


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo:
<https://youtu.be/lFPoT4Z2Ej4>
- Link slides (dạng .pdf đặt trên Github):
<https://github.com/PhamPhucHau/CS2205.APR2024.git/H%E1u%20Ph%E1m%20Ph%E1c%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">● Họ và Tên: Phạm Phúc Hậu● MSSV: 230101042 	<ul style="list-style-type: none">● Lớp: CS2205.APR2024● Tự đánh giá (điểm tổng kết môn): 7.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 3● Link Github: https://github.com/PhamPhucHau/CS2205.APR2024.git
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN TÍCH CẢM XÚC TRÊN TIN TỨC TÀI CHÍNH SỬ DỤNG MÔ HÌNH NGÔN NGỮ LỚN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

FINANCIAL SENTIMENT ANALYSIS USING LARGE LANGUAGE MODEL

TÓM TẮT (Tối đa 400 từ)

Phân tích quan điểm (cảm xúc) tài chính là một nhiệm vụ khó khăn và nhiều thách thức vì lĩnh vực tài chính sử dụng thuật ngữ riêng của nó và thiếu dữ liệu về lĩnh vực này. Các mô hình ngôn ngữ thường sẽ kém hiệu quả trong bối cảnh này vì các thuật ngữ được sử dụng trong ngữ cảnh về tài chính. Chúng tôi đề xuất một mô hình được huấn luyện sẵn để có thể giải quyết được vấn đề này mà nó yêu cầu bộ dữ liệu ít hơn và có thể huấn luyện lại trên một miền dữ liệu cụ thể. Chúng tôi chọn FinBERT, một mô hình dựa trên BERT để xử lý các tác vụ NLP trong lĩnh vực tài chính.

GIỚI THIỆU

Phân tích các văn bản tài chính như là tin tức, báo cáo tài chính, hoặc các thông báo chính thức của công ty là một công việc tiềm năng trong lĩnh vực tài chính. Mỗi ngày có hàng ngàn văn bản như vậy được tạo ra, việc phân tích và đưa ra các nhận định các văn bản tài chính như vậy là một thách thức lớn cho con người chúng ta xử lý. Do đó yêu cầu về việc phân tích và đưa ra nhận định tự động cho các văn bản tài chính bằng phương pháp xử lý ngôn ngữ tự nhiên đã được các nhà nghiên cứu đưa ra trong suốt thập kỷ qua.

Nhiệm vụ chính của của luận văn này là phân loại văn bản đó là tích cực, tiêu cực hay

là trung tính cho văn bản tài chính. Nhiệm vụ này có hai thách thức lớn, thứ nhất đó là yêu cầu một lượng lớn dữ liệu về tài chính và được gán nhãn sẵn khiến chi phí cho chuyên gia cao. Thứ hai đó là sử dụng các mô hình bình thường không phù hợp vì văn bản tài chính có những từ ngữ chuyên biệt, các công thức để diễn đạt sự tích cực hoặc tiêu cực và khó nhận biết.

Bằng cách sử dụng các bộ từ điển về phân tích văn bản tài chính như là bộ của Loughran và McDonald(2011) có lẽ là một giải pháp bởi vì họ tích hợp các kiến thức tài chính vào việc phân tích văn bản. Tuy nhiên, phương pháp đó dựa trên việc đếm từ, nó sẽ không thể phân tích được các ngữ nghĩa sâu xa hơn của một văn bản cụ thể.

Phương pháp học chuyển tiếp (NLP transfer learning) là một giải pháp tốt và hiệu quả hơn cho cả hai thách thức trên. Ý tưởng của mô hình này là thay vì random các tham số của mô hình một cách ngẫu nhiên thì chúng ta sẽ sử dụng các mô hình ngôn ngữ được train sẵn trên dữ liệu rất lớn để lấy được các tham số thích hợp thay vì ngẫu nhiên, sau đó sử dụng mô hình đã train sẵn và train lại tương ứng với từng bài toán cụ thể sẽ giúp cho hiệu suất tốt hơn nhiều. Với các tiếp cận này sẽ giải quyết được vấn đề khan hiếm dữ liệu. Các mô hình ngôn ngữ không yêu cầu nhãn bởi vì nhiệm vụ của chúng là dự đoán các nhãn tiếp theo. Chúng có thể học các biểu diễn các thông tin ngữ nghĩa, từ đó việc tinh chỉnh(fine-tuning) trên dữ liệu đã có nhãn là học các sử dụng các thông tin, ngữ nghĩa để dự đoán nhãn.

Vì vậy, để có thể xử lý được nhiệm vụ phân tích tình cảm, quan điểm của văn bản tài chính, chúng tôi sử dụng mô hình ngôn ngữ FinBert để Fine tune lại dữ liệu để có thể có được một mô hình dự đoán tình cảm, quan điểm đối với văn bản tài chính.

MỤC TIÊU

Mục tiêu 1 : Xây dựng và tinh chỉnh mô hình FinBERT để phân tích cảm xúc tài chính.

Mục tiêu 2: Đánh giá hiệu suất của FinGPT trên các tập dữ liệu tài chính thực tế.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung :

Tìm hiểu Mô hình BERT và FinBERT

BERT (Bidirectional Encoder Representations from Transformers): Giới thiệu về mô hình BERT, kiến trúc và cơ chế hoạt động.

FinBERT: Giới thiệu về mô hình FinBERT, sự khác biệt và cải tiến so với BERT để phù hợp với phân tích cảm xúc tài chính.

2.2. Thu thập và Chuẩn bị Dữ liệu

Nguồn dữ liệu: Sử dụng dataset Financial Sentiment Analysis của Kaggle và StockEmotions Dataset.

Tiền xử lý dữ liệu: Làm sạch và chuẩn bị dữ liệu để đảm bảo chất lượng cao cho huấn luyện và đánh giá mô hình. Bao gồm việc loại bỏ các ký tự đặc biệt, xử lý các từ viết tắt và chuẩn hóa văn bản.

2.3. Xây dựng Mô hình

Kiến trúc mô hình: Sử dụng kiến trúc BERT ban đầu và điều chỉnh các tham số cho phù hợp với dữ liệu tài chính.

Huấn luyện mô hình: Huấn luyện FinBERT trên tập dữ liệu tài chính lớn để học các biểu diễn ngữ nghĩa phù hợp với ngữ cảnh tài chính.

2.4. Tinh chỉnh Mô hình

Fine-tuning: Sử dụng tập dữ liệu đã gắn nhãn cảm xúc (positive, negative, neutral) để tinh chỉnh FinBERT cho nhiệm vụ phân tích cảm xúc.

Đánh giá mô hình: Sử dụng các chỉ số đánh giá như độ chính xác, độ chính xác trung bình, F1-score để đánh giá hiệu suất của FinBERT.

Phương pháp :

3.1. Thu thập và Chuẩn bị Dữ liệu

Tập hợp dữ liệu: Sử dụng dataset Financial Sentiment Analysis của Kaggle và StockEmotions Dataset.

Tiền xử lý: Sử dụng các thư viện Python như NLTK, SpaCy để tiền xử lý văn bản.

3.2. Huấn luyện FinBERT

Sử dụng thư viện Hugging Face Transformers: Sử dụng thư viện này để xây dựng và huấn luyện FinBERT.

Tối ưu hóa tham số: Sử dụng Grid Search hoặc Random Search để tìm ra các tham số tối ưu cho mô hình.

3.3. Tinh chỉnh và Đánh giá

Fine-tuning trên tập dữ liệu có gắn nhãn: Sử dụng tập dữ liệu có gắn nhãn cảm xúc để tinh chỉnh FinBERT.

Đánh giá mô hình: Sử dụng các kỹ thuật cross-validation để đánh giá hiệu suất mô

hình trên các tập dữ liệu kiểm tra.

KẾT QUẢ MONG ĐỢI

1. Hiệu suất Mô hình Cải thiện

Độ chính xác cao: Mô hình FinBERT dự kiến sẽ đạt được độ chính xác cao trong việc phân loại cảm xúc tài chính (positive, negative, neutral) trong các văn bản tài chính.

Điều này sẽ được đo lường bằng các chỉ số như độ chính xác (accuracy), độ chính xác trung bình (precision), độ hồi phục (recall), và F1-score.

Khả năng tổng quát hóa tốt: Mô hình không chỉ hoạt động tốt trên tập dữ liệu huấn luyện mà còn thể hiện hiệu suất tốt trên các tập dữ liệu kiểm tra và đánh giá khác nhau, cho thấy khả năng tổng quát hóa tốt.

2. Phân Tích Cảm Xúc Chính Xác

Nhận diện cảm xúc tinh tế: FinBERT sẽ có khả năng nhận diện các cảm xúc tài chính phức tạp và tinh tế trong các văn bản, giúp phân biệt rõ ràng giữa các loại cảm xúc khác nhau, ngay cả khi chúng xuất hiện trong các ngữ cảnh khác nhau.

Phân tích sâu sắc: Mô hình có thể phân tích và đưa ra các thông tin cảm xúc sâu sắc từ các văn bản tài chính, giúp người dùng hiểu rõ hơn về tâm lý thị trường và các yếu tố ảnh hưởng đến cảm xúc nhà đầu tư.

TÀI LIỆU THAM KHẢO

- [1] Araci, D. (2019). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:1908.10063.
- [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).
- [3] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 328-339).