	<p><b><u>Thủ tục:</u></b> <b>NGHIÊN CỨU KHOA HỌC SINH VIÊN</b></p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: <b>06</b> Ngày hiệu lực:</p>
--	--	--

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

(Nguyễn Quang Huy, Lớp 23H50301, Khoa Công Nghệ Thông Tin)

(Nguyễn Trần Nhật AnHuy, Lớp 23H50301, Khoa Công Nghệ Thông Tin)

(Nguyễn Thanh Tùng, Lớp 23H50301, Khoa Công Nghệ Thông Tin)

### I. TÓM TẮT CÔNG TRÌNH

Đề tài nhằm phát triển một ứng dụng hỗ trợ nhà đầu tư cá nhân và tổ chức trong việc đưa ra quyết định đầu tư chính xác và kịp thời thông qua kết hợp các mô hình trí tuệ nhân tạo (AI) và phân tích cơ bản. Hệ thống cho phép hiển thị giá chứng khoán streaming real-time, dự báo giá cổ phiếu bằng mô hình LSTM kết hợp chỉ số sentiment từ Logistic Regression trên dữ liệu tin tức tài chính, đồng thời cung cấp nền tảng blog phân tích mã cổ phiếu và chia sẻ kiến thức đầu tư.

### II. QUÁ TRÌNH NGHIÊN CỨU VÀ KẾT QUẢ


#### 2.1. Mở đầu

##### 2.1.1. Mở đầu

Thị trường chứng khoán Việt Nam ngày càng thu hút nhưng nhiều nhà đầu tư cá nhân thiếu kinh nghiệm, dễ ra quyết định cảm tính trong bối cảnh biến động cao. AI có khả năng xử lý lượng dữ liệu lớn, hỗ trợ dự báo xu hướng với độ chính xác cao hơn phương pháp truyền thống.

##### 2.1.2. Lý do chọn đề tài

Phần lớn công cụ hiện nay tập trung phân tích kỹ thuật, ít khai thác phân tích cơ bản và sentiment analysis; đề tài kết hợp hài hòa giữa hai hướng để nâng cao hiệu quả đầu tư.

	<p><b><u>Thủ tục:</u></b> <b>NGHIÊN CỨU KHOA HỌC SINH VIÊN</b></p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	---

### ***2.1.3. Đối tượng và phạm vi nghiên cứu***

Nghiên cứu tập trung vào cổ phiếu nhóm VNINDEX và HNXINDEX giai đoạn 2011–2025, mô hình AI (Logistic Regression, LSTM) và phân tích cơ bản của các doanh nghiệp niêm yết.

## **2.2. Cơ sở lý thuyết**

- Logistic Regression dùng để phân loại cảm xúc văn bản tài chính (positive/neutral/negative) trên cơ sở TF-IDF, đóng vai trò auxiliary signal cung cấp thông tin định tính bổ sung cho mô hình giá.

- LSTM (Long Short-Term Memory) khắc phục vanishing gradient của RNN truyền thống, ghi nhớ thông tin dài hạn trên chuỗi thời gian giá và sentiment index, phù hợp với dữ liệu chuỗi không đều và phi tuyến.

## **2.3. Phương pháp đề xuất**

### ***2.3.1. Kiến trúc hệ thống***

Gồm hai micro-services chính – sentiment analysis service (TF-IDF + Logistic Regression) và price prediction service (LSTM) – tích hợp qua MERN stack (MongoDB, Express.js, React.js, Node.js).

### ***2.3.2. Tiền xử lý dữ liệu***


- Tin tức tài chính: thu thập tự động mỗi 30 phút (CafeF, Bloomberg, Reuters...), lowercase, loại stop-words, giữ bigram/trigram tài chính, biểu diễn TF-IDF với 40k từ

- Giá cổ phiếu: loại bỏ bản ghi lỗi, điều chỉnh chia tách/cổ tức, chuẩn hóa Min-Max, chia train/test 80/20.

### ***2.3.3. Huấn luyện mô hình***

- Sentiment model: Logistic Regression đa lớp (C=1.0, penalty=L2, class\_weight=balanced).

- Price model: LSTM 32 units + Dropout 0.2 → Dense(1), optimizer Adam lr=0.001, batch\_size=32, epochs=20 (early stopping).

	<p><b><u>Thủ tục:</u></b> <b>NGHIÊN CỨU KHOA HỌC SINH VIÊN</b></p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	---

#### **2.3.4. Gộp dữ liệu**

- Kết hợp sentiment index (S\_pos, S\_neu, S\_neg) với chuỗi giá 60 ngày làm đầu vào cho LSTM BM18\_Nhom28\_TrinhHungCu....

### **2.4 Kết quả thực nghiệm**

#### **2.4.1. Mô tả dữ liệu**

- Giá VCB giai đoạn 01/01/2020–31/12/2023, 6 trường (Date, Open, High, Low, Close, Volume).
- Tin tức: Financial PhraseBank (train), CafeF/VnExpress/Vietstock.

#### **2.4.2. Kết quả sentiment analysis**

- Accuracy=0.763, Precision\_macro=0.835, Recall\_macro=0.622, F1\_macro=0.673; recall lớp negative đạt 98%.

#### **2.4.3. Kết quả dự báo giá VCB**


- Biểu đồ Predicted vs Real cho thấy mô hình LSTM kết hợp sentiment cải thiện độ chính xác dự báo, trung bình sai số giảm so với baseline.

#### **2.4.4. Ứng dụng thực tế**

- Giao diện streaming real-time giá và giá dự đoán, chức năng tìm kiếm, tra cứu cơ bản, blog chia sẻ phân tích, chịu tải 200 concurrent users, độ trễ  $\leq 30s$ .

### **III. KẾT LUẬN**

- Đã xây dựng thành công pipeline kết hợp sentiment analysis và LSTM, giúp nâng cao độ chính xác dự báo và cung cấp tín hiệu đầu tư kịp thời.
- Mô hình đáp ứng yêu cầu real-time và khả năng mở rộng, dễ tích hợp micro-services vào hệ thống giao dịch.
- Thực nghiệm chứng minh sentiment index đóng góp đáng kể cho độ chính xác dự báo.
- Đề xuất mở rộng: thu thập dữ liệu mạng xã hội, bổ sung biến vĩ mô; fine-tune PhoBERT/FinBERT; khảo sát TFT, Graph Neural Networks; áp dụng Reinforcement Learning cho chiến lược giao dịch.

	<p><b><i>Thủ tục:</i></b> <b>NGHIÊN CỨU KHOA HỌC SINH VIÊN</b></p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: <b>06</b> Ngày hiệu lực:</p>
--	--	--

#### IV. 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](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/>