NGHIỆM THU ĐỀ TÀI NCKHSV CẤP TRƯỜNG

Xây dựng ứng dụng hỗ trợ đầu tư chứng khoán bằng trí tuệ nhân tạo và phân tích cơ bản

Thực hiện bởi:

Nguyen Quang Huy - 523H0140

Nguyen Tran Nhat An - 523H0115

Nguyen Thanh Tung - 523H0192

GVHD: TS. Trịnh Hùng Cường

NỘI DUNG THUYẾT TRÌNH

Chương 1. Mở đầu

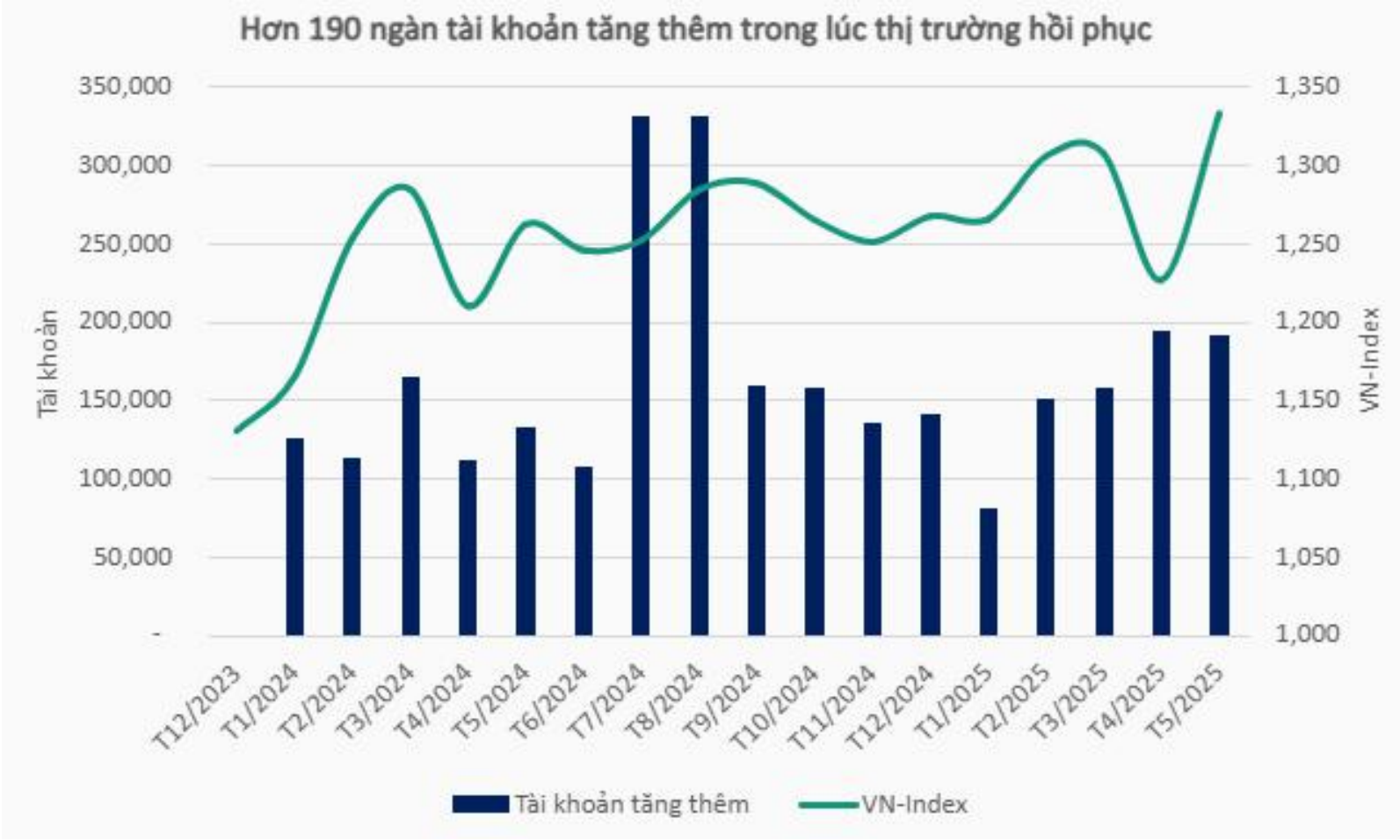
Chương 2. Cơ sở lý thuyết

Chương 3. Phương pháp đề xuất

Chương 4. Kết qủa thực nghiệm

Chương 1
Mở đầu

1.1 Đặt vấn đề và lý do chọn đề tài



Số lượng tài khoản giao dịch được đăng ký và chỉ số VN-Index hàng tháng

1.1 Đặt vấn đề và lý do chọn đề tài

- Thị trường chứng khoán là kênh đầu tư hấp dẫn nhưng tiềm ẩn nhiều rủi ro, đặc biệt trong bối cảnh kinh tế toàn cầu biến động. Nhà đầu tư cá nhân thường thiếu kinh nghiệm và công cụ phân tích hiệu quả, dẫn đến quyết định mang tính cảm tính.
- Tại Việt Nam, thị trường chứng khoán đang phát triển mạnh với hàng trăm nghìn tài khoản mới mỗi tháng, nhưng sự hiểu biết của nhà đầu tư còn hạn chế.
- Trí tuệ nhân tạo (AI) nổi lên như một giải pháp hỗ trợ xử lý dữ liệu lớn, nâng cao độ chính xác trong phân tích và dự báo thị trường. Do đó, việc xây dựng ứng dụng hỗ trợ đầu tư chứng khoán kết hợp AI và phân tích cơ bản là cần thiết và mang tính thực tiễn cao.

1.1 Đặt vấn đề và lý do chọn đề tài

- Hiện nay, ứng dụng trí tuệ nhân tạo trong đầu tư tài chính ngày càng phổ biến, nhưng tại Việt Nam vẫn còn nhiều hạn chế, đặc biệt là chưa kết hợp hiệu quả giữa phân tích kỹ thuật và phân tích cơ bản.
- Đề tài hướng đến việc tích hợp hai phương pháp này thông qua các mô hình AI, nhằm hỗ trợ nhà đầu tư ra quyết định chính xác hơn, giảm rủi ro và tối ưu hóa lợi nhuận.

1.2 Đối tượng và phạm vi nghiên cứu

- Nghiên cứu các mô hình trí tuệ nhân tạo (AI) ứng dụng trong dự báo và phân tích xu hướng thị trường chứng khoán.
- Kết hợp phân tích cơ bản, bao gồm:
  - Phân tích các bài báo tài chính, nhận định chuyên gia, báo cáo thị trường.
- Phạm vi nghiên cứu: Thị trường chứng khoán Việt Nam. Tập trung vào các mã thuộc VNINDEX và HNXINDEX giai đoạn từ 2011 – 2025.

1.3 Những mô hình AI hiện có và hạn chế

- Logistic Regression:
  - Ưu điểm: Đơn giản, dễ triển khai, giúp xác định yếu tố ảnh hưởng đến biến mục tiêu.
  - Hạn chế: Khó xử lý các mối quan hệ phi tuyến, hiệu quả kém trên dữ liệu tài chính phức tạp.
- LSTM (Long Short-Term Memory):
  - Ưu điểm: Ghi nhớ dài hạn, phù hợp với dữ liệu chuỗi thời gian và biến động tài chính.
  - Hạn chế: Cần nhiều dữ liệu, khó huấn luyện và tốn tài nguyên tính toán

1.3 Những mô hình AI hiện có và hạn chế

- Hiểu biết bối cảnh thị trường còn hạn chế:
  - Các mô hình truyền thống khó tích hợp dữ liệu định tính như tin tức, chính sách, cảm xúc nhà đầu tư.
  - Hướng cải tiến: Kết hợp dữ liệu định tính từ tin tức, báo cáo tài chính, và phân tích tâm lý.
- Tính ổn định và nhất quán chưa cao:
  - Mô hình phản ứng chậm với biến động bất thường, gây khó khăn trong thực thi chiến lược đầu tư.
  - Hướng cải tiến: Áp dụng mô hình ensemble, kết hợp nhiều thuật toán để tăng độ chính xác và ổn định.

1.4 Phương pháp nghiên cứu

- Thu thập và xử lý dữ liệu:
  - Thu thập dữ liệu lịch sử giá cổ phiếu từ các sàn HOSE, HNX, SSI Corporation và thư viện vnstock.
  - Tiến hành làm sạch, chuẩn hóa và xử lý dữ liệu để đảm bảo tính chính xác, đầy đủ và nhất quán.
- Xây dựng mô hình học máy:
  - Áp dụng các mô hình Logistic Regression và LSTM để phân tích, dự báo xu hướng giá cổ phiếu dựa trên dữ liệu lịch sử, tài chính và vĩ mô.
  - Sử dụng ngôn ngữ lập trình Python với các thư viện: TensorFlow, PyTorch, scikit-learn, pandas và numpy.

1.4 Phương pháp nghiên cứu

- Đánh giá mô hình:
  - Sử dụng các chỉ số đánh giá: accuracy, precision, recall, F1-score và Mean Squared Error (MSE).
  - Kiểm thử mô hình trên dữ liệu thực tế và so sánh với hiệu quả đầu tư thực tiễn để xác định tính ứng dụng.

Chương 2
Cơ sở lý thuyết

2.1.1 Logistic Regression và TF-IDF trong phân tích cảm xúc văn bản tài chính

- Bài toán cảm xúc tài chính:
  - Phân loại văn bản tài chính theo cảm xúc: tích cực, tiêu cực, trung tính.
- Biểu diễn văn bản với TF-IDF:
  - TF đo tần suất từ trong văn bản: $TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$
  - IDF giảm trọng số của từ phổ biến trong toàn bộ tập văn bản: $IDF(t) = \log\left(\frac{N}{n_t}\right)$
  - TF-IDF giúp mô hình nhận diện từ khóa quan trọng, giảm nhiễu.

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

2.1.1 Logistic Regression và TF-IDF trong phân tích cảm xúc văn bản tài chính

- Logistic Regression:
  - Mô hình phân loại nhị phân sử dụng hàm sigmoid.

$$P(y=1|x) = \frac{1}{1+e^{-z}} \quad \text{trong đó } z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

  - Có thể áp dụng các hình thức regularization như L1, L2, Elastic Net.
  - Ưu điểm: dễ triển khai, kết quả dễ giải thích, nhanh.
- Vai trò trong hệ thống:
  - Cung cấp tín hiệu cảm xúc từ tin tức, bổ trợ mô hình chính.
  - Phản ánh tâm lý nhà đầu tư theo thời gian thực.

2.1.2 LSTM

- Tổng quan:
  - LSTM là biến thể của RNN, giải quyết vấn đề vanishing gradient.
  - Sử dụng các cổng: Forget, Input, Output để điều khiển dòng thông tin.
- Cấu trúc cell LSTM:
  - Nhận đầu vào xt, trạng thái ẩn ht-1 và bộ nhớ Ct-1.
  - Cập nhật bộ nhớ và trạng thái mới tại mỗi bước thời gian.

2.1.2 LSTM

- Ưu điểm:
  - Nhớ được thông tin dài hạn trong chuỗi thời gian.
  - Xử lý được chuỗi không đều, không cần chu kỳ rõ ràng.
  - Phù hợp với dữ liệu phi tuyến và nhiều nhiễu.
- Nhược điểm:
  - Thời gian huấn luyện dài.
  - Dễ overfitting nếu dữ liệu hạn chế.
  - Khó giải thích hơn mô hình tuyến tính.

Chương 3
Phương pháp đề xuất

3.1 Kiến trúc tổng thể

- Gồm hai thành phần chính:
  - Mô hình phân loại cảm xúc từ tin tức: TF-IDF + Logistic Regression.
  - Mô hình dự đoán giá cổ phiếu: LSTM dựa trên dữ liệu giá và chỉ số cảm xúc.
- Quy trình:
  - Tiền xử lý dữ liệu giá và tin tức.
  - Huấn luyện mô hình cảm xúc.
  - Gộp dữ liệu cảm xúc với dữ liệu giá.
  - Huấn luyện mô hình LSTM trên tập dữ liệu kết hợp.

3.2 Tiền xử lý dữ liệu

- Giá cổ phiếu
  - Thuộc tính: Date, Open, High, Low, Close, Volume.
  - Làm sạch, chuẩn hóa (MinMaxScaler), tính MA và EMA.
  - Tách tập train/test theo thứ tự thời gian (80/20).
- Tin tức tài chính:
  - Nguồn train: Financial PhraseBank → dịch và kiểm tra chuyên ngành.
  - Nguồn thực tế: CafeF, Vietstock, Bloomberg...
  - Kết quả đầu ra: 3 cột tỷ lệ cảm xúc (positive, neutral, negative).
  - Xử lý văn bản: lowercase, stop-word removal, giữ bigram/trigram tài chính.

3.3 Mô hình phân loại cảm xúc

3.3.1 TF-IDF:

- Biểu diễn văn bản bằng tần suất và độ đặc trưng của từ.

3.3.2 Logistic Regression đa lớp:

- Sử dụng softmax và hàm mất mát cross-entropy.
- Tham số: C=1.0, penalty=L2, class_weight=balanced.

3.3.3 Chỉ số cảm xúc theo ngày:

- Tính tỷ lệ cảm xúc (S_pos, S_neu, S_neg) làm đặc trưng đầu vào cho mô hình giá.

3.3 Mô hình dự đoán giá cổ phiếu LSTM

3.4.1 Thiết lập chuỗi dữ liệu:

- Window size: 60 ngày.
- Đặc trưng đầu vào: [O, H, L, C, V, MA/EMA, S_pos, S_neu, S_neg].

3.4.2 Kiến trúc mạng:

- LSTM(32) → Dropout(0.2) → Dense(1).
- Loss: Mean Squared Error.
- Optimizer: Adam (lr=0.001).

3.4.3 Chiến lược huấn luyện:

- batch_size: 32, epochs: 20 (early stopping = 3).
- shuffle: False.
- validation: walk-forward (mỗi 10 ngày cập nhật mô hình).

Chương 4
Kết quả thực nghiệm

4.1 Mô tả tập dữ liệu

- Giá cổ phiếu (VCB):
  - Thời gian: 01/01/2020 – 31/12/2023
  - Trường dữ liệu: Date, Open, High, Low, Close, Volume
  - Tiền xử lý: loại trùng/lỗi, điều chỉnh chia tách/cổ tức, chuẩn hóa Min-Max, chia train/test 80:20
- Tin tức và cảm xúc:
  - Nguồn train: Financial PhraseBank (dịch + hậu kiểm)
  - Nguồn kiểm thử: CafeF, Bloomberg, Reuters (crawler 30 phút/lần)
  - Xử lý: lowercase → stop-word removal → TF-IDF → Logistic Regression

4.2 Thiết lập thực nghiệm và kết quả đầu ra

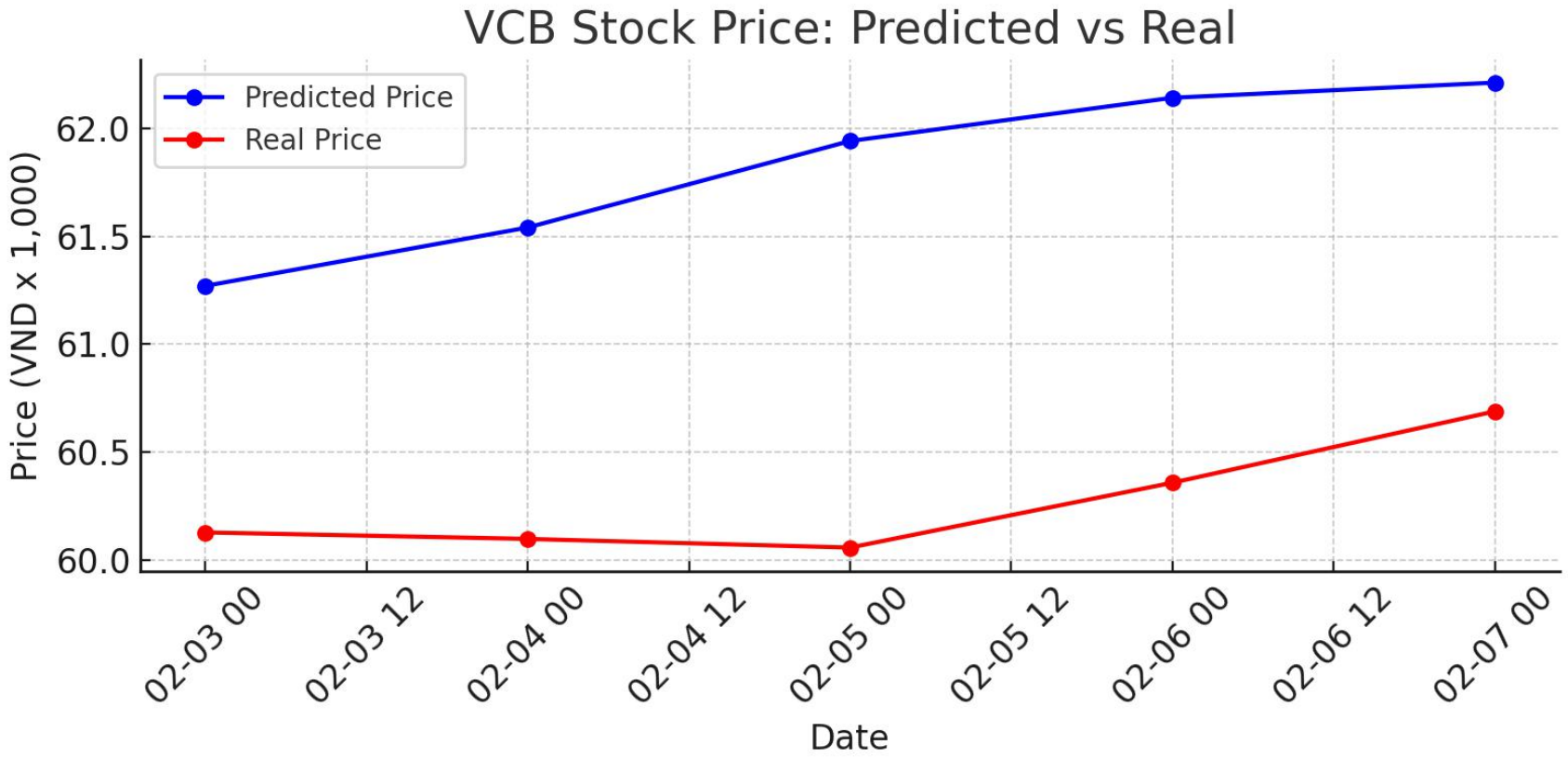
- **Mô hình cảm xúc:** TF-IDF (40k từ) + Logistic Regression (C=1, L2, balanced)
- Kết quả mô hình cảm xúc
  - Accuracy: 0.763
  - Precision (macro): 0.835
  - Recall (macro): 0.622
  - F1-macro: 0.673

	Pred\,0	1	2
Actual\,0	561	9	1
1	153	132	4
2	42	11	47

- Nhận xét:
  - Recall lớp tiêu cực đạt 98%
  - Lớp tích cực & trung tính thấp → cần cải thiện bằng class weight/SMOTE

4.2 Thiết lập thực nghiệm và kết quả đầu ra

- **Mô hình LSTM:** window = 60, LSTM 32, Dropout 0.2, Dense(1), Adam 0.001, epochs 20, batch 32



4.3 Ứng dụng hỗ trợ đầu tư chứng khoán

- Chức năng chính:
  - Hiển thị biểu đồ giá thực & dự đoán theo thời gian thực
  - Tra cứu cổ phiếu, xem phân tích cơ bản
  - Hệ thống blog tài chính: viết bài, bình luận, chia sẻ
- Công nghệ:
  - MongoDB: lưu cổ phiếu, blog, người dùng
  - Express.js + Node.js: backend/API
  - React.js: frontend, biểu đồ, giao diện blog

4.4 Ý nghĩa nghiên cứu

- Khoa học:
  - Kết hợp Logistic Regression + LSTM trên cùng pipeline
  - Tạo tập dữ liệu nhiều chiều (giá + cảm xúc)
  - Mã nguồn và quy trình mở, dễ tái lập
- Thực tiễn:
  - Công cụ hỗ trợ quyết định đầu tư theo thời gian thực
  - Cảm xúc tin tức phản ánh nhanh tâm lý thị trường
  - Mô hình dễ tích hợp vào hệ thống tài chính (micro-service)

4.5 Định hướng nghiên cứu tiếp theo

- Mở rộng dữ liệu:
  - Thêm nguồn từ mạng xã hội (Twitter, StockTwits, Facebook)
  - Bổ sung dữ liệu vĩ mô: lãi suất, CPI, tỷ giá
- Nâng cấp mô hình cảm xúc:
  - Fine-tune PhoBERT, FinBERT
  - Aspect-based sentiment theo tiêu chí (lợi nhuận, ESG...)
- Cải tiến mô hình giá:
  - Nghiên cứu TFT, N-Beats
  - Ứng dụng GNN để mô hình hóa quan hệ cổ phiếu
- Học tăng cường (RL):
  - DQN, PPO cho chiến lược giao dịch
  - Đánh giá qua paper-trading: Sharpe ratio, max drawdown

Cảm ơn thầy cô đã lắng nghe