GPT-INVESTAR: Nâng cao chiến lược đầu tư cổ phiếu thông qua phân tích báo cáo thường niên bằng mô hình ngôn ngữ lớn

Người thực hiện báo cáo:
Vũ Bảo Chinh – 23020341
Phạm Tiến Dũng – 23020345
Lê Hồng Anh – 23020327

Ngày 18 tháng 6 năm 2025

Tóm tắt nội dung

Báo cáo thường niên của các công ty niêm yết công khai đóng vai trò then chốt trong việc cung cấp thông tin quan trọng về tình hình tài chính của doanh nghiệp, từ đó có thể giúp đánh giá tác động tiềm năng đến giá cổ phiếu của công ty. Những báo cáo này có tính chất toàn diện và chi tiết, thường có độ dài từ 100 trang trở lên, thậm chí có thể vượt quá con số này đáng kể. Việc phân tích các báo cáo này là một nhiệm vụ cực kỳ phức tạp và tốn thời gian, ngay cả đối với một công ty duy nhất, chưa kể đến toàn bộ vũ trụ các doanh nghiệp tồn tại trên thị trường. Trong suốt nhiều năm qua, các chuyên gia tài chính đã trở nên thành thạo trong việc trích xuất thông tin có giá trị từ những tài liệu này một cách tương đối nhanh chóng. Tuy nhiên, khả năng này đòi hỏi nhiều năm thực hành và kinh nghiệm tích lũy. Nghiên cứu này nhằm mục đích đơn giản hóa quy trình đánh giá báo cáo thường niên của tất cả các công ty bằng cách tận dụng khả năng của các Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs). Các thông tin chi tiết được tạo ra bởi LLM được biên soạn thành một bộ dữ liệu theo phong cách Quant (định lượng) và được bổ sung bằng dữ liệu giá cổ phiếu lịch sử. Sau đó, một mô hình Học máy (Machine Learning) được huấn luyện với các đầu ra của LLM làm đặc trưng (features). Kết quả kiểm tra walk-forward cho thấy hiệu suất vượt trội đầy hứa hẹn so với lợi nhuận của chỉ số SP 500. Nghiên cứu này có ý định cung cấp một khung làm việc cho các nghiên cứu tương lai theo hướng này. Để tạo điều kiện thuận lợi cho việc này, mã nguồn đã được phát hành dưới dạng mã nguồn mở

Từ khoá: ChatGPT, LLM, Stocks, Investing, Quantitative Finance

Github: https://github.com/UditGupta10/GPT-InvestAR.

Mục lục

1 Giới thiệu

Trong khuôn khổ nghiên cứu này, tác giả tập trung phân tích các cổ phiếu được niêm yết trên thị trường chứng khoán Hoa Kỳ – một trong những thị trường tài chính lớn nhất và minh bạch nhất trên thế giới. Cụ thể, tập dữ liệu được lựa chọn bao gồm 1.500 công ty đại diện cho ba nhóm vốn hóa chính trên thị trường. Trong thị trường tài chính, vốn hóa thị trường (market capitalization) của một công ty là tổng giá trị thị trường của tất cả cổ phiếu đang lưu hành của công ty đó.

- SP 500 : Top 500 công ty vốn hóa lớn
- SP 400 : Top 400 công ty vốn hoá trung bình
- SP 600 : Top 600 công ty vốn hoá nhỏ

Việc lựa chọn này không chỉ đảm bảo tính đa dạng về quy mô doanh nghiệp mà còn tạo điều kiện để mô hình có thể học được từ các mẫu báo cáo với phong cách và cấu trúc khác nhau, từ đó nâng cao khả năng khái quát hóa trong các ứng dụng thực tiễn.

Tại Hoa Kỳ, các công ty đại chúng có nghĩa vụ pháp lý phải nộp báo cáo thường niên toàn diện lên Ủy ban Chứng khoán và Giao dịch Hoa Kỳ (U.S. Securities and Exchange Commission – SEC), thông qua biểu mẫu gọi là hồ sơ 10-K. Đây là một yêu cầu bắt buộc theo quy định của Đạo luật Giao dịch Chứng khoán năm 1934, với mục tiêu đảm bảo tính minh bạch và bảo vệ quyền lợi của nhà đầu tư. Hồ sơ 10-K đóng vai trò như một bản hồ sơ định kỳ mô tả chi tiết tình hình tài chính, kết quả kinh doanh, chiến lược hoạt động và các yếu tố rủi ro mà doanh nghiệp đang hoặc có thể sẽ đối mặt. Đặc biệt, vì được chuẩn hóa theo một khuôn mẫu chung do SEC quy định, các báo cáo này giúp nhà đầu tư dễ dàng so sánh giữa các công ty khác nhau trong cùng ngành hoặc giữa các thời kỳ khác nhau.

Một báo cáo 10-K điển hình bao gồm các phần tài chính định lượng như: bảng cân đối kế toán (balance sheet), báo cáo kết quả hoạt động kinh doanh (income statement), báo cáo lưu chuyển tiền tệ (cash flow statement) – những tài liệu cốt lõi cho việc đánh giá sức khỏe tài chính của doanh nghiệp. Tuy nhiên, giá trị của hồ sơ 10-K không chỉ dừng lại ở các con số kế toán. Các phần nội dung định tính như: Tổng quan về doanh nghiệp (Business Overview), Các yếu tố rủi ro (Risk Factors), Phân tích và thảo luận của ban lãnh đạo (Management's Discussion and Analysis – MDA), hay Các tranh chấp pháp lý (Legal Proceedings) thường chứa đựng những thông tin chiến lược và vi mô có ảnh hưởng lớn đến đánh giá dài hạn của doanh nghiệp. Điều đáng chú ý là các nội dung này thường mang tính chủ quan, phụ thuộc vào văn phong diễn đạt, ngữ cảnh ngành nghề và cách giải thích của người đọc. Chính vì vậy, việc phân tích toàn diện hồ sơ 10-K đòi hỏi không chỉ kỹ năng tài chính, mà còn là khả năng đọc hiểu ngôn ngữ tự nhiên và nhận diện hàm ý ẩn sau văn bản.

Trong bối cảnh đó, sự phát triển nhanh chóng của các Mô hình Ngôn ngữ Lớn (Large Language Models – LLMs) như GPT-3.5 (còn được biết đến qua sản phẩm ChatGPT của OpenAI) đã mở ra khả năng mới trong việc xử lý và hiểu các tài liệu phi cấu trúc dài dòng như hồ sơ 10-K. Những mô hình này, được huấn luyện trên hàng trăm tỷ token ngôn ngữ tự nhiên, có thể thực hiện các tác vụ như tóm tắt, trả lời câu hỏi, phân loại nội dung và trích xuất thông tin một cách linh hoạt và chính xác [1]. Trong lĩnh vực tài chính, nghiên cứu của Lopez-Lira và Tang (2023) [2] đã cung cấp bằng chứng thực nghiệm cho thấy các mô hình LLM có thể phân tích các báo cáo tài chính và dự đoán được phản ứng của thị trường chứng khoán trước thông tin mới.

Dựa trên những tiến bộ này, nghiên cứu hiện tại đề xuất áp dụng mô hình ngôn ngữ để hỗ trợ nhà đầu tư và chuyên gia phân tích tài chính trong việc hiểu và khai thác nội dung từ hồ sơ 10-K. Thay vì yêu cầu mô hình đọc toàn bộ văn bản hoặc xử lý các chỉ số đơn lẻ, chúng tôi khai thác khả năng của LLM để trả lời các câu hỏi định tính phức tạp – tương tự như những gì một chuyên viên phân tích có thể đặt ra khi đánh giá một doanh nghiệp. Ví dụ, một câu hỏi có thể là: "Công ty có chiến lược tăng trưởng và đổi mới rõ ràng không? Có bất kỳ sáng kiến chiến lược hoặc quan hệ đối tác nào gần đây không?". Đây là những câu hỏi đòi hỏi sự tổng hợp, suy luận và hiểu sâu văn bản – những năng lực mà các mô hình truyền thống hoặc rule-based khó có thể thực hiện hiệu quả.

2 Bối cảnh và Các Nghiên cứu Liên Quan

2.1 Mô hình Ngôn ngữ Lớn GPT-3.5 và Khả năng của nó

Bài báo này sử dụng phiên bản GPT-3.5 của OpenAI, cũng là phiên bản hiện đang được ứng dụng trong ChatGPT. GPT-3.5 thuộc nhóm các Mô hình Ngôn ngữ Lớn (Large Language Models – viết tắt là LLMs), được huấn luyện trên một lượng dữ liệu cực kỳ đa dạng và khổng lồ, bao gồm sách, bài viết, báo cáo và nhiều nguồn ngôn ngữ tự nhiên khác.

Nhờ quy mô huấn luyện lớn, GPT-3.5 có khả năng hiểu và xử lý ngôn ngữ tự nhiên ở mức độ rất cao, từ việc phân tích cú pháp (cách sắp xếp và cấu trúc câu), ngữ nghĩa (ý nghĩa của từ và câu) đến việc suy luận dựa trên ngữ cảnh. Nhờ đó, GPT-3.5 có thể trả lời các câu hỏi, thực hiện các tác vụ như tóm tắt văn bản, phân loại nội dung, hoặc đưa ra các dự đoán dựa trên văn bản đầu vào một cách hiệu quả.

2.2 Nghiên cứu của A. Lopez-Lira et al. và Ứng dụng LLM trong Tài chính

Trong nghiên cứu của mình, A. Lopez-Lira cùng các cộng sự đã sử dụng GPT-3.5 để phân tích các tiêu đề tin tức liên quan đến từng công ty. Cụ thể, họ thiết kế một lời nhắc (prompt) chứa tiêu đề tin tức, yêu cầu mô hình gán nhãn cảm xúc (sentiment labeling) cho từng tiêu đề, ví dụ như tích cực, trung tính hoặc tiêu cực. Việc gán nhãn cảm xúc ở đây nhằm mục đích sử dụng các nhãn này để dự đoán xu hướng giá cổ phiếu của công ty trong ngày tiếp theo.

Kết quả nghiên cứu cho thấy việc lựa chọn cổ phiếu dựa trên phân tích cảm xúc từ GPT-3.5 đem lại sức mạnh dự đoán có ý nghĩa thống kê rõ rệt, đồng nghĩa với việc LLM có tiềm năng trong việc hỗ trợ ra quyết định đầu tư tài chính.

2.3 Hạn chế của Phương pháp dựa trên Phân tích Tin tức ngắn hạn

Mặc dù thành công, phương pháp của Lopez-Lira và cộng sự vẫn tồn tại một số hạn chế quan trọng:

- **Đơn giản hóa phản hồi:** Họ hướng dẫn LLM trả lời dưới dạng phản hồi ba giá trị (positive, neutral, negative) cho mỗi tiêu đề, và trực tiếp sử dụng phản hồi này như biến dự báo duy nhất để quyết định mua hoặc bán cổ phiếu. Trong thuật ngữ học máy (Machine Learning), việc dựa vào một biến duy nhất được xem là chưa tận dụng hết tiềm năng khai thác các tương tác phức tạp giữa nhiều yếu tố khác nhau.
- **Chi phí giao dịch cao:** Phương pháp lựa chọn cổ phiếu hàng ngày dựa trên các tin tức tức thời khiến danh mục đầu tư phải liên tục thay đổi, từ đó phát sinh chi phí giao dịch lớn. Trong nghiên cứu, khi tính chi phí giao dịch khoảng 0,25% mỗi lần mua/bán, lợi nhuận ròng giảm từ 500% xuống còn 50%.
- **Khả năng sinh lời giảm khi xem xét chi phí:** Nghiên cứu của D. Blitz và cộng sự cũng chỉ ra rằng các tín hiệu giao dịch trong khung thời gian ngắn thường không tạo ra lợi nhuận sau khi trừ chi phí giao dịch, trong khi các dự báo dài hạn (6–12 tháng) có thể giúp duy trì lợi nhuận ròng đáng kể.

2.4 Hướng tiếp cận mới: Phân tích Báo cáo Thường Niên và Ứng dụng Mô hình Kết hợp

Lấy cảm hứng từ nghiên cứu trên, đồng thời nhằm khắc phục nhược điểm của việc giao dịch ngắn hạn, bài báo này đề xuất một cách tiếp cận mới như sau:

- Thay vì dựa vào tin tức hàng ngày, nghiên cứu tập trung phân tích các báo cáo thường niên toàn diện của công ty (hồ sơ 10-K), được công bố mỗi năm một lần.
- Do tần suất công bố thấp hơn, các tín hiệu trích xuất từ báo cáo này có thể áp dụng trong thời gian dài hơn, giảm thiểu chi phí giao dịch và tránh sự dao động quá lớn trong danh mục đầu tư.
- Ngoài ra, nghiên cứu còn kết hợp đầu ra từ mô hình LLM với một mô hình học máy hiện đại. Mô hình học máy là thuật toán có khả năng tự học từ dữ liệu để nhận diện các đặc trưng quan trọng, loại bỏ các yếu tố không liên quan hoặc gây nhiễu.
- Mô hình này còn có thể mô hình hóa các tương tác phức tạp giữa các đặc trưng, từ đó nâng cao độ chính xác trong việc dự báo hiệu suất cổ phiếu trong năm tiếp theo.

3 Dữ liệu

3.1 Nguồn dữ liệu và phạm vi thu thập

Một trong những bước quan trọng nhất trong nghiên cứu này là việc thu thập dữ liệu báo cáo thường niên từ các công ty. Báo cáo thường niên được gọi là *hồ sơ 10-K* (10-K filings) là tài liệu chi tiết do các công ty đại chúng tại Hoa Kỳ phải nộp lên Ủy ban Chứng khoán và Giao dịch Hoa Kỳ (Securities and Exchange Commission – SEC). Đây là nguồn dữ liệu chính cung cấp thông tin toàn diện về hoạt động tài chính, hoạt động kinh doanh, rủi ro và các yếu tố khác của công ty trong một năm tài chính.

Trong nghiên cứu, tác giả thu thập báo cáo 10-K từ năm 2002 đến năm 2023 của 1.500 công ty hàng đầu theo vốn hóa thị trường. Vốn hóa thị trường (market capitalization) là giá trị tổng cộng của tất cả cổ phiếu đang lưu hành của một công ty, được tính bằng cách nhân giá cổ phiếu hiện tại với tổng số cổ phiếu phát hành. Tổng số báo cáo thu thập được là 24.200 tài liệu, với dung lượng lưu trữ khoảng 85 GB.

3.2 Lưu trữ và xử lý dữ liệu

Các báo cáo 10-K được tải xuống từ cơ sở dữ liệu EDGAR và lưu trữ cục bộ trên hệ thống máy tính để phục vụ phân tích. Việc lưu trữ cục bộ (local storage) giúp giảm thiểu độ trễ khi truy xuất dữ liệu và cho phép xử lý nhanh hơn trong các bước tiếp theo. Mỗi báo cáo được gắn kèm thông tin về ngày nộp, giúp xác định chính xác khung thời gian của dữ liệu.

Thông tin trong các báo cáo này được sử dụng như bối cảnh (context) để mô hình ngôn ngữ lớn (Large Language Model – LLM) có thể hiểu và trả lời các câu hỏi tài chính một cách chính xác. Bối cảnh ở đây nghĩa là thông tin nền tảng được cung cấp để mô hình hiểu sâu hơn về vấn đề trước khi đưa ra câu trả lời. Các câu trả lời này tiếp tục được trích xuất và chuyển đổi thành các đặc trưng đầu vào (features) cho mô hình học máy (Machine Learning – ML). Phần chi tiết kỹ thuật sẽ được trình bày trong phần 4.3.

3.3 Định nghĩa giá trị mục tiêu (Target Values)

Trong nghiên cứu này, mục tiêu chính là dự đoán lợi nhuận cổ phiếu dựa trên thông tin trích xuất từ báo cáo 10-K. Giá trị mục tiêu (target value) là thông số mà mô hình học máy cần học để dự đoán. Giá trị này được tính dựa trên phần trăm lợi nhuận cổ phiếu giữa hai thời điểm nộp báo cáo thường niên liên tiếp.

Cụ thể, để đảm bảo tính chính xác và tránh ảnh hưởng của các biến động ngắn hạn, điểm bắt đầu của khoảng thời gian được lấy là giá cổ phiếu sau hai ngày kể từ ngày nộp hồ sơ. Điểm kết thúc là giá cổ phiếu trước hai ngày ngày nộp hồ sơ tiếp theo. Khoảng thời gian này phản ánh hiệu quả hoạt động của công ty trong khoảng thời gian giữa hai báo cáo.

Ngoài việc tính phần trăm lợi nhuận tổng thể, lợi nhuận tại các điểm trung gian (25%, 50%, 75% của khoảng thời gian) cũng được ghi nhận. Đồng thời, mức lợi nhuận tối đa và tối thiểu trong giai đoạn này được xác định để cung cấp thêm thông tin về biến động giá cổ phiếu. Quá trình này giúp mô hình học máy có thể dự đoán không chỉ lợi nhuận trung bình mà còn dự đoán biến động rủi ro của cổ phiếu. Chi tiết về tính toán giá trị mục tiêu sẽ được trình bày ở phần 4.4.

3.4 So sánh với chỉ số chuẩn

Ngoài việc tính toán lợi nhuận riêng của từng cổ phiếu, nghiên cứu còn tính lợi nhuận của chỉ số S&P 500 trong cùng khoảng thời gian. S&P 500 là một chỉ số chứng khoán phổ biến, bao gồm 500 công ty lớn nhất niêm yết tại Hoa Kỳ, thường được xem như thước đo sức khỏe chung của thị trường chứng khoán.

Việc so sánh lợi nhuận của danh mục cổ phiếu do mô hình học máy lựa chọn với lợi nhuận của S&P 500 giúp đánh giá hiệu quả của mô hình so với thị trường chung, từ đó đưa ra kết luận về khả năng tạo ra lợi nhuận vượt trội hay không.

3.5 Phân chia tập dữ liệu và quy trình huấn luyện

Để phát triển và đánh giá mô hình học máy, bộ dữ liệu được chia thành hai phần chính: tập huấn luyện (training set) và tập thử nghiệm (testing set). Tập huấn luyện là dữ liệu được dùng để xây dựng và tối ưu hóa mô hình, trong khi tập thử nghiệm là dữ liệu chưa từng được mô hình nhìn thấy trước đó, dùng để kiểm tra khả năng dự đoán của mô hình trong thực tế.

Nghiên cứu sử dụng dữ liệu từ năm 2002 đến 2017 cho tập huấn luyện, và từ 2018 đến 2023 cho tập thử nghiệm. Lưu ý, các báo cáo thường niên năm 2023 chỉ được dùng trong giai đoạn đánh giá mô hình, không tham gia huấn luyện. Việc phân chia này tuân theo thực tiễn phổ biến nhằm tránh hiện tượng quá khớp (overfitting), khi mô hình chỉ "nhớ" dữ liệu huấn luyện mà không thể áp dụng tốt cho dữ liệu mới.

3.6 Chi phí và giới hạn trong việc sử dụng mô hình ngôn ngữ lớn

Sử dụng các mô hình LLM như GPT-3.5 của OpenAI phát sinh chi phí tính theo số lượng từ (token) được gửi đến và nhận về từ mô hình. Mỗi câu hỏi và câu trả lời đều tiêu tốn một lượng token nhất định, đồng nghĩa với chi phí tài chính.

Ngoài ra, việc gửi thêm phần ngữ cảnh — tức là các đoạn trích từ báo cáo 10-K — cũng làm tăng chi phí sử dụng mô hình. Vì vậy, trong nghiên cứu này, một số lượng điểm dữ liệu đã được lấy mẫu nhằm giảm chi phí và thời gian xử lý: 1.000 điểm từ tập huấn luyện và 500 điểm từ tập thử nghiệm.

Tổng chi phí ước tính cho toàn bộ nghiên cứu khoảng 60 USD. Thời gian xử lý để tạo embeddings và trả lời câu hỏi cũng khá lớn, khoảng 50 giờ sử dụng GPT-3.5. Nếu mở rộng thí nghiệm toàn bộ tập dữ liệu, thời gian và chi phí sẽ tăng lên đáng kể.

4 Phương pháp

4.1 Truy cập báo cáo thường niên (10-K)

Để xây dựng bộ dữ liệu phục vụ cho nghiên cứu, bước đầu tiên là thu thập các báo cáo tài chính thường niên theo mẫu 10-K của 1500 công ty hàng đầu tại Hoa Kỳ, được lựa chọn dựa trên mức *vốn hóa thị trường* (market capitalization) lớn nhất. Danh sách các công ty và mã cổ phiếu tương ứng (*ticker symbol*) được lấy từ Wikipedia.

Mặc dù các báo cáo 10-K đã được công khai trên hệ thống EDGAR của Ủy ban Chứng khoán và Giao dịch Hoa Kỳ (SEC), việc truy xuất hàng loạt các đường dẫn (URL) chính xác cho từng báo cáo là một thách thức do cấu trúc dữ liệu phân tán. Để giải quyết vấn đề này, nghiên cứu sử dụng API của nền tảng Financial Modeling Prep[1] để lấy danh sách các URL của báo cáo 10-K trong quá khứ. Cần lưu ý rằng phiên bản miễn phí của API này có thể bị giới hạn tốc độ truy cập (*rate limit*).

4.2 Nhúng Tài liệu (Document Embedding)

Để Mô hình Ngôn ngữ Lớn (LLM) có thể trả lời các truy vấn liên quan đến báo cáo 10-K, cần xác định các đoạn nội dung có liên quan trong tài liệu. Điều này được thực hiện thông qua kỹ thuật *document embeddings* — phương pháp mã hóa mỗi đoạn văn bản thành một vector có chiều cố định phản ánh ngữ nghĩa trong không gian nhiều chiều.

Độ tương đồng giữa truy vấn và các đoạn văn bản được đo bằng chỉ số *cosine similarity*, được định nghĩa như sau:

$$\text{similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

Trong đó $\mathbf{a}$ và $\mathbf{b}$ là các vector embedding của truy vấn và đoạn văn bản.

Trong nghiên cứu này, mô hình `all-mpnet-base-v2` từ thư viện Sentence Transformers được chọn vì đạt điểm cao trong benchmark MTEB (Massive Text Embedding Benchmark)[2] và có thể xử lý nhanh trên máy tính xách tay thông thường. Mô hình `text-embedding-ada-002` của OpenAI tuy có độ chính xác cao nhưng chi phí cao hơn khi áp dụng cho văn bản dài.

Các vector embedding sau đó được lưu trữ trong *vector database* để truy vấn nhanh. Trong nghiên cứu này, `ChromaDB`[3] được sử dụng vì tương thích tốt với khung `LlamaIndex`[4] – một hệ thống mã nguồn mở hỗ trợ tích hợp giữa LLM và dữ liệu cấu trúc.

4.3 Sử dụng LLM để Tạo Đặc trưng (Feature Generation)

Mô hình LLM chính được sử dụng trong nghiên cứu là `GPT-3.5-Turbo` của OpenAI, được tích hợp thông qua API. Quá trình tạo đặc trưng (feature generation)

[1] https://financialmodelingprep.com/
[2] https://huggingface.co/spaces/mteb/leaderboard
[3] https://www.trychroma.com
[4] https://www.llamaindex.ai

được thực hiện như sau:

1. Tìm các đoạn văn bản liên quan từ báo cáo 10-K thông qua tìm kiếm theo embedding.
2. Tạo *System Prompt* để yêu cầu mô hình LLM trả lời câu hỏi với ngữ cảnh đã chọn.
3. LLM trả lời câu hỏi và gán một *điểm tin cậy* (confidence score) trong khoảng từ 0 đến 100.

Ví dụ một câu hỏi được dùng để tạo đặc trưng:

"Công ty có chiến lược rõ ràng cho tăng trưởng và đổi mới không? Có những sáng kiến chiến lược hoặc quan hệ đối tác nào gần đây không?"

Tổng cộng 27 câu hỏi như vậy được xây dựng theo phương pháp *prompt engineering*, và điểm số từ các câu trả lời sẽ tạo thành các đặc trưng để đưa vào mô hình học máy.

4.4 Tạo Nhãn (Label Creation)

Mỗi mã cổ phiếu sẽ được gán nhãn dựa trên hiệu suất đầu tư sau ngày nộp báo cáo thường niên. Các chỉ số chính bao gồm:

- **target_12m:** lợi nhuận phần trăm sau 12 tháng kể từ ngày nộp báo cáo.
- **sp500_12m:** lợi nhuận tương ứng của chỉ số S&P 500 để làm cơ sở so sánh.
- **target_max:** phần trăm lợi nhuận cao nhất (98th percentile) trong giai đoạn này.

Các bước xử lý để chuẩn hóa giá trị nhãn dựa trên phương pháp từ tài liệu của Numerai:

1. Tính lợi nhuận hàng năm của từng cổ phiếu.
2. Xếp hạng theo từng năm và chuẩn hóa kết quả trong khoảng $[0, 1]$.
3. Cắt giá trị ngoại lai và phân chia thành phần trăm tương ứng.

4.5 Mô hình Học máy (Machine Learning Model)

Tập dữ liệu được chia thành hai phần:

- **Tập huấn luyện:** dữ liệu từ năm 2002 đến 2017.
- **Tập kiểm tra:** dữ liệu từ năm 2018 đến 2023.

Mục tiêu là xây dựng bài toán hồi quy, nơi đầu ra là một số thực phản ánh mức độ lợi nhuận kỳ vọng. Mô hình được sử dụng là **Linear Regression với ràng buộc hệ số không âm**:

$$\min_{\mathbf{w} \geq 0} \|\mathbf{y} - X\mathbf{w}\|^2$$

Mô hình này đảm bảo rằng ảnh hưởng của mỗi đặc trưng đến lợi nhuận không mang giá trị âm, phù hợp với kỳ vọng rằng điểm số từ LLM nên phản ánh mối quan hệ tích cực với kết quả đầu tư.

Dù các mô hình mạnh hơn như **Gradient Boosted Decision Trees (GBDT)** có khả năng bắt các quan hệ phi tuyến giữa đặc trưng và đầu ra, nghiên cứu này ưu tiên mô hình tuyến tính do tính đơn giản và dễ giải thích. GBDT có thể được nghiên cứu trong các bước tiếp theo.

5 Kết quả

Phân tích các dự đoán trên tập kiểm tra cho thấy mô hình học máy (Machine Learning) sử dụng đặc trưng ngôn ngữ được trích xuất từ mô hình GPT-3.5 có hiệu suất vượt trội trong việc lựa chọn cổ phiếu. Kết quả này được thể hiện qua việc danh mục đầu tư do mô hình đề xuất mang lại lợi nhuận cao hơn so với chỉ số tham chiếu S&P 500. Toàn bộ quy trình huấn luyện mô hình và ước lượng lợi nhuận được thực hiện trong môi trường `Jupyter Notebook` với tiêu đề *"modeling and return estimation.ipynb"*.

5.1 Xác định số lượng cổ phiếu tối ưu để đầu tư

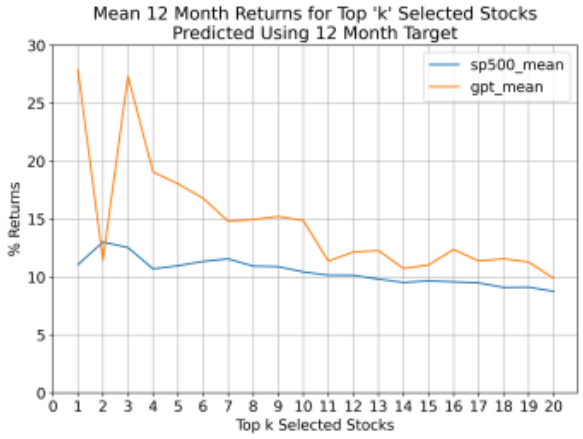


Fig. 1: Comparison of 12 Month Returns for various K values

Hình 1: So sánh lợi nhuận 12 tháng giữa các giá trị k khác nhau (biến mục tiêu: lợi nhuận 12 tháng)

Hình 1 cho thấy hiệu suất đầu tư của mô hình GPT khi chọn ra k cổ phiếu hàng đầu dựa trên điểm số dự đoán. Kết quả cho thấy lợi nhuận trung bình cao hơn khi chọn số lượng cổ phiếu nhỏ (k nhỏ), và giảm dần khi k tăng lên. Điều này cho thấy mô hình có khả năng phân biệt tốt giữa các cổ phiếu có tiềm năng sinh lời cao và thấp.

Giải thích thuật ngữ:

- **Top-k**: là phương pháp lựa chọn k cổ phiếu có điểm số dự đoán cao nhất theo mô hình.
- **Biến mục tiêu 12 tháng**: là lợi nhuận cổ phiếu sau 12 tháng kể từ ngày công bố báo cáo thường niên.

Lưu ý: Lợi nhuận S&P 500 ở các giá trị k khác nhau có dao động nhỏ do thời điểm công bố báo cáo thường niên giữa các công ty là khác nhau. Do đó, lợi nhuận của từng cổ phiếu được tính trên cùng một khoảng thời gian với S&P 500 để đảm bảo công bằng trong so sánh.

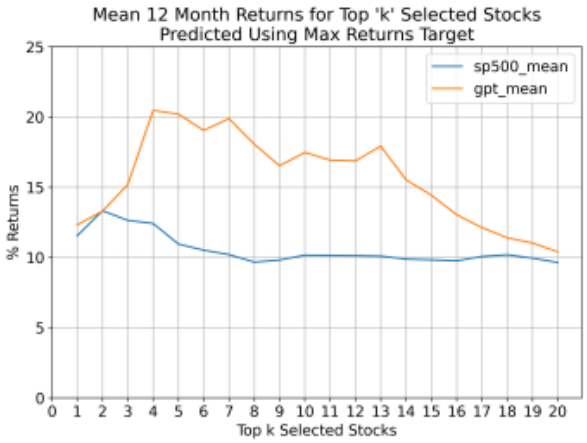


Fig. 2: Comparison of 12 Month Returns for various K values

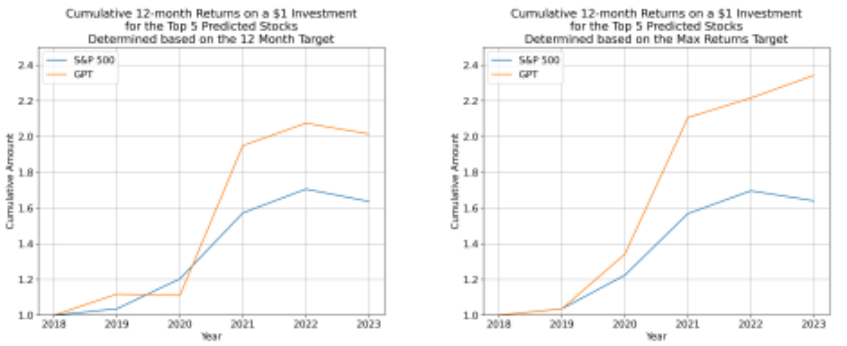
Hình 2: So sánh lợi nhuận theo chiến lược Max Returns với các giá trị k khác nhau

Hình 2 thể hiện kết quả khi thay đổi biến mục tiêu sang **Max Returns**, tức là lợi nhuận tối đa đạt được trong vòng 12 tháng sau ngày công bố báo cáo. Kết quả vẫn giữ xu hướng tương tự, trong đó $k = 5$ mang lại hiệu suất đầu tư tốt nhất.

Nhận xét quan trọng:

- Nên chọn k nhỏ (ví dụ: $k = 5$) để tối ưu hóa lợi nhuận.
- $k = 5$ tương đương với 5% tổng số cổ phiếu (vì mỗi năm có 100 cổ phiếu, chọn 5 cổ phiếu là 5%).

5.2 Phân tích lợi nhuận tích lũy cho các chiến lược đầu tư khác nhau



(a) Modelled using 12-Month Returns Target (b) Modelled using Max Returns Target

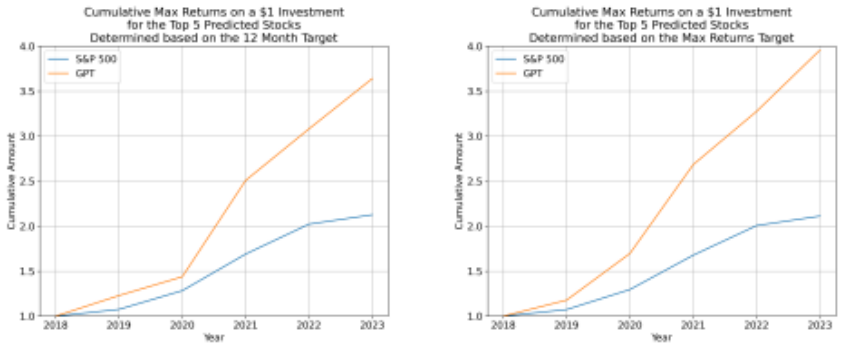
Fig. 3: Cumulative 12-month Returns on a $1 Investment for the Top 5 Predicted Stocks

Hình 3: Lợi nhuận tích lũy từ khoản đầu tư $1 với top 5 cổ phiếu theo hai biến mục tiêu khác nhau

Hình 3 minh họa quá trình tích lũy lợi nhuận từ năm 2018 đến đầu năm 2023 nếu mỗi năm đầu tư $1 vào 5 cổ phiếu được mô hình GPT lựa chọn. Có hai chiến lược được so sánh:

- **Chiến lược A**: Sử dụng biến mục tiêu là lợi nhuận 12 tháng.
- **Chiến lược B**: Sử dụng biến mục tiêu là Max Returns.

Kết quả cho thấy cả hai chiến lược đều vượt trội so với chỉ số S&P 500, nhưng chiến lược B cho hiệu suất tích lũy tốt hơn đáng kể.



(a) Modelled using 12-Month Returns Target (b) Modelled using Max Returns Target

Fig. 4: Cumulative Max Returns on a $1 Investment for the Top 5 Predicted Stocks

Hình 4: Chiến lược Max Returns: lợi nhuận tích lũy từ khoản đầu tư $1 với top 5 cổ phiếu

Chiến lược Max Returns giả định rằng nhà đầu tư mua cổ phiếu ngay sau khi báo cáo thường niên được công bố và bán tại thời điểm giá cổ phiếu đạt đỉnh trong năm kế tiếp (lấy mốc phân vị thứ 98 để tránh nhiễu do outlier).

Giải thích thuật ngữ:

- **Max Returns**: Lợi nhuận tối đa trong vòng 12 tháng sau công bố báo cáo, đo bằng tỉ lệ tăng giá tối đa so với thời điểm mua.
- **98th percentile**: Mức giá cao nhất nằm trong top 2% của toàn bộ các mức giá giao dịch trong năm.

Kết quả đáng chú ý:

- Theo hình 4, nếu áp dụng chiến lược Max Returns, khoản đầu tư $1 mỗi năm có thể tích lũy thành hơn $4 sau 5 năm.
- Trong khi đó, cùng khoảng thời gian, chỉ số S&P 500 chỉ đạt khoảng gấp đôi, cho thấy chiến lược đầu tư dựa trên mô hình ngôn ngữ mang lại lợi nhuận vượt trội.

6 Kết luận

Như đã trình bày trong Mục 3 và 4, các điểm số được trích xuất từ mô hình ngôn ngữ lớn (Large Language Model - LLM), cụ thể là GPT-3.5, có thể đóng vai trò

như các đặc trưng (feature) quan trọng trong mô hình học máy (Machine Learning) cho bài toán lựa chọn cổ phiếu. Tùy thuộc vào mục tiêu đầu tư (ví dụ: lợi nhuận sau 12 tháng hoặc lợi nhuận tối đa), mô hình có thể được tùy chỉnh để tối ưu hóa hiệu suất kỳ vọng.

Trong nghiên cứu này, hai loại biến mục tiêu đã được kiểm chứng:

- Lợi nhuận sau 12 tháng kể từ thời điểm công bố báo cáo thường niên.
- Lợi nhuận tối đa (Max Returns), đo bằng mức tăng giá cao nhất trong vòng 12 tháng.

Cả hai biến đều mang lại kết quả khả quan, trong đó Max Returns cho thấy lợi thế rõ rệt hơn trong đầu tư dài hạn.

Các phát hiện chính

- Mô hình học máy kết hợp với đặc trưng ngôn ngữ từ LLM có thể đạt hiệu suất đầu tư vượt trội so với chỉ số S&P 500.
- Cách xác định biến mục tiêu có ảnh hưởng lớn đến hiệu suất của mô hình, cho phép linh hoạt trong việc điều chỉnh chiến lược đầu tư.
- Đầu tư dài hạn dựa trên phân tích LLM không đòi hỏi tần suất giao dịch cao, giúp giảm chi phí giao dịch và phù hợp với nhà đầu tư cá nhân.

Hướng phát triển tương lai

- Nghiên cứu thêm các phương pháp xác định biến mục tiêu phù hợp hơn với từng loại ngành hoặc điều kiện thị trường.
- Thử nghiệm mô hình trên tập dữ liệu lớn và đa dạng hơn để kiểm tra tính tổng quát và khả năng mở rộng.
- Tích hợp thêm các kỹ thuật xử lý ngôn ngữ tự nhiên nâng cao (như nhận diện thực thể, trích xuất quan hệ hoặc tóm tắt tài liệu) để cải thiện độ chính xác trong việc hiểu nội dung báo cáo thường niên.

Tổng kết

Nghiên cứu này khẳng định tiềm năng ứng dụng các mô hình ngôn ngữ lớn trong phân tích tài chính, đặc biệt là trong chiến lược đầu tư dài hạn. Việc tận dụng đặc trưng từ văn bản giúp mô hình học máy khai thác thêm chiều sâu ngữ nghĩa trong báo cáo thường niên, mở ra cơ hội cho các phương pháp đầu tư thông minh và hiệu quả hơn trong tương lai.

Tài liệu

[1] T.T. Guang Lu, Sylvia B. Larcher, *Hybrid long document summarization using c2f-far and ChatGPT: A practical study*, arXiv e-prints (2023). https://arxiv.org/abs/2306.01169. arXiv:2306.01169

[2] A. Lopez-Lira, Y. Tang, *Can ChatGPT forecast stock price movements? Return predictability and large language models* (2023). https://ssrn.com/abstract=4412788. 10.2139/ssrn.4412788

[3] D. Blitz, M.X. Hanauer, T. Hoogteijling, C. Howard, *The term structure of machine learning alpha* (2023). https://ssrn.com/abstract=4474637. 10.2139/ssrn.4474637

[4] Financial Modeling Prep, https://site.financialmodelingprep.com/. Accessed: 2023-09-01

[5] Massive Text Embedding Benchmark, https://huggingface.co/blog/mteb. Accessed: 2023-09-01

[6] Sentence transformers - all-mpnet-base-v2. https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: 2023-09-01.

[7] Chromadb - vector database. https://docs.trychroma.com/. Accessed: 2023-09-01.

[8] J. Liu. LlamaIndex (2022). https://doi.org/10.5281/zenodo.1234. https://github.com/jerryjliu/llama_index

[9] OpenAI - GPT 3.5 Turbo. https://platform.openai.com/docs/models/gpt-3-5. Accessed: 2023-09-01.

[10] Numerai data documentation. https://docs.numer.ai/numerai-tournament/data. Accessed: 2023-09-01.

[11] M.H. Martin Slawski, Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. arXiv e-prints (2014). https://arxiv.org/abs/1205.0953. arXiv:1205.0953.