

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA

TRẦN QUANG PHÚC

KHAI PHÁ QUAN ĐIỂM VỚI CÁC PHƯƠNG PHÁP QUẢN THỂ

Ngành: Khoa học máy tính

Mã số: 9480101

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

ĐÀ NẴNG – NĂM 2026

Công trình được hoàn thành tại
TRƯỜNG ĐẠI HỌC BÁCH KHOA

Người hướng dẫn khoa học:

1. GS.TS. Huỳnh Xuân Hiệp

2. TS. Lê Thị Mỹ Hạnh

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp cơ sở tại

Trường Đại học Bách khoa

Vào hồi ... giờ ... ngày ... tháng ... năm.....

Có thể tìm hiểu luận án tại:

- Thư viện quốc gia Việt Nam.

- Trung tâm Thông tin - Học liệu & Truyền thông, Đại học Đà Nẵng

PHẦN MỞ ĐẦU

1. Tính cấp thiết của luận án

Sự phát triển mạnh mẽ của World Wide Web và các nền tảng mạng xã hội như Twitter, Facebook, TripAdvisor, Yelp và Amazon đã tạo ra khối lượng khổng lồ dữ liệu văn bản chứa đựng quan điểm, đánh giá của người dùng về sản phẩm, dịch vụ và các vấn đề kinh tế – xã hội. Ví dụ điển hình: Amazon Food Reviews có hơn 568.000 đánh giá thực phẩm với phong phú ngữ cảnh về sản phẩm; TripAdvisor Hotel Reviews chứa 108.891 đánh giá của 1.850 khách sạn, được chấm điểm trên 7 khía cạnh riêng biệt (giá trị, phòng, vị trí, sạch sẽ, nhận phòng/lễ tân, dịch vụ, dịch vụ doanh nhân). Các quan điểm trực tuyến này ngày càng đóng vai trò quan trọng trong việc hỗ trợ ra quyết định của cá nhân và tổ chức.

Ví dụ: trên các đánh giá Amazon Food Reviews, hơn 568.000 đánh giá thực phẩm chứa đựng quan điểm đa chiều, không chỉ mức độ hài lòng mà còn cả thực thể (sản phẩm cụ thể), khía cạnh (hương vị, giá cả, chất lượng), người đánh giá và thời điểm đăng tải. Tương tự, TripAdvisor Hotel Reviews gồm 108.891 đánh giá của 1.850 khách sạn với 7 khía cạnh riêng biệt (giá trị, phòng, vị trí, sạch sẽ, nhận phòng/lễ tân, dịch vụ, dịch vụ doanh nhân). Việc tự động khai phá quan điểm từ những nguồn dữ liệu này không chỉ giúp doanh nghiệp cải thiện sản phẩm/dịch vụ mà còn hỗ trợ người tiêu dùng ra quyết định mua hàng thông minh hơn. Tuy nhiên, các phương pháp hiện có chủ yếu phân tích ở mức văn bản tổng thể, bỏ qua cấu trúc bộ năm phong phú (thực thể, khía cạnh, tình cảm, người đánh giá, thời điểm), đây là khoảng trống mà luận án hướng đến giải quyết.

Tuy nhiên, các hệ thống khai phá quan điểm hiện nay vẫn đối mặt với ba thách thức cốt lõi chưa được giải quyết triệt để: (1) Dữ liệu mất cân bằng: trên Amazon Food Reviews, đánh giá 5★ chiếm 63,88% trong khi 1★–2★ chỉ chiếm 14,42%; trên Yelp, lớp positive (4★–5★) chiếm khoảng 63–65%, khiến mô hình học thiên lệch về lớp đa số; (2) Mối quan hệ phi tuyến giữa các thành phần quan điểm: quan điểm của người dùng phụ thuộc đồng thời vào thực thể, khía cạnh, người đánh giá và thời điểm theo cách phi tuyến phức tạp mà các phương pháp vector hóa thông thường chưa nắm bắt được; (3) Thiếu khả năng diễn giải: các mô hình học sâu như BERT, RoBERTa đạt độ chính xác cao nhưng hoạt động như hộp đen, không thể giải thích tại sao một đánh giá được phân loại là tích cực hay tiêu cực theo từng khía cạnh cụ thể, hạn chế ứng dụng trong các lĩnh vực yêu cầu tính minh bạch.

Đặc biệt, phần lớn các nghiên cứu hiện nay xử lý quan điểm như một nhãn đơn giản, bỏ qua cấu trúc phong phú của dữ liệu đánh giá. Trong thực tế, người dùng và doanh nghiệp

muốn biết: ai đánh giá, khía cạnh gì của đối tượng nào, với tình cảm ra sao, vào thời điểm nào. Đây là bài toán trích xuất và phân tích bộ năm quan điểm chưa được nghiên cứu hệ thống trong khuôn khổ học quần thể.

Trong bối cảnh đó, các phương pháp học quần thể, đặc biệt là các mô hình rừng dựa trên cây quyết định, có tiềm năng lớn trong việc nâng cao đồng thời độ chính xác, khả năng xử lý mất cân bằng và tính diễn giải. Xuất phát từ những lý do trên, luận án với đề tài “Khai phá quan điểm với các phương pháp quần thể” được thực hiện nhằm đề xuất các mô hình khai phá quan điểm dựa trên học quần thể, góp phần khắc phục các hạn chế tồn tại của các phương pháp hiện nay.

2. Mục tiêu, phương pháp, đối tượng và phạm vi nghiên cứu của luận án

2.1 Mục tiêu nghiên cứu

Mục tiêu tổng quát là nghiên cứu và đề xuất các mô hình khai phá quan điểm dựa trên học quần thể nhằm khắc phục hạn chế của các mô hình truyền thống. Các mục tiêu cụ thể:

- Khảo sát tổng quan các nghiên cứu và hệ thống khai phá quan điểm, làm rõ những thành tựu và hạn chế hiện nay.
- Nghiên cứu mô hình quan điểm có cấu trúc dưới dạng bộ năm $o=(e, a, s, h, t)$ và các kỹ thuật trích xuất bộ năm từ dữ liệu văn bản.
- Đề xuất mô hình khai phá quan điểm dựa trên rừng phân loại, rừng mật độ và rừng đa tạp, phù hợp với đặc trưng dữ liệu quan điểm có cấu trúc.
- Xây dựng và chuyển đổi tập dữ liệu quan điểm sang dạng bộ năm quan điểm từ các bộ dữ liệu chuẩn phục vụ thực nghiệm.
- Đánh giá và so sánh các mô hình đề xuất với các phương pháp khai phá quan điểm hiện có trên các tập dữ liệu thực tế.

2.2 Đối tượng nghiên cứu

Đối tượng nghiên cứu bao gồm: (1) Dữ liệu văn bản chứa quan điểm của người dùng; (2) Quan điểm ở mức độ khía cạnh, cấp độ chi tiết nhất và có ứng dụng thực tiễn cao nhất; (3) Kỹ thuật trích xuất và biểu diễn dữ liệu quan điểm dưới dạng bộ năm; (4) Các phương pháp học quần thể với mô hình cơ sở đồng nhất là cây quyết định.

2.3 Phương pháp nghiên cứu

Luận án sử dụng kết hợp: (1) Phương pháp nghiên cứu tài liệu để phân tích, tổng hợp hơn 150 công trình liên quan đến khai phá quan điểm, học quần thể và xử lý ngôn ngữ tự nhiên; (2) Phương pháp thực nghiệm để xây dựng, cài đặt bằng ngôn ngữ R và Python, đánh

giá các mô hình đề xuất trên sáu tập dữ liệu đánh giá thực tế (bốn tập tiếng Anh và hai tập tiếng Việt VLSP-2018) nhằm kiểm chứng hiệu quả.

2.4 Phạm vi nghiên cứu

Luận án giới hạn trong: (1) Trích xuất và chuẩn hóa bộ năm quan điểm từ dữ liệu văn bản; (2) Nghiên cứu học quần thể với mô hình cơ sở đồng nhất là cây quyết định; (3) Áp dụng rừng phân loại, rừng mật độ và rừng đa tạp cho khai phá quan điểm mức độ khía cạnh; (4) Đánh giá trên dữ liệu đánh giá dịch vụ và sản phẩm thực tế.

3. Vấn đề nghiên cứu và các đóng góp của luận án

Câu hỏi nghiên cứu trung tâm của luận án được đặt ra như sau:

“Các phương pháp học quần thể dựa trên cây quyết định có thể khai thác cấu trúc bộ năm thành phần quan điểm $o = (e, a, s, h, t)$ để đồng thời nâng cao độ chính xác phân loại, giải quyết vấn đề mất cân bằng lớp và cung cấp khả năng giải thích minh bạch cho mô hình hay không?”

Luận án có bốn đóng góp chính.

Thứ nhất, luận án đề xuất mô hình Rừng phân loại quan điểm (F_{CF}) dựa trên ma trận bộ năm thành phần quan điểm kích thước $N \times 5$ và entropy Shannon có giám sát $H(O)$. Mô hình này góp phần giải quyết khoảng trống nghiên cứu trong việc biểu diễn có cấu trúc thông tin cảm xúc cho phân loại quan điểm ở mức khía cạnh, đồng thời bảo đảm khả năng giải thích của mô hình qua tầm quan trọng đặc trưng và biểu đồ phụ thuộc từng phần (Bài báo 2, 3).

Thứ hai, luận án phát triển mô hình Rừng mật độ quan điểm (F_{DF}) bằng cách tích hợp thuật toán phân cụm DBSCAN, entropy Gaussian $H_D(O)$ và cơ chế giải thích dựa trên SHAP. Mô hình này xử lý dữ liệu mất cân bằng theo mật độ phân bố của các cụm quan điểm, đồng thời dùng SHAP và chỉ số tương tác Shapley để diễn giải tầm quan trọng và tương tác giữa các thành phần bộ năm (Bài báo 4, 5).

Thứ ba, luận án đề xuất mô hình Rừng đa tạp quan điểm (F_{MF}), trong đó sử dụng ma trận quan hệ đa tạp $W \in R^{N \times N}$, khoảng cách đa tạp $dist_{\mathcal{M}}$ và bộ học siêu phân tầng (stacking meta-learner) $\mathbf{M}$. Cấu trúc đa tạp được học trực tiếp từ tính chất đồng xuất hiện tại nút lá của rừng ngẫu nhiên không giám sát, không cần giải bài toán trị riêng toàn cục như ISOMAP hay UMAP, nhằm khai thác các mối quan hệ phi tuyến giữa các thành phần bộ năm; thực nghiệm cho thấy tính khả thi của hướng tiếp cận, đồng thời chỉ ra chi phí của ma trận W là rào cản khi mở rộng quy mô (Bài báo 6).

Cuối cùng, luận án xây dựng quy trình chuyển đổi dữ liệu đánh giá sang cấu trúc bộ năm thành phần và áp dụng cho hai tập dữ liệu tiếng Anh do Mukherjee và cộng sự công bố là yelpHotelData (hơn 880 nghìn bộ năm) và yelpResData (hơn 14 triệu bộ năm), cùng hai tập dữ liệu tiếng Việt VLSP-2018 Restaurant (hơn 15 nghìn cặp đánh giá – khía cạnh) và VLSP-2018 Hotel (hơn 23 nghìn cặp), phục vụ huấn luyện và đánh giá các mô hình đề xuất (Bài báo 4, 5, 6, 9); mô hình trích xuất bộ năm quan điểm từ văn bản ($F1 = 0,78$ trên Twitter US Airline Sentiment) được công bố tại Bài báo 7, 8.

Theo số liệu đã công bố, tính theo quy trình chia ngẫu nhiên theo dòng bộ năm (Bảng 4.4, §4.8 của Luận án toàn văn), độ chính xác trung bình trên hai tập dữ liệu Yelp là khoảng 0,77 đối với F_{CF} , 0,78 đối với F_{DF} và 0,80 đối với F_{MF} . Ngoài kết quả trên hai tập Yelp (Bảng 4.4), mô hình rừng phân loại quan điểm còn được kiểm chứng riêng trên Amazon Food Reviews và TripAdvisor Hotel Reviews (Chương 2), đạt Accuracy 0,80-0,82 và 0,91 tương ứng (kèm hạn chế Recall chỉ 0,50 trên TripAdvisor do mất cân bằng nhãn). Các mô hình rừng đề xuất chạy trên CPU; riêng mô hình F_{MF} không yêu cầu sử dụng GPU trong quá trình huấn luyện hoặc suy luận và vẫn bảo đảm khả năng giải thích thông qua ma trận quan hệ đa tập W và độ lợi thông tin trên đa tập ($IG_{\mathcal{M}}$).

Khi đánh giá lại theo quy trình chia theo đánh giá gốc (mọi dòng bộ năm của một đánh giá chỉ nằm trong một tập; Phụ lục J của Luận án toàn văn, 5 hạt giống $\times 10.000$ đánh giá), Balanced Accuracy của mọi mô hình chỉ dùng bộ năm giảm về khoảng 0,51–0,57: F_{CF} và F_{DF} tương đương Random Forest (chênh lệch không có ý nghĩa thống kê), F_{MF} thấp hơn Random Forest khoảng 0,02–0,03 do bước phân cụm đa tập chỉ học trên mẫu con; các mô hình đọc văn bản gốc (BERT, RoBERTa) đạt 0,76–0,82. Kiểm tra có kiểm soát với cùng Random Forest và cùng 112.275 bộ năm cho Accuracy 0,946 khi chia theo dòng và 0,611 khi chia theo đánh giá gốc; vì vậy luận án đề xuất chia theo đánh giá gốc khi đánh giá mô hình trên dữ liệu bộ năm.

4. Bố cục của luận án

Ngoài phần mở đầu, kết luận và hướng phát triển, luận án được kết cấu gồm 4 chương: Chương 1 (Cơ sở lý luận về khai phá quan điểm và học quần thể); Chương 2 (Khai phá quan điểm với rừng phân loại); Chương 3 (Khai phá quan điểm với rừng mật độ); Chương 4 (Khai phá quan điểm với rừng đa tập).

CHƯƠNG 1. CƠ SỞ LÝ LUẬN VỀ KHAI PHÁ QUAN ĐIỂM VÀ HỌC QUẦN THỂ

Nội dung chính của chương bao hàm kết quả nghiên cứu của Bài báo số 1, 7, 8, 9. Chương này đặt nền móng lý luận cho toàn bộ luận án.

Tóm lược Chương 1 – Bài toán: xây dựng cơ sở lý luận và trích xuất tự động bộ năm quan điểm (e, a, s, h, t) từ văn bản. Phương pháp: tổng quan khai phá quan điểm và học quần thể; đề xuất khung xác suất trích xuất bộ năm kết hợp phân tích phụ thuộc cú pháp. Kết quả: trên Twitter US Airline Sentiment (4.500 mẫu gán nhãn), mô hình trích xuất đạt $F1 = 0,78$ ($P = 0,81, R = 0,75$), thời gian 8,6 ms/tweet, bộ nhớ 210 MB. Đóng góp: nền tảng biểu diễn và dữ liệu bộ năm cho ba mô hình rừng ở Chương 2–4 (Bài báo 7, 8, 9).

1.1. Khai phá quan điểm

1.1.1. Mô hình bộ năm quan điểm

Theo [1], một quan điểm có thể được biểu diễn dưới dạng bộ năm thành phần quan điểm:

$$o = (e, a, s, h, t),$$

trong đó e là thực thể được đánh giá, a là khía cạnh hoặc thuộc tính của thực thể, s biểu thị tình cảm/cảm xúc được thể hiện đối với khía cạnh a của thực thể e , h là chủ thể bày tỏ quan điểm, tức cá nhân hoặc tổ chức đưa ra nhận xét, và t là thời điểm phát biểu quan điểm.

Đối với một tài liệu hoặc câu đánh giá,

$$d_i,$$

thông tin quan điểm được trích xuất được biểu diễn dưới dạng một tập hợp các bộ năm thành phần

$$O = \{o_1, o_2, \dots, o_m\},$$

Trong bài toán phân loại tình cảm/cảm xúc, điểm đánh giá của người dùng được chuyển đổi thành các nhãn tình cảm/cảm xúc theo chiến lược ánh xạ ba mức. Cụ thể, các đánh giá từ 1★ đến 2★ được gán vào lớp tiêu cực (negative), các đánh giá từ 4★ đến 5★ được gán vào lớp tích cực (positive), trong khi các đánh giá 3★ được loại bỏ (đối với Amazon Food Reviews và TripAdvisor Hotel Reviews) do thường biểu thị quan điểm trung lập hoặc không rõ ràng và sẽ được khai thác trong phần mở rộng của luận án; riêng hai tập Yelp (Chương 3, 4) áp dụng ngưỡng bất đối xứng, gộp 3★ vào lớp negative.

So với các phương pháp biểu diễn tình cảm/cảm xúc truyền thống, mô hình bộ năm thành phần quan điểm cung cấp một biểu diễn ngữ nghĩa phong phú hơn bằng cách đồng thời mô hình hóa thực thể, khía cạnh, tình cảm, chủ thể phát biểu và thông tin thời gian. Tuy nhiên, không gian biểu diễn năm chiều này cũng làm phát sinh các mối quan hệ phụ thuộc phức tạp giữa các thành phần của quan điểm. Đặc biệt, tình cảm không chỉ phụ thuộc vào từng thành phần riêng lẻ mà còn chịu ảnh hưởng đồng thời của nhiều yếu tố ngữ cảnh. Những mối quan hệ phi tuyến này khó có thể được mô hình hóa hiệu quả bằng các phương pháp trích chọn đặc trưng truyền thống hoặc các biểu diễn véc-tơ tuyến tính. Đây chính là động lực nghiên cứu để

luận án đề xuất ba mô hình học quần thể mới, bao gồm Rừng phân loại quan điểm (Opinion Classification Forest), Rừng mật độ quan điểm (Opinion Density Forest) và Rừng đa tạp quan điểm (Opinion Manifold Forest).

1.1.2. Các tác vụ khai phá quan điểm

Khai phá quan điểm được thực hiện qua 8 tác vụ [1]: (1) Trích xuất và phân giải thực thể; (2) Trích xuất và phân giải khía cạnh; (3) Trích xuất biểu thức quan điểm; (4) Phân loại tình cảm cấp khía cạnh; (5) Xác định người giữ quan điểm; (6) Xác định thời gian; (7) Tổng hợp quan điểm; (8) So sánh quan điểm đối lập. Luận án tập trung vào tác vụ phân loại tình cảm cấp khía cạnh (ABSA) vì cung cấp thông tin chi tiết nhất và có giá trị ứng dụng cao nhất trong thương mại điện tử.

1.1.3. Các cấp độ khai phá quan điểm

Khai phá quan điểm có thể được thực hiện ở ba cấp độ phân tích khác nhau:

Cấp độ tài liệu: nhằm xác định cảm xúc tổng thể được thể hiện trong toàn bộ văn bản đối với một thực thể cụ thể.

Cấp độ câu: tập trung xác định cực tính cảm xúc của từng câu bằng cách phân loại mỗi câu thành tích cực, tiêu cực hoặc trung lập.

Cấp độ khía cạnh: còn được gọi là phân tích tình cảm dựa trên khía cạnh (Aspect-Based Sentiment Analysis - ABSA), thực hiện phân tích cảm xúc ở cấp chi tiết bằng cách xác định đối tượng được nhận xét và xác định cảm xúc tương ứng đối với từng khía cạnh cụ thể của thực thể.

Luận án tập trung nghiên cứu khai phá quan điểm ở cấp độ khía cạnh, vì đây là mức phân tích cung cấp thông tin tình cảm chi tiết và có giá trị thực tiễn cao nhất. So với phân tích ở cấp tài liệu và cấp câu, khai phá quan điểm ở cấp độ khía cạnh cho phép hiểu rõ hơn sự hài lòng hoặc không hài lòng của người dùng đối với từng thuộc tính cụ thể của sản phẩm hoặc dịch vụ. Nhờ đó, phương pháp này đã trở thành một kỹ thuật nền tảng trong các hệ thống thương mại điện tử thông minh, hệ thống gợi ý, phân tích phản hồi khách hàng và hỗ trợ ra quyết định trong doanh nghiệp.

1.2. Học quần thể và các mô hình rừng

1.2.1. Cây quyết định - mô hình cơ sở đồng nhất

Cây quyết định có cấu trúc dạng cây trong đó mỗi nút đại diện cho một đặc trưng, mỗi nhánh là một điều kiện và mỗi nút lá là kết quả phân loại. Trong phân loại với C lớp, tiêu chí phân tách dựa trên Entropy Shannon:

$$H(O) = -\sum_{c \in C} p(c) \log p(c)$$

Độ lợi thông tin tại nút v được tính bằng:

$$IG = H(O) - \sum_{i \in \{L, R\}} \frac{|O^i|}{|O|} H(O^i)$$

Nút tốt nhất tại mỗi bước phân tách là nút tối đa hóa độ lợi thông tin IG . Cây quyết định có ưu điểm là diễn giải được qua quy tắc IF-THEN, nhưng dễ bị quá khớp (overfitting) với

dữ liệu huấn luyện; đây là động lực chính để sử dụng cây quyết định như mô hình cơ sở trong học quần thể.

1.2.2. Bootstrap và các kỹ thuật học quần thể

Bootstrap [13] là kỹ thuật lấy mẫu có hoàn lại tạo B tập con $O^*_1, O^*_2, \dots, O_B^*$ từ tập gốc O , mỗi tập chứa khoảng 63,2% mẫu khác nhau. Bagging (Bootstrap Aggregating) [62] huấn luyện B mô hình độc lập trên B tập Bootstrap và kết hợp bằng majority voting làm giảm phương sai mà không tăng thiên lệch. Boosting [94] xây mô hình tuần tự, mỗi mô hình tập trung vào các mẫu khó từ mô hình trước. Stacking [63] huấn luyện bộ học siêu cấp (meta-learner) M tổng hợp kết quả từ K mô hình cơ sở trên tập xác thực độc lập.

1.2.3. Rừng phân loại quan điểm, Rừng mật độ quan điểm và Rừng đa tập quan điểm

Các phương pháp học quần thể dựa trên rừng đã phát triển từ các mô hình phân loại có giám sát sang các mô hình học mật độ và học biểu diễn trên đa tập nhằm khai thác tốt hơn cấu trúc tiềm ẩn của dữ liệu.

Rừng phân loại quan điểm (Opinion Classification Forest) [65][66] mở rộng mô hình Classification Forest truyền thống bằng cách tích hợp biểu diễn bộ năm thành phần quan điểm (Opinion Quintuple) vào khung học Bagging sử dụng cây quyết định làm bộ học cơ sở. Trong quá trình xây dựng cây, mỗi nút được phân chia dựa trên entropy Shannon có giám sát

$$H(O),$$

đồng thời áp dụng cơ chế lựa chọn ngẫu nhiên các thuộc tính tại mỗi lần phân chia nhằm tăng tính đa dạng giữa các cây và cải thiện khả năng tổng quát hóa của mô hình.

Tiếp theo, Rừng mật độ quan điểm (Opinion Density Forest) [66][68][69][70] mở rộng mô hình trên sang bài toán học không giám sát bằng cách thay thế entropy Shannon bằng entropy Gaussian

$$H_D(O),$$

giúp mô hình học được cấu trúc phân bố mật độ nội tại của dữ liệu quan điểm mà không cần sử dụng nhãn lớp. Nhờ đó, mô hình đặc biệt phù hợp để xử lý dữ liệu mất cân bằng và khám phá các cụm quan điểm tiềm ẩn.

Trên cơ sở đó, Rừng đa tập quan điểm (Opinion Manifold Forest) [66][71] tiếp tục tích hợp học đa tập vào cấu trúc rừng bằng cách xây dựng ma trận quan hệ đa tập

$$W \in R^{N \times N},$$

nhằm biểu diễn các mối quan hệ hình học tiềm ẩn giữa các mẫu quan điểm thông qua cơ chế đồng xuất hiện tại các nút lá (co-leaf occurrence). Từ ma trận này, mô hình ước lượng khoảng cách đa tập

$$dist_{\mathcal{M}},$$

qua đó bảo toàn cấu trúc phi tuyến nội tại của không gian bộ năm thành phần quan điểm và nâng cao hiệu quả phân loại.

Ba mô hình rừng được đề xuất trong luận án thể hiện quá trình phát triển liên tục của các phương pháp học quần thể. Rừng phân loại quan điểm sử dụng entropy Shannon có giám sát

$$H(O),$$

để thực hiện phân loại quan điểm ở mức khía cạnh. Rừng mật độ quan điểm thay thế tiêu chí này bằng entropy Gaussian

$$H_D(O),$$

nhằm mô hình hóa cấu trúc phân bố mật độ của dữ liệu quan điểm không có nhãn. Cuối cùng, Rừng đa tập quan điểm sử dụng entropy Gaussian trên đa tập

$$H_{\mathcal{M}}(O),$$

để khai thác cấu trúc đa tập tiềm ẩn của các bộ năm thành phần quan điểm. Quá trình phát triển này phản ánh sự chuyển đổi từ học có giám sát truyền thống sang học dựa trên mật độ và cuối cùng là học biểu diễn trên đa tập nhằm xử lý hiệu quả các mối quan hệ phi tuyến phức tạp trong dữ liệu quan điểm.

1.3. Tiếp cận học quần thể

Câu hỏi thực tiễn quan trọng đặt ra: khi nào nên dùng phương pháp học quần thể (học máy truyền thống như rừng ngẫu nhiên), khi nào nên dùng học sâu (LSTM, BERT, LLM)? Luận án đề xuất bốn tiêu chí lựa chọn:

(1) *Dữ liệu có cấu trúc bảng (tabular data)*: khi dữ liệu đã được trích xuất thành đặc trưng có cấu trúc như bộ năm quan điểm (e, a, s, h, t) thì các phương pháp cây/rừng vẫn là lựa chọn mạnh, đặc biệt trên tập dữ liệu vừa và nhỏ [141];

(2) *Khả năng diễn giải*: khi ứng dụng yêu cầu giải thích lý do phân loại (y tế, tài chính, quản lý), mô hình rừng + SHAP cung cấp diễn giải minh bạch mà học sâu end-to-end khó cung cấp [144][145];

(3) *Tài nguyên tính toán*: khi không có GPU chuyên dụng hoặc cần triển khai trên phần cứng hạn chế, rừng ngẫu nhiên và rừng mật độ là lựa chọn phù hợp do không yêu cầu GPU [143];

(4) Ngược lại, học sâu Transformer phù hợp hơn khi làm việc với văn bản thô không qua tiền xử lý, xử lý mĩa mai, phủ định hoặc quan điểm ẩn, và khi có đủ dữ liệu huấn luyện lớn.

Luận án áp dụng học quần thể theo tiêu chí (1), (2) và (3): dữ liệu đã cấu trúc hóa thành bộ năm, ưu tiên diễn giải và chi phí tính toán hợp lý.

1.4. Nghiên cứu liên quan

Khảo sát hơn 150 công trình thuộc năm nhóm chính được tóm tắt trong bảng sau:

Bảng 1.1. Tóm tắt các nhóm nghiên cứu liên quan

Nhóm nghiên cứu	Đại diện tiêu biểu	Điểm mạnh	Hạn chế chính
Học máy truyền thống	SVM[86], NB[204], LR[203]	Hiệu quả trên dữ liệu nhỏ; đơn giản	Không xử lý mất cân bằng, thiếu biểu diễn cấu trúc bộ năm

Học quần thể	RF [65], GBC [84], Stacking [63]	Độ chính xác cao; ổn định	Chưa tích hợp phân cụm mật độ, chưa áp dụng bộ năm
Học sâu RNN/CNN	LSTM [85], BiLSTM [88]	Nắm bắt ngữ cảnh chuỗi	Hộp đen, yêu cầu dữ liệu lớn, khó diễn giải
Transformer/LLM	BERT [81], RoBERTa [83], GPT [127]	Độ chính xác cao nhất	Hộp đen, yêu cầu GPU, chi phí cao, không diễn giải được
Đa tạp/GCN	GCN [137] Manifold [118][129]	Khai thác cấu trúc phi tuyến	Phức tạp, chưa áp dụng bộ năm 5 chiều

Thông qua khảo sát các công trình nghiên cứu liên quan, luận án xác định ba khoảng trống nghiên cứu chính. Thứ nhất, chưa có mô hình học quần thể tích hợp phân cụm dựa trên mật độ với rừng ngẫu nhiên trên biểu diễn bộ năm thành phần quan điểm nhằm đồng thời xử lý dữ liệu mất cân bằng và khai thác cấu trúc cụm của dữ liệu quan điểm. Thứ hai, các phương pháp học đa tạp chưa được khai thác trên biểu diễn bộ năm thành phần quan điểm để mô hình hóa các quan hệ phi tuyến giữa các thành phần của quan điểm. Thứ ba, các mô hình phân tích cảm xúc dựa trên khía cạnh (Aspect-Based Sentiment Analysis - ABSA) hiện đại mặc dù đạt độ chính xác cao nhưng vẫn còn hạn chế về khả năng giải thích minh bạch đối với từng quyết định phân loại ở mức khía cạnh. Trên cơ sở ba khoảng trống nghiên cứu này, luận án lần lượt đề xuất mô hình Rừng phân loại quan điểm (F_{CF}), Rừng mật độ quan điểm (F_{DF}) và Rừng đa tạp quan điểm (F_{MF}), được trình bày tương ứng trong Chương 2, Chương 3 và Chương 4.

1.5. Đánh giá mô hình

Luận án sử dụng bộ 4 độ đo chuẩn (Bảng 1.2) trong thực nghiệm 3 chương:

Bảng 1.2. Bộ độ đo đánh giá hiệu năng mô hình

Độ đo	Công thức	Ý nghĩa
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Tỷ lệ phân loại đúng. Thiếu tin cậy khi mất cân bằng ($\rho > 3$)
Precision (P)	$TP/(TP+FP)$	Độ chính xác dự đoán lớp positive. Quan trọng khi FP có chi phí cao
Recall (R)	$TP/(TP+FN)$	Khả năng phát hiện lớp positive. Quan trọng khi FN có chi phí cao
F1-score	$2PR/(P+R)$	Trung bình điều hòa P và R. Độ đo chính khi dữ liệu mất cân bằng

Ngoài ra, hệ số Cohen's Kappa [79] $\kappa = (P_o - P_e)/(1 - P_e)$ được sử dụng để đánh giá độ đồng thuận chú thích dữ liệu, trong đó P_o là tỷ lệ đồng thuận thực tế và P_e là tỷ lệ kỳ vọng. Giá trị $\kappa \geq 0,80$ biểu thị đồng thuận rất tốt, $\kappa \in [0,60, 0,80)$ biểu thị đồng thuận tốt. Phụ lục F

bổ sung độ đo F1-macro, F1-pos, F1-neg đánh giá mô hình trên tập VLSP-2018 ($\rho \approx 10,5$ với Restaurant và $\approx 3,2$ với Hotel).

1.6. Trích xuất bộ năm quan điểm

1.6.1. Mô hình trích xuất

Mô hình trích xuất tự động bộ năm dựa trên xác suất kết hợp:

$$P(e, a, s, h, t|x) = P_e(e|x) \cdot P_a(a|e, x) \cdot P_s(s|e, a, x) \cdot P_h(h|s, G) \cdot P_t(t|x)$$

trong đó P_e, P_a, P_s, P_h, P_t là xác suất trích xuất từng thành phần theo điều kiện lẫn nhau; $G=(V, E)$ là đồ thị phụ thuộc cú pháp. Mô hình gồm sáu bước: tiền xử lý văn bản $\rightarrow$ phân tích cú pháp phụ thuộc $\rightarrow$ trích xuất thực thể e theo PMI và từ điển miền $\rightarrow$ trích xuất khía cạnh a theo quan hệ cú pháp $\rightarrow$ phân tích tình cảm s qua VADER và SentiWordNet $\rightarrow$ trích xuất người giữ quan điểm h và thời gian $t \rightarrow$ lọc theo độ tin cậy $Conf(o) \geq \tau$. Chỉ giữ lại bộ năm có điểm tin cậy đủ cao, đảm bảo chất lượng đầu vào cho ba mô hình khai phá quan điểm.

1.6.2. Kết quả thực nghiệm trích xuất bộ năm quan điểm

Tập dữ liệu Twitter US Airline Sentiment [80] với 14.640 tweets về 6 hãng hàng không Hoa Kỳ (United, US Airways, American, Southwest, Delta, Virgin America) và 4.500 tweets được chú thích thủ công bằng Label Studio bởi ba người chú thích độc lập.

Bảng 1.3. Kết quả thực nghiệm trích xuất bộ năm quan điểm

Kịch bản	Mô tả	Kết quả chính
KB1: Chú thích	4.500 tweets được chú thích thủ công	Kappa $\kappa > 0,79$ (entity, aspect, sentiment); $\kappa = 0,76$ (holder); $\kappa = 0,79$ (time)
KB2: Trích xuất	Trích xuất bộ năm trên 14.640 tweets; đánh giá trên 4.500 tweets đã chú thích (chia 70/15/15)	$F_1 = 0,78$ ($P=0,81, R=0,75$); Entity đạt $F_1 = 0,86$ cao nhất; Holder đạt $F_1 = 0,68$ thấp nhất
KB3: So sánh	So sánh với phương pháp quy tắc và spaCy	Mô hình đề xuất ($0,78$) $>$ spaCy ($0,70$) $>$ Quy tắc ($0,66$); thời gian 8,6 ms/tweet, bộ nhớ 210 MB

Mô hình trích xuất đạt $F_1 = 0,78$, vượt phương pháp quy tắc ($0,66$) và spaCy ($0,70$), với thời gian xử lý 8,6 ms/tweet và bộ nhớ 210 MB, phù hợp ứng dụng thời gian thực. Entity đạt $F_1 = 0,86$ cao nhất vì phụ thuộc vào từ điển miền; Holder đạt $F_1 = 0,68$ thấp nhất vì người giữ quan điểm thường ngầm định, khó xác định trong văn bản mạng xã hội. Kappa $\kappa > 0,79$ cho entity, aspect và tình cảm xác nhận độ đồng thuận chú thích rất tốt [79].

1.7. Kết luận chương

Chương 1 đã xây dựng cơ sở lý luận cho toàn bộ luận án thông qua việc xác lập mô hình bộ năm thành phần quan điểm $o = (e, a, s, h, t)$ làm đơn vị biểu diễn thống nhất của dữ liệu

quan điểm, hệ thống hóa các phương pháp học quần thể dựa trên cây quyết định và làm rõ sự khác biệt giữa ba tiêu chí entropy Shannon, entropy Gaussian và entropy Gaussian trên đa tạp được sử dụng trong ba mô hình đề xuất. Đồng thời, chương đã tổng quan các công trình nghiên cứu liên quan, xác định ba khoảng trống nghiên cứu về xử lý dữ liệu mất cân bằng, khai thác cấu trúc đa tạp của dữ liệu quan điểm và nâng cao khả năng giải thích của mô hình. Bên cạnh đó, quy trình trích xuất bộ năm thành phần quan điểm đạt $F1\text{-score} = 0,78$, tạo nguồn dữ liệu đầu vào có chất lượng và nền tảng cho việc xây dựng mô hình trích xuất bộ năm trên các bộ dữ liệu khác.

CHƯƠNG 2. KHAI PHÁ QUAN ĐIỂM VỚI RỪNG PHÂN LOẠI

Nội dung chính bao hàm kết quả nghiên cứu của Bài báo số 2, 3.

2.1. Giới thiệu

Chương 2 đề xuất mô hình Rừng phân loại quan điểm (Opinion Classification Forest) giải quyết khoảng trống nghiên cứu thứ nhất là biểu diễn và phân loại tình cảm có giám sát trên dữ liệu bộ năm với khả năng diễn giải. Đóng góp cốt lõi là biểu diễn câu quan điểm dưới dạng ma trận bộ năm $N \times 5$ thay vì vector đặc trưng đơn lẻ, bảo toàn cấu trúc ngữ nghĩa đa chiều.

Tóm lược Chương 2 – Bài toán: phân loại cực tính quan điểm có giám sát trên dữ liệu bộ năm và diễn giải kết quả phân loại. Phương pháp: rừng phân loại quan điểm trên ma trận bộ năm $N \times 5$ (Bagging, entropy Shannon), diễn giải bằng tầm quan trọng đặc trưng, độ sâu tối thiểu và PDP. Kết quả: Accuracy 0,80–0,82 trên Amazon Food Reviews; Accuracy 0,91 trên TripAdvisor Hotel Reviews nhưng Recall của lớp negative chỉ 0,50 do mất cân bằng nhãn. Đóng góp: đóng góp thứ nhất của luận án (Bài báo 2, 3).

2.2. Ma trận bộ năm quan điểm

Cho tập bộ năm $O = (o_1, o_2, o_3, \dots, o_N)$, với mỗi $o_i = (e_i, a_i, s_i, h_i, t_i)$, $i = 1, 2, \dots, N$.

Ma trận bộ năm quan điểm O_{mat} kích thước $N \times 5$ được định nghĩa:

$$O_{mat} = \begin{bmatrix} e_1 & a_1 & s_1 & h_1 & t_1 \\ e_2 & a_2 & s_2 & h_2 & t_2 \\ e_3 & a_3 & s_3 & h_3 & t_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_N & a_N & s_N & h_N & t_N \end{bmatrix},$$

trong đó mỗi hàng là một bộ năm quan điểm và mỗi cột là một đặc trưng. Các giá trị đặc trưng được mã hóa số (label encoding) trước khi đưa vào mô hình. Theo hiểu biết của nghiên cứu sinh, đây là một trong số ít biểu diễn trong lĩnh vực ABSA sử dụng toàn bộ 5 thành phần bộ năm đồng thời làm đặc trưng phân loại.

2.3. Mô hình rừng phân loại quan điểm

2.3.1. Cây phân loại quan điểm

Cây phân loại quan điểm được xây dựng bằng cách phân chia nút v thành hai nút con L và R . Tiêu chí phân tách sử dụng Entropy Shannon theo công thức (1.4) và độ lợi thông tin theo công thức (1.6): $s \in \{positive, negative\}$. Cây được biểu diễn dưới dạng luật IF-THEN: đường dẫn từ nút gốc đến nút lá tạo thành luật quyết định, ví dụ: “*IF Entity=Hotel AND Aspect=Room AND Score ≥ 4 THEN sentiment=positive*”.

2.3.2. Dựng rừng phân loại quan điểm

Rừng phân loại quan điểm được xây dựng qua 4 bước:

Bước 1 - Lấy mẫu Bootstrap. Tạo B tập con $O^*_1, O^*_2, \dots, O^*_B$ bằng lấy mẫu có hoàn lại từ O_{mat} . Mỗi tập chứa N mẫu (có trùng lặp), tương đương $\sim 63,2\%$ mẫu khác nhau.

Bước 2 - Lựa chọn đặc trưng ngẫu nhiên. Tại mỗi nút phân tách, chọn ngẫu nhiên $m_{try} = \lfloor \sqrt{4} \rfloor = 2$ đặc trưng từ 4 đặc trưng của bộ năm thay vì xét tất cả, nhằm tạo sự đa dạng giữa các cây.

Bước 3 - Xây dựng cây phân loại. Với mỗi tập Bootstrap O_b^* và m_{try} đặc trưng ngẫu nhiên tại mỗi nút, xây cây phân loại quan điểm T_b tối đa hóa I cho đến điều kiện dừng.

Bước 4 - Tổng hợp dự đoán. Tổng hợp $n = 100$ cây bằng đa số phiếu bầu (majority voting):

$$\hat{y} = \operatorname{argmax}_{c \in C} \frac{1}{n} \sum_{i=1}^n \mathbb{1}[T_i(o) = c]$$

2.3.3. Độ phức tạp tính toán

Độ phức tạp huấn luyện. Mỗi cây được xây dựng trên tập Bootstrap $N' \approx 0,63N$ mẫu, tại mỗi nút chọn ngẫu nhiên m_{try} đặc trưng và tìm ngưỡng tách tốt nhất, dẫn đến độ phức tạp trên dữ liệu cân bằng được biểu diễn bởi công thức:

$$\Theta(n \times N' \times m_{try} \times h) = O(n \times N \times m_{try} \times \log N)$$

Độ phức tạp dự đoán. Mỗi mẫu mới được phân loại qua n cây, mỗi cây duyệt theo chiều sâu $O(\log N)$, tổng cộng độ phức tạp của thuật toán đối với mỗi mẫu theo công thức:

$$\Theta(n \times \log N)$$

2.3.4. Mô hình tổng quan

Mô hình F_{CF} gồm ba giai đoạn: (1) Mô hình hóa quan điểm, xây dựng O_{mat} và Bootstrap; (2) Khám phá và biểu diễn quan điểm, xây dựng mô hình F_{CF} ; (3) Định hướng quan điểm, dự đoán tình cảm qua majority voting và diễn giải qua tầm quan trọng đặc trưng, biểu đồ phụ thuộc từng phần (PDP) và phân tích tương tác đặc trưng.

2.4. Diễn giải mô hình rừng phân loại quan điểm

Diễn giải ở hai cấp độ:

(1) *Cấp cây:* Quy tắc IF-THEN dọc theo đường từ gốc đến lá, ví dụ: “*IF Entity=Hotel AND Aspect=Room AND Score $\geq$ 4 THEN sentiment=positive*”; độ sâu tối thiểu và tầm quan trọng đặc trưng tại nút.

(2) *Cấp rừng:* Tầm quan trọng đặc trưng, tương tác đặc trưng, biểu đồ phụ thuộc từng phần (PDP) cho thấy quan hệ phi tuyến giữa từng đặc trưng và xác suất dự đoán.

Phát hiện Time có độ sâu tối thiểu trung bình nhỏ nhất (1,222, làm nút gốc 146 lần), thời điểm đánh giá là yếu tố phân loại mạnh nhất; Holder đứng thứ hai (1,286, 119 lần). Tương tác âm mạnh Entity:Aspect = -1,728 và Entity:Time = -0,898 phản ánh cấu trúc phi tuyến của không gian bộ năm 5 chiều, là cơ sở thực nghiệm trực tiếp cho các mô hình tiếp theo (Chương 3, Chương 4).

Bảng 2.1. Tầm quan trọng đặc trưng của mô hình F_{CF} trên tập TripAdvisor Hotel Reviews

(nguồn: Bảng 2.11 trong Luận án toàn văn)

Đặc trưng	Độ sâu TB nhỏ nhất	Số lần làm nút gốc	Độ lợi thông tin (IG)	Hoán vị	Vai trò trong phân loại
Time (t)	1,2220	146	89,3295	0,0000	Yếu tố phân loại mạnh nhất - mùa, ngày, xu hướng thay đổi theo thời gian
Holder (h)	1,2860	119	92,1901	0,0021	Phong cách đánh giá nhất quán của từng người dùng
Aspect (a)	1,4820	110	28,6544	0,0043	Khía cạnh cụ thể quyết định tình cảm đặc thù
Entity (e)	1,2900	125	—	0,0122	Thực thể nền tảng; tương tác mạnh với Aspect (−1,728)

Biểu đồ PDP hai chiều trên TripAdvisor Hotel Reviews (Hình 2.11–2.14 của Luận án toàn văn) cho thấy: (1) trên không gian {Time, Holder}, Holder có ảnh hưởng mạnh — tình cảm của một số người dùng ổn định, trong khi của một số người dùng khác thay đổi theo thời gian; (2) trên không gian {Entity, Time}, một số thực thể có xác suất tích cực cao ổn định bất kể thời gian, trong khi các thực thể khác biến động mạnh; (3) Time:Holder = 0,321 là tương tác dương duy nhất (Bảng 2.12 của Luận án toàn văn): khi Time ở nút gốc, Holder có xu hướng xuất hiện sớm hơn trong cây.

2.5. Thực nghiệm

2.5.1. Tập dữ liệu

Bảng 2.2. Thông tin tập dữ liệu Amazon Food và TripAdvisor Hotel

Tập dữ liệu	Quy mô gốc	Sau tiền xử lý	Tỷ lệ chia
Amazon Food Reviews[67]	568.454 đánh giá	8.020 mẫu bộ năm (stratified sampling 1:1)	80% huấn luyện /20% kiểm tra
TripAdvisor Hotel Reviews[2]	108.891 đánh giá	45.706 mẫu bộ năm (loại trung tính 3★)	80% huấn luyện /20% kiểm tra

2.5.2. Kết quả thực nghiệm trên Amazon Food Reviews

Chương 2 thực hiện ba kịch bản thực nghiệm trên hai tập dữ liệu. Tập Amazon Food Reviews [67] gồm 568.454 đánh giá với phân phối mất cân bằng (5★ chiếm 63,88%; 1★–2★ chỉ 14,42%); sau khi áp dụng lấy mẫu cân bằng (balanced sampling) thu được tập thực nghiệm với tỷ lệ positive:negative $\approx$ 1:1. Tập TripAdvisor Hotel Reviews sau tiền xử lý còn 45.706 mẫu bộ năm quan điểm.

Kịch bản 1. Phân loại theo khía cạnh thực phẩm (Bảng 2.8 của Luận án toàn văn): Mô hình F_{CF} đạt Accuracy = 0,80, Precision = 0,80, Recall = 0,76, F_1 = 0,78. Chênh lệch P-R nhỏ (0,80–0,76 = 0,04) cho thấy mô hình phân loại cân bằng hai lớp, kết quả của chiến lược lấy mẫu cân bằng.

Bảng 2.3. Kết quả F_{CF} trên Amazon Food Reviews (nguồn: Bảng 2.8 và 2.9 của Luận án toàn văn)

Kịch bản	Tập dữ liệu	Accuracy	Precision	Recall	F1
KB1. Khía cạnh	Amazon Food Reviews	0,80	0,80	0,76	0,78
KB2. Thời gian/Người	Amazon Food Reviews	0,82	0,85	0,67	0,75

Kịch bản 2. Phân loại theo thời gian và người đánh giá (Bảng 2.9 của Luận án toàn văn): Accuracy tăng lên 0,82 và Precision đạt 0,85 nhưng Recall giảm còn 0,67. Chênh lệch P-R lớn hơn ($0,85 - 0,67 = 0,18$) cho thấy đặc trưng Time và Holder nhận diện tốt lớp positive nhưng bỏ sót nhiều hơn lớp negative. Phân tích tầm quan trọng đặc trưng (Bảng 2.11 của Luận án toàn văn, trên TripAdvisor Hotel Reviews): *Time* có độ sâu tối thiểu trung bình nhỏ nhất (1,2220, làm nút gốc 146 lần), thời điểm đánh giá là yếu tố phân loại mạnh nhất; *Holder* đứng thứ hai (1,2860, 119 lần), phong cách đánh giá mỗi người nhất quán. Tương tác mạnh nhất $Time: Holder = 0,321$ (Bảng 2.12 của Luận án toàn văn); tương tác âm mạnh $Entity: Aspect = -1,728$ cho thấy quan hệ phi tuyến giữa các thành phần bộ năm, động lực trực tiếp dẫn đến F_{DF} (Chương 3) và F_{MF} (Chương 4).

2.5.3. Kết quả trên TripAdvisor Hotel Reviews

Bảng 2.4. Kết quả mô hình F_{CF} theo từng khía cạnh trên TripAdvisor (nguồn: Bảng 2.10 của Luận án toàn văn)

Tập dữ liệu	Accuracy	Precision	Recall	F1	Nhận xét
TripAdvisor Hotel Reviews (45.706 mẫu, 80/20)	0,91	0,96	0,50	0,66	Dữ liệu mất cân bằng với tỉ lệ 4:1

Kết quả Bảng 2.4 cho thấy hiện tượng nghịch lý về độ chính xác: Accuracy = 0,91 rất cao nhưng Recall = 0,50 và F1 = 0,66 thấp cho thấy mô hình bỏ sót 50% mẫu negative. Nguyên nhân là TripAdvisor có tỷ lệ mất cân bằng nhãn khoảng 4:1 (positive chiếm áp đảo), khiến mô hình có xu hướng dự đoán thiên về lớp đa số positive để tối đa hóa Accuracy. Điều này cho thấy chỉ riêng cơ chế Bootstrap chưa đủ để xử lý dữ liệu mất cân bằng nghiêm trọng, và mức mất cân bằng trong dữ liệu đánh giá khách sạn sâu hơn và khó xử lý hơn so với Amazon.

Kết quả thực nghiệm cho thấy mô hình F_{CF} đạt Accuracy 0,80 - 0,82 trên tập dữ liệu Amazon Food Reviews và 0,91 trên TripAdvisor Hotel Reviews. Tuy nhiên, con số 0,91 trên TripAdvisor phản ánh hiện tượng nghịch lý độ chính xác do mất cân bằng nhãn 4:1, Recall chỉ đạt 0,50 và F1 = 0,66. Đối chiếu mô hình F_{CF} với các mô hình học máy khác (RF, Gradient Boosting, SVM) cùng các mốc tham chiếu học sâu/Transformer (LSTM, BiLSTM-Attention,

BERT, RoBERTa – dùng văn bản gốc làm đầu vào) được trình bày tại Chương 3 và Chương 4 trên tập dữ liệu Yelp cùng bộ năm quan điểm.

Mô hình rừng phân loại quan điểm (F_{CF}) có độ phức tạp huấn luyện là $\Theta(n \times N \times m_{try} \times \log N)$ và dự đoán là $\Theta(n \times \log N)$ tương đương với Random Forest tiêu chuẩn [65], độ phức tạp mô hình F_{CF} ở mức thấp và chỉ yêu cầu CPU thông thường. F_{CF} là mô hình đơn giản và hiệu quả tính toán, phù hợp làm nền tảng cho các mô hình phức tạp hơn.

2.6. Kết luận chương

Chương 2 đề xuất mô hình rừng phân loại quan điểm (F_{CF}) với đóng góp cốt lõi là biểu diễn bộ năm quan điểm dưới dạng ma trận $N \times 5$. Mô hình F_{CF} đạt Accuracy 0,80–0,82 (Amazon) và Accuracy 0,91 (TripAdvisor, dù Recall chỉ 0,50 và $F1 = 0,66$ do mất cân bằng nhãn 4:1). Tuy nhiên, mô hình F_{CF} chưa xử lý tốt dữ liệu mất cân bằng và chưa khai thác cấu trúc cụm phi tuyến, hai hạn chế được hướng tới lần lượt trong Chương 3 và Chương 4. Kết quả trong Bảng 2.4 (Bảng 2.10 của Luận án toàn văn) cho thấy giới hạn của F_{CF} trên dữ liệu mất cân bằng thực tế và giải thích lý do cần F_{DF} (Chương 3). Ngoài ra, tương tác âm mạnh $Entity: Aspect = -1,728$ và $Entity: Time = -0,898$ (Bảng 2.12 của Luận án toàn văn) phản ánh cấu trúc phi tuyến của không gian bộ năm 5 chiều, gợi ý thực nghiệm cho hướng tiếp cận F_{MF} (Chương 4). Như vậy Chương 2 không chỉ đề xuất F_{CF} mà còn cung cấp bằng chứng định lượng cho khoảng trống cần giải quyết ở hai chương tiếp theo (Chương 3 và Chương 4).

CHƯƠNG 3. KHAI PHÁ QUAN ĐIỂM VỚI RỪNG MẬT ĐỘ

Nội dung chính bao hàm kết quả nghiên cứu của Bài báo số 4, 5.

3.1. Giới thiệu

Chương 3 đề xuất mô hình F_{DF} tích hợp DBSCAN và SHAP. Mô hình tự động phân cụm ($K = 7 - 9$ cụm), xây rừng mật độ với Entropy Gaussian $H_D(O)$, diễn giải qua SHAP φ_j và tương tác Shapley $\varphi_{i,j}$. Kết quả Accuracy trung bình 0,78 trên hai tập dữ liệu Yelp (số liệu đã công bố, chia theo dòng bộ năm), chạy trên CPU, không cần GPU.

Tóm lược Chương 3 – Bài toán: mất cân bằng lớp và cấu trúc cụm theo mật độ của dữ liệu bộ năm. Phương pháp: rừng mật độ quan điểm – phân cụm DBSCAN (7–9 cụm), mỗi cụm một rừng với entropy Gaussian, diễn giải bằng SHAP và chỉ số tương tác Shapley. Kết quả: Accuracy 0,75 (Hotel) và 0,80 (Restaurant), trung bình 0,78 (số liệu đã công bố, chia theo dòng bộ năm), F1 trung bình 0,68; Entity và Time có trung bình $|\text{SHAP}|$ lớn nhất. Khi chia theo đánh giá gốc (Phụ lục J của Luận án toàn văn), rừng mật độ đạt Balanced Accuracy tương đương Random Forest. Đóng góp: đóng góp thứ hai của luận án (Bài báo 4, 5).

3.2. Biểu diễn và chuẩn hóa bộ năm quan điểm

Mỗi bộ năm quan điểm $o_i = (e_i, a_i, s_i, h_i, t_i)$ được ánh xạ sang vector số trong không gian $\mathbb{R}^5$ qua label encoding. Chuẩn hóa Z-score: $\hat{x} = (x - \mu)/\sigma$ đảm bảo thang đo đồng nhất trước khi đưa vào DBSCAN.

3.3. Mô hình rừng mật độ quan điểm

3.3.1. Phân cụm mật độ quan điểm bằng DBSCAN

DBSCAN [72] phát hiện cụm dựa trên tham số ε (bán kính lân cận) và MinPts. Điểm o là điểm lõi nếu $|N_\varepsilon(o)| \geq \text{MinPts}$.

3.3.2. Cây mật độ quan điểm

Cây mật độ được xây dựng không giám sát sử dụng Entropy Gaussian, công thức:

$$H(O) = \frac{1}{2} \log((2\pi e)^d |\Lambda(O)|)$$

trong đó $d = 4$ là số chiều đặc trưng đầu vào (e, a, h, t); $|\Lambda(O)|$ là ma trận hiệp phương sai $d \times d$ của tập điểm O tại nút; $|\Lambda(O)|$ là định thức ma trận. Entropy mật độ đo độ phân tán của đám mây dữ liệu qua thể tích hình học. Tiêu chí phân tách là độ lợi thông tin mật độ:

$$IG_D(O_j, O_{jL}, O_{jR}) = H_D(O_j) - \frac{|O_{jL}|}{|O_j|} H_D(O_{jL}) - \frac{|O_{jR}|}{|O_j|} H_D(O_{jR})$$

Nút tốt nhất tối đa hóa IG_D phân chia làm hai nhóm co cụm chặt nhất.

3.3.3. Rừng mật độ quan điểm

Rừng mật độ gồm n cây mật độ được huấn luyện ngẫu nhiên trên mỗi cụm C_k . Mật độ tổng hợp: $p_{DF}(o) = \frac{1}{n} \sum_{k=1}^n p_k(o)$. Nhãn tình cảm theo cụm mật độ cao nhất.

3.4. Diễn giải rừng mật độ với SHAP và tương tác Shapley

Giá trị SHAP [125] của đặc trưng j , công thức:

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{j\}}(o_{S \cup \{j\}}) - f_S(o_S)]$$

Chỉ số tương tác Shapley $\varphi_{i,j}$ đo đóng góp kết hợp cặp đặc trưng (i, j) , công thức (3.28):

$$I_{i,j}(o) = \sum_{S \subseteq F \setminus \{i,j\}} \frac{|S|!(|F|-|S|-2)!}{2(|F|-1)!} \delta_{ij}(o, S)$$

3.5. Độ phức tạp tính toán

Độ phức tạp huấn luyện. Độ phức tạp huấn luyện của thuật toán xây dựng rừng mật độ được xác định bởi sự kết hợp giữa chi phí xây dựng cây và chi phí phân cụm mật độ DBSCAN tại các nút lá được biểu diễn theo công thức

$$\Theta(n \times m_{try} \times N \times \log N) + \Theta(n \times L \times N \times \log N)$$

Trong đó:

- n : số lượng cây trong rừng mật độ;
- N : số lượng mẫu huấn luyện;
- m_{try} : số đặc trưng được chọn ngẫu nhiên tại mỗi nút phân tách;
- L : số nút lá trung bình của một cây, xấp xỉ $N/\min_split$;

Độ phức tạp dự đoán. Độ phức tạp của thuật toán dự đoán và định hướng quan điểm được xác định bởi ba giai đoạn thực thi độc lập. Với n cây trong rừng, K khía cạnh cần định hướng và N mẫu huấn luyện, độ phức tạp tổng thể là:

$$\Theta(n \times \log N) + \Theta(K \times n \times \log N) + \Theta(n \times L \times d^2)$$

3.6. Thực nghiệm

3.6.1. Tập dữ liệu

Hai tập dữ liệu Yelp được sử dụng: yelpHotelData (đánh giá khách sạn) và yelpResData (đánh giá nhà hàng). Sau khi nhị phân hóa nhãn ($4\star-5\star \rightarrow \text{positive}$; $1\star-3\star \rightarrow \text{negative}$): yelpHotelData gồm 560.769 mẫu positive và 321.705 mẫu negative; yelpResData gồm 9.650.537 mẫu positive và 5.140.963 mẫu negative. Chia 80% huấn luyện/20% kiểm tra với stratified sampling.

3.6.2. Kết quả thực nghiệm

Thiết lập tham số phân cụm DBSCAN: $\varepsilon = 0,5$, MinPts=5, kết quả $K = 7$ cụm (yelpHotelData) và $K = 9$ cụm (yelpResData).

Bảng 3.1. Kết quả mô hình F_{DF} trên hai tập dữ liệu Yelp

Tập dữ liệu	Accuracy (Huấn luyện)	Accuracy (Kiểm tra)	Precision	Recall	F1-score
yelpHotelData	0,84	0,75	0,71	0,62	0,67

(7 cụm)					
yelpResData (9 cụm)	0,85	0,80	0,72	0,67	0,69

Phân tích SHAP: Entity và Time có tầm quan trọng cao nhất (trung bình $|\text{SHAP}| \approx 0,087$ và $0,085$), Holder đứng thứ ba ($\approx 0,077$), Aspect thấp nhất ($\approx 0,023$) do số lượng khía cạnh ít nên khả năng phân biệt thấp (Hình 3.5 của Luận án toàn văn). Phân tích tương tác SHAP (Hình 3.6 của Luận án toàn văn) cho thấy các tương tác của đặc trưng với chính nó và cặp Entity – Time xuất hiện gần gốc cây nhất, còn các tương tác liên quan đến Aspect chỉ xuất hiện ở các nhánh sâu. Mô hình có độ chính xác trên tập dữ liệu yelpResData đạt 0,80 cao hơn tập dữ liệu yelpHotelData (0,75) vì đặc trưng Entity nhà hàng đa dạng và rõ ràng hơn.

3.6.3. So sánh với các phương pháp tiêu biểu

Về tính công bằng so sánh: các phương pháp học máy truyền thống dựa trên đặc trưng có cấu trúc (Random Forest, Gradient Boosting dựa trên cây quyết định; SVM dựa trên siêu phẳng phân tách biên cực đại) và mô hình rừng mật độ quan điểm đề xuất đều sử dụng đặc trưng bộ năm quan điểm đã mã hóa số làm đầu vào; trong khi đó, các mô hình học sâu và Transformer (LSTM, BiLSTM-Attention, BERT, RoBERTa) nhận trực tiếp câu văn bản đánh giá gốc làm đầu vào theo đúng cách vận hành tiêu chuẩn của các kiến trúc này, trên cùng tập mẫu, cùng nhãn tình cảm và cùng tỉ lệ chia huấn luyện/kiểm tra.

Bảng 3.2. So sánh F_{DF} với 7 phương pháp tiêu biểu

Mô hình	Mô hình hóa cấu trúc	Độ chính xác trung bình	Điểm F1 trung bình	Khả năng giải thích	Độ phức tạp tính toán
Mô hình F_{DF} (Đề xuất)	Có	0,78	0,68	Rất cao	$\Theta(n \times m_{try} \times N \times \log N) + \Theta(n \times L \times N \times \log N)$
RoBERTa [83]	Không	0,79	0,70	Rất thấp	$\Theta(L^2 \times d)$ cho mỗi token (Rất cao - GPU)
BERT [81]	Không	0,78	0,69	Rất thấp	$\Theta(L^2 \times d)$ cho mỗi token (Rất cao - GPU)
BiLSTM-Attention [88]	Không	0,77	0,68	Thấp	$\Theta(n \times L \times d)$ (Trung bình)
Gradient Boosting [84]	Không	0,76	0,67	Cao	$\Theta(n \times N \times \log N)$ (Trung bình)
Random Forest [65]	Không	0,76	0,66	Cao	$\Theta(n \times N \times m_{try} \times \log N)$ (Thấp)
LSTM [85]	Không	0,74	0,65	Thấp	$\Theta(n \times L \times d^2)$ (Trung bình)
SVM [86]	Không	0,75	0,66	Trung bình	$\Theta(N^2 \sim N^3)$ (Cao)

Mô hình F_{DF} đạt Accuracy trung bình 0,78, cao hơn Random Forest (0,76) theo số liệu đã công bố, chia theo dòng bộ năm, không cần GPU và cung cấp diễn giải qua cấu trúc cụm và giá trị SHAP. Các mô hình học sâu và Transformer sử dụng văn bản đánh giá gốc làm đầu vào, đạt từ 0,74 (LSTM) đến 0,79 (RoBERTa), được trình bày như mốc tham chiếu. Số liệu theo Bài báo 5; kết quả chạy lại theo quy trình thống nhất được đối chiếu tại Phụ lục J của Luận án toàn văn: khi chia theo đánh giá gốc, rừng mật độ tương đương Random Forest (chênh lệch không có ý nghĩa thống kê).

Mô hình F_{DF} có độ phức tạp huấn luyện cao hơn Random Forest do thêm thành phần DBSCAN $\Theta(n \times L \times N \times \log N)$ tại các nút lá, nhưng chi phí thực tế thấp hơn lý thuyết vì DBSCAN chỉ chạy trên tập con tại nút lá. Các mô hình BERT/RoBERTa $\Theta(L^2 \times d)$ yêu cầu GPU chuyên dụng, trong khi mô hình F_{DF} chạy trên CPU với chi phí tính toán hợp lý.

3.7. Kết luận Chương

Chương 3 đề xuất mô hình F_{DF} tích hợp DBSCAN và SHAP cho khai phá quan điểm. Mô hình tự động phân cụm dữ liệu (7–9 cụm) theo phân phối mật độ thực tế, xây rừng mật độ không giám sát với Entropy Gaussian $H_D(O)$ và diễn giải kết quả qua giá trị SHAP và tương tác Shapley. F_{DF} đạt Accuracy trung bình 0,78 (số liệu đã công bố, chia theo dòng bộ năm; khi chia theo đánh giá gốc, tương đương Random Forest), chạy trên CPU với chi phí tính toán thấp. Tuy nhiên, mô hình F_{DF} sử dụng khoảng cách Euclidean trong DBSCAN chưa khai thác cấu trúc hình học phi tuyến sâu hơn, hạn chế này được hướng tới trong Chương 4.

CHƯƠNG 4. KHAI PHÁ QUAN ĐIỂM VỚI RỪNG ĐA TẬP

Nội dung chính bao hàm kết quả nghiên cứu của Bài báo số 6.

4.1. Giới thiệu

Chương 4 đề xuất mô hình Rừng đa tập quan điểm (F_{MF} - Opinion Manifold Forest) giải quyết quan hệ phi tuyến sâu hơn trong không gian bộ năm. Thay vì khoảng cách Euclidean, F_{MF} học khoảng cách đa tập $dist_{\mathcal{M}}$ từ chính dữ liệu qua ma trận quan hệ W , khai thác cấu trúc hình học d' -chiều tiềm ẩn ($d' < 5$) của không gian bộ năm Yelp.

Tóm lược Chương 4 – Bài toán: quan hệ phi tuyến giữa các thành phần bộ năm mà khoảng cách Euclidean chưa nắm bắt được. Phương pháp: rừng đa tập quan điểm – ma trận quan hệ W học từ rừng không giám sát, khoảng cách đa tập $\sqrt{1 - W}$, DBSCAN trên đa tập (3–4 cụm), tổng hợp K rừng bằng Stacking. Kết quả: Accuracy 0,78 (Hotel) và 0,82 (Restaurant), trung bình 0,80 (số liệu đã công bố, chia theo dòng bộ năm; khi chia theo đánh giá gốc, thấp hơn Random Forest khoảng 0,02–0,03 Balanced Accuracy); chiều nội tại ước lượng $d' \approx 2,78$ (Hotel) và 3,66 (Restaurant), nhỏ hơn số chiều quan sát $D = 4$ (trên Restaurant, phương pháp DANCo cho $d' = 4$ nên bằng chứng chưa nhất quán). Đóng góp: đóng góp thứ ba của luận án (Bài báo 6).

4.2. Phân cụm quan điểm trên không gian đa tập

4.2.1. Biểu diễn đặc trưng và giả thuyết đa tập

Giả thuyết đa tập: dữ liệu bộ năm Yelp thực chất nằm gần đa tập d' -chiều với $d' < 5$, do: (1) quan hệ phi tuyến thực thể - khía cạnh; (2) phong cách đánh giá nhất quán của người dùng; (3) phân phối điểm sao tập trung ở hai cực. Tính phù hợp được kiểm chứng gián tiếp qua kết quả thực nghiệm (§4.8 của Luận án toàn văn) và kiểm chứng trực tiếp bằng ước lượng chiều nội tại (Phụ lục K của toàn văn luận án).

4.2.2. Phân cụm DBSCAN trên không gian đa tập

Phân cụm DBSCAN được áp dụng trên không gian đa tập với khoảng cách

$$dist_{\mathcal{M}}(o_i, o_j) = \sqrt{1 - W_{ij}}$$

thay vì khoảng cách Euclidean thông thường. Hai quan điểm “gần nhau” theo nghĩa ngữ nghĩa thường đồng nút lá trong rừng.

4.3. Cây đa tập quan điểm

Cây đa tập được xây dựng sử dụng Entropy Gaussian đa tập ở công thức:

$$H_{\mathcal{M}}(O) = \frac{1}{2} \log[(2\pi e)^{d'} |\Lambda_{\mathcal{M}}(O)|]$$

trong đó d' là chiều nội tại đa tập ($d' < d = 5$); $\Lambda_{\mathcal{M}}(O)$ là ma trận hiệp phương sai có trọng số W_{ij} được tính bởi công thức:

$$\Lambda_{\mathcal{M}}(O) = \frac{\sum_{i,j \in O} W_{ij} (o_i - \mu(O))(o_j - \mu(O))^T}{\sum_{i,j \in O} W_{ij}},$$

Tiêu chí phân tách tối ưu hóa tại công thức (4.9):

$$IG_{\mathcal{M}} = \frac{1}{2} \log |\Lambda_{\mathcal{M}}(O_j)| - \sum_{i \in \{L,R\}} \frac{|O_{ji}|}{|O_j|} \log |\Lambda_{\mathcal{M}}(O_{ji})|$$

4.4. Rừng đa tập quan điểm

4.4.1. Ma trận quan hệ đa tập

Ma trận quan hệ đa tập

$$W \in \mathbb{R}^{N \times N}.$$

được xây dựng từ n_W cây không giám sát bằng cơ chế đếm đồng nút lá tại công thức:

$$W_{ij} = \frac{1}{n_W} \sum_{k=1}^{n_W} W_{ij}^{(k)} = \frac{1}{n_W} \sum_{k=1}^{n_W} \mathbf{1} \left[l_{T_k^{(w)}}(o_i) = l_{T_k^{(w)}}(o_j) \right]$$

trong đó $L(o_i, T_k)$ là chỉ số nút lá mà o_i rơi vào trong cây T_k . $W_{ij} \in [0,1]$: $W_{ij} = 1$ nghĩa là luôn đồng nút lá (rất gần về ngữ nghĩa); $W_{ij} = 0$ nghĩa là không bao giờ đồng nút lá.

Khoảng cách đa tập ở công thức:

$$dist_{\mathcal{M}}(o_i, o_j) = \sqrt{1 - W_{ij}}$$

Ma trận W phản ánh cấu trúc ngữ nghĩa tốt hơn Euclidean: cùng thực thể e và khía cạnh a sẽ có W_{ij} cao dù khoảng cách Euclidean có thể lớn.

4.4.2. Rừng đa tập quan điểm

Sau khi thực hiện phân cụm trên không gian đa tập như trình bày ở Mục 4.2.2, một Rừng đa tập quan điểm chuyên biệt

$$\mathcal{F}_k,$$

được xây dựng cho từng cụm đa tập nhằm học các đặc trưng và ranh giới quyết định đặc thù của từng vùng dữ liệu.

Để khai thác thông tin từ tất cả các cụm đa tập, luận án sử dụng mô hình học quần thể Stacking, trong đó bộ học siêu cấp (meta-learner)

$$M,$$

được xây dựng bằng Random Forest. Bộ học này tổng hợp các dự đoán của K rừng đa tập chuyên biệt trên tập xác thực để đưa ra quyết định phân loại cuối cùng theo công thức:

$$\hat{y} = M(RF_1(o), RF_2(o), \dots, RF_K(o)).$$

Cơ chế Stacking giúp tận dụng tính bổ sung giữa các rừng đa tập chuyên biệt, từ đó nâng cao khả năng tổng quát hóa và cải thiện hiệu năng phân loại so với việc sử dụng một mô hình đơn lẻ.

4.5. Độ phức tạp tính toán

Độ phức tạp thuật toán xây dựng rừng đa tập đơn lẻ. Độ phức tạp huấn luyện được tách thành hai giai đoạn: (1) xây ma trận quan hệ đa tập W (n_W cây $T_k^{(w)}$ không giám sát) và (2) xây rừng đa tập quan điểm dựa trên khoảng cách đa tập đã học:

$$\Theta(N \times \log N \times (n_w + n \times d'))$$

Độ phức tạp thuật toán dự đoán của rừng đa tập. Độ phức tạp của dự đoán rừng đa tập gồm hai bước: Tính giá trị ước lượng $\widehat{W}(o, o_i)$ cho k_{rep} cặp (o, o_i) mẫu mới và duyệt n cây để phân loại.

$$\Theta((k_{\text{rep}} + 1) \times n \times \log N)$$

Độ phức tạp thuật toán xây dựng kết hợp các rừng đa tập bao gồm bốn bước độc lập, thực hiện tuần tự: Xây dựng ma trận quan hệ đa tập W (Giai đoạn 1 Thuật toán 4.1); Phân cụm đa tập DBSCAN; Xây dựng K rừng $\mathcal{F}_k$ chuyên môn hóa và Huấn luyện meta-learner Stacking:

$$\Theta(N \times \log N \times (n_w + n \times d')) + \Theta(K \times N_{\text{val}})$$

4.6. Kết quả thực nghiệm

4.6.1. Kết quả trên hai tập dữ liệu Yelp

Thiết lập tham số phân cụm trên đa tập:

- $\varepsilon = 0,3$ (nhỏ hơn Chương 3 dùng $\varepsilon = 0,5$ vì $\text{dist}_{\mathcal{M}} \in [0,1]$ có phân phối dày đặc hơn khoảng cách Euclidean, cùng ngưỡng ε sẽ tạo quá nhiều cụm);
- $\text{MinPts} = 5$ (giữ nhất quán với Chương 3 để so sánh công bằng). Tham số ε được lựa chọn qua biểu đồ k -distance trên không gian đa tập: tìm điểm gãy (elbow point) của đường cong khoảng cách đến lân cận gần thứ MinPts .

Kết quả là $K = 3$ cụm trên yelpHotelData (ứng với ba rừng thành phần RF1, RF2, RF3) và $K = 4$ cụm trên yelpResData (RF1–RF4). Các cụm được đặc tả theo thống kê thực tế của các đặc trưng (e , a , h , t) trong từng cụm, không gán nhãn chủ đề ngữ nghĩa, vì tập yelpHotelData chỉ gồm ba khía cạnh AcceptsCreditCards, PriceRange và WiFi.

Số cụm $K = 3/4$ ít hơn đáng kể so với Chương 3 ($K=7/9$ cụm Euclidean) vì không gian đa tập nén các quan điểm ngữ nghĩa tương tự lại thành cụm chặt hơn, điều này dẫn đến các rừng chuyên biệt $\mathcal{F}_k$ học được ranh giới quyết định thuần nhất hơn, có thể góp phần giải thích vì sao, theo số liệu đã công bố, F_{MF} đạt Accuracy cao hơn F_{DF} .

Bảng 4.1. Kết quả F_{MF} trên yelpHotelData và yelpResData (nguồn: Bảng 4.2 và 4.3 của Luận án toàn văn)

Tập dữ liệu	Tập	Accuracy	Precision	Recall	F1	Nhận xét
yelpHotelData	Train	0,94	—	—	—	Mô hình học tốt cấu trúc khách sạn
yelpHotelData	Test	0,78	0,74	0,62	0,67	Recall thấp: vẫn còn mất cân bằng
yelpResData	Train	0,95	—	—	—	Dữ liệu nhà hàng phong phú hơn

Tập dữ liệu	Tập	Accuracy	Precision	Recall	F1	Nhận xét
yelpResData	Test	0,82	0,75	0,67	0,71	Cao nhất trong 3 mô hình đề xuất (số liệu công bố)

Ghi chú: Bảng chỉ báo cáo các độ đo trên tập kiểm tra, là cơ sở so sánh với các mô hình khác (Bảng 4.4 của Luận án toàn văn); Accuracy trên tập huấn luyện được trình bày tại Hình 4.7, 4.8 để đánh giá mức khớp quá mức. Kết quả chạy lại ba mô hình rừng với mã nguồn hiện tại được trình bày tại Phụ lục J (Bảng J.7).

Mô hình rừng đa tập quan điểm F_{MF} đạt Accuracy 0,78/0,82 (Hotel/ Restaurant) trên tập kiểm tra, cao nhất trong ba mô hình đề xuất. Khoảng cách Train - Test: Hotel (0,94 $\rightarrow$ 0,78 = 0,16), Res (0,95 $\rightarrow$ 0,82 = 0,13). Khoảng cách lớn hơn so với F_{DF} (0,09/0,05) có thể do kiến trúc phức tạp hơn (4 giai đoạn + Stacking), và kết quả test cao nhất trong ba mô hình theo số liệu đã công bố, chia theo dòng bộ năm; tuy nhiên khoảng cách Train – Test lớn đòi hỏi thận trọng với nguy cơ quá khớp, và khi chia theo đánh giá gốc (Phụ lục J của Luận án toàn văn), rừng đa tập thấp hơn Random Forest khoảng 0,02–0,03 Balanced Accuracy.

Chuỗi tiến triển 3 mô hình $F_{CF}(0,77) \rightarrow F_{DF}(0,78) \rightarrow F_{MF}(0,80)$. Theo số liệu đã công bố, chia theo dòng bộ năm, mức cải thiện: +1% từ rừng phân loại sang rừng mật độ và +2% (trung bình) từ rừng mật độ sang rừng đa tập (Hotel +3%: 0,78 so với 0,75; Restaurant +2%: 0,82 so với 0,80). Mức cải thiện này không được tái lập khi chia theo đánh giá gốc (Phụ lục J của Luận án toàn văn). Các lý do có thể của mức cải thiện theo số liệu công bố: (1) khoảng cách $dist_{\mathcal{M}}(o_i, o_j) = \overline{1 - W_{ij}}$ phản ánh ngữ nghĩa tốt hơn Euclidean; (2) phân cụm chính xác hơn trên không gian đa tập (K=3/4 cụm so với 7/9 cụm Euclidean); (3) Stacking meta-learner học cho từng rừng chuyên biệt thay vì trung bình đơn giản.

Recall trên tập kiểm tra nhất quán thấp (0,62–0,67) ở các mô hình rừng là thách thức chung của dữ liệu Yelp mất cân bằng nhãn (khoảng 1,7:1 và 1,9:1). Recall của rừng đa tập ngang với rừng mật độ (Restaurant: 0,67; Hotel: 0,62). Rừng đa tập không cần GPU và cung cấp diễn giải tường minh qua ma trận W và IG_M , phù hợp với yêu cầu AI có trách nhiệm.

4.6.2. So sánh tổng hợp tất cả mô hình

Tương tự phương pháp luận đã trình bày tại mục 3.6.3: các phương pháp học máy truyền thống dựa trên đặc trưng có cấu trúc (Random Forest, Gradient Boosting, SVM và ba mô hình rừng đề xuất) sử dụng đặc trưng bộ năm quan điểm đã mã hóa số làm đầu vào; các mô hình học sâu và Transformer (LSTM, BiLSTM-Attention, BERT, RoBERTa) sử dụng câu văn bản đánh giá gốc làm đầu vào theo đúng cách vận hành tiêu chuẩn của các kiến trúc này, đảm bảo tính công bằng so sánh ở cùng tập mẫu, cùng nhãn tình cảm và cùng tỉ lệ chia huấn luyện/kiểm tra.

Bảng 4.2. So sánh tổng hợp tất cả mô hình trên yelpHotelData và yelpResData

Mô hình	Độ chính xác trên tập test yelpHotelData	Độ chính xác trên tập test yelpResData	Độ chính xác Trung bình	Khả năng giải thích	Độ phức tạp tính toán
Rừng đa tạp (F_{MF})	0,78	0,82	0,80	Rất cao	$\Theta(N \times \log N \times (nW + n \times d')) + \Theta(K \times N_{val})$ (Kết hợp 2 giai đoạn)
Rừng mật độ (F_{DF})	0,75	0,80	0,78	Rất cao	$\Theta(n \times m_{try} \times N \times \log N) + \Theta(n \times L \times N \times \log N)$ (Trung bình)
Rừng phân loại (F_{CF})	0,76	0,77	0,77	Rất cao	$\Theta(n \times N \times m_{try} \times \log N)$ (Thấp)
BiLSTM-Attention [88]	0,76	0,78	0,77	Thấp	$\Theta(n \times L \times d)$ (Trung bình)
Random Forest [65]	0,74	0,77	0,76	Cao	$\Theta(n \times N \times m_{try} \times \log N)$ (Thấp)
Gradient Boosting [84]	0,74	0,78	0,76	Cao	$\Theta(n \times N \times \log N)$ (Trung bình)
SVM [86]	0,70	0,80	0,75	Trung bình	$\Theta(N^2 \sim N^3)$ (Cao)
LSTM [85]	0,71	0,76	0,74	Thấp	$\Theta(n \times L \times d^2)$ (Trung bình)
RoBERTa [83]	0,74	0,83	0,79	Rất thấp	$\Theta(L^2 \times d)$ cho mỗi token (Rất cao - GPU)
BERT [81]	0,74	0,82	0,78	Rất thấp	$\Theta(L^2 \times d)$ cho mỗi token (Rất cao - GPU)

Mô hình rừng đa tạp quan điểm (F_{MF}) đạt Average Accuracy 0,80 theo số liệu đã công bố, chia theo dòng bộ năm, cao nhất trong ba mô hình rừng đề xuất, cao hơn F_{DF} 2% và F_{CF} 3%. Các lý do có thể: (1) khai thác cấu trúc phi tuyến qua học đa tạp; (2) phân cụm chính xác hơn trong không gian đa tạp so với Euclidean; (3) ma trận W phản ánh ngữ nghĩa quan điểm sâu hơn khoảng cách Euclidean. Các mô hình học sâu và Transformer sử dụng văn bản đánh giá gốc làm đầu vào, đạt từ 0,74 (LSTM) đến 0,79 (RoBERTa), được trình bày như mốc tham chiếu (số liệu theo Bài báo 5, 6; đối chiếu kết quả chạy lại tại Phụ lục J của Luận án toàn văn).

Rừng đa tạp không cần GPU và cung cấp diễn giải tường minh qua ma trận đa tạp W và độ lợi thông tin trên đa tạp $IG_{\mathcal{M}}$.

Mô hình F_{MF} có độ phức tạp huấn luyện $\Theta(N \times \log N \times (nW + n \times d')) + \Theta(K \times N_{val})$ gồm 2 giai đoạn: xây dựng ma trận đa tạp W và xây rừng đa tạp, nhưng giai đoạn xây dựng ma trận đa tạp W chỉ chạy một lần và được chia sẻ. Mô hình F_{DF} thêm chi phí DBSCAN tại nút lá và mô hình F_{CF} là đơn giản nhất. Cả ba mô hình đề xuất không cần GPU, trong khi BERT/RoBERTa có độ phức tạp $\Theta(L^2 \times d)$ đòi hỏi phần cứng chuyên dụng.

4.7. Kết luận Chương

Chương 4 đề xuất mô hình F_{MF} khai thác giả thuyết đa tạp trong dữ liệu bộ năm quan điểm, một hướng tiếp cận trong phân tích tình cảm dựa trên khía cạnh (ABSA). Mô hình xây ma trận quan hệ W từ rừng không giám sát, phân cụm DBSCAN trên không gian đa tạp và tổng hợp qua Stacking. Kết quả Accuracy trung bình 0,80 (số liệu đã công bố, chia theo dòng bộ năm) và ước lượng chiều nội tại cho thấy tính khả thi của giả thuyết $d' < 5$ trên dữ liệu Yelp (rõ trên Hotel, chưa nhất quán trên Restaurant). Khi chia theo đánh giá gốc, rừng đa tạp thấp hơn Random Forest khoảng 0,02–0,03 Balanced Accuracy do bước phân cụm đa tạp chỉ học trên mẫu con; chi phí ma trận W là rào cản cần giải quyết khi mở rộng quy mô. Mô hình rừng đa tạp quan điểm (F_{MF}) đóng góp vào xu hướng học quần thể cho xử lý ngôn ngữ tự nhiên và phù hợp với yêu cầu mô hình AI phát triển ngày càng có trách nhiệm hơn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Khai phá quan điểm là lĩnh vực có nhiều ứng dụng trong kinh doanh, giải trí, y tế, giáo dục và quản lý xã hội. Luận án “Khai phá quan điểm với các phương pháp quần thể” đã đề xuất và đánh giá ba mô hình khai phá quan điểm dựa trên học quần thể với mô hình cơ sở đồng nhất là cây quyết định.

Kết quả nghiên cứu đã đạt được (các giá trị độ chính xác dưới đây là số liệu đã công bố, tính theo quy trình chia ngẫu nhiên theo dòng bộ năm; kết quả chia theo đánh giá gốc được tóm tắt ở cuối phần này):

- *Rừng phân loại quan điểm (F_{CF})*: Độ chính xác (Accuracy) là 0,77 trên tập dữ liệu yelpHotelData và yelpResData. Diễn giải tốt qua tầm quan trọng đặc trưng và PDP.

- *Rừng mật độ quan điểm (F_{DF})*: Độ chính xác trung bình (Average Accuracy) là 0,78 trên tập dữ liệu yelpHotelData và yelpResData, không cần GPU. Đồng thời, diễn giải bằng giá trị SHAP đã xác định thực thể (Entity) là đặc trưng chủ đạo (trung bình $|SHAP| \approx 0,087$, Hình 3.5 của Luận án toàn văn).

- *Rừng đa tập quan điểm (F_{MF})*: khai thác giả thuyết đa tập trong dữ liệu bộ năm quan điểm trên tập dữ liệu yelpHotelData và yelpResData. Mô hình xây ma trận quan hệ W , phân cụm DBSCAN trên không gian đa tập ($d' < 5$) và tổng hợp qua Stacking. Kết quả độ chính xác trung bình (Average Accuracy) là 0,80, cao nhất trong ba mô hình đề xuất.

- Quy trình chuyển đổi dữ liệu sang bộ năm quan điểm, áp dụng cho yelpHotelData (>880K bộ năm), yelpResData (>14M bộ năm) và hai tập tiếng Việt VLSP-2018 Restaurant (>15K cặp đánh giá - khía cạnh), VLSP-2018 Hotel (>23K cặp); trên VLSP-2018, kết quả của ba mô hình rừng được trình bày cùng các mô hình cơ sở và mốc tham chiếu PhoBERT (Phụ lục F của Luận án toàn văn).

- Ngoài ra, luận án còn đề xuất mô hình trích xuất bộ năm quan điểm, đạt $F_1 = 0,78$ ($P=0,81$, $R=0,75$) trên tập dữ liệu Twitter US Airline Sentiment. Mô hình này cung cấp nền tảng cho việc xây dựng các tập dữ liệu bộ năm quan điểm.

Chuỗi tiến triển độ chính xác (Accuracy) mô hình từ mô hình rừng phân loại quan điểm $F_{CF}((0,77))$ đến rừng mật độ quan điểm $F_{DF}(0,78)$ và rừng đa tập quan điểm $F_{MF}(0,80)$ theo số liệu đã công bố phản ánh việc khai thác cấu trúc dữ liệu ngày càng tinh tế, đồng thời cả ba mô hình duy trì khả năng diễn giải qua cấu trúc cây, cụm và ma trận quan hệ.

Khi đánh giá lại theo quy trình chia theo đánh giá gốc (Phụ lục J của Luận án toàn văn, 5 hạt giống $\times$ 10.000 đánh giá), Balanced Accuracy của mọi mô hình dùng bộ năm giảm về khoảng 0,51–0,57 trên cả hai tập dữ liệu: F_{CF} và F_{DF} tương đương Random Forest (chênh lệch không có ý nghĩa thống kê) và thấp hơn SVM tuyến tính khoảng 0,01; F_{MF} thấp hơn Random Forest khoảng 0,02–0,03 do bước phân cụm đa tập chỉ học trên mẫu con. Các mô hình đọc văn bản gốc (BERT, RoBERTa) đạt 0,76–0,82, nhưng khi mọi mô hình chỉ dùng bộ năm, F_{CF} và F_{DF} tương đương BiLSTM-Attention và cao hơn về trung bình so với LSTM, BERT, RoBERTa. Trên VLSP-2018, F_{DF} đạt Balanced Accuracy 0,724 trên tập Hotel (không khác biệt có ý nghĩa so với Logistic Regression 0,730) và 0,660 trên tập Restaurant (cao nhất trong các mô hình dùng bộ năm), trong khi mốc tham chiếu PhoBERT đọc văn bản gốc đạt 0,832 (Hotel) và 0,791 (Restaurant).

Như vậy, đóng góp của ba mô hình rừng nằm chủ yếu ở cách biểu diễn bộ năm có cấu trúc và khả năng diễn giải; độ chính xác bị giới hạn bởi lượng thông tin của bốn thành phần (e, a, h, t). Đối chiếu với ba khoảng trống nghiên cứu: khả năng diễn giải theo từng thành phần quan điểm được đáp ứng; xử lý mất cân bằng mới đáp ứng một phần (tác động rõ khi mất cân bằng mạnh như VLSP-2018 Restaurant); hướng khai thác cấu trúc đa tập cho thấy tính khả thi nhưng còn rào cản về quy mô của ma trận W . Ngoài ra, luận án định lượng mức rò rỉ dữ liệu khi chia theo dòng bộ năm (với cùng Random Forest và 112.275 bộ năm, Accuracy là 0,946 khi chia theo dòng so với 0,611 khi chia theo đánh giá gốc) và đề xuất chia theo đánh giá gốc khi đánh giá các mô hình trên dữ liệu bộ năm.

Tuy nhiên, luận án cũng nhận thức rõ hạn chế: các mô hình rừng xử lý đặc trưng đã trích xuất sẵn (bộ năm quan điểm), phụ thuộc chất lượng bước trích xuất ban đầu, trong khi các mô hình end-to-end Transformer học biểu diễn trực tiếp từ văn bản thô, có khả năng nắm bắt mỉa mai, phủ định và quan điểm ẩn.

2. Hướng phát triển

Ngoài các kết quả đạt được, luận án có thể tiếp tục mở rộng các hướng nghiên cứu sau:

Phát triển tập dữ liệu bộ năm quan điểm yelpHotelData và yelpResData thành tập dữ liệu đa miền, đa ngôn ngữ, bổ sung tiếng Việt từ VLSP-2018 và các ngôn ngữ khác.

Xây dựng bộ năm quan điểm dựa trên Sentiment Treebank, phát triển các thuật toán phân cụm, dựng cây, dựng rừng khác nhau để khai phá quan điểm. Đồng thời phát triển mô

hình khai phá quan điểm theo hướng diễn giải để tăng tính hiệu quả và độ tin cậy, chắc chắn của mô hình.

Kết hợp với biểu diễn ngữ nghĩa của các mô hình học sâu (BERT embedding, PhoBERT cho tiếng Việt) với cấu trúc rừng để cải thiện chất lượng trích xuất bộ năm và xử lý mỉa mai, phủ định, quan điểm ẩn.

Mở rộng sang phân tích quan điểm theo thời gian thực trên dữ liệu streaming và bài toán ABSA đa ngôn ngữ.

Giảm chi phí tính và lưu ma trận quan hệ W (lấy mẫu con, biểu diễn thưa theo chỉ số nút lá, tính W cục bộ theo cụm) để mở rộng rừng đa tạp quan điểm cho dữ liệu quy mô lớn.

Kết hợp diễn giải mô hình với suy luận nhân quả (phân tích phản thực tế, mô hình nhân quả có cấu trúc) để phân biệt các yếu tố thực sự tác động đến cực tính cảm xúc với các tương quan trong dữ liệu quan sát.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

TT	Tên công trình	Thể loại	Năm công bố
1	Bài báo số 1	Tạp chí Quốc tế International Journal of Advanced Computer Science and Applications (IJACSA)	2020
2	Bài báo số 2	Hội nghị Quốc tế The EAI International Conference on Context-Aware Systems and Applications (ICCASA)	2021
3	Bài báo số 3	Hội nghị Quốc tế The International Conference on Intelligent Information Technology (ICIIT)	2023
4	Bài báo số 4	Tạp chí Quốc tế Journal of EAI Endorsed Transactions on Context-aware Systems and Applications	2023
5	Bài báo số 5	Hội nghị Quốc tế The International Conference on Machine Learning and Soft Computing (ICMLSC)	2024
6	Bài báo số 6	Hội nghị Quốc tế The EAI International Conference on Context-Aware Systems and Applications (ICCASA)	2024
7	Bài báo số 7	Hội nghị Quốc tế The RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam	2025
8	Bài báo số 8	Hội nghị Quốc tế The International Conference on Intelligent Information Technology (ICIIT)	2026
9	Bài báo số 9	Tạp chí Quốc tế EAI Endorsed Transactions on Scalable Information Systems (đã chấp nhận xuất bản)	2026