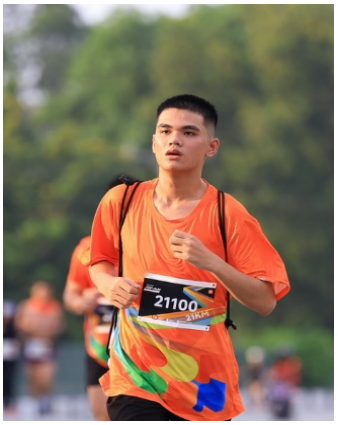
THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/RfCV91t4ecc
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/TwinHter/CS519.Q11.KHTN/blob/main/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*

• Họ và Tên: Nguyễn Hữu Đặng Nguyên
• MSSV: 23521045
 
	• Lớp: CS519.Q11.KHTN
	• Tự đánh giá (điểm tổng kết môn): 8.5/10
	• Số buổi vắng: 1
	• Số câu hỏi QT cá nhân: 7
	• Số câu hỏi QT của cả nhóm: 16
	• Link Github: https://github.com/TwinHter/CS519.Q11.KHTN
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng
	  ○ Viết nội dung
	  ○ Làm video
	  ○ Làm Slide

• Họ và Tên: Đặng Quốc Cường • MSSV: 23520192	• Lớp: CS519.Q11.KHTN • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 9 • Số câu hỏi QT của cả nhóm: 16 • Link Github: https://github.com/TwinHter/CS519.Q11.KHTN

	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: ○ Lên ý tưởng ○ Làm Slide ○ Làm Poster

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA) PHÁT HIỆN ĐÁNH GIÁ GIẢ DỰA TRÊN MẠNG NGỮ NGHĨA TÍCH HỢP THÔNG TIN KHÍA CẠNH
TÊN ĐỀ TÀI TIẾNG ANH (IN HOA) FAKE REVIEW DETECTION USING ASPECT-ENHANCED SEMANTIC NETWORKS
TÓM TẮT *(Tối đa 400 từ)* Đánh giá trực tuyến là yếu tố then chốt định hướng quyết định mua sắm và uy tín doanh nghiệp. Tuy nhiên, vấn nạn đánh giá giả ngày càng tinh vi đặt ra yêu cầu cấp thiết đối với các phương pháp phát hiện đánh giá giả hiệu quả. Các nghiên cứu hiện nay chủ yếu tập trung vào nội dung văn bản và cấu trúc quan hệ giữa người dùng, sản phẩm và đánh giá, trong khi thông tin ở mức độ khía cạnh và sự mâu thuẫn cảm xúc tiềm ẩn mà người viết đề cập đến vẫn chưa được khai thác một cách đầy đủ. Trong nghiên cứu này, chúng tôi đề xuất một hướng tiếp cận phát hiện đánh giá giả dựa trên đồ thị dị thể, tích hợp trực tiếp thông tin khía cạnh vào cấu trúc đồ thị. Cụ thể, các

khía cạnh được trích xuất từ nội dung đánh giá, sau đó được đưa về thành nhóm các khái niệm tổng quát và liên kết với mạng dưới dạng các node. Kiến trúc này cho phép mô hình học được các biểu diễn phức hợp giữa ngữ nghĩa, cấu trúc hành vi và tương quan khía cạnh - cảm xúc**.** Thực nghiệm trên tập dữ liệu chuẩn YelpChi để phân tích mức độ đóng góp của thông tin khía cạnh trong việc nâng cao hiệu quả phát hiện đánh giá giả so với các mô hình hiện có. Nghiên cứu hướng tới việc làm rõ vai trò của thông tin khía cạnh trong việc giúp mô hình nhận diện các mẫu bất thường tinh vi, đồng thời gợi mở một hướng tiếp cận kết hợp phân tích khía cạnh và học biểu diễn trên đồ thị cho bài toán phát hiện đánh giá giả.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Sự phổ biến của các nền tảng thương mại điện tử và hệ thống đánh giá trực tuyến như Yelp hay Amazon khiến các bài đánh giá của người dùng trở thành nguồn tham khảo quan trọng, ảnh hưởng trực tiếp đến quyết định mua sắm và uy tín của doanh nghiệp. Tuy nhiên, lợi ích này đang bị ảnh hưởng bởi sự gia tăng của các đánh giá giả mạo, trong đó nhiều nội dung được tạo một cách tinh vi hay được sinh tự động bởi các mô hình AI. Những đánh giá này có thể làm sai lệch nhận thức người dùng và gây nhiều ảnh hưởng tiêu cực qua đó đặt ra thách thức cho các hệ thống tự động trong việc phát hiện đánh giá giả một cách hiệu quả.

Các phương pháp truyền thống chủ yếu dựa vào đặc trưng văn bản hoặc thống kê hành vi [1]. Sự ra đời của các mô hình Transformer như BERT [2] đã nâng cao khả năng hiểu ngữ nghĩa, chúng vẫn gặp hạn chế lớn trước các đánh giá giả được ngụy trang tinh vi, khó phân biệt chỉ bằng nội dung.

Một hướng tiếp cận hiệu quả là mô hình hóa hệ sinh thái đánh giá dưới dạng đồ thị, trong đó người dùng, sản phẩm và đánh giá được biểu diễn như các thực thể có quan hệ với nhau [3]. Các mô hình mạng nơ-ron đồ thị cho phép lan truyền và tổng hợp thông tin giữa các thực thể liên quan, từ đó hỗ trợ phát hiện các mẫu hành vi bất

thường. Tuy nhiên, phần lớn các phương pháp hiện nay chưa tập trung khai thác đầy đủ thông tin ngữ nghĩa chi tiết trong nội dung đánh giá.

Trong thực tế, mỗi đánh giá thường đề cập đến nhiều khía cạnh khác nhau của thực thể. Các đánh giá giả có xu hướng thể hiện những bất thường ở mức độ khía cạnh, như tập trung quá mức vào một số khía cạnh hoặc sự mâu thuẫn giữa chúng. Các nghiên cứu cũng chỉ ra rằng việc đưa thông tin khía cạnh một cách tường minh vào mô hình giúp cải thiện khả năng học biểu diễn ngữ nghĩa [4]. Tuy nhiên, việc tích hợp thông tin này một cách có hệ thống vào các mô hình đồ thị vẫn còn hạn chế.

Từ những quan sát trên, nghiên cứu này đề xuất một phương pháp phát hiện đánh giá giả dựa trên đồ thị dị thể [5], trong đó thông tin khía cạnh được trích xuất, chuẩn hóa và tích hợp trực tiếp vào cấu trúc đồ thị dưới dạng các nút riêng biệt. Mô hình mạng đồ thị sau đó được sử dụng để học biểu diễn kết hợp giữa nội dung văn bản, hành vi người dùng và thông tin khía cạnh cho nhiệm vụ phân loại đánh giá giả.

Phương pháp được dự kiến đánh giá trên bộ dữ liệu YelpChi [6], một tập dữ liệu thực tế và được gán nhãn phổ biến trong nghiên cứu phát hiện đánh giá giả. Thông qua nghiên cứu này, chúng tôi kỳ vọng làm rõ vai trò của thông tin khía cạnh trong việc nâng cao hiệu quả phát hiện đánh giá giả, đồng thời đóng góp một hướng tiếp cận kết hợp giữa phân tích khía cạnh và học biểu diễn trên đồ thị dị thể cho bài toán này.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Xây dựng được một mô hình phát hiện đánh giá giả dựa trên đồ thị dị thể, trong đó khía cạnh và cảm xúc được tích hợp vào cấu trúc đồ thị để khai thác đồng thời ngữ nghĩa văn bản, hành vi người dùng và mối liên hệ giữa các khía cạnh.
2. Khảo sát mức độ đóng góp của thông tin khía cạnh đối với hiệu quả của mô hình đồ thị, thông qua việc so sánh kết quả khi có và khi không sử dụng các thông tin này trong quá trình huấn luyện.
3. Thực nghiệm và so sánh mô hình đề xuất với một số phương pháp hiện có trên

các bộ dữ liệu chuẩn, nhằm đánh giá hiệu quả của mô hình.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Tổng quan nội dung nghiên cứu

Nghiên cứu tập trung trả lời câu hỏi chính: "**Liệu việc tích hợp thông tin khía cạnh vào mô hình mạng đồ thị ngữ nghĩa có giúp cải thiện hiệu quả phát hiện đánh giá giả so với các phương pháp hiện có hay không?**"

Để trả lời câu hỏi trên, nghiên cứu thực hiện các nội dung sau:

- Khai thác thông tin khía cạnh và cảm xúc từ nội dung đánh giá: Trích xuất và chuẩn hóa thông tin khía cạnh - cảm xúc từ văn bản.
- Mô hình hóa người dùng, sản phẩm, đánh giá và khía cạnh trong đồ thị dị thể.
- Đánh giá mức độ đóng góp của thông tin khía cạnh thông qua thực nghiệm và so sánh với các phương pháp phát hiện đánh giá giả hiện có.

2. Thu thập dữ liệu và tiền xử lý

Nghiên cứu sử dụng các bộ dữ liệu cho bài toán phát hiện đánh giá giả, bao gồm:

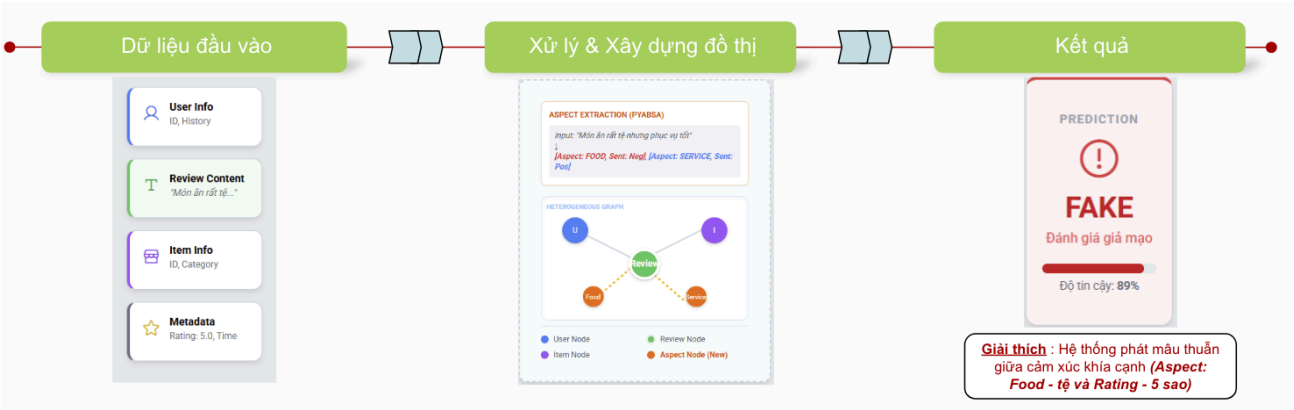
- Người dùng (User), sản phẩm/dịch vụ/cơ sở kinh doanh (Item).
- Nội dung đánh giá và các metadata như rating, số lượng đánh giá.
- Nhãn đánh giá: giả hoặc thật.

Xử lý khía cạnh: Từ nội dung đánh giá, các khía cạnh và cảm xúc tương ứng được trích xuất bằng phương pháp dựa trên pattern hoặc các mô hình trích xuất khía cạnh được huấn luyện sẵn như PyABSA [7]. Các khía cạnh sau đó được chuẩn hóa về các nhóm khái niệm tổng quát (ví dụ SERVICE_POS, FOOD_NEG, …) bằng các kỹ thuật clustering nhằm giảm nhiễu khi đưa vào mô hình đồ thị.

3. Thiết kế và xây dựng mô hình: Hệ thống được thiết kế dựa trên kiến trúc Mạng nơ-ron Đồ thị Dị thể (Heterogeneous Graph)

- **Đồ thị dị thể**:
  - Nút: User, Item, Review, Aspect.
  - Cạnh: User→Review, Review→Item, Review→Aspect, Item→Aspect…
- **Biểu diễn nút**:
  - Review: embedding từ mô hình ngôn ngữ như BERT.
  - Aspect: vector đại diện cho nhóm khía cạnh, lấy từ embedding token (GloVe, …) hoặc tổng hợp embedding ngữ cảnh từ BERT.
  - User, Item: đặc trưng hành vi đơn giản kết hợp với embedding học được.
- **Mô hình học**:
  - Sử dụng mạng đồ thị dị thể để xử lý nhiều loại nút và quan hệ, đồng thời học trọng số cho từng loại quan hệ.
  - Biểu diễn cuối của nút Review được đưa vào lớp phân loại để dự đoán.

4. Thực nghiệm và đánh giá: Mô hình được đánh giá bằng các chỉ số chuẩn đồng thời so sánh với các mô hình tham chiếu. Nghiên cứu cũng thực hiện ablation study nhằm đánh giá trực tiếp ảnh hưởng của thành phần khía cạnh đến hiệu quả mô hình.



KẾT QUẢ MONG ĐỢI

Phương pháp đánh giá:

- Đánh giá trên các dataset chuẩn cho bài toán Fake Review Detections.
- Đánh giá bằng: Accuracy, Precision, Recall, F1-score.
- So sánh với kết quả của các nghiên cứu liên quan và các baseline.

- Thực hiện ablation study để xác định đóng góp của các thành phần.

Kết quả dự kiến: Bảng so sánh chi tiết về hiệu năng giữa mô hình đề xuất và các baseline. Kết quả ablation study làm rõ tác động của từng thành phần đối với hiệu quả mô hình qua đó cung cấp các phân tích về hướng mở rộng mô hình.

Mô hình sau khi được hoàn thiện được kỳ vọng có thể ứng dụng vào thực tiễn như phát hiện đánh giá giả theo thời gian thực.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Shebuti Rayana, Leman Akoglu: Collective Opinion Spam Detection: Bridging Review Networks and Metadata. KDD 2015: 985-994

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

[3] Ao Li, Zhou Qin, Runshi Liu, Yiqun Yang, Dong Li: Spam Review Detection with Graph Convolutional Networks. CIKM 2019: 2703-2711

[4] Chi Sun, Luyao Huang, Xipeng Qiu: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. NAACL-HLT (1) 2019: 380-385

[5] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Yizhou Sun: Heterogeneous Graph Transformer. WWW 2020: 2704-2710

[6] Rami Mohawesh; Shuxiang Xu; Son N. Tran; Robert Ollington; Matthew Springer; Yaser Jararweh: Fake Reviews Detection: A Survey. IEEE Access, vol. 9, pp. 65771-65802, 2021

[7] Heng Yang, Chen Zhang, Ke Li: PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis. CIKM 2023: 5117-5122