TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

---

**Công trình Nghiên cứu khoa học sinh viên năm học 2024 - 2025**

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


**ĐƠN VỊ KHOA CÔNG NGHỆ THÔNG TIN**

**GIẢNG VIÊN HƯỚNG DẪN:**

*TIẾN SĨ TRỊNH HÙNG CƯỜNG*

**NHÓM SINH VIÊN THỰC HIỆN:**

**1.** *NGUYỄN QUANG HUY*

**2.** *NGUYỄN TRẦN NHẬT AN*

**3.** *NGUYỄN THANH TÙNG*


TP. Hồ Chí Minh, tháng 5 năm 2025

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
**TRƯỜNG ĐẠI HỌC TÔNG ĐỨC THẮNG**
**KHOA CÔNG NGHỆ THÔNG TIN**

# CÔNG TRÌNH NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM HỌC 2024-2025

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

*Người hướng dẫn:* **TS. TRỊNH HÙNG CƯỜNG**

*Người thực hiện:* **NGUYỄN QUANG HUY - 523H0140**

**NGUYỄN TRẦN NHẬT AN - 523H0115**

**NGUYỄN THANH TÙNG - 523H0192**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

# LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành nhất đến Ban giám hiệu trường Đại Học Tôn Đức Thắng vì đã tạo điều kiện cho chúng em hoàn thành công trình nghiên cứu khoa học này thông qua hệ thống thư viện đa dạng tài liệu hay và bổ ích.

Chúng em xin chân thành cảm ơn Thầy T.S. Trịnh Hùng Cường đã giảng dạy và truyền đạt kiến thức một cách tận tình, chi tiết đồng thời cung cấp tài liệu tham khảo giúp chúng em đủ nền tảng để vận dụng vào việc viết bài báo cáo này.

Trong quá trình làm bài báo cáo nghiên cứu, việc chúng em khó tránh khỏi thiếu sót là điều chắc chắn, chúng em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để kiến thức của chúng em được hoàn thiện hơn.

Sau cùng em xin chân thành cảm ơn và kính chúc quý thầy cô trong khoa Công Nghệ Thông Tin luôn khỏe mạnh và thành công trong sự nghiệp giảng dạy.

# LỜI CAM KẾT

Chúng em xin cam đoan đây là sản phẩm nghiên cứu của riêng chúng em và được sự hướng dẫn của TS. Trịnh Hùng Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét được chính chúng em thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong nghiên cứu còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng em xin hoàn toàn chịu trách nhiệm về nội dung nghiên cứu của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng em gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 25 tháng 5 năm 2025*
*Tác giả*
*(ký tên và ghi rõ họ tên)*

*Nguyễn Quang Huy*

*Nguyễn Trần Nhật An*

*Nguyễn Thanh Tùng*

# MỤC LỤC

# TÓM TẮT

Đề tài nghiên cứu khoa học này nhằm mục tiêu hỗ trợ nhà đầu tư chứng khoán nâng cao hiệu quả quyết định đầu tư thông qua việc ứng dụng các mô hình trí tuệ nhân tạo (AI) và phân tích cơ bản trong lĩnh vực đầu tư chứng khoán. Việc xây dựng ứng dụng này không chỉ đáp ứng nhu cầu ngày càng tăng về công nghệ trong lĩnh vực đầu tư tài chính, mà còn giúp các nhà đầu tư cá nhân và tổ chức đưa ra quyết định chính xác, nhanh chóng và hiệu quả hơn trong bối cảnh thị trường chứng khoán biến động liên tục.

Đề tài nghiên cứu gồm bốn chương:

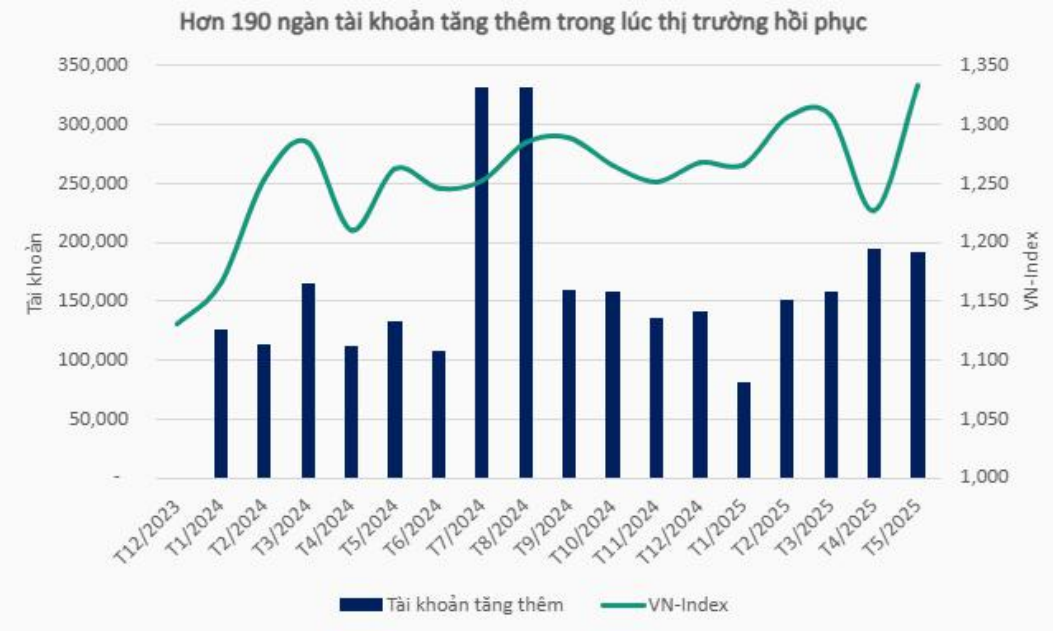
- **Chương 1**: Mở đầu, nội dung bao gồm việc đặt vấn đề, nêu lý do lựa chọn đề tài, sự cần thiết của ứng dụng AI và phân tích cơ bản trong lĩnh vực đầu tư chứng khoán, xác định phạm vi và đối tượng nghiên cứu, đánh giá các giải pháp hiện có và chỉ ra những hạn chế còn tồn tại để làm cơ sở cho đề xuất giải pháp mới.
- **Chương 2**: Cơ sở lý thuyết, trình bày các lý thuyết cơ bản về thị trường chứng khoán, phân tích cơ bản và các phương pháp học máy liên quan như Logistic Regression và LSTM.
- **Chương 3**: Phương pháp đề xuất, nêu rõ phương pháp xây dựng và triển khai mô hình ứng dụng AI hỗ trợ quyết định đầu tư chứng khoán.
- **Chương 4**: Kết quả và định hướng tương lai, tổng hợp các kết quả đạt được từ ứng dụng đã xây dựng, đánh giá hiệu quả của mô hình trong thực tế đầu tư, rút ra các kết luận về ý nghĩa khoa học và giá trị ứng dụng thực tiễn, đồng thời đưa ra các định hướng cải tiến và mở rộng phạm vi nghiên cứu trong tương lai.

# CHƯƠNG 1. MỞ ĐẦU

## 1.1 Đặt vấn đề

Thị trường chứng khoán luôn được coi là một trong những kênh đầu tư hấp dẫn nhưng cũng tiềm ẩn rất nhiều rủi ro. Trong bối cảnh kinh tế toàn cầu biến động phức tạp và khó lường như hiện nay, việc đầu tư vào thị trường chứng khoán ngày càng đòi hỏi nhà đầu tư phải có kỹ năng phân tích và dự đoán với độ chính xác cao để đưa ra những quyết định đúng đắn. Tuy nhiên, thực tế cho thấy phần lớn các nhà đầu tư cá nhân vẫn gặp rất nhiều khó khăn do thiếu kinh nghiệm, thông tin và khả năng phân tích kỹ thuật cũng như phân tích cơ bản về thị trường.

Bên cạnh đó, thị trường chứng khoán Việt Nam đang trong giai đoạn phát triển mạnh mẽ, số lượng nhà đầu tư mới tham gia ngày càng tăng. Dựa vào số liệu về số lượng tài khoản giao dịch chứng khoán của nhà đầu tư trong nước và ngoài nước của Tổng công ty Lưu ký và Bù trừ chứng khoán Việt Nam (VSDC) cung cấp, hàng tháng có từ 90,000 ngàn tài khoản cho tới 330,000 tài khoản được đăng ký, bên cạnh đó VN-Index cũng tăng từ 1,000 điểm lên gần 1,300 điểm với khoảng 1 tỷ Đô được giao dịch với mỗi phiên.




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

Tuy nhiên, sự hiểu biết của họ về thị trường thường hạn chế, dẫn đến các quyết định đầu tư cảm tính, thiếu cơ sở khoa học. Điều này không chỉ ảnh hưởng tới kết quả đầu tư mà còn có thể gây ra những biến động khó kiểm soát cho toàn thị trường.

Trong bối cảnh đó, trí tuệ nhân tạo (AI) nổi lên như một công cụ hữu ích, có khả năng xử lý lượng dữ liệu lớn, phức tạp và dự đoán xu hướng thị trường với độ chính xác cao hơn so với các phương pháp truyền thống. Việc tích hợp AI vào các hoạt động phân tích chứng khoán giúp nhà đầu tư mới có thể tiếp cận dễ dàng hơn với thị trường, giảm thiểu rủi ro và nâng cao hiệu quả đầu tư. Từ những lý do trên, nghiên cứu về việc xây dựng ứng dụng hỗ trợ đầu tư chứng khoán sử dụng AI kết hợp với phân tích cơ bản trở thành vấn đề cấp thiết và có ý nghĩa thực tiễn rất lớn.

## 1.2 Lý do chọn đề tài

Hiện nay, các ứng dụng trí tuệ nhân tạo trong đầu tư tài chính ngày càng được quan tâm, tuy nhiên việc ứng dụng trí tuệ nhân tạo trong đầu tư chứng khoán tại Việt Nam vẫn còn nhiều hạn chế. Đa phần các công cụ chỉ tập trung vào phân tích kỹ thuật mà chưa khai thác hiệu quả các phương pháp phân tích cơ bản. Xuất phát từ nhu cầu thực tế này, đề tài hướng tới việc kết hợp hài hòa giữa phân tích kỹ thuật và phân tích cơ bản thông qua việc sử dụng các mô hình trí tuệ nhân tạo, giúp các nhà đầu tư đưa ra quyết định đầu tư hợp lý, giảm thiểu rủi ro và gia tăng lợi nhuận.

## 1.3 Đối tượng nghiên cứu

Trong đề tài này, đối tượng nghiên cứu bao gồm:

- Các mô hình trí tuệ nhân tạo đang được áp dụng trong dự báo và phân tích xu hướng thị trường chứng khoán.
- Phân tích các bài báo, báo cáo tài chính và nhận định từ các chuyên gia ảnh hưởng tới xu hướng đầu tư của các nhà đầu tư.

## 1.4 Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào thị trường chứng khoán Việt Nam, đặc biệt là các mã cổ phiếu trong nhóm VNINDEX và HNXINDEX từ giai đoạn năm 2011 đến

năm thời điểm hiện tại (2025), nhằm đảm bảo tính cập nhật và khả năng ứng dụng thực tế của các mô hình dự báo.

## 1.5 Những mô hình AI hiện nay

### *1.5.1 Logistic Regression*

Logistic Regression dựa trên giả thuyết rằng mối quan hệ tuyến tính giữa các đặc trưng đầu vào và logit của biến. Ưu điểm nổi bật của mô hình này là sự đơn giản trong việc thể hiện trọng số của từng yếu tố đầu vào, giúp người dùng dễ dàng nhận biết các yếu tố nào tác động mạnh đến dự báo. Tuy nhiên, hạn chế lớn của Logistic Regression là khả năng mô hình hóa các mối quan hệ phi tuyến kém, đặc biệt khi áp dụng vào dữ liệu tài chính vốn rất đa dạng và phức tạp. Vì thế, Logistic Regression thường được sử dụng như một bước sàng lọc ban đầu hoặc bổ trợ cho các mô hình khác để khai thác tối đa ưu thế của từng phương pháp.

### *1.5.2 LSTM (Long Short-Term Memory)*

LSTM là phiên bản cải tiến của RNN, được xây dựng để khắc phục hạn chế "vanishing gradient" của các mô hình hồi quy thông thường. Với khả năng ghi nhớ thông tin dài hạn qua các input, output, và forget gates, LSTM cho phép xử lý tốt các chuỗi dữ liệu phức tạp, dữ liệu tài chính vốn có tính bất động và thay đổi theo thời gian. Các mô hình LSTM thường làm việc tốt khi dự báo những biến động nhỏ ngẫu nhiên, từ đó phát hiện ra các quy luật ẩn không thể nhận ra được thông qua các phương pháp tuyến tính. Mặc dù mạnh mẽ, việc huấn luyện mô hình LSTM đòi hỏi lượng dữ liệu lớn và sự điều chỉnh tinh vi các tham số, gây khó khăn trong việc triển khai và quản lý nguồn lực tính toán. Điều này thúc đẩy sự kết hợp LSTM với các kỹ thuật khác nhằm tối ưu hóa hiệu suất và độ ổn định.

## 1.6 Những vấn đề hạn chế trong phát triển AI

Dự đoán giá và xu hướng của thị trường chứng khoán luôn là một bài toán phức tạp do đặc thù của thị trường với nhiều yếu tố tác động đa chiều như sau:

- Tình trạng tài chính và kinh tế vĩ mô: Các chỉ số như GDP, lạm phát, lãi suất, tỷ giá hối đoái và chính sách tiền tệ có thể ảnh hưởng trực tiếp đến tâm lý nhà đầu tư và xu hướng chung của thị trường.
- Giá lịch sử: Dữ liệu giá cổ phiếu trong quá khứ là nền tảng quan trọng cho các phương pháp phân tích kỹ thuật và học máy trong việc dự đoán xu hướng tương lai.
- Tin tức tài chính: Những sự kiện như thay đổi chính sách của Nhà nước, biến động thị trường quốc tế, hoạt động M&A, hay kết quả kinh doanh đột biến của doanh nghiệp có thể tác động mạnh đến giá cổ phiếu trong ngắn hạn.
- Sản phẩm và dịch vụ của doanh nghiệp niêm yết: Mức độ cạnh tranh, sự đổi mới trong sản phẩm, và hiệu quả hoạt động kinh doanh của doanh nghiệp đều ảnh hưởng đến giá trị nội tại và kỳ vọng tăng trưởng của cổ phiếu.

Việc ứng dụng các AI nhằm tăng cường độ chính xác, thời gian dự báo và khả năng nắm bắt các xu hướng thị trường đang gặp phải một số hạn chế quan trọng. Dưới đây là một số vấn đề nổi bật và các hướng tiếp cận cải tiến.

*1.6.1 Hiểu Biết Và Xử Lý Bối Cảnh Thị Trường Vấn đề*

- Vấn đề: Thị trường chứng khoán bị ảnh hưởng bởi nhiều yếu tố bên ngoài như dữ liệu kinh tế vĩ mô, tin tức, sự kiện chính trị – những yếu tố không nằm trong khuôn khổ định lượng truyền thống. Các mô hình thống kê, dù đã được cải tiến để xử lý dữ liệu chuỗi thời gian, nhưng không thể nắm bắt trọn vẹn bối cảnh đa chiều, dẫn đến những dự đoán thiếu sót khi xảy ra biến động đột biến.
- Hướng nghiên cứu: Nghiên cứu hướng tích hợp thêm các nguồn dữ liệu định tính như tin tức thị trường, báo cáo tài chính, hay phân tích tâm lý nhà đầu tư (không dựa vào kỹ thuật học sâu) có thể là bước đột phá. Việc liên kết dữ liệu từ nhiều nguồn sẽ giúp mô hình hiểu được "bức tranh" tổng thể của thị trường hơn, từ đó nâng cao độ chính xác của dự đoán.

*1.6.2 Đảm Bảo Tính Ổn Định Và Nhất Quán Của Mô hình*

- Vấn đề: Các mô hình dự báo hiện nay thường cho ra kết quả không ổn định khi đối mặt với các biến động dữ liệu ngắn hạn hoặc những sự kiện độc lập. Điều này làm giảm độ tin cậy, khiến cho các chiến lược đầu tư bị ảnh hưởng khi mô hình phản hồi chậm kịp với những thay đổi mới.
- Hướng nghiên cứu: Áp dụng các mô hình ensemble – kết hợp ưu điểm của Logistic Regression và LSTM – nhằm tối ưu hóa khả năng tổng hợp và điều chỉnh dự báo. Cập nhật dữ liệu theo thời gian thực sẽ giúp duy trì sự nhất quán trong mọi điều kiện thị trường.

*1.6.3 Xử Lý Dữ Liệu Thời Gian Thực Và Đảm Bảo Tốc Độ Phản Hồi Vấn đề*

- Vấn đề: Thị trường chứng khoán luôn biến đổi không ngừng với khối lượng dữ liệu khổng lồ từ nhiều nguồn khác nhau. Việc xử lý nhanh chóng và cung cấp dự báo kịp thời là một thách thức lớn, nhất là khi các mô hình truyền thống chưa tối ưu cho xử lý dữ liệu thời gian thực.
- Hướng nghiên cứu: Phát triển các giải pháp về kiến trúc tính toán phân tán và thuật toán tối ưu hoá thực hiện song song. Nghiên cứu cách sử dụng công nghệ streaming data và áp dụng các phương pháp xử lý dữ liệu theo thời gian thực sẽ giúp các mô hình đưa ra phản hồi nhanh, đảm bảo hiệu quả ra quyết định của nhà đầu tư.

Nhìn chung, mặc dù các mô hình truyền thống như Logistic Regression và LSTM đã mang lại nhiều kết quả tích cực trong việc dự đoán xu hướng chứng khoán, song vẫn tồn tại những rào cản liên quan đến tính đa chiều và phi tuyến tính của dữ liệu thị trường. Việc tích hợp đa dạng nguồn dữ liệu, cải thiện quy trình cập nhật thông tin thời gian thực và xây dựng hệ thống bảo mật mạnh mẽ sẽ là những hướng nghiên cứu cần thiết để đưa hệ thống dự báo đạt hiệu quả cao hơn.

## 1.7 Phương pháp nghiên cứu

*1.7.1 Thu thập và xử lý dữ liệu*

- Nguồn dữ liệu: Thu thập dữ liệu lịch sử về giá cổ phiếu từ các sàn giao dịch chứng khoán như HOSE, HNX, SSI Corporation và thư viện vnstock Tin tức tài chính từ các trang báo vietstock, vnexpress, cafef.vn.
- Xử lý dữ liệu: Tiến hành làm sạch, chuẩn hóa và xử lý dữ liệu nhằm đảm bảo tính chính xác, đầy đủ và nhất quán, tránh nhiễu loạn và sai lệch ảnh hưởng tới mô hình.

*1.7.2 Xây dựng mô hình học máy*

- Huấn luyện mô hình: Ứng dụng các mô hình Logistic Regression và LSTM nhằm phân tích, dự báo các xu hướng biến động giá cổ phiếu dựa trên các yếu tố lịch sử, tài chính và vĩ mô.
- Công cụ và thư viện: Sử dụng ngôn ngữ lập trình Python kết hợp các thư viện phổ biến như TensorFlow, PyTorch, scikit-learn, pandas, và numpy để xây dựng và tối ưu mô hình.

*1.7.3 Đánh giá mô hình*

- Chỉ số đánh giá: Đánh giá hiệu suất mô hình dựa trên các chỉ số thống kê quan trọng như accuracy, precision, recall, F1-score và Mean Squared Error (MSE).
- Kiểm thử thực tế: Thực hiện đánh giá mô hình trên dữ liệu thực tế, so sánh với hiệu quả đầu tư thực tiễn nhằm xác định độ chính xác, khả năng ứng dụng của các mô hình trong giao dịch thực tế.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1 Mô hình AI hỗ trợ đầu tư chứng khoán

Trong bối cảnh thị trường chứng khoán ngày càng biến động phức tạp, việc ứng dụng **trí tuệ nhân tạo (AI)** trong phân tích và dự đoán giá cổ phiếu trở thành một xu hướng tất yếu. Các mô hình học máy có khả năng học từ dữ liệu lịch sử và phát hiện các mẫu (pattern) nhằm đưa ra dự đoán cho tương lai. Trong đề tài này, nhóm nghiên cứu đã ứng dụng hai mô hình tiêu biểu gồm: Logistic Regression, LSTM.

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

*a. Bài toán phân loại cảm xúc*

Phân loại cảm xúc văn bản tài chính là quá trình tự động xác định thái độ hoặc quan điểm được biểu đạt trong các văn bản liên quan đến thị trường tài chính. Đây là một bài toán phân loại có giám sát (supervised classification), trong đó mỗi văn bản được gán một nhãn cảm xúc:

- Tích cực (positive)
- Tiêu cực (negative)
- Trung tính (neutral)

*b. Biểu diễn văn bản bằng TF-IDF*

Văn bản là dạng dữ liệu phi cấu trúc, cần được chuyển đổi sang dạng số để có thể xử lý bằng các thuật toán học máy. Một trong những phương pháp phổ biến nhất là TF-IDF (Term Frequency – Inverse Document Frequency).

Thành phần của TF-IDF bao gồm:

- **Term Frequency (TF)**: Đo tần suất xuất hiện của một từ trong một văn bản cụ thể.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

trong đó:

*t*: từ cần tính tần suất

*d*: văn bản cụ thể

$f_{t,d}$: số lần xuất hiện của từ $t$ trong văn bản $d$

- **Inverse Document Frequency (IDF)**: Giảm trọng số của các từ phổ biến bằng cách đo lường mức độ quan trọng của một từ trong toàn bộ tập văn bản.

$$IDF(t) = \log\left(\frac{N}{n_t}\right)$$

trong đó:

$N$: từ cần tính tần suất

$n_t$: văn bản cụ thể

TD-IDF Score:

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

Ưu điểm của TF-IDF trong phân tích tài chính:

- Nhấn mạnh các từ đặc trưng có giá trị phân biệt cao.
- Giảm ảnh hưởng của các từ thông dụng như "cổ phiếu", "giá", "thị trường".
- Phù hợp với các văn bản chứa nhiều thuật ngữ tài chính chuyên môn.

*c. Mô hình Logistic Regression*

- **Cơ sở toán học**: Mô hình Logistic Regression sử dụng hàm sigmoid để ánh xạ đầu vào thành xác suất:

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

- **Hàm mất mát và tối ưu hóa**: Hàm log-likelihood được dùng để tối ưu hóa:

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} [y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

trong đó:

$p_i$: Xác xuất dự đoán văn bản thứ $i$ thuộc lớp 1

- **Regularization**: Nhằm tránh overfitting, mô hình có thể áp dụng các hình thức

  + L1 Regularization (Lasso):

$$\lambda \sum_{j=1}^{n} |\beta_j|$$

  + L2 Regularization (Ridge):

$$\lambda \sum_{j=1}^{n} \beta_j^2$$

  + Elastic Net:

$$\lambda_1 \sum_{j=1}^{n} |\beta_j| + \lambda_2 \sum_{j=1}^{n} \beta_j^2$$

**Ưu điểm:**

- Khả năng giải thích tốt: hệ số $\beta_j$ cho biết **mức độ ảnh hưởng của từ** $x_j$ đến cảm xúc dự đoán.
- Thực thi nhanh, dễ triển khai.

*d. Vai trò bổ trợ trong hệ thống dự đoán chứng khoán*

Mô hình phân tích cảm xúc đóng vai trò như một tín hiệu phụ (auxiliary signal), hỗ trợ mô hình chính dự đoán xu hướng thị trường. Cảm xúc từ các bài báo, tin tức tài chính có thể:

- Cung cấp thông tin định tính bổ sung cho các chỉ số kỹ thuật.
- Phản ánh tâm lý nhà đầu tư và biến động thị trường theo thời gian thực.

*2.1.2 LSTM*

*a. Tổng quan*

LSTM là một kiến trúc thuộc nhóm mạng nơ-ron hồi tiếp (RNN – Recurrent Neural Networks), được thiết kế nhằm khắc phục nhược điểm của RNN truyền thống trong việc học mối quan hệ dài hạn trong chuỗi thời gian.

RNN truyền thống gặp phải vấn đề vanishing gradient khiến việc học thông tin từ quá khứ xa trở nên không hiệu quả. LSTM khắc phục điều này thông qua cơ chế bộ nhớ có kiểm soát với ba "cổng":

- Forget gate
- Input gate
- Output gate

*b. Cấu trúc của một cell LSTM*

Tại mỗi bước thời gian *t*, LSTM nhận đầu vào $x_t$ , trạng thái ẩn trước đó $h_{t-1}$ và trạng thái bộ nhớ $C_{t-1}$, sau đó tính toán như sau:

- Forget Gate: Xác định phần nào của thông tin cũ cần bị "quên" khỏi cell memory:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- Input Gate: Xác định phần nào của thông tin mới sẽ được thêm vào

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- Cập nhật trạng thái bộ nhớ

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- Output Gate: Xác định trạng thái ẩn $h_t$:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

*c. Ưu điểm của LSTM trong dự báo tài chính*

- Nhớ thông tin dài hạn: Giúp mô hình dự đoán xu hướng cổ phiếu dựa trên chuỗi dữ liệu dài hạn như giá, khối lượng, tin tức.

- Linh hoạt với chuỗi không đều: LSTM không yêu cầu dữ liệu có chu kỳ rõ ràng như Prophet.
- Tương thích với dữ liệu phi tuyến: Phù hợp với các mô hình thị trường tài chính có quan hệ phi tuyến và nhiều yếu tố nhiễu.

*d. Nhược điểm*

- Yêu cầu thời gian huấn luyện dài.
- Dễ bị **overfitting** nếu dữ liệu ít.
- Khó giải thích hơn so với các mô hình tuyến tính như Logistic Regression.

# CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

## 3.1 Kiến trúc tổng thể hệ thống

Hệ thống dự đoán giá cổ phiếu được xây dựng gồm hai thành phần chính:

- Mô hình phân loại cảm xúc từ tin tức: Sử dụng kỹ thuật TF-IDF kết hợp Logistic Regression để gán nhãn cảm xúc (tích cực, tiêu cực hoặc trung tính) cho các bài viết tài chính.
- Mô hình LSTM dự đoán giá cổ phiếu: Dự đoán giá mở cửa của cổ phiếu ngày tiếp theo dựa trên dữ liệu chuỗi thời gian (giá lịch sử) và chỉ số cảm xúc của tin tức trong những ngày trước đó.

Quy trình tổng thể bao gồm:

- Tiền xử lý dữ liệu giá cổ phiếu và dữ liệu văn bản từ tin tức.
- Huấn luyện mô hình phân loại cảm xúc bằng TF-IDF và hồi quy tuyến tính.
- Gộp dữ liệu cảm xúc với dữ liệu giá để tạo tập dữ liệu hợp nhất.
- Huấn luyện mô hình LSTM trên tập dữ liệu kết hợp để dự đoán giá cổ phiếu.

## 3.2 Tiền xử lý dữ liệu

*3.2.1 Giá cổ phiếu*

- Bộ thuộc tính: Date, Open, High, Low, Close, Volume.
- Làm sạch: loại bản ghi trùng/lỗi; điều chỉnh giá theo chia tách/cổ tức nếu có.

- Chuẩn hóa: *MinMaxScaler* 0→10 → 10→1 để ổn định gradient cho LSTM.
- Tính toán các chỉ số kỹ thuật bao gồm:
  + Moving Average: MA_5, MA_10, MA_20
  + Exponential Moving Average: EMA_5, EMA_10
- Tách tập: Train 80 %, Test 20 % theo thứ tự thời gian (không xáo trộn).

*3.2.2 Tin tức tài chính (Dữ liệu cảm xúc)*

- Nguồn train: bộ *Financial PhraseBank* (Kaggle) → dịch sang Tiếng Việt bằng Google Translate API, sau đó hậu kiểm thủ công các thuật ngữ chuyên ngành.
- Nguồn thực tế: CafeF, Vietstock, Bloomberg, Reuters…; thu thập tự động 10 bài báo gần nhất.
- Phân tích cảm xúc tạo ra 3 cột dữ liệu: neutral, positive, negative chứa tỉ lệ các cảm xúc
- Xử lý: lower-case, bỏ stop-words, ký tự đặc biệt; giữ lại bigram/trigram chứa thuật ngữ tài chính (vd. "lãi suất", "room tín dụng").

## 3.3 Mô hình phân tích (phân loại) cảm xúc

*3.3.1 Biểu diễn văn bản TF-IDF*

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

$$IDF(t) = \log\left(\frac{N}{n_t}\right)$$

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

*3.3.2 Classifier – Logistic Regression đa lớp*

- ***Softmax***:

$$\hat{p}_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad z = Wx + b$$

- **Loss**: Cross-Entropy; tối ưu bằng *liblinear* (sklearn).
- **Hyper-parameters**: `C=1.0`, `penalty=L2`, `class_weight=balanced`.

*3.3.3 Tính Sentiment Index theo ngày*

$$S_{\text{pos}}(d) = \frac{1}{m_d} \sum_{i=1}^{m_d} \hat{p}_{i,\text{pos}} \qquad S_{\text{neu}}(d) = \frac{1}{m_d} \sum_{i=1}^{m_d} \hat{p}_{i,\text{neu}}$$

$$S_{\text{neg}}(d) = \frac{1}{m_d} \sum_{i=1}^{m_d} \hat{p}_{i,\text{neg}}$$

Trong đó mdm_dmd là số tin ngày ddd. Bộ $S_{pos}$,$S_{neu}$,$S_{neg}$ được xem là đặc trưng định tính nhập vào mô hình giá.

## 3.4 Mô hình dự đoán giá cổ phiếu LSTM

*3.4.1 Thiết lập dữ liệu chuỗi*

- **Window size**: 60 ngày (t – 59 → t).
- **Đặc trưng**: `[Open, High, Low, Close, Volume, indicators, S_pos, S_neu, S_neg]`.

*3.4.2 Kiến trúc mạng*

*Input (None, 60, F) → LSTM(32) → Dropout(0.2) → Dense(1, linear)*

**Loss**: Mean Squared Error **Optimizer**: Adam (lr = 0.001)

*3.4.3 Chiến lược huấn luyện*

| ***Tham số*** | ***Giá trị*** |
|---|---|
| *batch_size* | *32* |

| | |
|---|---|
| *epochs* | *20 (early stopping patience = 3)* |
| *shuffle* | *False* |
| *Validation* | *walk-forward (mỗi 10 ngày đánh giá, cập nhật mô hình)* |

# CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

## 4.1 Mô tả tập dữ liệu

### *4.1.1 Giá cổ phiếu (VCB)*

- Khoảng thời gian: 01/01/2020 – 31/12/2023
- Trường dữ liệu: Date, Open, High, Low, Close, Volume
- Tiền xử lý:

1. Loại bản ghi trùng hoặc thiếu
2. Điều chỉnh giá theo chia tách / cổ tức.
3. Sắp xếp thời gian tăng dần, chuẩn hóa Min - Max
4. Chia tập dữ liệu Train : Test theo tỉ lệ 80 : 20

### *4.1.2 Dữ liệu tin tức và cảm xúc tin thị trường*

- Nguồn dữ liệu training: Financial PhraseBank (Kaggle) → dịch vi-VI qua Google Translate, hậu kiểm thuật ngữ.
- Nguồn dữ liệu kiểm thử thực tế: CafeF, Bloomberg, Reuters, Vietstock (crawler 30 phút/lần).
- Tiền xử lý: lowercase → bỏ stop-word → TF-IDF → Logistic Regression đa lớp → nhãn {0,1,2}\{0,1,2\}{0,1,2}.

## 4.2 Thiết lập thực nghiệm

| Thành phần | Cấu hình chính |
|---|---|
| *Sentiment model* | TF-IDF 40 k từ; Logistic Regression (C=1, penalty=L2, class_weight=balanced) |
| *Chỉ số cảm xúc* | $\bar{p}_c(d) = \frac{1}{m_d} \sum_{i=1}^{m_d} p_{i,c}, \qquad c \in \{\text{neg}, \text{neu}, \text{pos}\}$ |
| *LSTM* | window = 60, LSTM 32, Dropout 0.2, Dense(1), Adam 0.001, epochs 20, batch 32 |

### 4.3 Kết quả mô hình phân tích cảm xúc

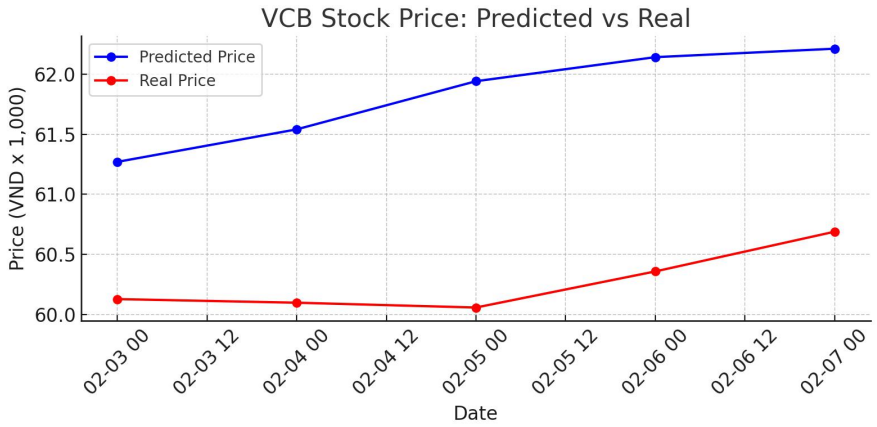
| Metric | Giá trị |
|---|---|
| *Accuracy* | *0.763* |
| *Precision (macro)* | *0.835* |
| *Recall (macro)* | *0.622* |
| *F1-macro* | *0.673* |

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

Nhận xét:

- Recall lớp *tiêu cực* 98 % → tin xấu được phát hiện tốt.
- Lớp *trung tính* & *tích cực* bị hụt recall do mất cân bằng; sẽ cải thiện nếu dùng **class weight** hoặc **SMOTE**.

### 4.4 Kết quả mô hình dự đoán giá cổ phiếu (VCB)




## 4.5 Xây dựng ứng dụng

Ứng dụng được xây dựng nhằm trực quan hóa kết quả mô hình AI trong việc dự đoán giá cổ phiếu và cung cấp nền tảng blog chia sẻ kiến thức đầu tư. Hệ thống bao gồm các chức năng chính:

- Hiển thị biểu đồ streaming giá thực (Delay tùy thuộc vào rate limit của API) và giá dự đoán theo thời gian thực.
- Tìm kiếm và tra cứu cổ phiếu.
- Xem phân tích cơ bản.
- Đọc các bài viết, bình luận và chia sẻ trong hệ thống blog tài chính.

### *4.5.1 MERN Stack:*

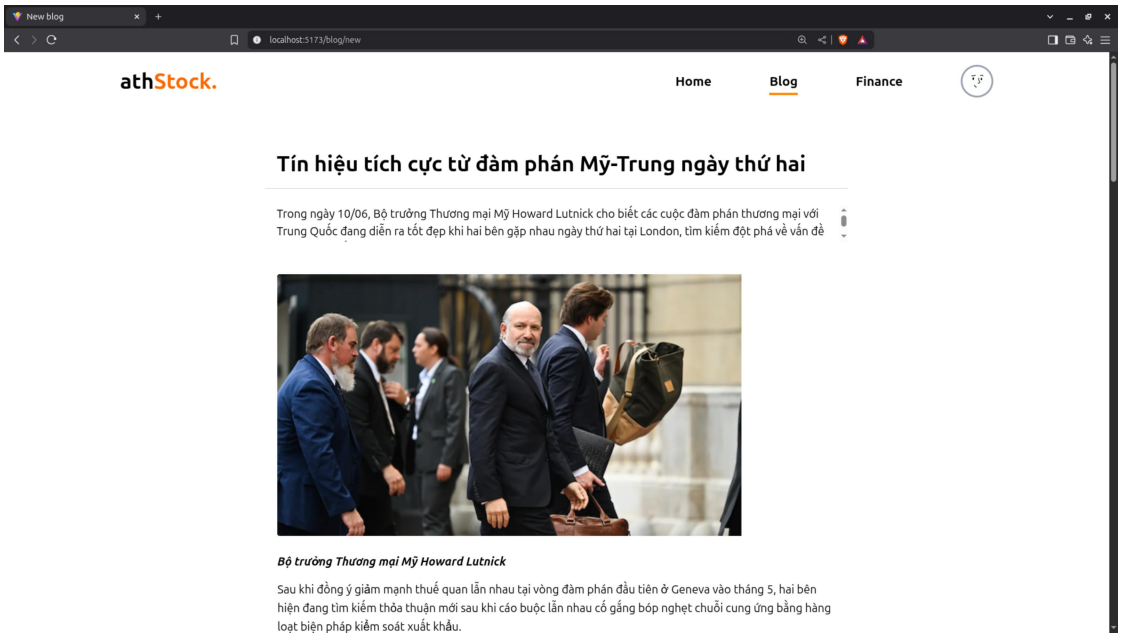
Để đảm bảo tính linh hoạt, khả năng mở rộng và hiệu suất cao, hệ thống được phát triển dựa trên **MERN Stack**, bao gồm:

| ***Thành phần*** | ***Mô tả*** |
|---|---|
| MongoDB | Cơ sở dữ liệu NoSQL lưu trữ thông tin cổ phiếu, bài viết blog và người dùng |
| *Express.js* | Framework backend chạy trên Node.js, cung cấp API phục vụ frontend và xử lý dữ liệu |

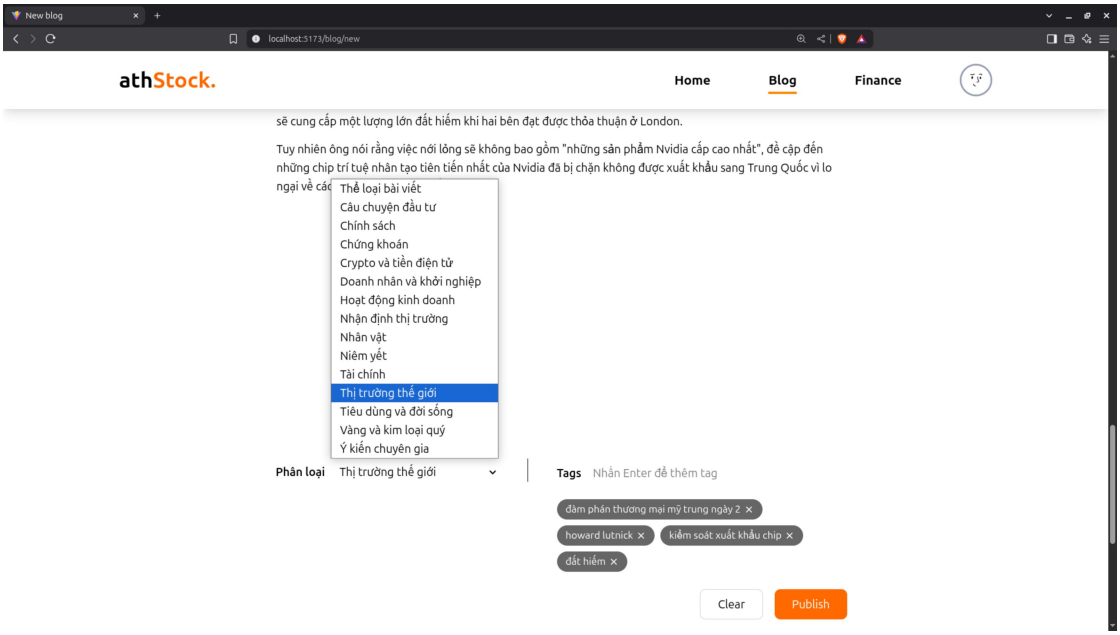
| | |
|---|---|
| *React.js* | Thư viện xây dựng giao diện người dùng, hiển thị biểu đồ, dữ liệu thị trường và nội dung blog |
| *Node.js* | Nền tảng JavaScript phía server, chịu trách nhiệm vận hành Express.js và quản lý các kết nối |

**Một số giao diện của ứng dụng:**

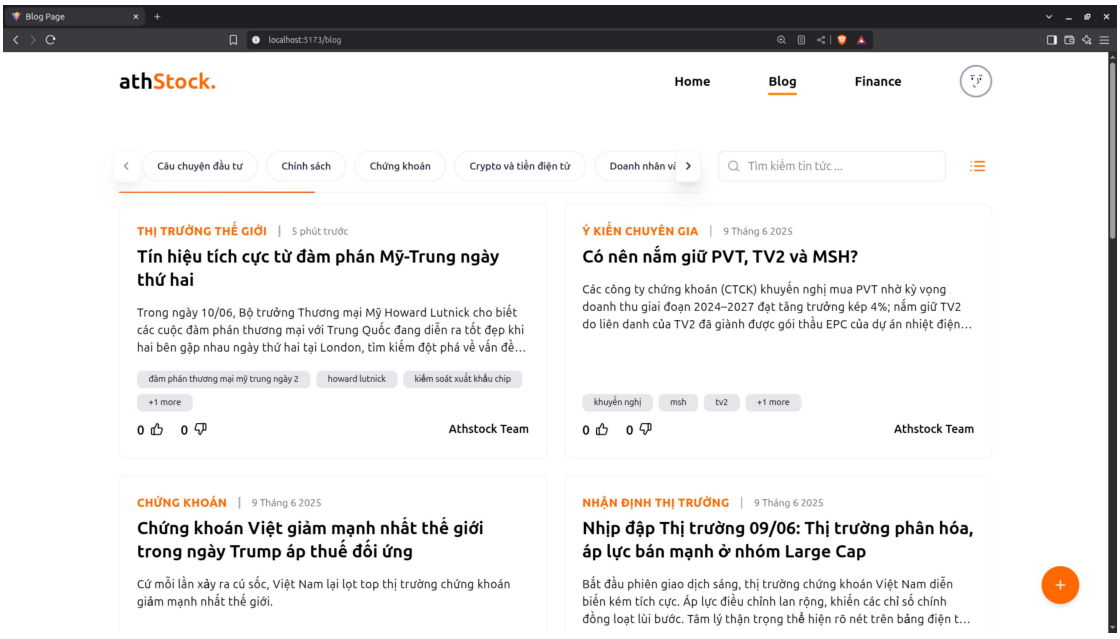
- Cho phép người dùng đăng các bài báo, phân tích tài chính hoặc bình luận về các sự kiện liên quan tới tài chính việt name và toàn cầu

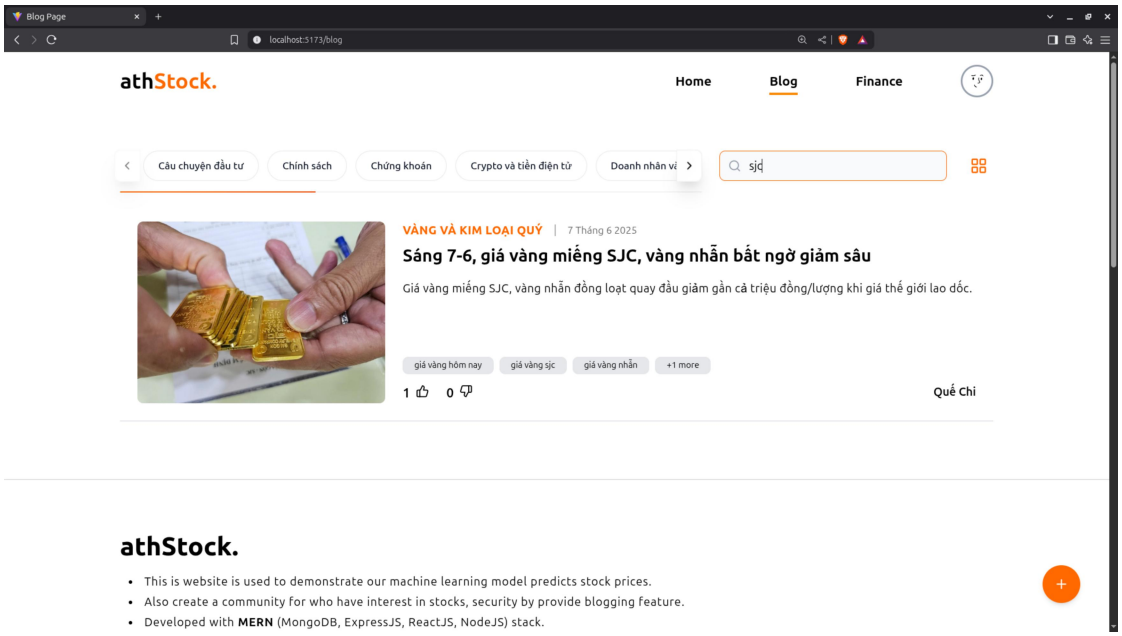


- Với mỗi bài đăng sẽ được tùy chọn phân loại vào các danh mục bài viết được cung cấp sẵn và gắn thẻ để người dùng có thể dễ dàng tìm kiếm bài viết theo tag (keyword).

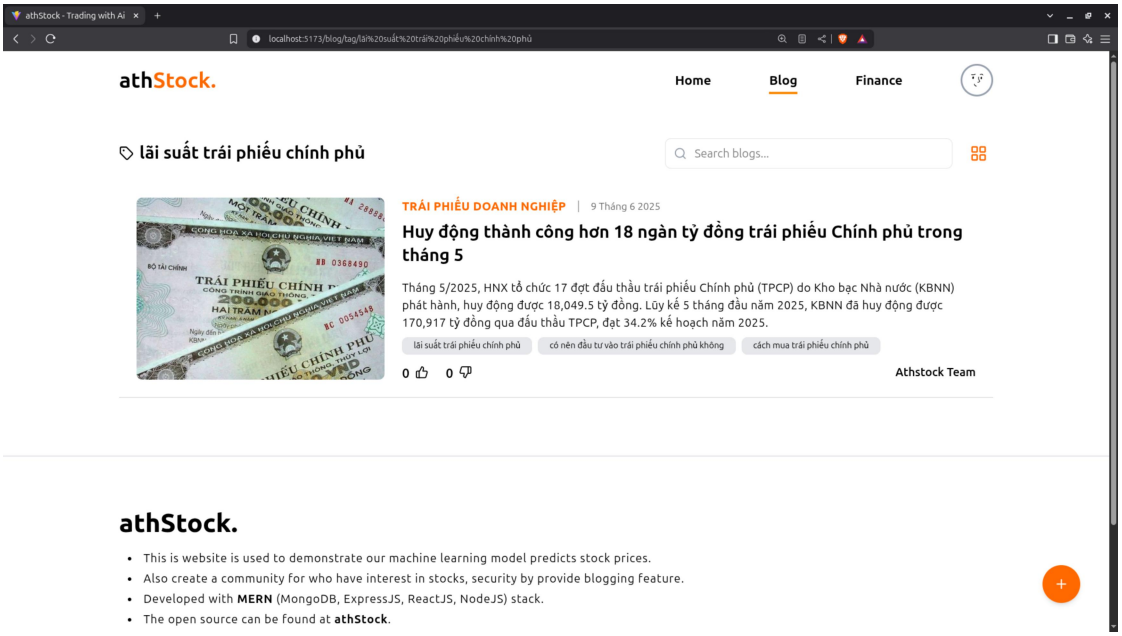



- Giao diện tại trang tin tức bao gồm các thành phần như tùy chọn danh mục bài viết, ô tìm kiếm bài viết theo từ khóa hoặc tiêu đề. Và tùy chọn hiển thị bài viết dưới dạng lưới (grid) hoặc danh sách (list).

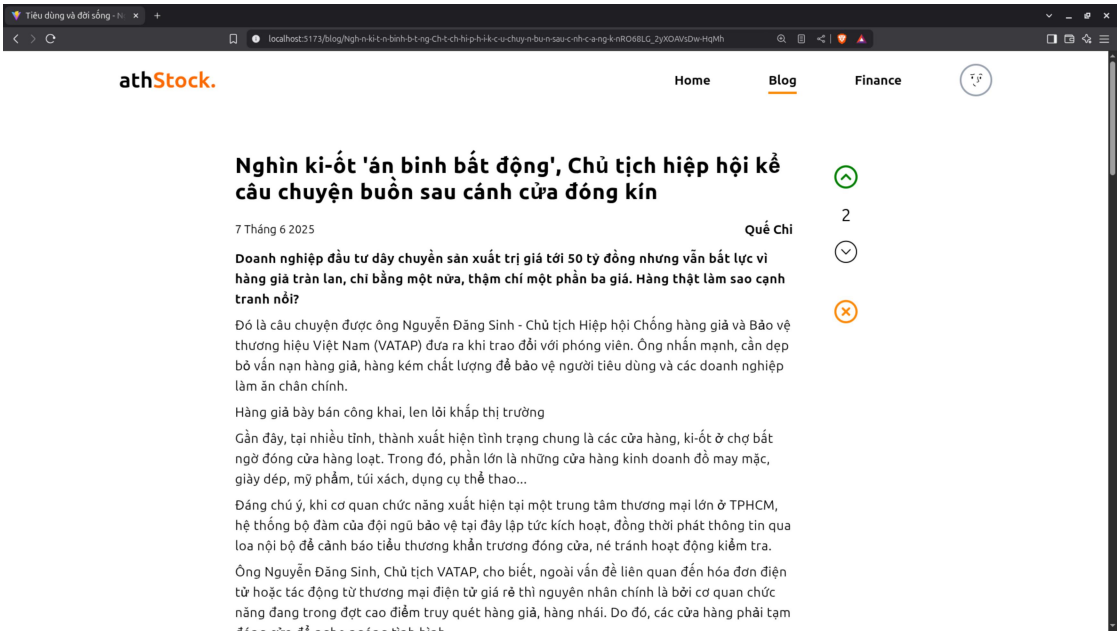




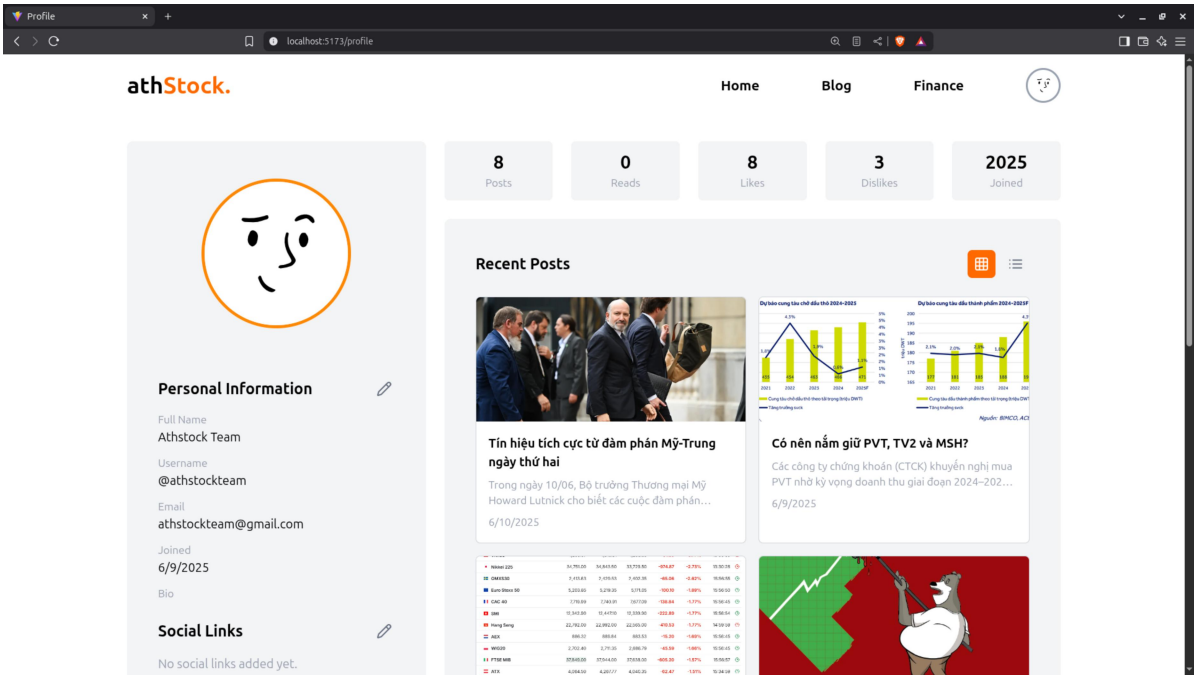

- Người dùng có thể sử dụng tag như một công cụ phân loại và tìm kiếm bài viết



- Với mỗi bài đăng người dùng hứng thú hoặc cảm thấy hữu ích thì có thể upvote (like) để bài viết có thể được đề xuất cho các người dùng khác và ngược lại downvote (dislike) nếu như cảm thấy nội dung không phù hợp với bản thân.




- Cung cấp giao diện quản lý thông tin người dùng.



## 4.6 Ý nghĩa kết quả nghiên cứu

### *4.6.1 Ý nghĩa khoa học*

- Kết hợp hai mô hình machine learning (Logistic Regression, LSTM) trên cùng một pipeline dự báo; tăng độ chính xác cho các dự đoán.
- Ghép **dòng thời gian giá** với **cảm xúc văn bản tài chính** qua mô hình Logistic Regression + TF-IDF ⇒ tạo tập dữ liệu nhiều chiều, có giá trị cho các nghiên cứu thị trường hành vi (*behavioral finance*).
- Công khai code, tham số, quy trình tiền xử lý → các nhóm nghiên cứu khác có thể tái lập (reproducible) hoặc mở rộng sang thị trường khác.

### *4.6.2 Ý nghĩa thực tiễn*

- Công cụ hỗ trợ quyết định: Website streaming giá thực & giá dự đoán giúp nhà đầu tư theo dõi biến động và nhận cảnh báo xu hướng tức thì.
- Phân tích cảm xúc tin tức phản ánh nhanh tâm lý đám đông; nhà đầu tư cá nhân tiếp cận sớm hơn so với chỉ quan sát giá.
- Mô hình triển khai dưới dạng *micro-service* dễ tích hợp vào cổng giao dịch, ứng dụng di động của công ty chứng khoán.

## 4.7 Những điều đạt được

| **Kết quả** | **Nội dung cụ thể** |
|---|---|
| Độ chính xác của mô hình | Khi kết hợp LSTM và phân tích cảm xúc thì độ chính tăng nhẹ. |
| Sản phẩm hệ thống | Hoàn thiện website MERN, streaming ≤ 60000 ms, chịu tải 200 concurrent users<br>Dashboard hiển thị giá thực–giá dự đoán, biểu đồ sai số, blog chia sẻ kiến thức |

## 4.8 Định hướng nghiên cứu tiếp theo

**Mở rộng nguồn dữ liệu**

- Thu thập tin mạng xã hội (Twitter, StockTwits, Facebook Group) để phản ánh tâm lý tức thời.
- Bổ sung dữ liệu vĩ mô (lãi suất, CPI, tỷ giá) làm biến giải thích.

**Nâng cấp mô hình cảm xúc**

- Fine-tune PhoBERT, FinBERT-multilingual giúp tăng recall lớp *tích cực/trung tính*.
- Khai thác aspect-based sentiment: tách cảm xúc theo từng tiêu chí (lợi nhuận, quản trị, ESG).

**Cải tiến mô hình giá**

- Khảo sát Temporal Fusion Transformer (TFT) và N-Beats cho chuỗi thời gian.
- Tận dụng Graph Neural Networks để mô hình hóa quan hệ ngành nghề và sở hữu chéo giữa cổ phiếu.

**Học tăng cường (RL) cho chiến thuật giao dịch**

- Xây dựng *agent* DQN/Proximal-Policy-Optimization tối ưu lợi nhuận điều chỉnh rủi ro.
- Thực hiện *paper-trading* (tài khoản ảo) để đánh giá Sharpe ratio, max drawdown.

# TÀI LIỆU THAM KHẢO


1. Bộ nhớ dài-ngắn hạn. Xem ngày 25.5.2025 Wikipedia, https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=March%202022),and%20other%20sequence%20learning%20methods
2. Hồi quy tuyến tính trong Machine Learning NguyenDuong. 2017. Linear Regression - Hồi quy tuyến tính trong Machine Learning, xem ngày 25.5.2025 Viblo, https://viblo.asia/p/hoi-quy-tuyen-tinh-trong-machine-learning-1VgZvaw7KAw
3. Predictive Modeling of Stock Prices Using Transformer Model Mozaffari, L. and Zhang, J. 2024. Predictive Modeling of Stock Prices Using Transformer Model. ICMLT 2024, Oslo, Norway. ACM Digital Library. Xem ngày 25.5.2025 ACM, https://dl.acm.org/doi/fullHtml/10.1145/3674029.3674037
4. Multioutput Regression in Machine Learning GeeksforGeeks. Xem ngày 25.5.2025 GeeksforGeeks, https://www.geeksforgeeks.org/multioutput-regression-in-machine-learning/
5. TF-IDF và vai trò của TF-IDF trong SEO VietMoz. Xem ngày 25.5.2025 VietMoz, https://vietmoz.edu.vn/tf-idf/