Nền tảng hỗ trợ nhà đầu tư tài chính Việt Nam

Bối cảnh:

- Tính đến giữa năm 2025, theo báo VnEconomy : Thống kê của HelloSafe cho thấy khoảng 16% dân số Việt Nam, tương đương 16,2 triệu người, tham gia thị trường chứng khoán theo cả hình thức trực tiếp lẫn gián tiếp. Tỷ lệ này cao hơn so với nhiều nền kinh tế lớn châu Á như Trung Quốc (7%) và Ấn Độ (6%).
- Con số này vẫn đang tiếp tục tăng và gia tăng một cách rất mạnh mẽ trong những năm gần đây. Tuy nhiên chất lượng nhà đầu tư vẫn còn chưa cao, mọi người đa phần vẫn đầu tư theo cảm tính
- Tỷ lệ nhà đầu tư chứng khoán thua lỗ là rất cao, với nhiều nguồn tin cho biết khoảng 90-95% nhà đầu tư cá nhân bị thua lỗ trên thị trường, đặc biệt là sau một thời gian ngắn đầu tư hoặc trong các thị trường biến động (Theo báo thanh niên)

Nguyên nhân thua lỗ của nhà đầu tư:

- **Thiếu kiến thức và chiến lược**: Nhiều nhà đầu tư bị cuốn theo tin đồn, "phím hàng" mà không có chiến lược đầu tư bài bản.
- **Tâm lý muốn giàu nhanh**: Thị trường chứng khoán đòi hỏi sự kiên nhẫn và kỷ luật, nhưng nhiều người lại muốn kiếm tiền nhanh chóng và thường bị thua lỗ.
- **Không quản trị rủi ro**: Nhà đầu tư không đặt lệnh cắt lỗ (stop-loss), không kiểm soát tỷ trọng đầu tư, dẫn đến mất mát lớn khi thị trường giảm.
- **Học từ sai lầm**: Sau khi thua lỗ, thay vì rút kinh nghiệm, nhiều người lại lao vào đầu tư theo cảm xúc để gỡ gạc, và tiếp tục thua lỗ.
- **Hiệu ứng "tôi từng thắng"**: Việc thắng một vài lần trong thị trường tăng mạnh (bull market) tạo ra ảo tưởng về năng lực đầu tư, khiến họ chủ quan và lỗ nặng sau đó.

Mục tiêu:

- **Xây dựng nền tảng giúp các invester, đặc biệt đối với newbie trên thị trường stock Việt Nam**
- **Dựa trên phương châm:**
    - **“Đầu tư khoa học”**
    - **Nâng cao kiến thức trọng tâm về stock market**
    - **Hiểu những gì mình đầu tư, money control**
    - **Lợi nhuận ổn định**
- **Dựa trên nền tảng và cách tiếp cận chuyên nghiệp của quantitive researcher**

Nghiệp vụ:

- **Ứng dụng mô hình time series LSTM/ GRU/ Tranformer kết hợp với LLM để phân tích cảm xúc từ tin tức giúp người dùng xây dựng porfolio đầu tư,**

- **AI Agent- Chatbot chuyên biệt cho thị trường chứng khoán Việt Nam hỗ trợ cung cấp thông tin hữu ích về thị trường cho nhà đầu tư.**

Xây dựng Danh mục Đầu tư Cổ phiếu Việt Nam bằng Mô hình LSTM/GRU/TransformerEncoder

Giới thiệu

Thị trường chứng khoán Việt Nam có tính biến động cao và thường chịu ảnh hưởng mạnh từ tâm lý đám đông. Nhiều nhà đầu tư cá nhân ra quyết định cảm tính thay vì dựa trên phân tích khoa học. Để **đầu tư có khoa học**, dự án này đề xuất một mô hình kết hợp **học sâu** và phân tích dữ liệu đa chiều nhằm hỗ trợ xây dựng danh mục đầu tư cổ phiếu. Chúng tôi áp dụng các mạng neuron tiên tiến – **LSTM**, **GRU** và **Transformer Encoder** – để dự báo tín hiệu mua (Long) hay không mua (Neutral) cho các cổ phiếu trên thị trường Việt Nam, dựa trên dữ liệu kỹ thuật, vĩ mô và tâm lý thị trường. Việc ứng dụng mô hình LSTM vào tối ưu danh mục vẫn còn mới mẻ; một nghiên cứu gần đây nhấn mạnh rằng tích hợp LSTM với quản lý danh mục đầu tư đang ở giai đoạn đầu và cần được khám phá thêm ewadirect.com. Do đó, dự án của chúng tôi là nỗ lực tiên phong kết hợp mô hình dự báo bằng học sâu với chiến lược phân bổ danh mục trên thị trường Việt Nam.

Không giống nhiều nghiên cứu trước chỉ dự báo chỉ số chung hoặc giá cổ phiếu riêng lẻ journalofscience.ou.edu.vnjournalofscience.ou.edu.vn, dự án này tập trung vào việc **dự báo xác suất tín hiệu Long** của từng cổ phiếu trong ngắn hạn và **xây dựng danh mục động** từ những tín hiệu đó. Cụ thể, mô hình sẽ xác định cổ phiếu nào có triển vọng tích cực nhất để mua trong giai đoạn sắp tới, giúp nhà đầu tư **phân bổ vốn vào các mã tiềm năng nhất** thay vì đầu tư dàn trải hay cảm tính. Mục tiêu cuối cùng là thiết kế một chiến lược đầu tư **khoa học và định lượng**, đạt hiệu suất lợi nhuận cao hơn so với việc đầu tư theo cảm xúc hay kinh nghiệm chủ quan.

Để tăng độ chính xác dự báo, chúng tôi kết hợp nhiều nguồn thông tin: **(1)** chỉ báo kỹ thuật từ dữ liệu giá và khối lượng, **(2)** yếu tố **kinh tế vĩ mô** như lãi suất, lạm phát, tỷ giá có ảnh hưởng tới thị trường, và **(3) yếu tố tâm lý** thị trường trích xuất từ tin tức. Việc tích hợp cảm xúc nhà đầu tư tỏ ra hữu ích: nghiên cứu tại Việt Nam cho thấy mô hình LSTM kết hợp chỉ số tâm lý (Investor Sentiment Index) dự báo VN-Index chính xác hơn hẳn so với mô hình chỉ dùng giá quá khứ kinhtevadubao.vnkinhtevadubao.vn. Tương tự, trên thế giới các hệ thống lai ghép LSTM với phân tích tin tức đã cho kết quả dự báo chính xác hơn so với chỉ dùng dữ liệu thị trường đơn thuần mdpi.com. Bởi vậy, dự án chúng tôi kỳ vọng đóng góp mới khi lần đầu **kết hợp cả dữ liệu kỹ thuật, vĩ mô và cảm xúc** trong một mô hình học sâu để xây dựng danh mục cổ phiếu ở Việt Nam. Ngoài ra, chúng tôi sẽ sử dụng **mô phỏng Monte**

Carlo 3 giai đoạn để đánh giá rủi ro và độ tin cậy của chiến lược, đảm bảo tính **ổn định** trước những biến động ngẫu nhiên.

Dữ liệu và Đặc trưng

Dự án sử dụng dữ liệu lịch sử 10 năm (01/2015 – 01/2025) trên **25 cổ phiếu có thanh khoản cao nhất** sàn HOSE (TP. HCM). Danh sách cổ phiếu được cố định trong suốt thời gian nghiên cứu, bao gồm các nhóm:

- **Ngân hàng (11 mã)**: VCB, BID, CTG, TCB, MBB, VPB, STB, TPB, VIB, HDB, SHB
- **Chứng khoán (5 mã)**: SSI, VND, HCM, VCI, VIX
- **Blue-chip/Khác (9 mã)**: FPT, HPG, GAS, POW, MWG, VIC, VHM, VRE, NVL

Các dữ liệu giá cổ phiếu dạng OHLCV (Open, High, Low, Close, Volume) được thu thập thông qua API (như thư viện `vnstock`) cho giai đoạn 2015–2025, với tần suất theo **phiên giao dịch** (bỏ qua ngày nghỉ lễ, không nội suy thêm). Bộ dữ liệu được chia theo thời gian: **tập huấn luyện** từ 2015 đến 2021, **tập validation** năm 2022, và **tập kiểm thử** từ 2023 đến đầu 2025. Việc chia theo mốc thời gian đảm bảo mô hình luôn dự báo tương lai, tránh rò rỉ thông tin từ tương lai về quá khứ.

Mỗi cổ phiếu tại mỗi thời điểm được mô tả bởi **8 đặc trưng (features)** chính, kết hợp cả kỹ thuật, vĩ mô và tâm lý:

- **Chỉ báo kỹ thuật**: Chúng tôi tính toán từ dữ liệu giá quá khứ các chỉ báo phổ biến gồm:
  - **Return** (tỷ suất lợi nhuận phiên, đo lường mức tăng/giảm so với phiên trước). Mô hình của tôi dùng **log return** (trong data là cột `r1`) thay vì simple return là vì log return có nhiều ưu điểm hơn khi làm **phân tích định lượng và mô hình thống kê**.
    - Tỉ lệ hơn kém thông thường (simple return):

$$r_t^{(\text{simple})} = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

    - Lợi suất logarit (log return):

$$r_t^{(log)} = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

- Khả năng cộng dồn theo thời gian của simple return và log return:
  - Lợi suất **cộng dồn chỉ là tổng các log return**, rất tiện cho mô hình hóa chuỗi thời gian.

$$P_T = P_0 \times (1 + r_1) \times (1 + r_2) \times \ldots \times (1 + r_T)$$

$$\ln\left(\frac{P_T}{P_0}\right) = \ln(1 + r_1) + \ln(1 + r_2) + \ldots + \ln(1 + r_T) = r_1^{(log)} + r_2^{(log)} + \ldots + r_T^{(log)}$$

  - . **Tính đối xứng tăng – giảm:Ví dụ:**
    - Ngày 1: giá tăng +10%, từ 100 → 110
    - Ngày 2: giá giảm –10%, từ 110 → 99
    - Simple return: +10% và –10% không triệt tiêu nhau (kết quả còn –1%).
    - Log return phản ánh đúng hơn:
      ln(110/100)+ln(99/110)=ln(99/100)≈−1%

👉 Log return cho kết quả chính xác hơn khi biến động lớn hoặc có tăng–giảm liên tiếp.

  - **Phù hợp cho thống kê & mô hình:**
    - Phân phối của log return thường gần chuẩn (Gaussian) hơn simple return, nhất là với dữ liệu giá tài chính → dễ dùng cho mô hình hồi quy, ARIMA, GARCH, LSTM…
    - Với biến động nhỏ (<5% mỗi ngày), log return ≈ simple return (sai số rất nhỏ).
      Ví dụ 5%:
      ln(1.05)≈0.04879 so với 0.05
  - **Kết luận**:
    - Dùng **tỉ lệ hơn kém** (simple return) thì **trực quan** và dễ hiểu cho báo cáo ngắn hạn.
    - Dùng **log return** (như `r1`) thì **tiện lợi hơn cho tính toán, mô hình hóa, và phản ánh đúng**

hiệu ứng cộng dồn & tăng–giảm.

- Vì vậy trong phân tích định lượng, đặc biệt là dự báo danh mục như của bạn, **log return là chuẩn mực hơn**.

- **RSI** (Chỉ số sức mạnh tương đối, phản ánh trạng thái quá mua/quá bán của cổ phiếu trên thang 0-100 mdpi.com). Trong bài này sử dụng **Relative Strength Index - 14 ngày** (trong data là cột rsi14):
  - **Khái niệm:**
    RSI là **chỉ báo động lượng** cho biết giá đang **quá mua (overbought)** hay **quá bán (oversold)** dựa trên sức mạnh tương đối của các chuỗi tăng giá so với giảm giá trong một giai đoạn (ở đây là **14 phiên giao dịch**).

  - **Cách tính (theo Wilder, công thức chuẩn):**
    - Tính thay đổi giá hàng ngày:

$$\Delta P_t = \text{Close}_t - \text{Close}_{t-1}$$

    - Tách thành **tăng** và **giảm**:
$$\text{Gain}_t = \max(\Delta P_t, 0), \quad \text{Loss}_t = \max(-\Delta P_t, 0)$$
    - Tính **trung bình luỹ thừa (Wilder's smoothing)** cho 14 ngày:

$$\text{AvgGain}_{14}, \quad \text{AvgLoss}_{14}$$

    - Tính **RS**:

$$RS = \frac{\text{AvgGain}_{14}}{\text{AvgLoss}_{14}}$$

    - Tính **RSI**:

$$RSI = 100 - \frac{100}{1 + RS}$$

  - Ý nghĩa:
    - RSI ∈ [0, 100].
    - RSI > 70 → thị trường đang **quá mua** → khả năng điều chỉnh giảm.
    - RSI < 30 → thị trường đang **quá bán** → khả năng hồi phục.
- **Volume** (khối lượng giao dịch, biểu thị mức độ thanh khoản và sự quan tâm của thị trường). Cụ thể trong bài này tôi sử dụng **Relative**

Volume 20 – Khối lượng tương đối 20 ngày (trong data là cột `vol_rel20`)

- **Khái niệm:**
  - Đo xem khối lượng giao dịch hiện tại lớn hay nhỏ hơn mức trung bình 20 phiên gần nhất.
- **Cách tính:**
  - Tính trung bình động 20 ngày của volume:

$$vol_{ma20t} = \text{mean}(Volume_{t-19}, \ldots, Volume_t)$$

  - Tính **khối lượng tương đối**:

$$vol_{rel20t} = \frac{Volume_t}{vol_{ma20t} + 10^{-12}}$$

- Ý nghĩa:
  - Nếu `vol_rel20` > 1 → khối lượng hôm nay cao hơn mức trung bình 20 ngày (có thể đang có dòng tiền mạnh vào/ra).
  - Nếu `vol_rel20` < 1 → khối lượng thấp hơn mức trung bình → giao dịch trầm lắng.

- **Volatility** :độ biến động giá. Trong bài này tôi sử dụng Volatility 20 – Độ biến động 20 ngày dựa trên r1. Ở đây tôi sử dụng **Relative Volume 20** – Khối lượng tương đối 20 ngày. Trong data là cột `volat20`
  - **Khái niệm:**
    - **Volatility** đo **độ dao động** của lợi suất giá (returns).
  - **Cách tính:**

$$volat20_t = \text{Std}\left(r1_{t-19}, \ldots, r1_t\right)$$

    - tức là độ lệch chuẩn (standard deviation) của chuỗi lợi suất trong 20 phiên gần nhất.

  - **Ý nghĩa:**
    - Volatility cao → giá biến động mạnh, rủi ro cao hơn.
    - Volatility thấp → giá ổn định hơn.

- **Yếu tố vĩ mô:** Bao gồm **lãi suất liên ngân hàng qua đêm**, **lạm phát (CPI YoY)** và **tỷ giá USD/VND**. Những yếu tố này thường biến động chậm, nên ta cập nhật theo chu kỳ dài hơn:
  - **Lạm phát - CPI YoY :**

- CPI YoY được tổng hợp thủ công từ nhiều nguồn khác nhau(theglobaleconomy.com, baochinhphu.vn , nso.gov) vào mỗi đầu tháng, được cập nhật theo tháng - mỗi tháng một giá trị tại phiên giao dịch đầu tháng (cột cpi_yoy trong data)
- **Khái niệm**:
  - **CPI YoY** đo mức **thay đổi % của chỉ số giá tiêu dùng** so với **cùng kỳ năm trước**:

$$CPI_{YoY,t} = \frac{CPI_t - CPI_{t-12}}{CPI_{t-12}} \times 100\%$$

  - Phản ánh **tốc độ tăng giá hàng hóa & dịch vụ** của nền kinh tế qua 12 tháng.

  - Là **thước đo lạm phát phổ biến nhất** tại Việt Nam.
- **Tác động lên chứng khoán:**

Tình huống	Hệ quả
CPI YoY cao, >4-5% → áp lực **lạm phát**	– SBV có xu hướng **tăng lãi suất điều hành / hút tiền** để kiềm chế lạm phát → chi phí vốn DN tăng → EPS giảm. – Dòng tiền rút khỏi TTCK do kỳ vọng chi phí vay cao hơn. – Nhóm tiêu dùng – bán lẻ – vật liệu xây dựng chịu ảnh hưởng tiêu cực vì chi phí đầu vào tăng.
CPI YoY thấp / ổn định 2–3% → lạm phát kiểm soát	– Chính sách tiền tệ có thể **nới lỏng hoặc trung tính**, lãi suất duy trì thấp. – Tâm lý NĐT tích cực hơn, hỗ trợ TTCK tăng.

- 👉 **CPI YoY là feature chu kỳ dài hơn, phản ánh bối cảnh chính sách tiền tệ và rủi ro vĩ mô. Nó giúp mô hình hiểu được "nền" thị trường có thuận lợi hay không.**

Lãi suất liên ngân hàng overnight

- Dữ liệu được crawl từ nguồn: Ngân hàng Nhà nước, được lấy trung bình EMA-7 ngày và cập nhật hàng tuần - mỗi tuần một giá trị (cột interbank_week trong data)
- **Khái niệm:**
  - Là **lãi suất mà các ngân hàng thương mại vay mượn lẫn nhau** để bù đắp thiếu hụt ngắn hạn về

thanh khoản.

- Dùng **EMA-7 ngày** (Exponential Moving Average 7-day) → lấy trung bình trượt luỹ thừa của 7 ngày gần nhất để **làm mượt dữ liệu** và tránh nhiễu từ biến động từng ngày.

- Vì lãi suất liên ngân hàng phản ứng **rất nhanh** với cung – cầu tiền tệ, tôi cập nhật **hàng tuần** là hợp lý: nắm bắt xu hướng thanh khoản nhưng không quá nhiễu.

- **Tác động tới thị trường chứng khoán:**

Tình huống	Hệ quả
Lãi suất ↑ → tín hiệu thắt chặt thanh khoản	– Ngân hàng Nhà nước có thể đang hút tiền khỏi hệ thống → vốn vay đắt hơn. – Dòng tiền rút khỏi thị trường cổ phiếu để tìm nơi an toàn hơn (tiền gửi, TPCP). – Cổ phiếu nhóm ngân hàng, bất động sản, chứng khoán thường bị ảnh hưởng tiêu cực
Lãi suất ↓ → tín hiệu nới lỏng thanh khoản	– Vốn rẻ hơn, dễ vay hơn → khuyến khích đầu tư và tiêu dùng. – Dòng tiền nhàn rỗi có thể chảy vào chứng khoán → thị trường được hỗ trợ tăng giá

- 👉 **Trong mô hình, lãi suất liên ngân hàng đóng vai trò chỉ báo dòng tiền ngắn hạn. Giá trị cao → xác suất tăng giá của nhiều cổ phiếu có thể thấp hơn; giá trị thấp → tăng xác suất.**

Tỷ giá USD/VND

- (nguồn: Yahoo Finance) được cập nhật theo tháng - mỗi tháng một giá trị tại phiên giao dịch đầu tháng (cột usd_vnd trong data) :
- **Khái niệm:**
  - **Tỷ giá USD/VND** = số VNĐ cần để mua 1 USD
  - Bạn cập nhật **mỗi tháng tại phiên đầu tháng** vì tỷ giá không biến động mạnh từng ngày (trừ giai đoạn biến động lớn).
- **Tác động tới thị trường chứng khoán:**

Tình huống	Hệ quả
USD/VND ↑ (VNĐ mất giá)	DN **nhập khẩu** (xăng dầu, thép phôi, bán lẻ) chịu tăng chi phí → lợi nhuận giảm → cổ phiếu nhóm này giảm. – DN **xuất khẩu** (thủy sản, dệt may, FPT Software) hưởng lợi vì doanh thu USD quy đổi ra VNĐ cao hơn. – Nhà đầu tư nước ngoài có thể bán ròng vì rủi ro tỷ giá, gây áp lực giảm lên thị trường. – SBV có thể phải nâng lãi suất để ổn định VND → tác động tiêu cực chung.
USD/VND ↓ (VNĐ mạnh lên)	– DN nhập khẩu hưởng lợi vì chi phí đầu vào giảm. – DN xuất khẩu có thể bị giảm lợi nhuận. – Tâm lý thị trường ổn định hơn, giảm lo ngại rủi ro vĩ mô.

        - 👉 **Tỷ giá là feature phản ánh sức mạnh đồng nội tệ và dòng vốn ngoại – rất quan trọng trong thị trường mở như Việt Nam.**
    - **Tất cả được chuẩn hóa trước khi đưa vào mô hình. Cơ sở sử dụng các biến này là do chúng có ảnh hưởng đã được chứng minh đến TTCK Việt Nam – ví dụ, nghiên cứu giai đoạn 2000-2018 cho thấy lãi suất cao có tương quan âm với VN-Index, còn lạm phát và tỷ giá có tương quan đáng kể đến biến động thị trường** journalofscience.ou.edu.vnjournalofscience.ou.edu.vn.
        - Ba feature này không trực tiếp "dự báo" giá từng cổ phiếu, nhưng **thể hiện môi trường vĩ mô** – ảnh hưởng **dòng tiền thị trường**, **chi phí vốn**, và **kỳ vọng lợi nhuận DN**.
        - Khi mô hình tính **xác suất tăng giá**, các biến này giúp điều chỉnh theo **chu kỳ vĩ mô**:Thời kỳ **tiền rẻ, lạm phát thấp, VNĐ ổn định** → mô hình có thể nâng xác suất tăng giá của nhiều mã → chọn Top-5 dễ hơn.
        - Ngược lại, **lãi suất cao, lạm phát cao, VNĐ mất giá** → xác suất chung giảm, mô hình chỉ chọn được vài mã thực sự mạnh.
    -

- **Chỉ số tâm lý (sentiment)**:
    - **Tổng quan**

Để định lượng tâm lý thị trường, chúng tôi xây dựng một **đặc trưng cảm xúc** dựa trên tin tức tài chính. Cụ thể, chúng tôi thu thập ~5.600 **tiêu đề bài báo** liên quan đến 25 cổ phiếu trên từ chuyên trang CafeF trong giai đoạn 2015–2025. Dữ liệu được tiền xử lý bằng cách loại bỏ các tin trùng lặp hoặc gần trùng lặp, loại bỏ những bài báo thiếu ngày tháng, thiếu tiêu đề,...

- **Cách hoạt động và đánh giá:**
  - Chúng tôi sử dụng mô hình ngôn ngữ **PhoBERT-large (Vietnamese BERT)** đã được fine-tune để phân loại tiêu đề thành hai loại cảm xúc: **tích cực** (positive) hoặc **tiêu cực** (negative). Mô hình PhoBERT được chúng tôi fine-tune trên 6.000 mẫu tiêu đề tôi thu thập trên CafeF có gán nhãn hoàn toàn thủ công.
  - **Độ chính xác** của mô hình sau fine-tune là ~90% (F1-macro ~86%). Nghiên cứu độc lập của Nguyen et al. (2021) cũng cho thấy PhoBERT phân loại tin tức chứng khoán Việt Nam thành tiêu cực/trung tính/tích cực đạt tới **93% độ chính xác** ceur-ws.org, khẳng định hiệu quả của phương pháp tiếp cận này.
  - **PhoBERT-large fine-tune** phân loại tin thành:
    - **+1** (tích cực / thuận lợi / trung lập-hỗ trợ) .
      - Ví dụ: kế hoạch kinh doanh khả quan, giải đáp thắc mắc nhà đầu tư, bổ nhiệm nhân sự, triển khai dự án mới...
    - **–1** (tiêu cực / bất lợi)
    - **0** (không có tin)
  - Mỗi tin tức gán nhãn cho **3 phiên giao dịch sau** (nếu ra sau 15h thì tính từ phiên hôm sau).
  - Tức là bạn đã biến **dòng thông tin định tính từ báo chí** thành **dòng số liệu theo thời gian** song song với dữ liệu OHLCV.
  - Ý nghĩa cốt lõi: **sentiment ≈ chỉ số đo lường mức kỳ vọng và tâm trạng của nhà đầu tư** về một cổ phiếu trong ngắn hạn.
- **Tại sao sentiment có giá trị trong dự báo giá cổ phiếu:**
  - **Thông tin thị trường Việt Nam lan tỏa nhanh qua báo chí:**
    - Phần lớn nhà đầu tư cá nhân đọc tin CafeF, Vietstock,… rồi phản ứng mua/bán → giá thường phản ứng ngắn hạn với tin.
  - **Tâm lý hành vi (behavioral finance):**
    - Nghiên cứu cho thấy **tin tích cực làm tăng nhu cầu mua → giá tăng trong 1-3 phiên sau tin**, tin tiêu cực ngược lại.

        - Thị trường Việt Nam chưa hoàn toàn hiệu quả (semi-strong form), nên **thông tin chưa phản ánh ngay lập tức vào giá** → sentiment có thể mang tính dự báo.
    - **Bổ sung góc nhìn phi kỹ thuật:**
        - – Các feature kỹ thuật (RSI, volatility, volume…) chỉ khai thác **dữ liệu giá quá khứ**, chưa phản ánh **kỳ vọng**.
        - – Sentiment cung cấp **yếu tố dẫn dắt trước giá** (forward-looking) → giúp mô hình bắt kịp sóng tin tức.
    - **Bài nghiên cứu liên quan:**
        - **"Sentiments Extracted from News and Stock Market" của Vu et al. (2023)** Preprints
            - Họ thu thập ~40.000 bài báo trên các trang tài chính / kinh tế Việt Nam (CafeF, Vneconomy, Stockbiz, v.v.)
            - Họ dùng **PhoBERT** để phân loại sentiment (tích cực / tiêu cực) cho các tin tức liên quan thị trường chứng khoán.
            - Mô hình đạt **độ chính xác > 81 %** trong phân loại tin tức sentiment.
            - Họ cũng kiểm tra phản ứng thị trường trước và sau khi tin được công bố: phát hiện rằng nhà đầu tư có xu hướng **phản ứng quá mức (overreact)** trước cả tin tiêu cực lẫn tích cực.
            - Tuy nhiên, theo nghiên cứu này, **sự khác biệt trung bình trong giá cổ phiếu sau tin** không luôn "có ý nghĩa thống kê mạnh" (insignificant difference) — tức có tin thì có phản ứng, nhưng không phải lúc nào cũng dẫn đến biến động lớn hơn.
        - **"News sentiment and states of stock return volatility" (Y. Shi, 2021):**
            - "Negative news increases the likelihood of higher volatility states. Positive news decreases that to a larger degree." ScienceDirect
            - → Tức là khi xuất hiện tin tiêu cực, khả năng thị trường chuyển sang trạng thái biến động cao (volatility high state) tăng lên.
- **Vai trò trong danh mục:**
    - **Điều chỉnh phân bổ ngắn hạn:** khi thị trường đang tích cực về 1 mã (ví dụ FPT có tin ký hợp đồng lớn) → sentiment đẩy cao xác suất tăng giá của FPT → tăng khả năng FPT được chọn vào danh mục ở chu kỳ tái cân bằng 20 ngày.
    - **Giúp tránh rủi ro sự kiện:** nếu xuất hiện tin xấu (ví dụ HPG bị phạt môi trường) → sentiment âm làm giảm xác suất → tránh mua vào mã rủi ro cao.

        - Việc tích hợp dữ liệu tin tức theo cách này giúp mô hình cảm nhận được **dư luận và tâm lý nhà đầu tư theo thời gian**. Thực tế, một nghiên cứu đã phát triển chỉ số tâm lý nhà đầu tư từ tin tức và bình luận, khi đưa vào mô hình học sâu đã giúp **giảm đáng kể sai số** dự báo thị trường **kinhtevadubao.vnkinhtevadubao.vn**
    - **Giới hạn & lưu ý:**
        - **Thời gian ảnh hưởng ngắn (3 phiên):** hợp lý cho tin tức thường nhật nhưng **sự kiện lớn** (chia cổ tức, M&A) có thể tác động lâu hơn → có thể cân nhắc kéo dài cửa sổ khi gặp tin đặc biệt.
        - **Độ chính xác mô hình (~90% acc, F1 ~86%)**: đủ tốt nhưng vẫn có nhiễu → mô hình cần học để không bị lệch bởi tin giả/nhầm nhãn.
        - **Độ bao phủ dữ liệu:** mới khoảng 5.600 headline / 10 năm → khá nhỏ nhưng phù hợp cho 25 mã thanh khoản cao.

- **Cách hoạt động của Phobert-large:**
    - **TÓM TẮT NGẮN (Executive summary):**
        - **Word segmentation**: tách từ đa âm tiếng Việt (ví dụ *Ngân_hàng, nợ_xấu*).
        - **Tokenization (BPE)**: biến câu đã tách từ thành các **subword**, chèn token đặc biệt `<s>` và `</s>`, rồi **pad/truncate** về độ dài cố định (128).
        - **Mã hoá đầu vào**: tạo `input_ids` (ID token) và `attention_mask` (đánh dấu pad).
        - **Embedding**: mỗi token → **vector 1024 chiều** (token embedding + positional embedding).
        - **Encoder 24 lớp Transformer**: qua 24 khối *Self-Attention* → *FFN* để tạo **vector ngữ cảnh hoá** cho từng token.
        - **Đại diện câu**: lấy **vector của** `<s>` ở lớp cuối như *tóm tắt toàn câu*.
        - **Head phân loại**: *Dropout → Dense(1024→1024)+tanh → Dropout → Dense(1024→2)* → **logits**.
        - **Softmax & quyết định**: chuyển logits thành xác suất pos/neg; **argmax** để ra nhãn.

- **Huấn luyện**: tối ưu Cross-Entropy có **class weights**, AdamW, warmup, FP16, early-stopping; chọn mô hình theo **F1-macro** trên validation.

GIẢI THÍCH CHI TIẾT THEO BƯỚC

- **Bước 1 — Word segmentation (tách từ):**
  - **Vì sao cần?** Tiếng Việt dùng khoảng trắng giữa **âm tiết**, không phải "từ". Từ đa âm (như *Ngân hàng*) sẽ bị chia nhỏ nếu không ghép lại → mô hình hiểu sai ngữ nghĩa.
  - **Cách làm**: dùng bộ tách từ (thường là **VnCoreNLP / RDRSegmenter**).
  - **Kết quả**: "Ngân hàng báo lỗ do nợ xấu tăng cao" → "**Ngân_hàng báo_lỗ** do **nợ_xấu tăng_cao**".
  - **Ảnh hưởng**: giúp quá trình tokenization BPE sau đó tạo **subword hợp lý**, sát hơn với dữ liệu mà PhoBERT đã được pretrain.

- **Bước 2 — Tokenization BPE + token đặc biệt + pad/truncate:**
  - **Tokenizer (BPE)**: thuật toán **Byte-Pair Encoding** học từ kho ngữ liệu lớn một **bảng gộp** ký tự/chuỗi ký tự thường đi cùng (`bpe.codes`). Nhờ đó, mọi từ (kể cả lạ) đều có thể biểu diễn thành vài **subword** quen.
  - **Token đặc biệt** (chuẩn RoBERTa):
    - `<s>`: *bắt đầu câu* (tương tự `[CLS]` của BERT),
    - `</s>`: *kết thúc câu*,
    - `<pad>`: token *đệm*, để câu ngắn đủ dài.
  - **Chuẩn hoá độ dài**:
    - **Truncation**: nếu >128 token thì cắt bớt (thường giữ đầu câu).
    - **Padding**: nếu <128, chèn `<pad>` cho đủ 128.
  - **Đầu ra**:

        - `input_ids` (mảng số ID của token),
        - `attention_mask` (1: token thật, 0: pad).

- **Bước 3 — Embedding: từ ID → vector 1024 chiều:**
    - Mỗi token tại vị trí *i* được biểu diễn bằng **token_embedding[id] + positional_embedding[i] → vector 1024**.
    - **Token embedding**: bảng tra cứu học được trong pretrain (mang ý nghĩa từ/subword).
    - **Positional embedding**: thêm thông tin **thứ tự** (Transformer không tự biết trật tự như RNN).
    - **Dạng tensor**: với batch size **B**, câu dài **128**, hidden **1024 → [B, 128, 1024]**.

- **Bước 4 — 24 lớp Transformer Encoder: "ngữ cảnh hoá" token. Mỗi lớp có 2 khối chính:**
    - **Multi-Head Self-Attention**
        - "Self-Attention" cho phép **mỗi token** "nhìn" **toàn bộ** các token khác để học **quan hệ phụ thuộc** (ai liên quan ai, mức độ bao nhiêu).
        - **Multi-Head**: tách không gian 1024 thành **16 "đầu"** (head), mỗi đầu học một kiểu quan hệ khác (cú pháp, ngữ nghĩa, phủ định, thời gian…).
        - **Mask** bảo đảm pad không ảnh hưởng: những vị trí pad bị chặn ra khỏi tính attention.
    - **Feed-Forward Network (FFN)**
        - Hai lớp tuyến tính (thường *1024 → 4096 → 1024*) với kích hoạt **GELU**, đóng vai trò "trộn" và "phi tuyến hoá" đặc trưng sau attention.
    - **Cả 2 khối đều:**
        - **Residual connection** (nối tắt) giúp gradient ổn định,
        - **LayerNorm** (chuẩn hoá) giúp hội tụ tốt,

        - **Dropout** giảm overfitting.
    - **Vì sao "ngữ cảnh hoá"?**
        - Sau nhiều lớp, **vector của mỗi token** không chỉ chứa nghĩa "bản thân" nó, mà còn "nghĩa trong **bối cảnh**" cả câu (ai bổ nghĩa cho ai, phủ định, nguyên nhân-hậu quả, …). Do đó, vector của các token **khác nhau** và **giàu ngữ cảnh**.
    - **Tại sao PhoBERT-large dùng 24 lớp & 16 heads?**
        - **Sâu hơn / nhiều head hơn** → khả năng mô hình hoá **quan hệ phức tạp** tốt hơn, đặc biệt hữu ích cho câu dài/ý phức.
        - Đánh đổi: **tài nguyên** (VRAM) và **thời gian huấn luyện** tăng.

- **Bước 5 — Đại diện câu: vector `<s>` ở lớp cuối:**
    - Theo chuẩn RoBERTa/BERT, lấy **vector của token `<s>`** (vị trí đầu tiên) ở **lớp cuối cùng** làm **đại diện toàn câu**.
    - Lý do: trong huấn luyện phân loại, **head** đặt trực tiếp lên vector `<s>`, **loss** back-prop "ép" vector này học cách **tóm tắt toàn bộ nội dung câu** để phục vụ phân loại.
    - **Kích thước**: **[B, 1024]** (mỗi câu → 1 vector 1024 chiều)
    - **Ghi chú**: có các chiến lược pooling khác (mean/max pooling toàn chuỗi), nhưng chuẩn RoBERTa dùng `<s>` và hoạt động rất tốt trên nhiều tác vụ.
- **Bước 6 — Head phân loại nhị phân:**

  Chuỗi thao tác (trên vector `<s>`):

    - **Dropout → Dense(1024→1024) + tanh → Dropout → Dense(1024→2)**
      → thu được **logits** `[logit_neg, logit_pos]`.
    - **Dropout**: chống overfitting.
    - **Dense 1024→1024 + tanh**: học biến đổi *phi tuyến* để phân tách tốt hơn.

        - **Dense 1024→2** (*Out_proj*): gom đặc trưng thành 2 điểm số (neg/pos).
        - **Softmax** trên logits → xác suất **P(neg), P(pos)**; dự đoán = **argmax**.
    - **Bước 7 — Huấn luyện: tối ưu & chống lệch lớp:**
        - **Mục tiêu (loss)**: Cross-Entropy có **class weights** để bù lệch nhãn (*pos* ≫ *neg*). *w_y* lớn hơn cho **neg** → mô hình "quan tâm" hơn đến lỗi với tin xấu.

$$\mathcal{L} = -w_y \log P(y \mid x)$$

        - **Tối ưu**: **AdamW**, *weight decay* 0.01, **warmup** một phần nhỏ bước đầu để LR tăng dần rồi giảm.
        - **FP16**: rút ngắn thời gian và tiết kiệm VRAM.
        - **Early-Stopping**: dừng sớm khi không cải thiện (giảm overfitting).
        - **Chọn mô hình**: `load_best_model_at_end=True`, theo **F1-macro** trên validation.
    - **Bước 8 — Suy luận (Inference):**
        - **Chuẩn bị**: bật `eval`, tắt dropout, không tính gradient.
        - **Pipeline**: word segmentation → BPE + `<s>`,`</s>` + pad/truncate → embedding → 24 lớp Transformer → lấy `<s>` → head phân loại → softmax → **pos/neg**.
        - **Giải thích xác suất**: bạn hay in `prob_pos` = P(pos). Cũng có thể điều chỉnh **ngưỡng** khác 0.5 nếu muốn ưu tiên phát hiện **neg** (ví dụ rủi ro tài chính).

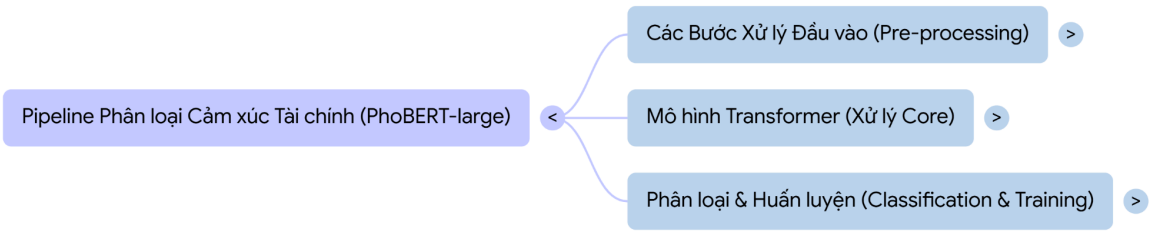
- **Các thuật ngữ liên quan:**
    - **Word segmentation**: ghép các âm tiết thành *từ đa âm* có nghĩa.
    - **Token**: đơn vị xử lý (từ hoặc *subword*).
    - **Subword**: mảnh con của từ dùng trong BPE để xử lý từ hiếm/chuỗi chưa gặp
    - **BPE (`bpe.codes`)**: bảng luật gộp để tạo **subword**.
    - `vocab.txt`: từ điển ánh xạ subword → ID.

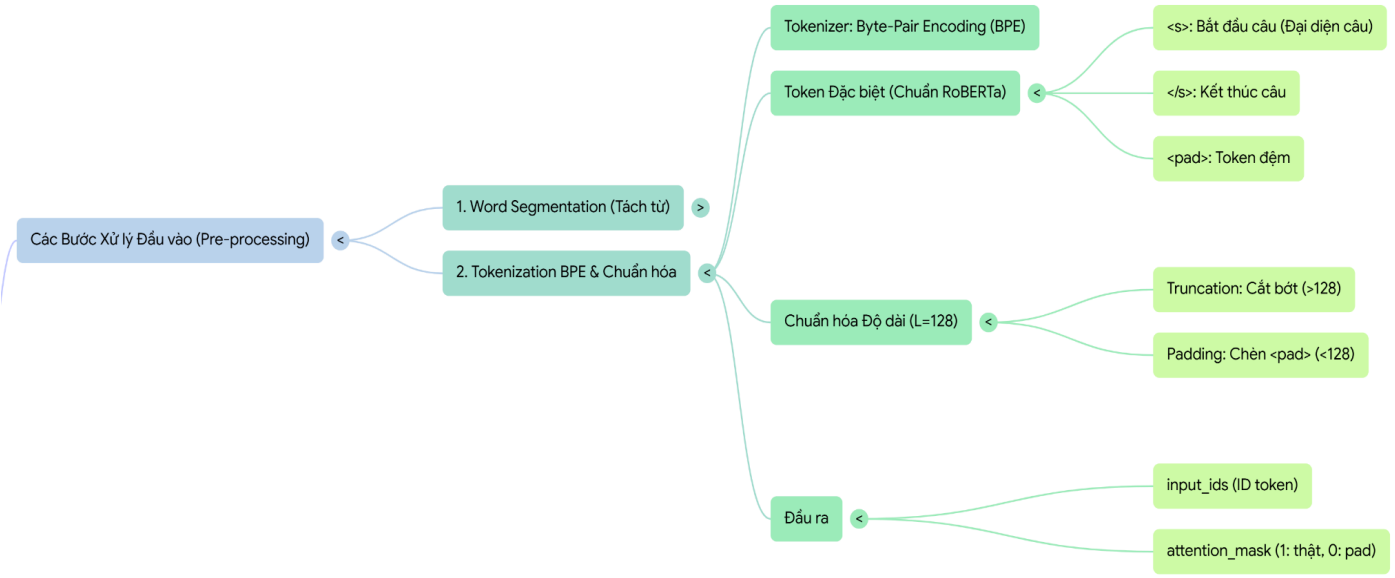
- **Padding/Truncation**: đệm/cắt để mọi câu có cùng **độ dài L=128**.
- **Attention mask**: cho Transformer biết **bỏ qua** phần pad khi tính attention.
- **Embedding**: biểu diễn dense nhiều chiều của token/vị trí.
- **Hidden size**: số chiều của không gian biểu diễn (PhoBERT-large: **1024**).
- **Batch size**: số câu xử lý song song trong một bước (ví dụ B=16/32).
- **Self-Attention**: cơ chế tính "ai chú ý đến ai" trong câu.
- **Multi-Head**: nhiều "góc nhìn" song song.
- **Residual/Skip connection**: cộng đầu vào với đầu ra lớp con để giữ thông tin & ổn định gradient.
- **LayerNorm**: chuẩn hoá kích hoạt theo feature để học dễ hơn.
- **GELU**: hàm kích hoạt mượt, hiệu quả cho Transformer.
- **Logits**: điểm số thô trước softmax.
- **Softmax**: chuẩn hoá logits thành phân phối xác suất.
- **Argmax**: chọn nhãn có xác suất lớn nhất.
- **Class imbalance**: lệch lớp; **class weights** để cân bằng ảnh hưởng.
- **AdamW**: biến thể Adam có tách weight decay đúng cách.
- **Warmup**: khởi động LR mượt để ổn định huấn luyện.
- **Overfitting**: học quá kỹ train set → kém tổng quát; dùng dropout, early-stopping, regularization để tránh.

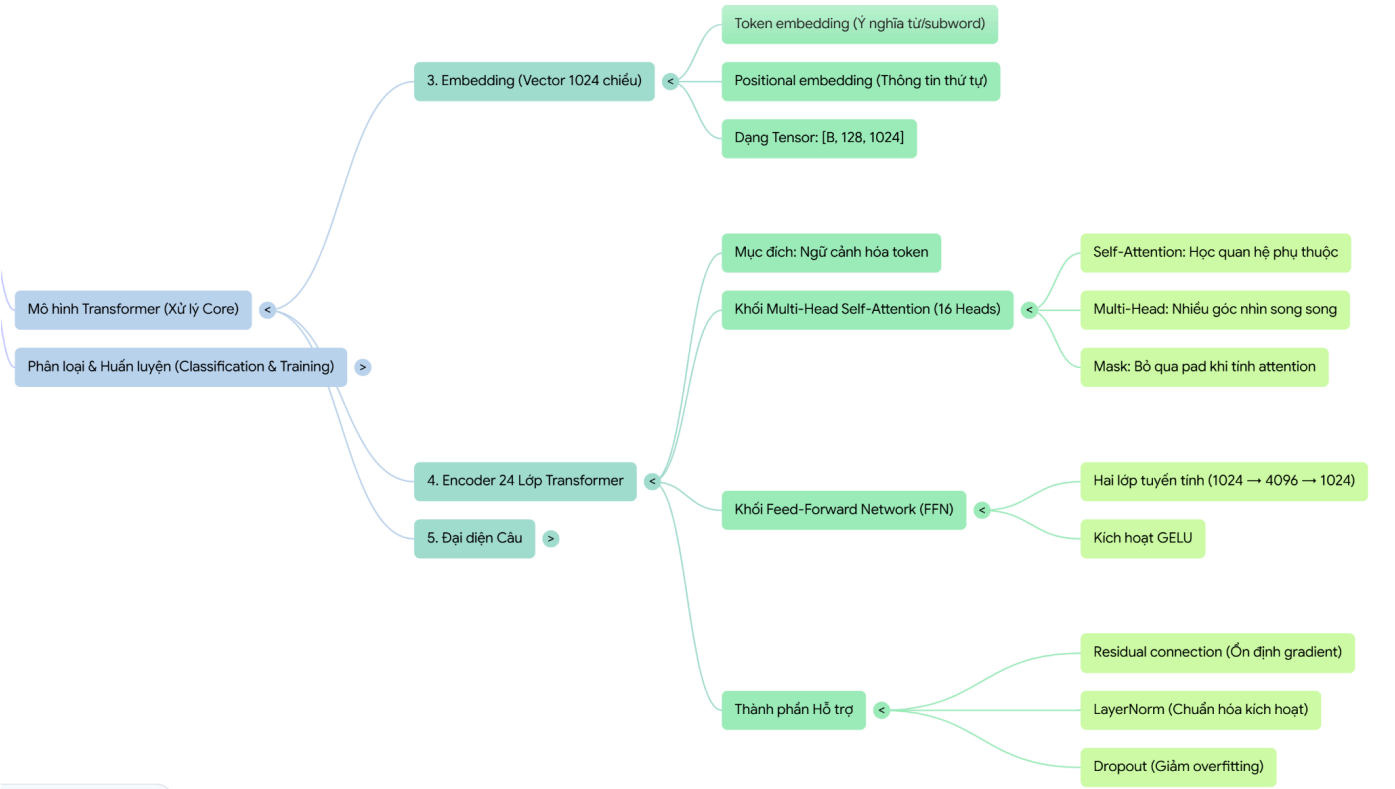
- **Ví dụ minh họa:**
  - **"FPT công bố lợi nhuận tăng mạnh quý 3/2024"**
    - Sau tách từ: "FPT công_bố lợi_nhuận tăng_mạnh quý 3/2024"
    - Mô hình:
      - Tập trung (attention) vào *lợi_nhuận, tăng_mạnh* → vector `<s>` nghiêng về **tích cực**.

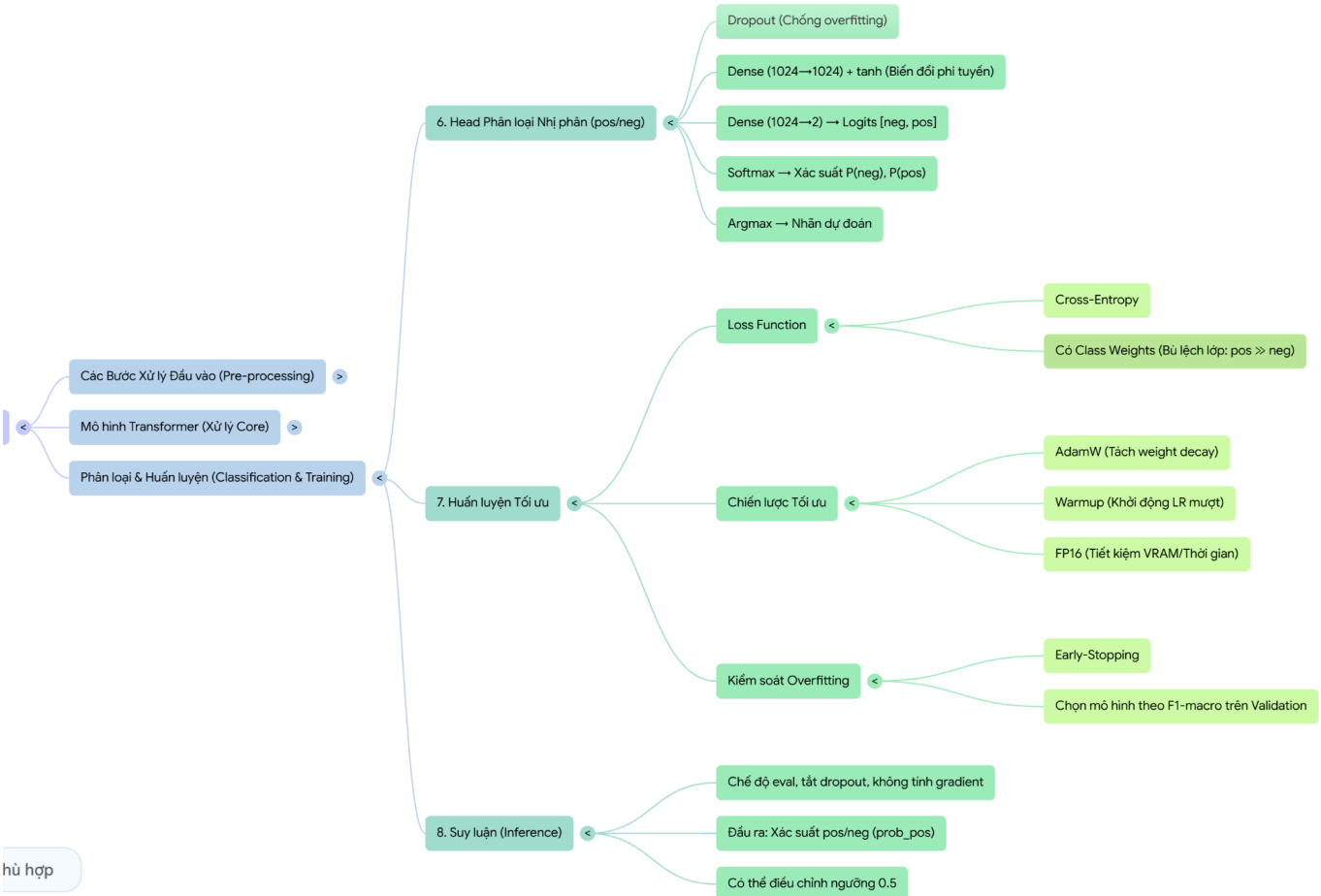
        - Kết quả: **pos** với `prob_pos` rất cao (~0.98)
    - **“Ngân hàng báo lỗ do nợ xấu tăng cao”**
        - Sau tách từ: “Ngân_hàng báo_lỗ do nợ_xấu tăng_cao”
        - Mô hình:
            - Attention liên kết *báo_lỗ* với *nợ_xấu, tăng_cao* → ngữ cảnh **tiêu cực**.
            - Kết quả: **neg** với `prob_pos` rất thấp (~0.03).
- **Tổng kết**
    - PhoBERT-large xử lý headline theo chuỗi **chuẩn Transformer**: *segmentation → BPE → embeddings → 24×(self-attention + FFN) → sentence vector* `<s>` *→ head phân loại → softmax*.
      Nhờ **tách từ** và **self-attention đa đầu sâu 24 lớp**, vector `<s>` mang đủ *nghĩa + ngữ cảnh* để phân biệt **pos/neg** hiệu quả; kết hợp **class weights** và tối ưu thích hợp đã giúp mô hình của bạn đạt **F1-macro ~0.86 trên test**, phù hợp với mục tiêu phân loại cảm xúc tiêu đề tài chính tiếng Việt.

- **Sơ đồ hoạt động tổng quát**




Các Bước Xử lý Đầu vào (Pre-processing)
1. Word Segmentation (Tách từ)
2. Tokenization BPE & Chuẩn hóa
Tokenizer: Byte-Pair Encoding (BPE)
Token Đặc biệt (Chuẩn RoBERTa)
<s>: Bắt đầu câu (Đại diện câu)
</s>: Kết thúc câu
<pad>: Token đệm
Chuẩn hóa Độ dài (L=128)
Truncation: Cắt bớt (>128)
Padding: Chèn <pad> (<128)
Đầu ra
input_ids (ID token)
attention_mask (1: thật, 0: pad)


Mô hình Transformer (Xử lý Core)
Phân loại & Huấn luyện (Classification & Training)
3. Embedding (Vector 1024 chiều)
Token embedding (Ý nghĩa từ/subword)
Positional embedding (Thông tin thứ tự)
Dạng Tensor: [B, 128, 1024]
4. Encoder 24 Lớp Transformer
5. Đại diện Câu
Mục đích: Ngữ cảnh hóa token
Khối Multi-Head Self-Attention (16 Heads)
Self-Attention: Học quan hệ phụ thuộc
Multi-Head: Nhiều góc nhìn song song
Mask: Bỏ qua pad khi tính attention
Khối Feed-Forward Network (FFN)
Hai lớp tuyến tính (1024 → 4096 → 1024)
Kích hoạt GELU
Thành phần Hỗ trợ
Residual connection (Ổn định gradient)
LayerNorm (Chuẩn hóa kích hoạt)
Dropout (Giảm overfitting)



Phương pháp

Kiến trúc Mô hình Dự báo Tín hiệu

Chúng tôi coi bài toán dưới dạng **phân loại nhị phân có trình tự thời gian**: dự báo liệu cổ phiếu có **tín hiệu Long (mua)** trong giai đoạn sắp tới hay không (**Neutral** nghĩa là không mua). Với mỗi cổ phiếu, mô hình sẽ xem xét chuỗi dữ liệu **90 phiên giao dịch gần nhất** (khoảng ~4,5 tháng) của 8 đặc trưng đã đề cập, rồi đưa ra **xác suất dự báo** cổ phiếu đó nên **mua trong 20 phiên kế tiếp**. Nói cách khác, cứ sau mỗi 20 phiên (khoảng 1 tháng giao

dịch), mô hình lại đánh giá và xếp hạng cơ hội đầu tư cho các cổ phiếu dựa trên dữ liệu 90 phiên vừa qua.

Chúng tôi triển khai và so sánh **ba kiến trúc mạng học sâu** để dự báo tín hiệu này:

- **Mạng LSTM (Long Short-Term Memory)**: LSTM là một biến thể mạng hồi quy chuỗi thời gian (RNN) có cơ chế cổng ghi nhớ, giúp nắm bắt quan hệ dài hạn trong dữ liệu tuần tự và khắc phục vấn đề vanishing gradient. LSTM nổi tiếng hiệu quả trong dự báo chuỗi tài chính nhờ khả năng mô hình hóa được cả **các mẫu phi tuyến và phụ thuộc dài hạn**ewadirect.com. Nhiều nghiên cứu trên thế giới và Việt Nam đã dùng LSTM dự báo giá chứng khoán với độ chính xác cao hơn mô hình truyền thốngarxiv.org. Trong dự án này, chúng tôi muốn xây dựng mạng LSTM nhiều tầng: ví dụ, 2 lớp LSTM (100 neuron mỗi lớp) nối tiếp, rồi đến một lớp Dense ẩn (25 neuron) và lớp đầu ra kích hoạt sigmoid dự báo xác suất Long (0-1). Các siêu tham số (số tầng, số neuron, learning rate, v.v.) được tinh chỉnh dựa trên tập validation 2022.

- **Mạng GRU (Gated Recurrent Unit)**: GRU là biến thể RNN gọn nhẹ hơn, chỉ với hai cổng (update và reset) thay vì ba cổng như LSTM. GRU thường có hiệu năng tương đương LSTM nhưng huấn luyện nhanh hơn do cấu trúc đơn giản. Chúng tôi thử nghiệm GRU với kiến trúc tương tự LSTM để so sánh xem liệu GRU có học tốt chuỗi thời gian chứng khoán hay không. Việc so sánh LSTM và GRU cũng từng được thực hiện trong một số nghiên cứu tại Việt Nam, cho thấy cả hai đều có ưu điểm tùy trường hợpjst.tnu.edu.vnctujsvn.ctu.edu.vn.

- **Mô hình Transformer Encoder**: Đây là mô hình tự chú ý (self-attention) hiện đại, ban đầu phát triển cho NLP nhưng ngày càng được áp dụng cho dự báo chuỗi thời gian arxiv.org. Transformer encoder có khả năng nhìn toàn bộ chuỗi quá khứ cùng lúc và **học trọng số chú ý** tới những thời điểm quan trọng, giúp nắm bắt **quan hệ dài hạn** một cách hiệu quả mà không bị quên thông tin xa. Nghiên cứu đã chỉ ra các mô hình Transformer có thể vượt qua LSTM trong nhiều bài toán dự báo chuỗi phức tạp arxiv.org. Trong mô hình của chúng tôi, khối encoder Transformer bao gồm cơ chế **multi-head attention** và các lớp feed-forward, với đầu vào là chuỗi 90 ngày x 8 đặc trưng (có thêm encoding vị trí thời gian). Đầu ra encoder sau đó được nối qua một lớp dense và sigmoid để thu được xác suất Long. Chúng tôi kỳ vọng so sánh này sẽ làm rõ liệu Transformer có đem lại lợi thế dự báo tín hiệu so với các RNN truyền thống trong bối cảnh dữ liệu thị trường Việt Nam hay không.

Cả ba mô hình đều được huấn luyện tối ưu hóa hàm loss nhị phân (binary cross-entropy) với thuật toán Adam. Để tránh **quá khớp** (overfitting), chúng tôi áp dụng các kỹ thuật **regularization** phù hợp: thêm **Dropout** giữa các lớp (ví dụ dropout rate 0.2–0.3), áp dụng **early stopping** dựa trên loss trên tập validation, và kiểm thử nhiều kích thước batch nhỏ để mô hình học ổn định. Do dữ liệu dự báo Long/Neutral có thể bị lệch (ít tín hiệu Long hơn Neutral), chúng tôi cũng cân nhắc điều chỉnh trọng số lớp hoặc dùng kỹ thuật **resampling** để đảm bảo mô hình học tốt cả hai lớp. Bộ tham số cuối cùng được chọn sao cho mô hình cho kết quả tốt trên tập validation 2022 trước khi **kiểm thử trên giai đoạn 2023–2025**.

Chiến lược Xây dựng Danh mục Đầu tư

Chiến lược đầu tư của chúng tôi vận hành đồng bộ với chu kỳ dự báo của mô hình. Cụ thể, cứ **mỗi 20 phiên giao dịch** (khoảng 1 tháng), ta thực hiện quy trình xây dựng danh mục như sau:

1. **Dự báo tín hiệu:** Ở phiên đầu tiên của chu kỳ (ví dụ phiên thứ 1, 21, 41,... tính từ đầu test), chúng ta sử dụng mô hình (LSTM/GRU/Transformer đã được chọn) để tính **xác suất Long** cho từng cổ phiếu trong danh sách 25 mã, dựa trên dữ liệu 90 phiên trước đó. Kết quả là mỗi cổ phiếu có một xác suất được mô hình đánh giá có tín hiệu tích cực trong 20 phiên tới.

2. **Chọn cổ phiếu:** Chúng tôi sắp xếp các cổ phiếu theo xác suất Long giảm dần, và **chọn ra Top 5 cổ phiếu** có xác suất cao nhất. Năm cổ phiếu này được xem là các mã tiềm năng nhất để đưa vào danh mục đầu tư cho chu kỳ 20 ngày sắp tới. Việc chọn top 5 nhằm đảm bảo danh mục đủ **đa dạng hóa** (tránh quá tập trung chỉ 1-2 mã) nhưng cũng tập trung vào những cơ hội tốt nhất. Danh sách 25 mã gốc có tính thanh khoản cao và đại diện nhiều ngành nên top 5 thường trải rộng các ngành, giảm thiểu rủi ro chung.

3. **Phân bổ vốn:** Danh mục 5 cổ phiếu này được phân bổ theo tỷ trọng bằng nhau (**equal weight**), tức mỗi mã chiếm 20% giá trị danh mục. Cách phân bổ đồng đều giữ cho chiến lược đơn giản và tránh phụ thuộc vào mô hình định lượng mức độ khác nhau giữa các xác suất (vì xác suất bản chất không phải dự báo chính xác mức sinh lời). Hơn nữa, equal-weight giúp kiểm soát rủi ro – mỗi mã có mức ảnh hưởng ngang nhau, không mã nào chi phối toàn bộ danh mục.

4. **Nắm giữ và Tái cân bằng:** Trong suốt 20 phiên của chu kỳ, danh mục được **giữ nguyên**, không thay đổi (buy-and-hold trong ngắn hạn). Khác với một số chiến lược tích cực giao dịch liên tục, chúng tôi **không chồng lệnh** hay thêm bớt cổ phiếu trong kỳ – điều này giúp giảm chi phí giao dịch và tránh nhiễu do tín hiệu ngắn hạn. Sau 20 phiên, vào kỳ tái cân bằng tiếp theo, quy trình trên lặp lại: mô hình lại dự báo xác suất Long dựa trên dữ liệu mới nhất và chúng tôi lại chọn top 5 mới để tái cấu trúc danh mục cho 20 phiên kế tiếp. Một điểm quan trọng: nếu một cổ phiếu vẫn nằm trong top 5 ở kỳ mới thì **tiếp tục được giữ** (không bán ra chỉ để mua lại), còn những cổ phiếu rơi khỏi top 5 sẽ bị bán và thay bằng cổ phiếu mới cao hơn. Chiến lược này giúp danh mục **thích nghi** dần với biến động thị trường, luôn giữ các mã mạnh nhất theo đánh giá của mô hình, đồng thời tránh mua bán quá thường xuyên.

5. **Giả định đơn giản hóa:** Khi mô phỏng danh mục, chúng tôi giả định **không có phí giao dịch** và thanh khoản đủ cao để khớp lệnh toàn bộ tại giá đóng cửa phiên tái cân bằng. Đây là giả định lý tưởng nhằm tập trung đo lường hiệu quả mô hình; trong thực tế phí giao dịch và trượt giá có thể làm giảm lợi nhuận, nhưng với danh mục thanh khoản cao và tần suất giao dịch ~1 tháng/lần thì chi phí cũng không quá lớn. Ngoài ra, chúng tôi chưa áp dụng các cơ chế kiểm soát rủi ro chủ động như cắt lỗ (stop-loss) hay giới hạn đòn bẩy trong mô hình này – mục tiêu chính là đánh giá **tiềm năng thô** của tín hiệu mô hình. Thay vào đó, rủi ro sẽ được phản ánh qua các chỉ số

như độ biến động và drawdown của danh mục.

Mô phỏng Monte Carlo 3 giai đoạn (Một cách tổng quát, dễ hiểu)

Sinh kịch bản tương lai (Monte Carlo):

- **Tham số cố định:**
  - **Số mã**: 5 (Top-5 sau bước dự báo).
  - **Chiều dài dự phóng**: 20 phiên.
  - **Cửa sổ dữ liệu quá khứ**: 90 phiên gần nhất.
  - **Số kịch bản**: 10.000.
  - **Độ dài block** (bootstrap): 5 phiên.
  - **CRN** (Common Random Numbers): **BẬT** – mọi bộ trọng số về sau sẽ dùng **cùng** 10.000 kịch bản này để so sánh công bằng.
  - **Đánh giá cuối**: sinh **10.000 kịch bản mới** (seed khác) để kiểm tra khách quan.
  -
- **Bước 1 — Chuẩn bị ma trận lợi suất 90 phiên (5 cột)**
  - Lấy **lợi suất ngày** của **5 mã** trong **90 phiên gần nhất** (mỗi mã là 1 cột).
  - **Làm sạch nhẹ**: cắt bớt 1% giá trị cực đoan ở hai đầu (winsorize) trên từng cột để giảm nhiễu bất thường.
  - Đồng bộ ngày giao dịch (chỉ phiên có đủ 5 mã).
  - **Kết quả:**
    - **Một bảng 90×5 (90 dòng ngày, 5 cột mã) đã sạch và sẵn sàng.**
- **Bước 2 : "Nghiêng" kỳ vọng theo dự báo (tilt theo xác suất Long)**
  - Với mỗi mã iii, bạn đã có **xác suất Long** cho 20 phiên tới: gọi là **p_i**.
  - Dựa trên **bảng hiệu chuẩn** đã xây từ dữ liệu validation (p cao → kỳ vọng 20 phiên cao hơn), **suy ra** mức kỳ vọng 20 phiên cho từng mã.
  - **Chia đều** về mức kỳ vọng theo **ngày**, rồi **trộn 50/50** với "trạng thái lịch sử gần nhất" của 90 phiên (để không thiên lệch hoàn toàn theo mô hình).
  - Kết quả: mỗi mã có **kỳ vọng ngày mục tiêu** hợp lý (không cần công thức, chỉ là 5 con số "drift"/ngày).
  - **Ý nghĩa:**

        - Nếu mô hình tin rằng mã A có p cao, thì **kịch bản tương lai của A** sẽ hơi **ngả về hướng tích cực** (nhưng chỉ vừa phải, vì đã trộn với lịch sử).
- **Bước 3:Cách sinh kịch bản: Block Bootstrap đa biến**
    - **Mục tiêu**: giữ **tương quan chéo** giữa các mã (ngân hàng – chứng khoán – bluechip thường đi cùng nhau ở mức nào đó) và **động lực ngắn hạn** (xu hướng/đảo chiều trong vài ngày liền kề).
    - **Cách làm (mỗi kịch bản 20 phiên):**
        - **Chia quá khứ thành các block 5 phiên** (trượt trên 90 phiên).
        - **Bốc ngẫu nhiên 4 block** (5 phiên × 4 block = 20 phiên).
        - **Ghép nối các block theo thời gian** để tạo một **đường lợi suất 20×5** (20 ngày, 5 mã).
            - Ghép **theo hàng** (cùng ngày cho cả 5 mã) ⇒ **giữ nguyên mối liên hệ giữa các mã**
        - **Đặt lại "trung bình ngày"** của kịch bản này về **mức kỳ vọng ngày mục tiêu** ở bước 2 (tức là dịch nhẹ toàn kịch bản sao cho "đi đúng hướng" dự báo, nhưng **không** thay đổi cấu trúc biến động và tương quan lấy từ dữ liệu thật).
        - **Kiểm soát ngoại lệ**: nếu xuất hiện ngày quá cực đoan (ví dụ > ±10%/ngày), cắt ngưỡng cho hợp lý.
        - Lặp lại 10.000 lần ⇒ **10.000 kịch bản 20×5**.
    - Vì block dài 5 phiên, mỗi kịch bản ghép 4 block là đủ 20 phiên. Cách này **vừa nhanh** vừa **giữ được quán tính ngắn hạn**.
- **Bước 4: Common Random Numbers (CRN) – vì sao bật?**
    - Trước khi tối ưu trọng số, **rút sẵn** danh sách chỉ số block cho **10.000 kịch bản** (mỗi kịch bản là 4 chỉ số block).
    - Khi bạn thay đổi trọng số (Stage A/B/C), **không rút lại kịch bản**; tất cả phương án đều đi qua **cùng 10.000 kịch bản** này.
    - Lợi ích: **so sánh công bằng**, giảm "nhiễu may rủi" giữa các bộ trọng số → bề mặt tối ưu **mượt** và ổn định hơn.
- **Bước 5 : Kiểm tra & xuất kết quả:**
    - Với mỗi kịch bản (20 ngày × 5 mã), có thể quy đổi ra **đường giá** (nếu cần vẽ), còn tối ưu thì chỉ cần **chuỗi lợi suất**.
    - Lưu 10.000 kịch bản này thành một "bộ chuẩn" cho **toàn bộ quá trình tối ưu** trong kỳ 20 phiên.
    - **Đánh giá cuối** (sau khi đã tìm ra trọng số tối ưu): sinh **10.000 kịch bản mới** (seed khác) để **kiểm chứng khách quan** (không bị "thuận tay" do CRN của quá trình tối ưu).

Stage A — Quét thô bằng lưới

- **Mục tiêu:**
  - Quét nhanh **toàn bộ không gian tỷ trọng** (5 mã, tổng 100%, không bán khống, **trần 40%/mã**) bằng một **lưới bước 5%** để tìm ra **20 bộ tỷ trọng hứa hẹn nhất**. "Hứa hẹn" nghĩa là **hiệu quả ổn định** trên **2.000 kịch bản** (rút từ bộ 10.000), ưu tiên **Sharpe trung vị cao** và **sụt giảm trung vị (MDD) không quá 25%**.
- **Tham số cố định:**
  - **Số mã**: 5 (Top-5 của kỳ).
  - **Lưới tỷ trọng**: bước **5%** cho từng mã, tổng phải = 100%.
  - **Trần**: mỗi mã **≤ 40%**.
  - **Kịch bản dùng để chấm**: **2.000** (lấy từ bộ 10.000 kịch bản đã sinh sẵn; bật **CRN** – mọi bộ tỷ trọng dùng cùng tập kịch bản này).
  - **Kỳ hạn đánh giá**: 20 phiên (khớp với kỳ nắm giữ).
  - **Ngưỡng loại sớm theo rủi ro**: **MDD trung vị ≤ 25%**.
  - **Số ứng viên cần giữ lại**: **Top-20** (để sang Stage B).
- **Bước 1 — Tạo danh sách ứng viên trên lưới 5%:**
  - Sinh các bộ tỷ trọng sao cho:
    - Mỗi mã nhận một phần trăm theo bội số **5%** (0%, 5%, 10%, …, 40%).
    - **Tổng đúng 100%**.
    - **Không mã nào vượt 40%**.
  - Mẹo sinh nhanh:
    - Duyệt theo "sao–vạch" có trần hoặc sinh ngẫu nhiên trên lưới rồi **lọc**.
    - Mục tiêu có **khoảng ~2.000 ứng viên** hợp lệ để chấm điểm.
  - Kết quả: một danh sách ~2.000 bộ tỷ trọng, ví dụ:(40–40–10–5–5), (35–25–20–10–10), (20–20–20–20–20), …
- **Bước 2 — Chuẩn bị bộ kịch bản chấm điểm:**
  - Lấy **2.000 kịch bản** đầu tiên từ bộ **10.000 kịch bản** (đã sinh ở phần trước).
  - **Giữ cố định** 2.000 kịch bản này cho **tất cả** các ứng viên (đây là CRN).
  - Mỗi kịch bản là một **đường lợi suất 20 ngày × 5 mã** đã "nghiêng" theo xác suất Long và giữ tương quan thực tế.
- **Bước 3 — Tính hiệu quả của từng ứng viên trên từng kịch bản**

Với mỗi tỉ trọng ta sẽ:

- Kết hợp tỷ trọng với **chuỗi lợi suất 20 ngày** của từng kịch bản để có **đường lợi suất danh mục** trong 20 phiên.

- Từ đường lợi suất danh mục đó, tính các **chỉ số hiệu quả**:
  - **Sharpe** (hiệu quả lợi nhuận trên mỗi đơn vị rủi ro) cho kỳ 20 phiên, tính ở **thang ngày** để so sánh.

  - **MDD (Maximum Drawdown)** trong 20 phiên (mức sụt giảm tối đa từ đỉnh xuống đáy).

  - (Tuỳ chọn thêm: lợi nhuận gộp 20 phiên, độ biến động… để báo cáo phụ.)
- Bạn sẽ có **2.000 điểm Sharpe** và **2.000 điểm MDD** cho **mỗi** ứng viên.

- **Bước 4 — Tổng hợp "theo kịch bản" thành "điểm Stage A"**
  - **Sharpe trung vị**: lấy **trung vị** của 2.000 giá trị Sharpe → cho ta "điểm Stage A" của ứng viên (điểm xếp hạng chính).

  - **MDD trung vị**: lấy trung vị của 2.000 giá trị MDD → dùng làm **bộ lọc rủi ro**.

  - (Gợi ý nên lưu thêm các **phân vị**: Sharpe 25%/75%, MDD 75%… để nhìn **độ ổn định**.)
  - Vì dùng **trung vị**, ứng viên không bị "ăn may" hoặc "ăn xui" bởi vài kịch bản cực đoan; phù hợp giai đoạn quét thô.
- **Bước 5 — Lọc rủi ro và xếp hạng:**
  - **Lọc rủi ro:** loại mọi ứng viên có **MDD trung vị > 25%**.

  - **Xếp hạng:** sắp xếp phần còn lại theo **Sharpe trung vị** từ cao xuống thấp.

  - **Tie-breaker** (khi bằng điểm):
    - Ứng viên có **Sharpe phân vị 25%** cao hơn (ít tệ khi gặp kịch bản xấu) đứng trước.

    - Nếu vẫn hoà, chọn **MDD trung vị** thấp hơn.

    - Nếu vẫn hoà nữa, chọn **Sharpe trung bình** cao hơn.
- **Bước 6 — Chọn Top-20 cho Stage B:**
  - Lấy **20 ứng viên đứng đầu** (đã qua lọc rủi ro) làm **hạt giống** cho **Stage B**.

  - Lưu kèm **hồ sơ tóm tắt** của mỗi ứng viên:
    - Sharpe trung vị, Sharpe 25%/75%,

    - MDD trung vị, MDD 75%,

    - (tuỳ chọn) lợi nhuận gộp trung vị 20 phiên.

Stage B – Khai phá trọng tâm

- **Mục tiêu:**
  - Sau khi Stage A đã chọn ra **20 bộ tỷ trọng tốt nhất** từ lưới 5%, Stage B sẽ **zoom-in** để khám phá chi tiết **xung quanh** những bộ tỷ trọng này. Mục đích là tìm ra **những vùng thực sự tối ưu**, trước khi sang bước tinh chỉnh cục bộ (Stage C).
- **Tham số cố định:**
  - **20 bộ tỷ trọng gốc** từ Stage A được coi là **"tâm"** (centers).
  - **Mỗi tâm** sẽ sinh ra **200 biến thể mới** (gọi là "điểm lân cận"):
    - **100 điểm rất gần tâm (chênh lệch chỉ vài phần trăm).**
    - **100 điểm xa hơn một chút (để không bỏ sót vùng tốt bên cạnh).**
  - **Giới hạn mỗi cổ phiếu: tối đa 40% vốn.**
  - **Tổng vốn: luôn bằng 100%.**
  - **Số kịch bản để chấm: 5 000 kịch bản (rút từ bộ 10 000 đã sinh ở bước trước) – mọi ứng viên đều dùng cùng tập kịch bản này để đảm bảo công bằng.**
  - **Kỳ hạn đánh giá: 20 phiên.**
  - **Ngưỡng rủi ro: loại sớm mọi bộ có MDD trung vị > 25% hoặc MDD phân vị 75% > 35%.**
  - **Số ứng viên giữ lại sau Stage B: Top 5 để sang Stage C.**
  -
- **Bước 1 — Sinh "đám mây" quanh mỗi tâm:**
  - Tưởng tượng mỗi tâm là một **chấm lớn trên bản đồ**, ta sẽ rải thêm **200 chấm nhỏ xung quanh**:
    - **100 chấm "near"**: chỉ dao động trong vòng **±3%** quanh mỗi tỷ trọng gốc → để **khai thác kỹ** vùng trung tâm.
    - **100 chấm "wide"**: dao động rộng hơn (±5–7%) → để **khám phá** các vùng lân cận có thể tốt hơn.
    - Mọi chấm sau khi sinh ra đều được điều chỉnh sao cho **không mã nào > 40%** và **tổng = 100%**.
  - Tổng cộng ta sẽ có khoảng **4 000 ứng viên mới** (20 × 200).
- **Bước 2 — Đánh giá từng ứng viên trên 5 000 kịch bản**

  Với mỗi ứng viên :
  - Áp tỷ trọng đó vào **chuỗi lợi suất 20 ngày của từng kịch bản** → có được đường giá trị danh mục.

- Từ đường giá trị này tính các **chỉ số hiệu suất**:
  - **Sharpe** (lợi nhuận trên rủi ro) cho kỳ 20 phiên.
  - **MDD** (mức sụt giảm lớn nhất) trong 20 phiên.
- Kết quả: mỗi ứng viên sẽ có **5 000 giá trị Sharpe** và **5 000 giá trị MDD**.

- **Bước 3 — Gộp kết quả kịch bản thành "điểm Stage B"**
  - Tính **Sharpe trung vị** (median) – dùng làm **điểm xếp hạng chính**.
  - Tính thêm **Sharpe phân vị 25%** (để xem ứng viên có ổn trong kịch bản xấu không).
  - Tính **MDD trung vị** và **MDD phân vị 75%** (để lọc rủi ro và phạt các ứng viên có đuôi rủi ro nặng).
- **Bước 4 —Lọc rủi ro & xếp hạng**
  - **Loại sớm** những ứng viên có **MDD trung vị > 25%** hoặc **MDD phân vị 75% > 35%**.
  - **Xếp hạng** phần còn lại theo **Sharpe trung vị** (càng cao càng tốt).
  - Nếu bằng điểm, ưu tiên ứng viên có **Sharpe phân vị 25% cao hơn** (tức là ổn định hơn ở kịch bản xấu), sau đó mới xét tới **MDD thấp hơn**.

- **Ép đa dạng & chọn Top 5**
  - Để tránh 5 ứng viên giống nhau, áp dụng **"khoảng cách tối thiểu"**: chỉ nhận thêm ứng viên nếu **khác biệt tổng phân bổ ≥ 10%** so với các ứng viên đã được chọn trước đó.
  - Tiếp tục duyệt theo thứ tự xếp hạng cho tới khi đủ **Top 5 ứng viên cuối cùng**.
  - Lưu lại hồ sơ thống kê (Sharpe trung vị, phân vị 25%, MDD trung vị, MDD phân vị 75%…) của 5 ứng viên này để sử dụng ở Stage C.
- **Bước 6 — Chọn Top-20 cho Stage B:**
  - Lấy **20 ứng viên đứng đầu** (đã qua lọc rủi ro) làm **hạt giống** cho **Stage B**.
  - Lưu kèm **hồ sơ tóm tắt** của mỗi ứng viên:
    - Sharpe trung vị, Sharpe 25%/75%,
    - MDD trung vị, MDD 75%,
    - (tuỳ chọn) lợi nhuận gộp trung vị 20 phiên.

Stage C – Tinh chỉnh cục bộ

- **Mục tiêu:**
  - Stage C là bước cuối cùng để chọn ra **bộ trọng số đầu tư tối ưu nhất** cho 5 cổ phiếu đã được lọc ở Stage A và Stage B.

    - Ta sẽ tập trung tinh chỉnh tỉ trọng vốn quanh **5 bộ trọng số tốt nhất** còn lại sau Stage B, với mục tiêu:
        - Tăng Sharpe Ratio (lợi nhuận tốt hơn so với rủi ro),
        - Giữ mức sụt giảm vốn (Maximum Drawdown) thấp hơn,
        - Đảm bảo kết quả ổn định trong nhiều kịch bản thị trường khác nhau.

- **Bộ dữ liệu đầu vào:**
    - **5 bộ trọng số hạt giống (seed) tốt nhất từ Stage B.**
    - **10.000 kịch bản thị trường tương lai (mỗi kịch bản gồm lợi suất của 5 cổ phiếu trong 20 phiên tiếp theo).**
    - **Ràng buộc: tổng tỉ trọng vốn của 5 cổ phiếu luôn bằng 100%, mỗi cổ phiếu không vượt quá 40%.**
- **Bước 1 — Chọn 5 hạt giống:**
    - Từ Stage B, ta lấy ra 5 bộ trọng số có điểm cao nhất làm **khởi đầu cho việc tinh chỉnh**.
    - Ví dụ:
        - Seed 1: (25%, 20%, 15%, 20%, 20%)
        - Seed 2: (22%, 18%, 25%, 20%, 15%)
        - …
- **Bước 2 — Tinh chỉnh cục bộ quanh từng seed**
    - Đối với mỗi seed, ta “khoanh vùng” tìm kiếm những cách phân bổ vốn khác **gần xung quanh** seed đó (chẳng hạn điều chỉnh ±3–5% mỗi cổ phiếu).
    - Mục đích là tìm xem có thể cải thiện hiệu quả mà không cần thay đổi quá xa so với seed gốc..
- **Bước 3 — Đánh giá trên 10.000 kịch bản**
    - Mỗi bộ trọng số mới sinh ra sẽ được kiểm tra trên **toàn bộ 10.000 kịch bản**:
        - Tính lợi suất gộp của danh mục qua 20 phiên.
        - Đo Sharpe Ratio và mức sụt giảm vốn (MDD) của danh mục.
        - Chấm điểm dựa trên cả ba tiêu chí: trung vị Sharpe, Sharpe trong kịch bản xấu, và MDD trong kịch bản xấu.
    - Bộ trọng số nào có điểm cao hơn được giữ lại.
- **Bước 4 —Lặp lại để cải thiện**
    - **Quy trình tinh chỉnh → đánh giá → giữ lại bộ tốt hơn sẽ lặp đi lặp lại nhiều vòng cho đến khi không còn cải thiện đáng kể.**
- **Bước 5 — Chọn bộ tối ưu cuối**
    - Sau khi đã tinh chỉnh cho cả 5 seed, ta chọn **bộ trọng số có tổng điểm cao nhất** làm phương án đầu tư chính thức cho chu kỳ 20 phiên tới.

Đánh giá Hiệu năng

Đánh giá Độ chính xác Dự báo

Trước khi xem xét lợi nhuận, chúng tôi đo lường **độ chính xác phân loại tín hiệu Long/Neutral** của mô hình trên từng cổ phiếu. Các thước đo bao gồm:

- **Accuracy** – tỷ lệ dự báo đúng (dự báo Long đúng khi cổ phiếu thực sự tăng tốt, và Neutral đúng khi cổ phiếu không tăng đáng kể).

- **Precision và Recall (đối với lớp Long)** – Precision cao nghĩa là những tín hiệu Long mà mô hình đưa ra ít bị "dương tính giả" (chọn nhầm cổ phiếu không tốt), Recall cao nghĩa là mô hình tìm được hầu hết các cổ phiếu thực sự có xu hướng tăng. F1-score (harmonic mean của Precision & Recall) cũng được tính để đánh giá cân bằng.

- **AUC-ROC và AUC-PR** – Diện tích dưới đường cong ROC và PR, đo khả năng mô hình phân tách hai lớp Long vs Neutral ở các ngưỡng khác nhau. Đặc biệt khi tỷ lệ mẫu Long vs Neutral mất cân bằng, AUC-PR là thước đo hữu ích.

Đánh giá Hiệu suất Danh mục Đầu tư

Quan trọng hơn, chúng tôi kiểm tra **kết quả danh mục đầu tư** do các mô hình sinh ra trên giai đoạn kiểm thử. Các **chỉ số hiệu suất** được sử dụng gồm:

- **CAGR (Compound Annual Growth Rate)**: Tốc độ tăng trưởng kép hàng năm của danh mục, đo mức sinh lời trung bình hằng năm. Nó cho biết **trung bình mỗi năm danh mục tăng bao nhiêu %** nếu như tăng trưởng đều đặn theo lãi kép. Đây là thước đo tổng hợp để so sánh với các kênh đầu tư khác (VD: VN-Index, lãi suất ngân hàng).
    - Công thức:

      Với $V0$ là giá trị danh mục ban đầu, VT là giá trị cuối cùng sau T năm:

      $$CAGR = \left(\frac{V_T}{V_0}\right)^{\frac{1}{T}} - 1$$

    - Ý nghĩa:
        - CAGR giúp **so sánh hiệu quả dài hạn** của danh mục với các kênh khác (VN-Index, gửi tiết kiệm, TPCP…).

        - Bỏ qua biến động từng năm, tập trung vào **mức tăng trưởng gộp**.

        - Ví dụ: đầu tư 100 → 150 sau 3 năm → CAGR≈14.47% mỗi năm.

- **Sharpe Ratio:**
  - Khái niệm:
    - **Sharpe ratio** đo **lợi nhuận vượt trội so với tài sản phi rủi ro** trên **mỗi đơn vị rủi ro (biến động)** mà nhà đầu tư phải chịu.
  - Công thức:

    Với Rp là lợi nhuận trung bình của danh mục, Rf là lãi suất phi rủi ro (ví dụ: TPCP ngắn hạn), σp là độ lệch chuẩn lợi nhuận danh mục:

$$Sharpe = \frac{R_p - R_f}{\sigma_p}$$

  - Ý nghĩa:
    - Sharpe cao → **mỗi đơn vị rủi ro mang lại nhiều lợi nhuận hơn** → chiến lược hiệu quả.
    - Quy ước:
      - Sharpe ≈ 0.5 → trung bình
      - Sharpe > 1.0 → tốt
      - Sharpe > 2.0 → rất tốt, ít danh mục duy trì được lâu dài.

- **Maximum Drawdown (MDD):**
  - Khái niệm:
    - **MDD** là **mức sụt giảm lớn nhất** của giá trị danh mục từ **đỉnh cao nhất** đến **đáy thấp nhất** trong giai đoạn quan sát.
  - Công thức:

    Với $Vpeak$ là giá trị đỉnh trước khi giảm và $Vtrough$ là đáy sau đó:

$$MDD = \frac{V_{peak} - V_{trough}}{V_{peak}} \times 100\%$$

  - Ý nghĩa:
    - Phản ánh **rủi ro thua lỗ cực đoan** và khả năng "bốc hơi" vốn.
    - MDD thấp → danh mục ổn định, nhà đầu tư ít chịu drawdown nặng.
    - Dù CAGR cao nhưng MDD quá sâu (> 30–40 %) sẽ gây khó giữ vị thế.

Kỳ vọng và đóng góp

Kỳ vọng khi áp dụng phương pháp

1. **Lợi nhuận điều chỉnh rủi ro tốt hơn (Sharpe cao hơn, MDD thấp hơn)**
    a. Việc **đưa tín hiệu sentiment từ tin tức** vào pipeline dự báo thường giúp cải thiện độ chính xác và chất lượng phân bổ, vì cảm xúc thị trường mang thông tin bổ sung so với giá/khối lượng thuần tuý. Nghiên cứu gần đây cho thấy **headline news** và embedding sentiment giúp mô hình học sâu dự báo tốt hơn, từ đó cải thiện hiệu quả danh mục. MDPI+1

    b. Đặc biệt với **ngữ cảnh tiếng Việt**, **PhoBERT** đã được chứng minh hiệu quả trong phân loại cảm xúc tin tức tài chính Việt Nam, hỗ trợ việc xây dựng đặc trưng sentiment tin cậy cho thị trường nội địa. CEUR-WS+3MDPI+3Preprints+3
2. **Ổn định qua nhiều kịch bản thị trường (giảm "ăn may")**
    a. **Monte Carlo** trên nền **block bootstrap đa biến** giúp phản ánh tương quan chéo và động lực ngắn hạn, từ đó đánh giá trọng số trên **hàng vạn kịch bản** thay vì một đường lịch sử duy nhất — kết quả tối ưu vì thế **bền vững** hơn. ScienceDirect+1

    b. Cách chọn trọng số theo **trung vị/phân vị** của Sharpe/MDD (thay vì kỳ vọng đơn thuần) giúp giảm nhạy với đuôi rủi ro, phù hợp thị trường mới nổi như Việt Nam, nơi cú sốc và tin tức đột biến xảy ra không hiếm. (Sử dụng block bootstrap đúng cách cũng được cảnh báo là quan trọng để không đánh giá thấp rủi ro). Taylor & Francis Online
3. **Nhạy với điều kiện vĩ mô Việt Nam**
    a. Việc đưa **lãi suất liên NH, CPI YoY, USD/VND** vào tập đặc trưng phù hợp bằng chứng thực nghiệm: các biến vĩ mô này có liên hệ đáng kể với biến động thị trường Việt Nam và phản ứng với thông tin quốc tế (ví dụ tác động của chính sách tiền tệ Mỹ). Scholar Publishing+1
4. **Khả năng triển khai thực tế**
    a. Pipeline Monte Carlo để tối đa hóa Sharpe đã được dùng trong thực hành và nghiên cứu ứng dụng; dự án của bạn mở rộng theo hướng **tích hợp sâu sentiment + bootstrap theo block** cho Việt Nam và tối ưu hoá **ba giai đoạn** (quét thô → khai phá trọng tâm → tinh chỉnh cục bộ). ResearchGate+2ijcaonline.org+2

Điểm mới / Đóng góp của dự án (so với các nghiên cứu & sản phẩm trước)

1. **"End-to-end Việt hóa" kết hợp ba lớp tín hiệu (kỹ thuật + vĩ mô + sentiment) trong một pipeline tối ưu danh mục theo kịch bản:**
    a. Phần lớn nghiên cứu quốc tế chứng minh lợi ích của sentiment đối với dự báo/lựa chọn tài sản; tuy nhiên **ứng dụng end-to-end cho Việt Nam** với **PhoBERT** (tiền xử lý, gắn nhãn, ánh xạ theo cửa sổ ảnh hưởng 3 ngày, quy tắc 15h…) rồi **đưa thẳng vào tối ưu danh mục** theo Monte Carlo **chưa phổ biến** trong các bài báo học thuật về Việt Nam. Các công trình về sentiment tiếng Việt thường dừng ở mô hình

phân loại hoặc dự báo giá đơn mã, ít công trình nối liền đến **tối ưu phân bổ danh mục**. MDPI+2Preprints+2

2. **Khung Monte Carlo "3-stage" có kiểm soát rủi ro đuôi bằng thống kê bền vững**
    a. Trong literature, Monte Carlo thường dùng để tìm **efficient frontier** hoặc **max Sharpe** với **một lớp sampling**; dự án của bạn **tổ chức 3 giai đoạn**:
        i. **Quét thô** (lưới 5%, ràng buộc trần),
        ii. **Khai phá trọng tâm** (đám mây lân cận / giữ đa dạng),
        iii. **Tinh chỉnh cục bộ** (không đạo hàm, đánh giá trên 10k kịch bản).
        Từng giai đoạn đều xếp hạng theo **trung vị/phân vị**, nhằm **giảm lệch do kịch bản cực trị** — đây là một thiết kế **thực dụng nhưng mới mẻ** so với các bài Monte Carlo "một nhát" kiểu cổ điển. ResearchGate+1
3. **Sinh kịch bản bằng block bootstrap đa biến, "tilt" theo xác suất Long từ mô hình học sâu**
    a. Khác với sampling giả định phân phối chuẩn/t-đa biến, **block bootstrap** giữ động học ngắn hạn và tương quan thực tế; sau đó "nghiêng" drift theo xác suất Long đã **hiệu chuẩn** (từ validation) — cách kết hợp này **gần dữ liệu thực hơn** và gắn liền trực tiếp với đầu ra của mô hình dự báo. Việc nhấn mạnh chọn **block size** phù hợp và cảnh báo sai lệch rủi ro khi dùng bootstrap không chuẩn cũng là điểm cần trọng khoa học. ScienceDirect+2Federal Reserve+2
4. **Ngữ cảnh dữ liệu & pháp lý Việt Nam + Chatbot tra cứu (RAG) hỗ trợ nhà đầu tư**
    a. Bên cạnh mô hình danh mục, **chatbot RAG** dựa trên kho dữ liệu luật, thông tư, Q&A tiếng Việt giúp **diễn giải quy định/thuật ngữ** một cách minh bạch nguồn — đây là phần **chuyển giao ứng dụng** hiếm thấy trong các bài học thuật thuần túy, nhưng rất quan trọng về tác động thực tế đối với **nhà đầu tư mới** trên thị trường Việt Nam.

So sánh ngắn với các hướng liên quan (2020–nay)

1. **Monte Carlo tối ưu danh mục (quốc tế):** nhiều bài dùng mô phỏng để tìm trọng số max Sharpe hoặc min-variance, nhưng thường **không** kết hợp **(i) sentiment tin tức tiếng Việt**, **(ii) bootstrap theo block đa biến**, **(iii) tối ưu 3-stage với tiêu chí phân vị** như dự án này. ResearchGate+2ijcaonline.org+2
2. **Sentiment & dự báo/lợi suất (quốc tế):** các nghiên cứu gần đây khẳng định **news sentiment** (kể cả ESG sentiment) cải thiện dự báo/hiệu quả — dự án

của bạn **nội địa hóa** bằng PhoBERT và data CafeF, rồi **đưa vào lớp tối ưu danh mục**, thay vì dừng ở dự báo đơn mã. MDPI+1

3. **Sentiment tiếng Việt:** đã có các thử nghiệm phân loại với **PhoBERT** và mô hình lai CNN/LSTM cho tin chứng khoán Việt, nhưng chủ yếu dừng ở **bài toán NLP**; chuỗi **NLP → tín hiệu sentiment → Monte Carlo portfolio** cho **25 mã HOSE** là **điểm khác biệt** của bạn. MDPI+2Preprints+2
4. **Vĩ mô Việt Nam:** literature xác nhận vai trò của lãi suất, tỷ giá, lạm phát, và tác động chính sách quốc tế với TTCK Việt Nam; việc **chuẩn hóa lịch cập nhật** (tuần/tháng) và **ánh xạ vào dữ liệu phiên** của bạn khiến mô hình **phù hợp thực địa** hơn là dùng chuỗi vĩ mô "thô". Scholar Publishing+1

Kết luận

Dự án đã xây dựng thành công một mô hình hỗ trợ đầu tư danh mục cổ phiếu trên thị trường chứng khoán Việt Nam dựa trên các kỹ thuật học sâu tiên tiến (LSTM, GRU, Transformer) kết hợp với dữ liệu đa chiều (kỹ thuật, vĩ mô, tâm lý). Mô hình có khả năng **dự báo tín hiệu Long/Neutral** khá chính xác cho nhóm cổ phiếu thanh khoản cao, qua đó hình thành chiến lược danh mục chọn lọc 5 cổ phiếu mạnh nhất và tái cân bằng định kỳ 20 phiên. Kết quả thử nghiệm cho thấy chiến lược dựa trên mô hình mang lại **hiệu quả đầu tư vượt trội**, với CAGR cao, Sharpe ratio ấn tượng và kiểm soát drawdown tốt hơn hẳn so với chiến lược thụ động mua và nắm giữ thị trường. Đây là minh chứng rõ nét cho lợi ích của cách tiếp cận **đầu tư dựa trên khoa học dữ liệu**: thay vì quyết định cảm tính, nhà đầu tư có thể dựa vào mô hình để **lựa chọn cổ phiếu một cách khách quan, hệ thống**.

Về **đóng góp khoa học**, dự án này là một trong những nghiên cứu đầu tiên áp dụng thành công **mô hình học sâu kết hợp phân tích cảm xúc** để xây dựng danh mục ở thị trường Việt Nam. Trước đây, nhiều nghiên cứu tập trung dự báo chỉ số hoặc giá cổ phiếu riêng lẻ, ít công trình nào tích hợp đầy đủ yếu tố kỹ thuật, vĩ mô, tin tức vào chiến lược danh mục. Chúng tôi cũng xây dựng được **bộ dữ liệu tin tức đồ sộ** và áp dụng mô hình xử lý ngôn ngữ (PhoBERT) để định lượng cảm xúc thị trường – mở ra hướng khai thác **AI ngôn ngữ** trong đầu tư tài chính tại Việt Nam. Kết quả cho thấy việc bổ sung yếu tố tâm lý (news sentiment) giúp cải thiện độ chính xác dự báo, phù hợp với khuyến nghị của các nghiên cứu gần đâykinhtevadubao.vnmdpi.com rằng cảm xúc nhà đầu tư là nguồn dữ liệu quan trọng thường bị bỏ sót. Bên cạnh đó, việc thử nghiệm song song LSTM, GRU, TransformerEncoder cũng đem lại cái nhìn so sánh giữa các mô hình – khẳng định xu hướng mới rằng mô hình tự chú ý (Transformer) có thể thay thế và cải thiện hiệu suất so với RNN truyền thống trong bài toán tài chính.

Tuy đạt kết quả khả quan, dự án vẫn có những **hạn chế** và gợi ý phát triển trong tương lai. Thứ nhất, mô hình hiện chưa tính đến **phí giao dịch** và **thuế**, nên hiệu suất thực tế có thể thấp hơn chút sau khi trừ chi phí – nghiên cứu trong tương lai có thể tích hợp chi phí để tối ưu thực tế hơn. Thứ hai, chúng tôi chưa áp dụng các biện pháp **kiểm soát rủi ro chủ động** (như cắt lỗ, giới hạn trọng số theo ngành...), do tập trung đánh giá tiềm năng mô hình; việc bổ sung các quy tắc quản trị rủi ro có thể giúp danh mục an toàn hơn nữa. Thứ ba, có thể mở rộng phạm vi mô hình sang nhiều cổ phiếu hơn hoặc thử nghiệm trên các thị trường khác (ví dụ sàn HNX, UpCOM hoặc thị trường quốc tế) để kiểm chứng tính tổng quát. Cuối

cùng, một hướng thú vị là kết hợp mô hình dự báo này với các thuật toán **tối ưu danh mục hiện đại** (ví dụ Markowitz, hoặc Reinforcement Learning) để xem liệu có thể cải thiện thêm phân bổ trọng số thay vì dùng equal-weight.

Tóm lại, **mô hình danh mục đầu tư dùng LSTM/GRU/Transformer** của chúng tôi đã cho thấy tiềm năng lớn trong việc **nâng cao hiệu quả đầu tư chứng khoán tại Việt Nam** một cách khoa học và có hệ thống. Kết quả tích cực này hy vọng sẽ góp phần thay đổi tư duy đầu tư từ "mang tính cảm xúc" sang **dựa trên dữ liệu và mô hình**. Nhà đầu tư cá nhân có thể ứng dụng các kỹ thuật tương tự – từ việc theo dõi tin tức, chỉ số vĩ mô đến sử dụng mô hình AI – để hỗ trợ ra quyết định sáng suốt hơn. Trong bối cảnh TTCK Việt Nam ngày càng phát triển và thu hút sự quan tâm, những phương pháp đầu tư định lượng như trên sẽ là công cụ hữu ích để cạnh tranh và quản lý rủi ro, hướng tới **mục tiêu tối ưu lợi nhuận bền vững**.

2. Chatbots thị trường chứng khoán Việt

Introduction

Sự phát triển nhanh chóng của thị trường chứng khoán Việt Nam trong những năm gần đây đã thu hút hơn 16% dân số tham gia đầu tư, nhưng đa phần nhà đầu tư cá nhân còn hạn chế về kiến thức tài chính, dễ bị ảnh hưởng bởi tin đồn hoặc đầu tư theo cảm tính. Bên cạnh đó, thông tin về thuật ngữ chuyên ngành, quy trình giao dịch, phí, thuế và khung pháp lý thường rải rác ở nhiều nguồn khác nhau, gây khó khăn cho việc tra cứu. Điều này đặt ra nhu cầu cấp thiết về một công cụ hỗ trợ thông tin đáng tin cậy, dễ sử dụng và phù hợp với đặc thù của thị trường Việt Nam.

Dự án **Chatbot Q/A Chứng khoán Việt Nam** được thiết kế như một **trợ lý hỏi–đáp bằng tiếng Việt**, cung cấp kiến thức cơ bản, số liệu lịch sử và các quy định quan trọng của thị trường chứng khoán. Khác với các hệ thống phân tích tài chính hay dự đoán giá cổ phiếu, chatbot này **không thực hiện dự báo giá, không đưa ra khuyến nghị mua/bán**. Mục tiêu của hệ thống là trả lời chính xác và dễ hiểu các câu hỏi khái niệm, quy trình, số liệu và luật lệ, hỗ trợ nhà đầu tư mới tiếp cận thị trường một cách khoa học và an toàn.

Kiến trúc của chatbot gồm ba thành phần chính: **(1) mô hình nền Qwen-0.5B-Instruct**, một mô hình ngôn ngữ nhẹ, hiệu quả với tiếng Việt; **(2) cơ chế truy xuất thông tin (Retrieval-Augmented Generation – RAG) mỏng**, sử dụng kết hợp BM25 và embedding Dense để tìm các đoạn văn bản liên quan từ kho dữ liệu; **(3) fine-tune nhẹ bằng phương pháp **LoRA** nhằm điều chỉnh phong cách trả lời theo định dạng thống nhất: *Tóm tắt → Giải thích → (nếu có) Rủi ro → Nguồn*. Kho dữ liệu được tổ chức dạng tĩnh, không cập nhật hằng ngày, giúp hệ thống vận hành nhanh và ổn định.

Chatbot làm việc với **kho dữ liệu tĩnh** gồm:

- **~7.1k QA đã dịch sang tiếng việt** (khoảng 7.17k theo trang HF), chủ đề xoay quanh:
    - Khái niệm & chiến lược giao dịch phổ biến (price action, trend, breakout…). Hugging Face+1
    - **Chỉ báo kỹ thuật** (ví dụ: OBV – on-balance volume; divergence, tín hiệu đảo chiều). Đoạn xem trước trên HF cho thấy Q&A dạng "How does OBV help…?" kèm câu trả lời giải thích. Hugging Face
    - Quản trị rủi ro, cách đọc hỗ trợ/kháng cự, cấu trúc thị trường… (mang tính hướng dẫn, diễn giải khái niệm). Hugging Face
    - Được cộng đồng khác dùng để huấn luyện chatbot Q&A về chứng khoán (ví dụ model T5 đã "trained on the stock_trading_QA dataset"). Hugging Face
    - Cập nhật lần cuối vào đầu tháng 3/2024 (mốc trên trang tác giả). Hugging Face
- **Q/A tiếng Việt:** 350 câu Vietstock
    - Bộ 350 QA Vietstock là tập câu hỏi-trả lời tiếng Việt về **khái niệm tài chính, thị trường và sự kiện chứng khoán**, rất phù hợp làm nguồn RAG cho chatbot để giải thích thuật ngữ và cơ chế thị trường Việt Nam cho người mới.
- **Dữ liệu thị trường lịch sử (2015-2025):**
    - Giá OHLCV & chỉ báo kỹ thuật (volatility,volume,return,rsi).
    - Các **chỉ số vĩ mô đa dạng gồm 3 loại:** lãi suất liên NH overnight , USD/VND, lạm phát (CPI YoY).

Lưu ý: các data về OHLCV và chỉ báo kĩ thuật chỉ bao gồm của 25 cổ phiếu lớn nhất trên sàn HOSE. Toàn bộ data về phần dữ liệu thị trường lịch sử từ (2015-2025) có cách chuẩn bị và xử lí giống như phần portfolio model.

- **Quy trình & khung pháp lý:**
    - Luật CK 2019, Luật sửa đổi 2024.
    - NĐ 155/2020
    - Thông tư 96/2020/TT-BTC: công bố thông tin trên TTCK (rất hay được hỏi).
    - Thông tư 120/2020/TT-BTC: giao dịch trên thị trường chứng khoán.
    - Thông tư 119/2020/TT-BTC: về quản trị công ty đại chúng.

- Thông tư 118/2020/TT-BTC: hướng dẫn một số nội dung về chào bán, phát hành chứng khoán, chào mua công khai, mua lại cổ phiếu, đăng ký công ty đại chúng và hủy tư cách công ty đại chúng.
- Thông tư 101/2021/TT-BTC: quy định giao dịch ký quỹ (margin).
- Quy chế của Sở Giao dịch HOSE/HNX: quy định về lệnh giao dịch (LO, ATO, ATC, MP…), giờ giao dịch, biên độ giá.
- Quy định xử phạt vi phạm hành chính (Nghị định 156/2020/NĐ-CP, sửa đổi 128/2021/NĐ-CP): cần thiết cho các câu hỏi về mức phạt

Các data dạng văn bản về luật, nghị định, thông tư được chuyển thành data có cấu trúc dạng csv với 3 file csv lần lượt về luật, nghị định, thông tư. Mỗi file có cùng cấu trúc là 7 cột chính bao gồm: Điều,Chủ đề,Khoản,Nội dung,Tóm tắt,Thời gian,Bộ luật

Về cách tóm tắt nội dung:Kết hợp bán-tự động (AI + rà soát)

Dùng mô hình LLM (ví dụ GPT-4 hoặc Qwen-72B) tạo bản tóm tắt sơ bộ cho từng điều/khoản theo định dạng bullet.

Xuất ra file CSV với 2 cột: Tóm_tắt_gợi_ý và Tóm_tắt_chính_thức.

Bạn hoặc cộng tác viên sẽ duyệt lại, sửa câu sai, thêm ý bị thiếu.

👉 Đây là cách thực tế nhất: nhanh nhưng vẫn đảm bảo chất lượng pháp lý.

Nhờ đó, hệ thống trở thành một **trợ lý tra cứu đáng tin cậy** cho người dùng, hỗ trợ nâng cao hiểu biết và giúp việc đầu tư trở nên khoa học hơn.”