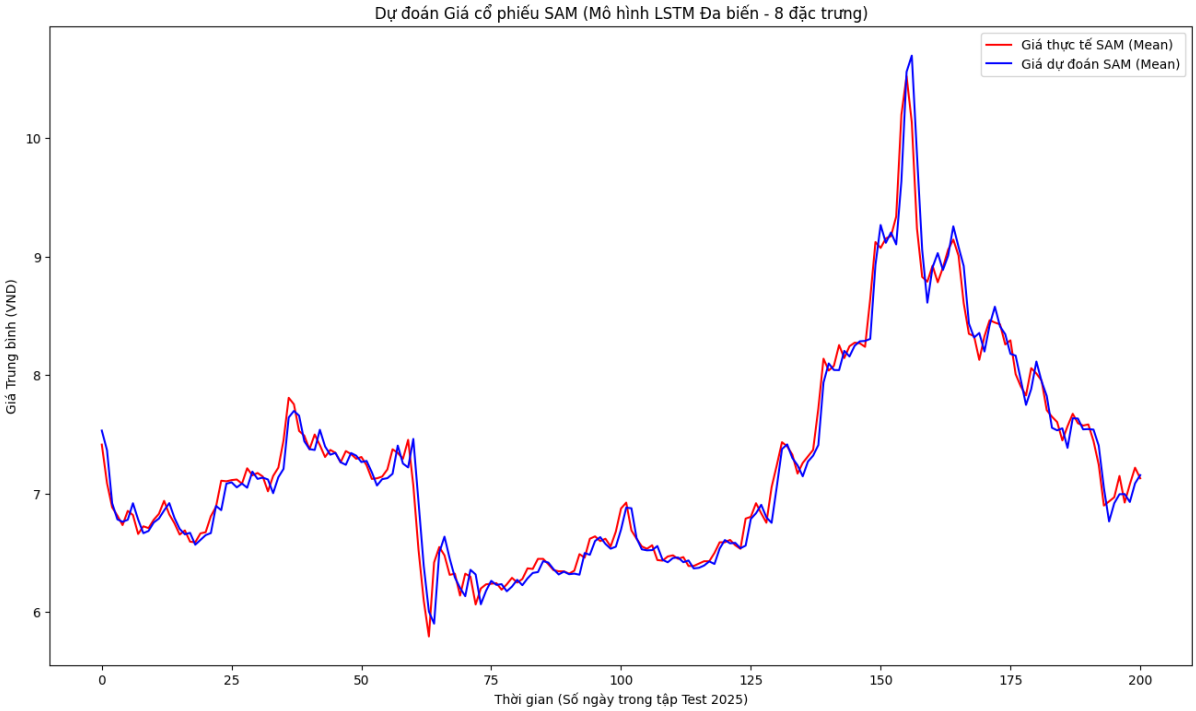
5.2 Mô hình LSTM đa biến

Để cải thiện độ chính xác và giải quyết vấn đề độ trễ (lag) quan sát được ở mô hình đơn biến, dự án đã xây dựng một mô hình LSTM đa biến (Multivariate LSTM).

Thay vì chỉ sử dụng một chuỗi "Mean Price"lịch sử, mô hình này được huấn luyện với một bộ dữ liệu đầu vào phong phú hơn, bao gồm 8 đặc trưng cho mỗi bước thời gian:

- Giá Mở cửa (Open)
- Giá Cao nhất (High)
- Giá Thấp nhất (Low)
- Giá Đóng cửa (Close)
- Khối lượng giao dịch (Volume)
- Giá trung bình (Mean)
- Trung bình động 10 ngày (SMA_10)
- Trung bình động hàm mũ 20 ngày (EMA_20)

Mục tiêu dự đoán vẫn là giá trị "Mean"của ngày tiếp theo.



Hình 10: So sánh giá trị thực tế (màu đỏ) và giá trị dự đoán (màu xanh) của mô hình LSTM đa biến với 8 đặc trưng.

Kết quả dự đoán của mô hình LSTM đa biến trên cùng tập dữ liệu thử nghiệm được trực quan hóa tại Hình 10.

Từ biểu đồ, có thể rút ra các nhận xét sau:

- Bám sát vượt trội: Đường dự đoán (màu xanh) gần như trùng khớp hoàn toàn với đường giá trị thực tế (màu đỏ). Mô hình đã học được các mối quan hệ phức tạp giữa 8 đặc trưng đầu vào để đưa ra dự đoán với độ chính xác rất cao.
- Khắc phục độ trễ: Vấn đề độ trễ tại các đỉnh và đáy nhọn đã được khắc phục gần như hoàn toàn. Quan sát tại các điểm đảo chiều quan trọng (ví dụ: cú sụt giảm mạnh ở mốc 60, hoặc đỉnh cao nhất ở mốc 155), đường dự đoán màu xanh đảo chiều gần như tức thời cùng lúc với giá thực tế.
- Dự đoán chính xác về biên độ: Không chỉ dự đoán đúng xu hướng, mô hình đa biến còn dự đoán rất chính xác về biên độ của giá. Đỉnh cao nhất (khoảng 10.5) và các đáy (quanh mốc 6.2) đều được mô hình dự đoán gần như chính xác.

5.3 Tổng kết

Việc chuyển đổi từ mô hình đơn biến sang đa biến đã mang lại một cải thiện rõ rệt về hiệu suất. Bằng cách cung cấp thêm "đặc trưng"cho mô hình (dưới dạng các đặc trưng OHLC, Khối lượng, và các chỉ báo kỹ thuật), mô hình LSTM có thể hiểu rõ hơn động lực của thị trường và đưa ra các dự đoán chính xác, kịp thời hơn.

6 Kết luận và hướng phát triển

6.1 Kết quả đạt được

Dự án đã mô phỏng thành công một quy trình Big Data cơ bản để phân tích dữ liệu chứng khoán, đạt được các mục tiêu chính sau:

- Kiến trúc: Một cụm Big Data hoàn chỉnh bao gồm HDFS (lưu trữ), YARN (quản lý tài nguyên), và Spark Standalone (xử lý) đã được triển khai thành công và ảo hóa bằng Docker.
- Thu thập và lưu trữ: Dữ liệu lịch sử (OHLCV) của thị trường chứng khoán Việt Nam đã được thu thập (sử dụng thư viện `vnstock`) và nạp thành công vào hệ thống tệp phân tán HDFS.
- Phân tích kỹ thuật: Đã thực thi phân tích kỹ thuật chuyên sâu trên 5 mã cổ phiếu vốn hóa lớn (VIC, HPG, FPT, VCB, MSN) bằng PySpark SQL. Các phân tích bao gồm:
  - Phân tích xu hướng (Mean Price, SMA 50, SMA 200, Golden/Death Cross).
  - Phân tích biến động (Daily Percentage Return, 30-Day Rolling Volatility).
  - Phân tích tương quan (Rủi ro vs Lợi nhuận, giữa các cổ phiếu)
  - Phân tích mối quan hệ Giá và Khối lượng.
- Học máy: Đã xây dựng và huấn luyện thành công mô hình LSTM đơn biến và đa biến, cho thấy kết quả trực quan tốt trong việc bám sát xu hướng giá của cổ phiếu SAM.

6.2 Hướng phát triển

Để xây dựng một hệ thống hoàn thiện hơn, dự án đề xuất các hướng phát triển trong tương lai:

- Xử lý thời gian thực: Tích hợp `Kafka` và `Spark Streaming` để xây dựng pipeline có khả năng phân tích và dự đoán dữ liệu giao dịch gần thời gian thực.
- Phân tích dữ liệu phi cấu trúc: Xây dựng pipeline `Spark NLP` để thu thập, xử lý và trích xuất phân tích cảm xúc (Sentiment Analysis) từ tin tức thị trường, sử dụng như một đặc trưng đầu vào quan trọng cho mô hình.

Tài liệu

[1] Apache Hadoop. https://hadoop.apache.org/

[2] Apache Spark. https://spark.apache.org/

[3] PySpark Documentation. https://spark.apache.org/docs/latest/api/python/index.html

[4] VnStock. https://vnstocks.com/docs

[5] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735-1780.

[6] Dự án tham khảo. https://github.com/thviet79/Stock-Price