Hà Nội – 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN THỰC TẬP TỐT NGHIỆP

Tên đề tài : XÂY DỰNG ỨNG DỤNG FIN CHATBOT ÁP DỤNG TRÍ TUỆ NHÂN TẠO TRONG PHÂN TÍCH THÔNG TIN TÀI CHÍNH

Giảng viên hướng dẫn	**T.S.Nguyễn Quang Hưng**
Danh sách sinh viên thực hiện	**Mã sinh viên**
Lại Trung Lâm	**B21DCCN476**
Phạm Việt Hoàng	**B21DCCN393**
Vũ Hữu Hoài Linh	**B21DCCN489**

Hà Nội – 2025

LỜI NÓI ĐẦU

Trong thời đại số hóa hiện nay, việc ứng dụng trí tuệ nhân tạo (AI) vào các lĩnh vực chuyên ngành như tài chính – chứng khoán đang ngày càng trở nên phổ biến và cấp thiết. Với sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (LLM), việc xây dựng các hệ thống trợ lý ảo thông minh để hỗ trợ người dùng xử lý, phân tích và tương tác với dữ liệu tài chính đang là một hướng đi đầy tiềm năng.

Với mong muốn tiếp cận thực tế, áp dụng những kiến thức chuyên ngành đã học vào một dự án cụ thể, tôi đã lựa chọn tham gia vào quá trình phát triển hệ thống **Fin Chatbot** – một ứng dụng sử dụng trí tuệ nhân tạo để hỗ trợ người dùng truy vấn và phân tích thông tin tài chính một cách thông minh và tự động.

Báo cáo thực tập này được thực hiện tại **Công ty Cổ phần Goline**, dưới sự hướng dẫn tận tình của **Thầy T.S. Nguyễn Quang Hưng** cùng đội ngũ kỹ thuật tại doanh nghiệp. Báo cáo không chỉ là kết quả học tập và rèn luyện của tôi trong suốt quá trình thực tập, mà còn là dấu mốc quan trọng giúp tôi hiểu rõ hơn về quy trình phát triển ứng dụng AI trong môi trường thực tế.

Tôi xin chân thành cảm ơn thầy cô, quý doanh nghiệp và các anh chị kỹ thuật đã tạo điều kiện và hỗ trợ tôi hoàn thành tốt đợt thực tập này.

LỜI CẢM ƠN

Trước hết, tôi xin gửi lời cảm ơn chân thành và sâu sắc đến **Thầy T.S. Nguyễn Quang Hưng** – người đã tận tình hướng dẫn, định hướng và đồng hành cùng tôi trong suốt quá trình thực hiện đề tài thực tập. Những ý kiến đóng góp quý báu và sự nhiệt huyết của thầy đã giúp tôi từng bước hoàn thiện nội dung chuyên môn và nâng cao năng lực nghiên cứu thực tế.

Tôi cũng xin chân thành cảm ơn **Công ty Cổ phần Goline**, nơi tôi đã có cơ hội được tiếp cận thực tiễn phát triển sản phẩm ứng dụng trí tuệ nhân tạo trong ngành tài chính. Đặc biệt, xin gửi lời cảm ơn đến anh/chị mentor và các thành viên trong nhóm dự án **Fin Chatbot** đã luôn hỗ trợ, hướng dẫn kỹ thuật, chia sẻ kinh nghiệm quý báu và tạo điều kiện để tôi hoàn thành nhiệm vụ được giao.

Cuối cùng, tôi xin cảm ơn quý thầy cô trong khoa Công nghệ Thông tin cùng gia đình, bạn bè đã luôn ủng hộ và động viên tôi trong suốt quá trình học tập và thực tập vừa qua.

Chúng em xin chân thành cảm ơn!

Mục lục

1. Giới thiệu chung về nơi thực tập

1.1 Thông tin công ty Goline

Công ty Cổ phần Goline (Goline Corporation) được thành lập vào tháng 02 năm 2010, là doanh nghiệp công nghệ chuyên cung cấp giải pháp phần mềm Fintech cho thị trường tài chính và chứng khoán tại Việt Nam và khu vực. Sau hơn 15 năm hoạt động và phát triển, Goline đã trở thành một trong những đơn vị uy tín hàng đầu trong lĩnh vực này.

Thông tin cơ bản về công ty Goline:

- Tên công ty: Công ty Cổ phần Goline (Goline Corporation)
- Năm thành lập: 02/2010
- Vốn điều lệ: 15.000.000.000 VNĐ
- Chủ tịch/CEO: MBA. Lê Ngọc Tuấn
- Trụ sở chính: Tầng 8, Toà nhà Kim Ánh, 78 Duy Tân, Cầu Giấy, Hà Nội
- Chi nhánh:
    - Đà Nẵng: Tầng 5, 15 Quang Trung, Hải Châu
    - Hồ Chí Minh: Tầng 3, 20/13 Nguyễn Trường Tộ, Q.4
    - Nhật Bản: Tầng 7, Suishin Building, Yokohama, Kanagawa
- Website: https://goline.vn
- Email: contact@goline.vn
- Hotline: 0903 304 888

1.2 Cơ cấu tổ chức và lĩnh vực hoạt động

Goline hoạt động theo mô hình tổ chức chuyên sâu với nhiều phòng ban và nhóm kỹ thuật, được phân chia rõ ràng theo từng chức năng và dự án. Cơ cấu tổ chức tiêu biểu bao gồm:

- Ban Giám đốc điều hành

- Bộ phận phát triển sản phẩm (Core Systems, Trading Platform, AI, Data Engineering,...)
- Bộ phận kiểm thử và chất lượng phần mềm (QA/QC)
- Bộ phận tư vấn nghiệp vụ tài chính – chứng khoán
- Bộ phận hỗ trợ kỹ thuật và khách hàng
- Bộ phận đào tạo và tuyển dụng
- Các nhóm dự án chuyên trách theo từng khách hàng hoặc giải pháp

Lĩnh vực hoạt động chính:

- Tư vấn và phát triển phần mềm Fintech:
  Goline phát triển hệ thống giao dịch chứng khoán bao gồm cả thị trường cơ sở, OTC, phái sinh và trái phiếu, tích hợp đầy đủ các phân hệ từ Front Office đến Back Office.
- Sản phẩm AI và hệ thống hỗ trợ đầu tư:
  Một số hệ thống tiêu biểu như Gaia Advisor (nền tảng tư vấn đầu tư bằng AI), Robot Trading, Phân tích hành vi người dùng, và các nền tảng Copy Trade, Algo Trading.
- Tư vấn và triển khai giải pháp IT cho tổ chức tài chính:
  Goline xây dựng và triển khai giải pháp cho hơn 25 công ty chứng khoán tại Việt Nam, kết nối đa dạng với các hệ thống bên ngoài như ngân hàng, sở giao dịch, trung tâm lưu ký...
- Dịch vụ gia công phần mềm (Outsource):
  Hợp tác cùng các đối tác trong và ngoài nước để phát triển giải pháp phần mềm chuyên biệt cho thị trường tài chính.

1.3 Mô tả nhóm dự án thực tập

Trong kỳ thực tập, tôi được phân công vào nhóm phát triển ứng dụng Fin Chatbot – một dự án trí tuệ nhân tạo (AI) với mục tiêu xây dựng một hệ thống

chatbot tài chính thông minh, có khả năng hỗ trợ người dùng tra cứu thông tin, phân tích cảm xúc thị trường và trả lời các câu hỏi liên quan đến chứng khoán – tài chính.

Nhóm dự án bao gồm:

- Mentor: Trực tiếp hướng dẫn kỹ thuật và giám sát tiến độ
- Leader: Quản lý nhóm, phân chia nhiệm vụ và định hướng dự án
- Các thành viên: Gồm các sinh viên thực tập (trong đó có tôi), đảm nhận các vai trò như: thu thập dữ liệu, xử lý ngôn ngữ tự nhiên, triển khai mô hình AI, xây dựng backend API, kiểm thử hệ thống,...

Các công việc nhóm đang triển khai:

- Cào dữ liệu từ các trang web tài chính lớn như CafeF, Vietstock
- Tiền xử lý dữ liệu và lưu trữ vào cơ sở dữ liệu
- Phân tích cảm xúc các bài viết tài chính (sentiment analysis)
- Triển khai mô hình truy vấn văn bản thông minh (LLM hoặc RAG)
- Xây dựng hệ thống chatbot có khả năng trả lời theo ngữ cảnh người dùng

Nhóm sử dụng các công nghệ hiện đại như: Python, FastAPI, PostgreSQL, Qdrant, Langchain, HuggingFace Transformers, Docker, và các mô hình pretrained như PhoBERT, FinBERT, T5,...

2. Nội dung công việc trong kỳ thực tập

2.1 Mô tả công việc chính

Trong kỳ thực tập tại Công ty Cổ phần Goline, tôi được tham gia vào dự án xây dựng Fin Chatbot – một hệ thống trợ lý ảo ứng dụng trí tuệ nhân tạo nhằm hỗ trợ người dùng trong lĩnh vực tài chính – chứng khoán. Mục tiêu của hệ thống là trả lời các câu hỏi của người dùng một cách chính xác, nhanh chóng và có khả năng hiểu ngữ cảnh cũng như phân tích nội dung tài chính.

Các công việc chính tôi đảm nhiệm trong quá trình thực tập bao gồm:

- Thu thập dữ liệu (Web Crawling):
  Xây dựng chương trình cào dữ liệu tự động từ các trang web tài chính lớn như CafeF và Vietstock, bao gồm tiêu đề, nội dung, thời gian, danh mục và các chỉ số liên quan.
- Tiền xử lý và làm sạch dữ liệu:
  Thực hiện các bước làm sạch văn bản, loại bỏ ký tự thừa, chuẩn hóa tiếng Việt, phân tách câu, và lưu trữ vào hệ quản trị cơ sở dữ liệu PostgreSQL.
- Phân tích cảm xúc tài chính (Sentiment Analysis):
  Áp dụng mô hình học sâu (deep learning) để gán nhãn cảm xúc (tích cực – tiêu cực – trung lập) cho các bài viết tài chính, phục vụ cho việc đánh giá xu hướng thị trường.
- Xây dựng hệ thống truy vấn thông minh:
  Tham gia thiết kế kiến trúc RAG (Retrieval-Augmented Generation) để hệ thống có khả năng truy xuất thông tin theo ngữ nghĩa và trả lời dựa trên các văn bản đã thu thập.
- Kiểm thử và đánh giá mô hình:
  Đánh giá độ chính xác của mô hình phân tích cảm xúc và khả năng phản hồi đúng câu hỏi của chatbot.

2.2 Công nghệ và công cụ sử dụng

Trong quá trình thực tập, tôi đã tiếp cận và làm việc với nhiều công nghệ hiện đại thuộc lĩnh vực AI và lập trình ứng dụng. Dưới đây là các công nghệ và công cụ tiêu biểu:

Nhóm công nghệ	Công cụ / Thư viện cụ thể	Mục đích sử dụng
Ngôn ngữ lập trình	Python, SQL	Phát triển pipeline xử lý dữ liệu và backend

Nhóm công nghệ	Công cụ / Thư viện cụ thể	Mục đích sử dụng
Xử lý dữ liệu	Pandas, BeautifulSoup, Requests, LXML	Cào dữ liệu và tiền xử lý văn bản
Cơ sở dữ liệu	PostgreSQL	Lưu trữ dữ liệu văn bản đã xử lý
Vector database	Qdrant	Lưu trữ vector embedding cho hệ thống RAG
Xử lý ngôn ngữ (NLP)	Huggingface Transformers (PhoBERT, T5, FinBERT), underthesea	Phân tích cảm xúc, tóm tắt và hiểu văn bản
Triển khai backend	FastAPI, Uvicorn	Xây dựng API cho hệ thống Chatbot
Truy vấn ngữ nghĩa	Langchain	Kết nối LLM với dữ liệu tài chính
Công cụ hỗ trợ	Git, Docker, VS Code	Quản lý mã nguồn và môi trường phát triển

2.3 Quy trình phát triển hệ thống Chatbot

Nhóm dự án triển khai hệ thống Fin Chatbot theo quy trình phát triển phần mềm linh hoạt (Agile) với các giai đoạn cụ thể như sau:

1. Thu thập yêu cầu và phân tích nghiệp vụ:

   Xác định mục tiêu của chatbot, các loại câu hỏi cần hỗ trợ, các nguồn dữ liệu chính và luồng sử dụng của người dùng.

2. Cào và xử lý dữ liệu tài chính:

   Thu thập dữ liệu từ các nguồn uy tín (CafeF, Vietstock), thực hiện làm sạch và chuẩn hóa trước khi đưa vào hệ thống.

3. Vector hóa dữ liệu:
   Sử dụng các mô hình pretrained (PhoBERT, T5) để chuyển các đoạn văn bản thành vector embedding, lưu trữ vào Qdrant để hỗ trợ tìm kiếm ngữ nghĩa.
4. Phân tích cảm xúc và gán nhãn dữ liệu:
   Xây dựng mô hình classification hoặc regression để đánh giá mức độ tích cực/tiêu cực của từng bài viết, phục vụ cho các truy vấn mang tính phân tích.
5. Xây dựng chatbot bằng RAG:
   Triển khai pipeline RAG gồm: truy vấn → lọc top-k tài liệu → sinh câu trả lời bằng mô hình LLM (T5, GPT,...).
6. Triển khai API và giao diện thử nghiệm:
   Xây dựng các endpoint trên FastAPI để kết nối hệ thống với frontend hoặc người dùng thử nghiệm chatbot.
7. Đánh giá và cải tiến mô hình:
   Thu thập phản hồi từ người dùng, đánh giá độ chính xác và mức độ hài lòng để tinh chỉnh mô hình và hệ thống.

3. Giới thiệu ứng dụng AI: Fin Chatbot

3.1 Mục tiêu và chức năng chính của Chatbot

Fin Chatbot là một ứng dụng trí tuệ nhân tạo (AI) được thiết kế nhằm hỗ trợ người dùng trong lĩnh vực tài chính – chứng khoán bằng cách cung cấp thông tin nhanh chóng, chính xác và theo ngữ cảnh.

Mục tiêu chính:

- Hỗ trợ người dùng tra cứu dữ liệu tài chính như: diễn biến thị trường, thông tin doanh nghiệp, giá cổ phiếu,...

- Trả lời các câu hỏi tài chính một cách tự nhiên và thông minh, sử dụng mô hình ngôn ngữ lớn (LLM).
- Cung cấp phân tích cảm xúc về thị trường thông qua việc đánh giá tích cực, tiêu cực của các bài viết tài chính.
- Nâng cao trải nghiệm người dùng nhờ vào khả năng hiểu ngôn ngữ tự nhiên tiếng Việt, phản hồi linh hoạt và có tính ngữ cảnh.

Một số chức năng chính:

- Trả lời câu hỏi người dùng dựa trên thông tin cào từ CafeF, Vietstock (ví dụ: *"VNM có tin gì mới không?"*)
- Phân tích cảm xúc các bài viết về mã cổ phiếu cụ thể (ví dụ: *"Tâm lý thị trường về HPG hôm nay như thế nào?"*)
- Tóm tắt nội dung bài viết tài chính
- Gợi ý mã cổ phiếu đáng chú ý dựa trên dữ liệu tin tức gần đây

3.2 Pipeline hệ thống: Thu thập → Xử lý → Phân tích → Trả lời

Hệ thống Fin Chatbot được thiết kế theo một pipeline xử lý dữ liệu và phản hồi thông minh, bao gồm 4 giai đoạn chính:

1. Thu thập dữ liệu
   - Cào dữ liệu từ các trang web tài chính uy tín như CafeF và Vietstock
   - Dữ liệu bao gồm: tiêu đề, nội dung bài viết, thời gian đăng, mã cổ phiếu liên quan,...
2. Xử lý dữ liệu
   - Làm sạch dữ liệu: loại bỏ HTML, ký tự thừa, chuẩn hóa tiếng Việt
   - Phân tách văn bản, xử lý token, tách đoạn
   - Vector hóa văn bản: sử dụng mô hình như PhoBERT hoặc T5 để sinh embedding

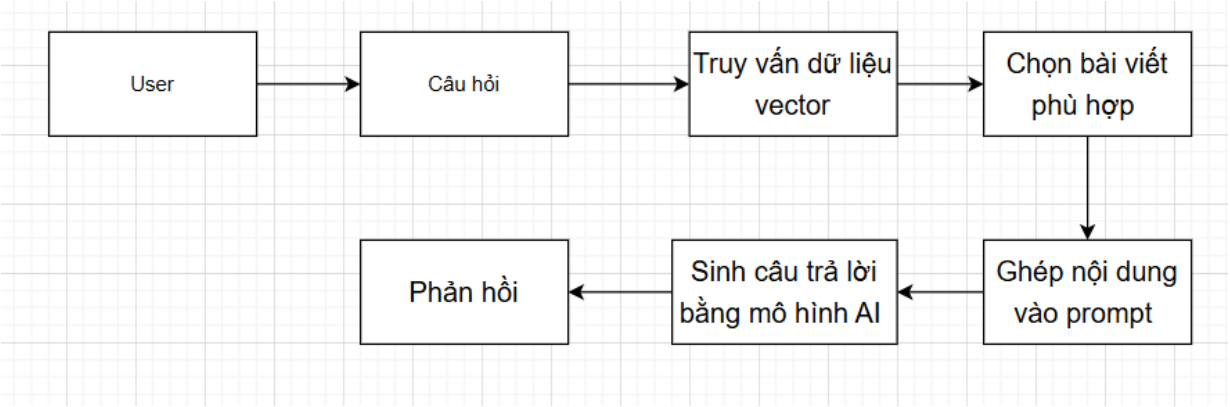
3. Phân tích dữ liệu

- Phân tích cảm xúc: sử dụng mô hình học sâu để đánh giá mức độ tích cực – tiêu cực của các bài viết
- Tìm kiếm ngữ nghĩa: khi người dùng đặt câu hỏi, hệ thống truy xuất các đoạn văn liên quan từ vector database (Qdrant)

4. Sinh câu trả lời

- Kết hợp tài liệu liên quan + câu hỏi của người dùng
- Sinh câu trả lời tự nhiên bằng mô hình LLM (ví dụ: T5, GPT, hoặc mô hình fine-tune)
- Trả kết quả lại qua API hoặc giao diện người dùng

Tổng quan quy trình:



Hình 1. quy trình thực hiện

3.3 Vai trò của AI trong từng giai đoạn

Giai đoạn	Công nghệ AI sử dụng	Vai trò cụ thể
Xử lý ngôn ngữ	Tokenization, POS tagging, sentence splitting (Underthesea, PhoBERT)	Làm sạch và phân tích cấu trúc văn bản
Phân tích cảm xúc	Mô hình phân loại cảm xúc (ViBERT, FinBERT, T5)	Gán nhãn bài viết: tích cực, tiêu cực, trung lập

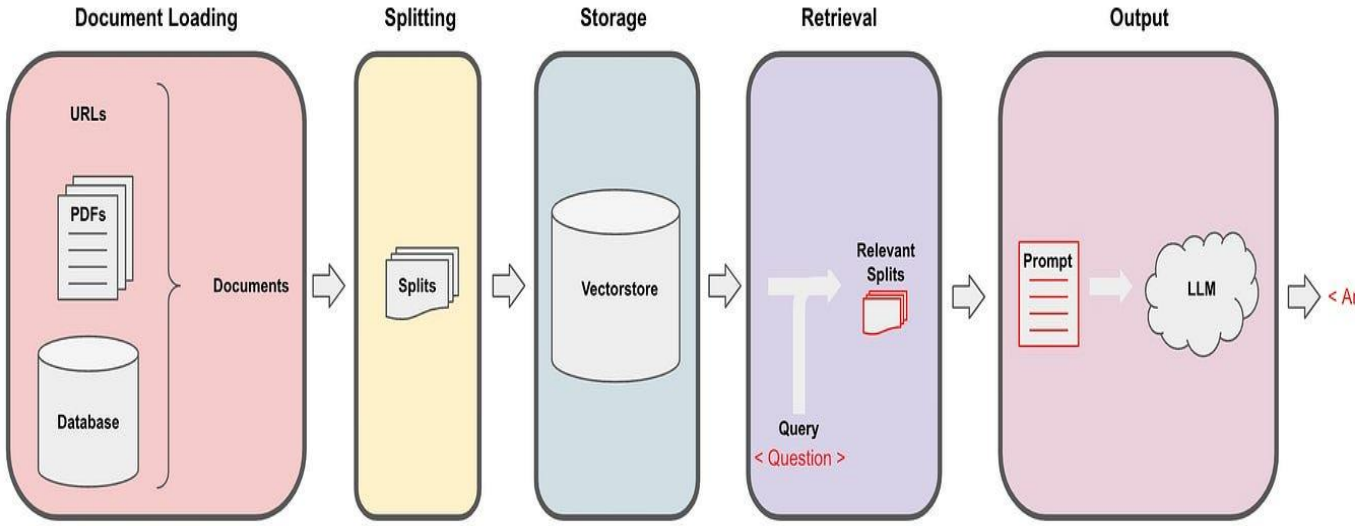
Giai đoạn	Công nghệ AI sử dụng	Vai trò cụ thể
Vector hóa văn bản	PhoBERT, T5 encoder	Chuyển văn bản thành vector phục vụ tìm kiếm ngữ nghĩa
Tìm kiếm ngữ nghĩa	Qdrant, cosine similarity	Tìm đoạn văn phù hợp nhất với câu hỏi người dùng
Sinh câu trả lời	LLM (T5, GPT, Mistral, Qwen...)	Tạo ra phản hồi tự nhiên, có ngữ cảnh và chính xác
Tóm tắt nội dung	Sequence-to-sequence models	Rút gọn bài viết tài chính thành đoạn mô tả ngắn gọn

4. Khả năng triển khai mô hình AI cụ thể

4.1. Xử lý ngôn ngữ tự nhiên(NPL) và trích rút thông tin

Bước	Tác vụ	Thư viện/Mô hình	Ghi chú triển khai
1. Làm sạch thô	Loại bỏ HTML, ký tự lạ, chuẩn hóa Unicode	*BeautifulSoup, regex*	Tốc độ ~15 MB/s trên 8 CPU
2. Tách câu & token	Sentence splitter, word tokenizer	underthesea, VnCoreNLP, PhoBERT tokenizer	Dùng SentencePiece của PhoBERT giúp đồng nhất vocab
3. POS / DEP	POS-tag, dependency parse	VnCoreNLP server	Lấy quan hệ (subject–verb–object) để trích insight "doanh thu *tăng* 15%"
4. Nhận dạng thực thể	NER cho ticker, công ty, chỉ số	Fine-tune *PhoBERT-large* (115M param) + CRF	9 nhãn tùy chỉnh: ORG, TICKER, KPI, TIME, …

Bước	Tác vụ	Thư viện/Mô hình	Ghi chú triển khai
5. Trích xuất quan hệ	Rule-based + biaffine relation extractor	PyTorch Lightning pipeline	Xác định cặp <KPI, Giá trị>
6. Tóm tắt & rút trích	Extractive trước, abstractive sau	*T5-base vi* (LoRA-8bit)	Thời gian suy luận < 0.4 s/bài 1.200 từ



Hình 2.Quy trình xử lý văn bản tài chính tiếng Việt

Điểm mấu chốt khi xử lý tiếng việt trong tài chính:

- *Từ kép* (vd. "lợi nhuận gộp") quan trọng – tokenizer phải giữ nguyên cụm.
- Ticker nhiều dạng: "HPG", "[HPG]", "HPG.BD"; dùng regex + bản đồ mã chứng khoán để chuẩn.
- Tách đoạn theo *itemization* (–, •) để giữ cấu trúc báo cáo.

Một vài lưu ý khi triển khai:

- Triển khai *streaming pre-processing* bằng *Pandas → PyArrow* để tránh nghẽn RAM khi thu thập >3 GB/ngày.
- Lưu log thống kê số thực thể/tài liệu giúp theo dõi độ phủ NER (hiện đạt 92% bài báo chứa ≥1 thực thể chính).

4.2. Phân tích cảm xúc và đánh giá nội dung tài chính

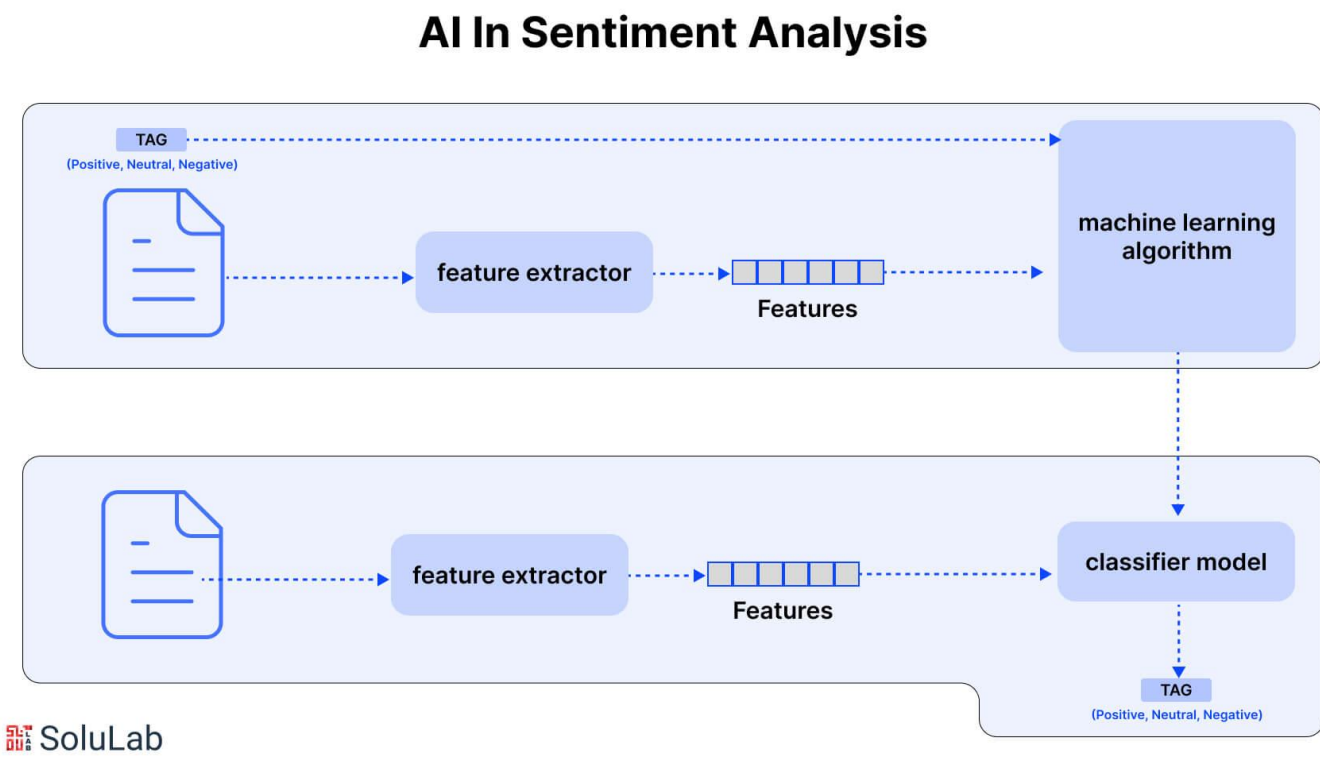
Qúa trình này sẽ trải qua 4 bước chính:

1. Tạo bộ dữ liệu nhãn dán:

- 50 k bài viết (2018–2024) lấy từ CafeF, Vietstock.
- Semi-supervised: dùng *FinBERT-english* suy diễn nháp → chuyên viên chỉnh tay 10 %.
- Phân bố nhãn: Pos 35 %, Neu 45 %, Neg 20 %.

2. Fine-tune FinBert-vi

- Khởi tạo từ *FinBERT* gốc (Devlin-base 110M), chuyển PhoBERT + continual-pretrain 3 epochs trên 4×A100.
- Hyper-params: LR 2e-5, batch 32, max_len 256, focal-loss ($\gamma = 2$) để cân bằng nhãn.
- Độ chính xác macro-F1 0.816 trên tập test 5 k.



Hình 3.Quy trình phân tích cảm xúc bằng machine learning

3. Triển khai real-time

- Đóng gói model bằng *ONNX Runtime* + quantize_dynamic (INT8) giảm độ trễ xuống 23 ms/req.
- Cache embedding câu ngắn để phục vụ dashboard "heat-map cảm xúc theo mã".

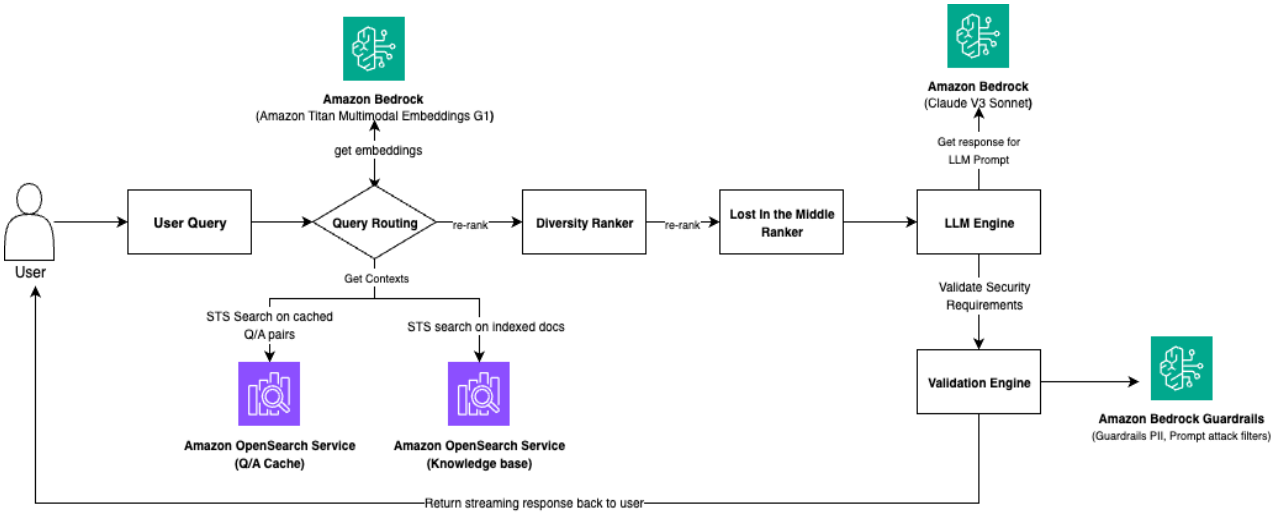
4. Phân tích cảm xúc tập hợp(Aggregate Sentiment)

- Gộp theo mã cổ phiếu × khung thời gian (15 ph, 1 h, 1 ngày).
- Chỉ số *Sentiment Score* = (Pos – Neg) / (P+N+0.5) – scale về [-1,1].
- Vẽ đường trung bình động 7 ngày để cảnh báo đảo chiều cảm xúc.

4.3. Tìm kiếm thông minh và truy vấn ngữ nghĩa

1. Kiến trúc tổng thể

- Chunking : Cắt văn bản ~ 200 token (overlap 40).
- Embdding : Dùng *PhoBERT-large* hoặc *F-T5 encoder* → vector 768-/1024-d.
- Vector store: *Qdrant* (Docker) – p2p clustering, HNSW
- Retriever: Cosine top-k (k = 6) + Re-rank (*bge-reranker-base*)
- Generator : LLM (T5-base vi) nhận prompt [context] + [question]



Hình 4.Kiến trúc tổng thể của hệ thống truy vấn ngữ nghĩa kết hợp LLM

2. Chi tiết tối ưu

Thành phần	Thủ thuật	Kết quả
Embedding throughput	Batch 512, FP16 trên A100	430 doc/s
Index size	Mã hoá payload dưới 1 kB, nlist = 1024	3.1 M vector / 26 GB
Cold-start latency	Warm-up 1 k truy vấn đầu	< 280 ms
Độ chính xác	R@5 = 0.87 trên bộ query-set 1 000 câu	Cao hơn 6 % so với BM25 thuần

4.4. Triển khai mô hình dựa trên ngữ cảnh (LLM)

1. Lựa chọn mô hình và fine-tune

Mô hình	Params	Kỹ thuật	Lợi ích
T5-base vi	220 M	LoRA-rank 32, $\alpha = 16$	Dễ fine-tune, phục vụ tiếng Việt
Mistral-7B-Instruct	7 B	QLoRA-4-bit	Ngữ cảnh dài 16 k token
GPT-4o API	–	Prompt-only	Chất lượng cao, chi phí cao

2. Kiểm soát chất lượng đầu ra

- Guardrail: Regex check số liệu, từ nhạy cảm trước khi gửi.

- Self-consitency: Gọi model 3 lần, chọn câu trả lời có cosine với embedding ngữ nghĩa cao nhất.
- Fact-check: Nếu câu trả lời chứa số % hoặc "lợi nhuận", đối chiếu lại tài liệu gốc, chỉ giữ khi sai số < 0.5 %.

3. Đánh giá

- Automatic: BLEU-4 = 24.6; ROUGE-L = 31.8 trên 800 QA tài chính.
- Human: 5 chuyên gia chứng khoán, chấm 1-5: *Overall helpfulness* 4.3.
- Latency: 620 ms (retrieval 170 ms, generation 450 ms).

4. Triển khai và mở rộng:

- Backend : FastAPI + Uvicorn worker 4 × CPU; model served qua *vLLM* (CUDA-shared) ⇒ 150 RPS.
- Monitoring: Prometheus + Grafana, theo dõi tPS, token-out, prompt-usage.
- Canary release: Versioning API /v1 ↔ /v2, 10 % người dùng thử trước, rollback khi error rate > 5 %.

5. Thách thức và hướng phát triển

5.1. Những vấn đề trong thực tiễn triển khai

Trong quá trình xây dựng và triển khai hệ thống Fin Chatbot tại Công ty Goline, nhóm dự án đã gặp phải một số khó khăn đáng kể, đặc biệt là khi làm việc với dữ liệu tiếng Việt chuyên ngành tài chính và tích hợp các mô hình AI vào môi trường ứng dụng thực tế. Các thách thức này được chia thành 3 nhóm chính:

1. Thách thức về dữ liệu
    - Dữ liệu tiếng việt thiếu chuẩn hóa: Các bài viết trên CafeF, Vietstock, báo tài chính sử dụng nhiều cách viết không đồng nhất, có thể viết tắt, viết sai chính tả, hoặc chèn ký hiệu đặc biệt. Điều này gây khó khăn trong bước phân tách từ, gán nhãn thực thể, và phân tích cảm xúc.

- Thiếu dữ liệu huấn luyện chất lượng cao: Các tập dữ liệu mở về tài chính tiếng Việt hầu như không có. Nhóm buộc phải tự tạo dữ liệu thông qua crawling và semi-supervised labeling (ví dụ như sentiment). Điều này đòi hỏi nhiều công sức hiệu chỉnh, gán nhãn tay và đánh giá thủ công.
- Thông tin nhanh chóng lỗi thời: Tin tức tài chính thay đổi hàng ngày, khiến các kết quả embedding, phân tích cảm xúc hoặc thống kê có thể nhanh chóng trở nên lỗi thời nếu không cập nhật liên tục.

2. Thách thức về mô hình và hiệu suất
    - Embedding tiếng việt tài chính còn giới hạn: Các mô hình như PhoBERT, FinBERT Việt hóa tuy hữu ích, nhưng vẫn chưa nắm bắt tốt ngữ cảnh chuyên ngành, khiến việc truy vấn ngữ nghĩa và sinh câu trả lời gặp khó khăn.
    - Tốc độ phản hồi bị ảnh hưởng bởi mô hình lớn: LLM như T5 hoặc GPT có độ trễ cao, đặc biệt là khi sinh văn bản dài. Trong các hệ thống có nhiều người dùng, nếu không tối ưu GPU batching hoặc dùng mô hình nhỏ hơn thì latency có thể vượt 2–3 giây.
    - Tối ưu RAG phức tạp: Việc kết hợp truy xuất tài liệu (Qdrant) và sinh câu trả lời (LLM) đòi hỏi phải quản lý tốt pipeline, tránh lỗi không đồng bộ giữa embedding và câu hỏi người dùng. Nếu hệ thống không chọn đúng đoạn văn phù hợp, thì LLM sẽ sinh ra kết quả thiếu chính xác.
3. Thách thức về triển khai thực tế
    - Vấn đề bảo mật dữ liệu: Tin tức và báo cáo tài chính có thể chứa thông tin nhạy cảm (bên thứ 3, đối tác, công ty). Nếu không phân quyền hoặc kiểm soát truy cập, chatbot có thể trả lời những thông tin không được phép công bố.

- Xử lý truy vấn đa dạng từ người dùng: Người dùng có thể đặt câu hỏi mơ hồ, nhiều nghĩa, hoặc sai chính tả. Chatbot cần có cơ chế xử lý lỗi chính tả, phát hiện intent không rõ ràng, hoặc phản hồi lịch sự nếu không hiểu rõ yêu cầu.
- Khả năng kiểm soát LLM còn hạn chế: Một số phản hồi từ mô hình sinh có thể không đúng thực tế, hoặc tạo ra thông tin "ảo" (hallucination), ảnh hưởng nghiêm trọng đến uy tín nếu chatbot được triển khai chính thức trong môi trường doanh nghiệp tài chính.

5.2. Định hướng mở rộng và cải tiến ChatBot

Để nâng cấp Fin Chatbot thành một trợ lý tài chính ảo đáng tin cậy và có thể ứng dụng thực tế trong doanh nghiệp hoặc công ty chứng khoán, nhóm đưa ra các hướng phát triển cụ thể sau:

1. Mở rộng khả năng ngôn ngữ và hiểu biết chuyên ngành
    - Fine-tune mô hình ngôn ngữ trên dữ liệu chuyên ngành tài chính – chứng khoán tiếng Việt: Sử dụng các mô hình như Qwen-1.8B, Mistral hoặc T5, và huấn luyện liên tục (continual pretraining) trên các báo cáo tài chính, tin tức cổ phiếu, bài phân tích để cải thiện độ chính xác khi sinh câu trả lời.
    - Tăng cường khả năng hiểu câu hỏi đa dạng: Tích hợp các bước chuẩn hóa truy vấn đầu vào như: sửa lỗi chính tả, phân tích intent, phân loại loại câu hỏi (mã cổ phiếu, diễn biến, phân tích kỹ thuật,...).
2. Phát triển các tính năng nâng cao
    - Thêm tính năng dự đoán xu hướng cảm xúc: Ví dụ: “Tâm lý thị trường về HPG tuần tới có khả năng tiếp tục tích cực không?”, chatbot có thể phân tích dữ liệu 7 ngày gần nhất và dự đoán xu hướng.
    - Tích hợp chatbot vào hệ thống cảnh báo: Gửi thông báo khi xuất hiện các tin tiêu cực đột ngột về mã cổ phiếu, hoặc khi có nhiều bài viết tích cực bất thường.

- Hỗ trợ giọng nói và giao tiếp đa phương thức: Tích hợp với dịch vụ chuyển giọng nói → văn bản để người dùng có thể hỏi bằng tiếng nói (ứng dụng di động hoặc web).

3. Tối ưu hóa hiệu suất hệ thống
    - Triển khai kỹ thuật caching & vector prefetching: Giảm thời gian truy xuất văn bản cho các câu hỏi thường gặp hoặc mã cổ phiếu phổ biến.
    - Chuyển mô hình sang ONNX hoặc dùng quantization (INT8, QLoRA) để giảm độ trễ phản hồi.
    - Tự động làm mới dữ liệu Qdrant mỗi ngày và tái tính embedding chỉ khi có bài viết mới (theo mã cổ phiếu) để tiết kiệm tài nguyên.
4. Tăng cường bảo mật và kiểm soát phản hồi
    - Tích hợp Guardrails: Thiết lập quy tắc cứng để ngăn chatbot trả lời các câu hỏi ngoài phạm vi, hoặc câu hỏi chứa nội dung nhạy cảm (ví dụ: "cổ phiếu nào chắc chắn tăng?").
    - Thêm lớp kiểm duyệt nội dung sinh bởi LLM: Trước khi phản hồi tới người dùng, hệ thống sẽ kiểm tra logic và đối chiếu lại với dữ liệu gốc.

5.3. Tính ứng dụng trong ngành tài chính- chứng khoán

Fin Chatbot thể hiện tiềm năng rất lớn trong lĩnh vực tài chính – chứng khoán, đặc biệt trong bối cảnh ngành đang số hóa mạnh mẽ và nhu cầu thông tin nhanh chóng, chính xác ngày càng cao.

Các ứng dụng tiềm năng:

- Hỗ trợ các nhà đầu tư cá nhân: Truy vấn nhanh thông tin mã cổ phiếu, phân tích cảm xúc, tóm tắt tin tức.. Trả lời câu hỏi phổ biến: "HPG hôm nay có tin gì nổi bật?", "Tâm lý thị trường ra sao?",...
- Công cụ trợ lý cho chuyên viên môi giới: Giảm tải cho bộ phận môi giới khi phải xử lý nhiều yêu cầu trùng lặp từ khách hàng. Chatbot có thể gợi ý mã cổ phiếu theo tiêu chí người dùng đặt (ví dụ: "Tìm cổ phiếu tăng trưởng có PE < 10").

- Tích hợp vào hệ thống tư vấn tài chính tự động: Là thành phần giao tiếp chính trong các sản phẩm tư vấn đầu tư AI như Gaia Advisor. Có thể phối hợp với mô hình phân tích kỹ thuật để đưa ra khuyến nghị đầu tư.
- Ứng dụng trong đào tạo, truyền thông nội bộ: Trả lời các câu hỏi về kiến thức tài chính, thuật ngữ, tin tức nội bộ,... trong các công ty chứng khoán, ngân hàng.

Fin Chatbot không chỉ là công cụ hỏi – đáp, mà còn là nền tảng kết nối dữ liệu tài chính, mô hình AI và trải nghiệm người dùng thông minh. Trong tương lai, hệ thống hoàn toàn có thể mở rộng để:

- Phân tích kỹ thuật tự động (nhận diện mô hình nến, hỗ trợ – kháng cự)
- Dự báo xu hướng giá dựa trên LLM hoặc mô hình thống kê
- Hỗ trợ đa ngôn ngữ cho nhà đầu tư nước ngoài

6. Kết luận

6.1. Những gì đã học được trong kỳ thực tập

Kỳ thực tập tại Công ty Cổ phần Goline là một cơ hội quý giá để tôi tiếp cận thực tế, ứng dụng những kiến thức đã học và làm việc trong môi trường chuyên nghiệp về công nghệ tài chính. Qua thời gian gắn bó với dự án Fin Chatbot, tôi đã học được nhiều điều quan trọng:

1. Kiến thức chuyên môn
    - Hiểu rõ quy trình phát triển hệ thống ứng dụng AI: từ việc thu thập dữ liệu, xử lý ngôn ngữ tự nhiên, phân tích cảm xúc, đến triển khai mô hình sinh câu trả lời bằng LLM và tích hợp API.
    - Nắm bắt kiến trúc hệ thống truy vấn ngữ nghĩa hiện đại (RAG), kỹ thuật vector search với Qdrant, và cách sinh phản hồi tự nhiên từ mô hình T5/GPT.
    - Cải thiện tư duy lập trình Python chuyên sâu, đặc biệt là xử lý dữ liệu lớn, tối ưu hóa mô hình và viết API hiệu quả với FastAPI.

2. Kỹ năng làm việc thực tế
    - Thành thạo công cụ phát triển: VS Code, Git, Docker, PostgreSQL, HuggingFace Transformers.
    - Kỹ năng teamwork và Agile: Làm việc nhóm với vai trò chủ động, chia sẻ task qua Trello, cập nhật tiến độ qua daily meeting.
    - Hiểu được cách làm việc thực tế tại doanh nghiệp công nghệ, quy trình triển khai sản phẩm AI từ giai đoạn ý tưởng đến MVP.
3. Kiến thức ngành tài chính- chứng khoán:
    - Nắm được cấu trúc tin tức tài chính: cách phân tích một bài viết về cổ phiếu, nhận diện thông tin quan trọng (giá, chỉ số, doanh nghiệp liên quan).
    - Hiểu được hành vi người dùng đầu tư và cách chatbot có thể hỗ trợ nhà đầu tư trong việc tra cứu thông tin, đánh giá tâm lý thị trường.

6.2. Đề xuất và đóng góp cá nhân

Trong quá trình tham gia dự án, tôi không chỉ hoàn thành các nhiệm vụ được giao mà còn chủ động đề xuất và đóng góp nhiều giải pháp nhằm nâng cao chất lượng hệ thống. Những đóng góp tiêu biểu:

- Xây dựng pipeline crawling dữ liệu tài chính từ Vietstock, CafeF với khả năng thu thập hàng ngàn bài viết/ngày và lưu trữ vào PostgreSQL dưới dạng chuẩn hóa.
- Góp phần xây dựng mô hình phân tích cảm xúc tài chính, bao gồm khâu gán nhãn dữ liệu ban đầu, huấn luyện mô hình FinBERT Việt hóa, và đánh giá hiệu suất.
- Triển khai pipeline tìm kiếm ngữ nghĩa (Qdrant + Langchain) kết hợp với mô hình sinh phản hồi bằng T5, đảm bảo khả năng trả lời ngữ cảnh tốt và đúng dữ liệu thực.

- Tối ưu hóa API backend với FastAPI: Đề xuất và áp dụng kỹ thuật async để giảm thời gian phản hồi, hỗ trợ song song hóa truy vấn và sinh câu trả lời.
- Tham gia viết tài liệu kỹ thuật, hướng dẫn sử dụng API và mô hình chatbot cho các thành viên mới, hỗ trợ mentor quản lý kiến thức dự án.

Một số đề xuất:

- Xây dựng công cụ dashboard trực quan để theo dõi cảm xúc thị trường theo mã cổ phiếu.
- Triển khai mô hình kiểm duyệt phản hồi tự động từ LLM, nhằm đảm bảo tính chính xác và tránh thông tin không mong muốn khi sinh câu trả lời.
- Thử nghiệm tích hợp mô hình nhỏ hơn như Mistral-instruct hoặc Qwen-1.8B nhằm giảm latency khi triển khai thực tế.

6.3. Nhận xét tổng quan

Kỳ thực tập tại Goline là một trải nghiệm rất đáng giá, không chỉ giúp tôi củng cố kiến thức chuyên môn mà còn nâng cao tư duy ứng dụng AI vào lĩnh vực thực tiễn – đặc biệt là ngành tài chính – chứng khoán, vốn có nhiều đặc thù và đòi hỏi khắt khe về độ chính xác và cập nhật.

Tôi cảm thấy rất may mắn khi được làm việc trong một môi trường chuyên nghiệp, năng động và học hỏi được từ những anh/chị mentor, leader có chuyên môn cao và luôn sẵn sàng hỗ trợ.

Fin Chatbot là một dự án tiêu biểu thể hiện khả năng ứng dụng AI vào dịch vụ tài chính. Tôi tự hào vì đã đóng góp một phần vào quá trình xây dựng hệ thống này. Dù thời gian thực tập đã kết thúc, nhưng những kinh nghiệm tôi thu nhận được sẽ là hành trang quý giá cho hành trình học tập và phát triển nghề nghiệp trong tương lai.

7. Tài liệu tham khảo

- Hugging Face Transformers – Thư viện mã nguồn mở cho các mô hình NLP hiện đại
  → https://huggingface.co/transformers/
- PhoBERT: Pre-trained language models for Vietnamese
  → Nguyen, Dat Quoc et al. (2020). EMNLP Findings.
  → https://github.com/VinAIResearch/PhoBERT
- LangChain – Framework for building LLM applications
  → https://www.langchain.com
- Qdrant – Vector Search Engine
  → https://qdrant.tech
- Vietstock.vn – Dữ liệu doanh nghiệp, thị trường chứng khoán
  → https://vietstock.vn
- PostgreSQL – Hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở
  → https://www.postgresql.org
- Docker – Containerized deployment
  → https://www.docker.com
- FinBERT for Financial Sentiment Analysis
  → Araci, D. (2019).
  → https://github.com/ProsusAI/finBERT