LỜI CAM ĐOAN

Tôi xin cam đoan tất cả các nội dung trong luận án: "*Nghiên cứu cải tiến một số phương pháp phân tích quan điểm mức khía cạnh dựa trên học máy*" là công trình nghiên cứu của riêng tôi, dưới sự hướng dẫn khoa học của PGS.TS.Nguyễn Việt Anh.

Tất cả các tài liệu tham khảo sử dụng trong luận án đều được nêu rõ nguồn gốc trong danh mục các tài liệu tham khảo.

Tất cả các kết quả, số liệu sử dụng trong luận án là trung thực và chưa được người khác công bố trong bất kỳ công trình khoa học nào.

Hà Nội, ngày 5 tháng 7 năm 2023

Nghiên cứu sinh

Nguyễn Thị Ngọc Tú

LỜI CẢM ƠN

Lời đầu tiên, tôi xin được bày tỏ lòng biết ơn sâu sắc nhất đến thầy PGS.TS Nguyễn Việt Anh, thầy đã luôn tận tình chỉ bảo, hướng dẫn tôi trong suốt quá trình định hướng nghiên cứu, phương pháp nghiên cứu, cho đến cách trình bày các bài báo khoa học, các báo cáo chuyên đề và luận án. Bên cạnh đó thầy còn là một một người bạn, một đồng nghiệp luôn động viên những lúc tôi gặp khó khăn trong chặng đường nghiên cứu của mình. Tôi cũng xin bày tỏ lòng biết ơn sâu sắc đến cô Nguyễn Thị Thu Hà, người cô đã luôn động viên giúp đỡ tôi trong quá trình nghiên cứu, viết các bài báo khoa học trong và ngoài nước.

Tôi xin chân thành cảm ơn Ban lãnh đạo Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học Việt Nam, các thầy cô Khoa Đào tạo Sau đại học của Học viện Khoa học và Công nghệ đã động viên, giúp đỡ và tạo điều kiện thuận lợi cho tôi trong suốt quá trình thực hiện luận án. Tôi cũng xin cảm ơn các thầy/cô Viện Công nghệ thông tin, Viện Hàn lâm Khoa học Việt Nam đã có nhiều đóng góp quý báu giúp tôi hoàn thiện luận án, sự tận tình hướng dẫn, động viên của các thầy/cô đã giúp tôi tự tin hơn trong con đường nghiên cứu khoa học. Tôi xin cảm ơn thầy PGS.TS Nguyễn Long Giang và thầy TS Vũ Văn Hiệu đã có những đóng góp quý báu cho các công bố nghiên cứu của tôi.

Tôi cũng xin gửi lời cảm ơn chân thành tới Ban giám hiệu trường Đại học Điện Lực, các đồng nghiệp/giảng viên tại khoa Công nghệ thông tin, trường Đại học Điện lực đã tạo điều kiện, giúp đỡ, động viên trong suốt quá trình học tập, nghiên cứu và hoàn thiện bảo vệ luận án.

Con xin cảm ơn bố mẹ hai bên gia đình, em xin cảm ơn chồng và hai con trai những người luôn ở bên, ủng hộ, động viên cho con/em có thời gian, điều kiện tốt nhất để nghiên cứu và hoàn thành luận án.

Hà Nội, ngày 5 tháng 7 năm 2023

Nghiên cứu sinh

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

Từ	Viết tắt của	Ý nghĩa
ACD	Aspect Category Detection	Phát hiện danh mục khía cạnh
ACP	Aspect Category Polarity	Phân cực danh mục khía cạnh
AOS	Aspect-based opinion summary	Tổng hợp quan điểm dựa trên khía cạnh
ATE	Aspect Term Extraction	Trích rút thuật ngữ khía cạnh
ATP	Aspect Term Polarity Identifier	Phân cực thuật ngữ khía cạnh
BOW	Bag of words	Túi từ
CNN	Convolutional Neural Network	Mạng nơ ron tích chập
CRF	Conditional Random Field	Trường ngẫu nhiên có điều kiện
DBN	Deep belief network	Mạng niềm tin sâu
DL	Deep learning	Học sâu
DM	Data Mining	Khai phá dữ liệu
DS	Dempster-Shafer	
FOS	Feature-based opinion summary	Tổng hợp quan điểm dựa trên đặc trưng
FS	Feature selection	Lựa chọn đặc trưng
HMM	Hidden Markov Model	Mô hình Markov ẩn
IE	Information Extraction	Trích rút thông tin
IG	Information Gain	Độ lợi thông tin
IR	Information Retrieval	Tra cứu thông tin
LDA	Latent Dirichlet Allocation	Phân bố Dirichlet ẩn
MI	Muatual Information	Thông tin tương hỗ
NB	Naive Bayes	
NER	Named entity recognition	Nhận dạng thực thể tên
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
OGBN	OR Gate Bayesian Network	Mạng Bayesian công OR
OM	Opinion Minning	Khai phá quan điểm
PMI	Pointwise mutual information	Điểm thông tin tương hỗ
POS	Part of Speech	Từ loại
PRM	Probabilistic Regression Model	Mô hình hồi quy xác suất
RNN	Recurrent Neural Network	Mạng nơ ron hồi quy
SVM	Support Vector Machine	Máy vector hỗ trợ
TF-IDF	Term Frequency – Inverse Document Frequency	Tần số từ - Tần số văn bản nghịch đảo
W2V	Word to Vector	Từ thành Vector

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

MỞ ĐẦU

Trong thời đại công nghệ thông tin phát triển hiện nay, lượng người dùng Internet ngày càng tăng. Theo thống kê của We Are Social and Hootsuite, tính đến tháng 1 năm 2022 có 4,95 tỉ người dùng Internet, với tỉ lệ 62,5% dân số trên toàn cầu. Trong đó, số người dùng mạng xã hội là 4,62 tỉ người dùng, bằng 58,4% tổng dân số thế giới. Kết quả khảo sát cũng cho thấy rằng đến hơn 77% người dùng trực tuyến mua hàng mỗi tháng. Như vậy, hầu hết các hoạt động của con người đã xuất hiện phổ biến trên mạng Internet và các phương tiện truyền thông trực tuyến. Đặc biệt, các trang thương mại điện tử ngày nay gia tăng hoạt động tương tác với người dùng thông qua việc khuyến khích họ chia sẻ các bài đánh giá về sản phẩm và thể hiện quan điểm của họ trên các trang web mua sắm (ví dụ Amazon, eBay v.v.) hoặc các trang mạng xã hội (ví dụ facebook.com, Twitter). Khai phá các bài đánh giá này có thể hiểu được quan điểm, tâm lý của người tiêu dùng từ đó giúp ích cho việc xây dựng các chiến lược của doanh nghiệp như: chiến dịch tiếp thị, sản phẩm ưu tiên, giám sát danh tiếng [1], nó cũng có thể được thực hiện để học hành vi của người tiêu dùng, thị trường mẫu, và dự đoán xu hướng tiêu dùng của xã hội [2].

Vì sự quan trọng của khai phá quan điểm mà trong thời gian hơn hai thập kỷ qua, các nhà nghiên cứu, các học giả, các tổ chức, và các doanh nghiệp quan tâm nghiên cứu lĩnh vực này [3–7]. Theo Bing Liu, các nhiệm vụ khai thác quan điểm được chia thành ba cấp độ chính: cấp độ văn bản, cấp độ câu và cấp độ cụm từ (cấp độ khía cạnh) [3]. Ở cấp độ văn bản, nhiệm vụ chính là xem xét toàn bộ văn bản như đầu vào và phân loại xem nó có thể hiện bất kỳ cảm xúc tổng thể nào hay không [8–10]. Cấp độ câu, đầu vào là các câu được tách ra từ văn bản có chứa quan điểm. Đây là một cấp độ phân tích chi tiết của mức văn bản, trong đó xác định tính phân cực cho mỗi câu và mỗi câu có thể chứa quan điểm khác nhau [11–14]. Cả hai việc phân tích quan điểm ở mức độ văn bản và mức độ câu chưa khám phá được rõ ràng điều gì được người dùng thích hay không thích. Ví dụ, trong câu đánh giá sau: "The laptop’s sound is good, but the battery life is very short", phân tích quan điểm ở mức văn bản và mức câu khó xác định được quan điểm thực sự mà người dùng đưa ra là gì. Khi xem xét đến các thuộc tính chất lượng loa (Speaker Quality) và thời lượng pin (Battery Life) của máy tính xách tay (laptop), các quan điểm được thể hiện cụ thể và rõ ràng hơn. Quan điểm trên khía cạnh chất lượng loa là tích cực (good), và quan điểm trên khía cạnh thời lượng pin là tiêu cực (very short). Mức độ phân tích này được gọi là phân tích quan điểm ở mức độ khía cạnh. Hiện nay, phân tích quan điểm dựa trên khía cạnh đang thu hút được nhiều sự quan tâm của cộng đồng nghiên cứu và các nhà phát triển ứng dụng [7]. Trong phân tích quan điểm dựa trên khía cạnh, việc tổng hợp hệ thống của các quan điểm về các thực thể và các thuộc tính của chúng có thể được tạo ra. Nhiệm vụ này

có thể biến văn bản phi cấu trúc thành dữ liệu có cấu trúc, đồng thời có thể sử dụng cho tất cả các loại phân tích định tính và phân tích định lượng. Mặc dù vậy, phân tích quan điểm mức độ văn bản và mức độ câu đều thực sự gặp thách thức lớn, song với mức độ khía cạnh thậm chí còn nhiều khó khăn hơn vì nó bao gồm nhiều vấn đề nhỏ [3–5, 15].

Hai vấn đề chính trong phân tích quan điểm dựa trên khía cạnh là *trích rút khía cạnh* (Aspect extraction) và *phân lớp cảm xúc khía cạnh* (Aspect sentiment classcification). Quá trình xác định chủ thể đối tượng của quan điểm và các từ thể hiện quan điểm trong các câu đưa ra được gọi là trích rút khía cạnh. Việc phân loại các từ quan điểm được trích rút vào một trong số các thang cực được gọi là phân lớp cảm xúc khía cạnh. Đã có nhiều nghiên cứu thực thi riêng rẽ bài toán trích rút khía cạnh [16–26] v.v, hoặc phân lớp cảm xúc khía cạnh [9, 27–34], tuy nhiên cũng có một số nghiên cứu giải quyết đồng thời cả hai bài toán của phân tích quan điểm dựa trên khía cạnh [35–38].

Một số thách thức chính trong phân tích quan điểm mức khía cạnh cần giải quyết:

- ***Đối với bài toán trích rút khía cạnh:***

  Hầu hết dữ liệu thế giới thực gắn với nhiệm vụ này đều không được gán nhãn [4].

  Nhiều câu đánh giá thiếu các thể hiện khía cạnh rõ ràng (danh từ thể hiện khía cạnh) dẫn đến vấn đề trích rút khía cạnh trở nên khó khăn hơn. Ngoài ra có nhiều cách thức ám chỉ các khía cạnh (đặc trưng ẩn) xuất hiện trong một câu khiến nhiệm vụ khai phá càng phức tạp, bởi phải xác định đặc trưng ẩn nào gán với khía cạnh nào. Ví dụ trong câu “Pictures taken can get blurred because of lack of image stabilizer but overall a great option for given budget”, hai khía cạnh khác nhau về chất lượng máy ảnh và giá cả được đề cập ngầm [39].

  Khi một từ xuất hiện trong câu thì cần xem xét nó xuất hiện cùng với những từ nào? Đối với nhiều từ cách giải thích của chúng phụ thuộc vào ngữ cảnh sử dụng chúng. Ví dụ từ “apple” xuất hiện trong hai câu: "Apple is a tasty fruit" và “Apple has just launched a new product” được hiểu theo hai nghĩa khác nhau. Điều này gây nên những khó khăn nhất định cho nhiệm vụ trích rút khía cạnh, đặc biệt các khía cạnh ẩn [40] .

  Một số khía cạnh có tần suất xuất hiện thấp dễ bị bỏ qua. Mặc dù vậy, những khía cạnh này có thể là những khía cạnh quan trọng. Làm thế nào có thể phát hiện được các khía cạnh như vậy cũng là một thách thức của nhiệm vụ trích rút khía cạnh.

- ***Đối với bài toán phân lớp cảm xúc khía cạnh:***

  Nhiệm vụ phân loại cảm xúc đa lớp có nhiều thách thức hơn. Sự hiện diện của

nhiều lớp làm cho một bộ phân loại nhất định khó xác định ranh giới giữa các lớp khác nhau hơn [41]. Hơn nữa trong thực tế, một từ có thể miêu tả nhiều trạng thái cảm xúc khác nhau (trong các ngữ cảnh khác nhau), ngay cả con người cũng khó phân biệt sự khác nhau này.

Khoảng cách giữa các lớp cảm xúc khác nhau nhỏ, giữa các lớp có cùng cực cảm xúc (ví dụ Emotional negative và Rational negative, hoặc Emotional positive và Rational positive) gần như là tương tự nhau và chúng rất dễ bị phân loại nhầm lẫn nhau [41].

Sự phụ thuộc vào ngữ cảnh, một từ có thể có các nghĩa khác nhau dựa trên ngữ cảnh và miền lĩnh vực được sử dụng. Nghĩa của cùng một từ có thể khác nhau đối với từng tình huống. Ví dụ: từ "long time" khi nói về thời lượng pin của điện thoại thì mang nghĩa tích cực, xong trong ngữ cảnh nói về tốc độ xử lý của CPU thì lại mang tính tiêu cực [40].

Sự hiện diện của phủ định, ví dụ các từ "not", "neither", "nor", v.v là rất quan trọng đối với phân tích quan điểm vì chúng có thể đảo ngược cực cảm xúc của một văn bản [7]. Tuy nhiên, không dễ để xử lý công việc này bằng cách đảo cực vì các từ phủ định có thể được tìm thấy trong một câu mà không ảnh hưởng đến cảm xúc thể hiện trong của văn bản. Đặc biệt trong phân loại cảm xúc đa lớp, phủ định không có nghĩa là cảm xúc chuyển đổi sẽ được chuyển thành lớp ngược với điều phủ định. Ví dụ câu "I do not simply love it.", nghĩa của câu này không thể khẳng định mang tính tiêu cực, mà ẩn ý có thể là rất tích cực.

Từ những khảo sát và đánh giá các kết quả nghiên cứu có được, tác giả cho rằng cần có một nghiên cứu đầy đủ trên tất cả các nhiệm vụ của phân tích quan điểm dựa trên khía cạnh để đem lại thông tin hữu ích nền cho các ứng dụng như hệ hỗ trợ ra quyết định, hệ thống phân tích đánh giá mối quan tâm và xu hướng tiêu dùng của thị trường, hệ thống hỗ trợ định hướng chiến lược sản phẩm của doanh nghiệp, v.v. Đồng thời cần tìm ra cách tiếp cận hiệu quả để vượt qua các thách thức trong lĩnh vực nghiên cứu, cải thiện hiệu suất của hệ thống phân tích quan điểm dựa trên khía cạnh.

Mục tiêu của luận án và nội dung nghiên cứu

Mục tiêu của luận án

Mục tiêu của luận án giải quyết ba bài toán sau:

- Thứ nhất, trích rút các khía cạnh (tính năng) của sản phẩm hoặc dịch vụ từ các bài đánh giá sản phẩm trực tuyến. Các bài đánh giá mà luận án tập trung giải quyết ở dạng chỉ đề cập đến một loại thực thể (một sản phẩm hoặc dịch vụ), các khía cạnh của sản phẩm hoặc dịch vụ là xác định trước, xem xét cả hai dạng thể hiện khía cạnh rõ ràng và thể hiện khía cạnh ẩn.

- Thứ hai, phân loại cảm xúc khía cạnh với đầu vào là các phần văn bản chứa các khía cạnh đã được trích rút từ văn bản gốc. Các quan điểm được luận án quan tâm là dạng thông thường (không xem xét dạng so sánh), quan điểm được xem xét ở cả hai dạng khách quan và chủ quan. Mức độ phân loại các quan điểm là đa mức dựa trên đánh giá cảm tính và đánh giá lý tính với 5 mức *tích cực cảm xúc* (5 sao), *tích cực lý trí* (4 sao), *trung tính* (3 sao), *tiêu cực lý trí* (2 sao), *tiêu cực cảm xúc* (1 sao) [3].
- Thứ ba, ước lượng trọng số khía cạnh xác định tầm quan trọng (mức độ quan tâm của người dùng) của từng khía cạnh trong tổng thể đánh giá của người dùng. Thông thường, bài toán này sẽ lấy kết quả đầu ra của bài toán thứ hai làm đầu vào để giải quyết vấn đề. Tuy nhiên, trong nghiên cứu đề xuất sẽ chỉ dựa trên nội dung bài đánh giá để giải quyết bài toán mà không yêu cầu đầu vào từ bài toán thứ hai.

Nội dung nghiên cứu của luận án

Dựa trên mục tiêu đã trình bày luận án tập trung giải quyết các bài toán sau đây:

- Bài toán trích rút khía cạnh, phân tích bài viết của người dùng trực tuyến thành các phân văn bản có chứa các khía cạnh cùng các quan điểm tương ứng dưới dạng trích rút câu. Các câu trước tiên được gán nhãn khía cạnh, sau đó chúng được gom nhóm khía cạnh và trích rút. Mô hình trích rút khía cạnh đề xuất trong luận án là mô hình dựa trên cách tiếp cận học bán giám sát với hai phương pháp cụ thể là phương pháp dựa trên xác suất có điều kiện kết hợp bootstrapping và phương pháp dựa trên Word to Vector (W2V) và độ đo hỗ trợ. Để cải thiện hiệu suất của mô hình trích rút khía cạnh đề xuất, luận án nghiên cứu và phân tích các đặc trưng liên quan đến thể hiện khía cạnh rõ ràng và thể hiện khía cạnh ẩn. Ngoài ra luận án còn nghiên cứu các kỹ thuật lựa chọn đặc trưng phù hợp để tạo ra tập các đặc trưng hữu ích nhất.
- Các khía cạnh và quan điểm được phân loại vào các lớp cảm xúc dựa trên các phương pháp học có giám sát. Cụ thể các phương pháp phân loại Naive Bayes (NB), OR Gate Bayesian Network (OGBN), Support Vector Machine (SVM) và một phương pháp kết hợp các mô hình cơ sở dựa trên lý thuyết Dempster-Shafer (DS) được áp dụng. Thêm vào đó, để cải thiện hiệu suất của các phương pháp phân lớp, luận án cũng tiến hành nghiên cứu và lựa chọn đặc trưng dựa trên các cơ sở lý thuyết thông tin (Information Gain (IG) và Muatual Information (MI)).
- Giải quyết bài toán ước lượng trọng số khía cạnh, luận án tiến hành nghiên cứu nội dung các bài viết, đề xuất một phương pháp ước lượng trọng số khía cạnh chỉ dựa trên nội dung bài viết cùng với sự xuất hiện của các từ liên quan khía

cạnh trong từng bài viết cá nhân người dùng và trong toàn bộ kho ngữ liệu.

Đối tượng nghiên cứu và phạm vi nghiên cứu

Đối tượng nghiên cứu

Với mục tiêu đã đề ra của luận án, đối tượng nghiên cứu của luận án bao gồm:

Các kỹ thuật và phương thức tiền xử lý cho các văn bản; Các kỹ thuật đặc trưng và lựa chọn đặc trưng trong phân tích quan điểm; Các mô hình và phương pháp trích rút khía cạnh, phân loại cảm xúc khía cạnh, ước lượng trọng số khía cạnh trong phân tích quan điểm mức khía cạnh;

Phạm vi nghiên cứu

- Nghiên cứu và phân tích các văn bản chứa quan điểm được sinh ra dựa trên hoạt động chia sẻ quan điểm của người dùng về các sản phẩm hoặc dịch vụ trên nền tảng trực tuyến.
- Nghiên cứu và phân tích các khía cạnh (tính năng) của các loại sản phẩm hoặc dịch vụ được người dùng chia sẻ trên nền tảng trực tuyến.
- Nghiên cứu và phân tích các cảm xúc mà người dùng thể hiện trong các bài viết chia sẻ về sản phẩm hoặc dịch vụ trên nền tảng trực tuyến.
- Hiện nay, dữ liệu mà người dùng chia sẻ về các quan điểm hoặc cảm xúc khi trải nghiệm các sảm phẩm hoặc dịch vụ là rất phong phú và đa dạng như dữ liệu văn bản (text), dữ liệu hình ảnh (image), dữ liệu phim (video), dữ liệu là các ký hiệu (symbol) v.v. Tuy nhiên, trong luận án này chỉ nghiên cứu dữ liệu dạng văn bản, còn các dạng dữ liệu khác không phải là phạm vi nghiên cứu trong luận án này.

Phương pháp nghiên cứu

- *Phương pháp nghiên cứu lý thuyết:* được sử dụng khi tìm hiểu các mô hình cơ sở về phân tích quan điểm, trích rút thông tin, tóm tắt văn bản. Từ đó tìm được các hạn chế, tồn tại trong các nghiên cứu đã có, đặt ra nhiệm vụ cho luận án, và đề xuất hướng thực hiện cho nhiệm vụ mới đặt ra.
- *Phương pháp so sánh:* được sử dụng để tìm ra điểm khác biệt của bài toán phân tích quan điểm so với các nhánh nghiên cứu khác của lĩnh vực khai phá dữ liệu. So sánh các phương pháp tiếp cận khác nhau trong khai phá quan điểm như phương pháp dựa trên từ điển với các phương pháp học máy, các phương pháp học máy không giám sát, bán giám sát, có giám sát trong phân tích quan điểm.
- *Phương pháp thiết kế:* xây dựng và kiểm nghiệm các mô hình đề xuất bằng thực nghiệm và đánh giá.
- *Phương pháp đánh giá bằng thực nghiệm:* thu thập dữ liệu, cài đặt các mô hình

đề xuất, xây dựng các bộ dữ liệu mẫu, thực hiện thử nghiệm trên các bộ dữ liệu mẫu và phân tích, đánh giá kết quả thử nghiệm.

Những đóng góp chính của luận án

Sau những nỗ lực nghiên cứu, luận án có những đóng góp chính như sau:

- Luận án đề xuất một hệ thống tổng thể thực hiện ba bài toán con trích rút khía cạnh, dự đoán điểm đánh giá khía cạnh, ước lượng trọng số khía cạnh của bài toán phân tích quan điểm dựa trên khía cạnh đối với các bài đánh giá sản phẩn trực tuyến. Để giải quyết bài toán trích rút khía cạnh, luận án đề xuất một phương pháp học bán giám sát dựa trên xác suất có điều kiện kết hợp thuật toán bootstrapping. Với bài toán phân lớp cảm xúc khía cạnh, phương pháp học có giám sát Naive Bayes được áp dụng. Cuối cùng, một phương pháp tiếp cận dựa vào nội dung bài viết với sự xuất hiện của các từ khía cạnh liên quan trong từng bài viết và trong toàn bộ kho ngữ liệu được đề xuất cho nhiệm vụ ước lượng trọng số khía cạnh. Hệ thống này được trình bày trong công bố [CT3] và [CT5].
- Luận án đã đề xuất một phương pháp học bán giám sát dựa trên Word to Vector kết hợp mô hình ngôn ngữ để trích rút khía cạnh. Phương pháp trích rút này đã phát huy ưu điểm của biểu diễn đặc trưng từ trong ngữ cảnh để cải thiện hiệu quả nhiệm vụ trích rút khía cạnh. Phương pháp đề xuất này được thể hiện trong [Ct2], và [CT4].
- Luận án đề xuất một phương pháp học có giám sát dựa trên sự kết hợp các thuật toán học giám sát cơ sở SVM và OGBN cùng cơ sở lý thuyết Dempster-Shafer để nâng cao hiệu quả phân loại cảm xúc khía cạnh. Phương pháp đề xuất đã kết hợp được các ưu điểm của hai phương pháp phân loại cơ sở và đem đến một cải thiện đáng kể về độ chính xác phân loại cảm xúc. Phương pháp này được công bố trong [CT5].

Bố cục của luận án

Luận án gồm phần mở đầu, 04 chương nội dung và phần kết luận:

Phần mở đầu: Trình bày về tính cấp thiết của đề tài và động lực nghiên cứu; mục tiêu, đối tượng, phạm vi nghiên cứu; phương pháp nghiên cứu; các đóng góp chính của luận án; bố cục luận án.

Chương 1: Tổng quan về phân tích quan điểm và phân tích quan điểm mức khía cạnh.. Chương này của luận án trình bày một số kiến thức nền tảng liên quan đến đề tài luận án như mô hình quan điểm (thực thể, khía cạnh, quan điểm, người sở hữu quan điểm, thời gian xuất hiện quan điểm), đối tượng nghiên cứu của phân tích quan điểm, các bài toán chính của phân tích quan điểm, các loại quan điểm, các mức độ phân tích quan điểm, các đặc trưng trong xử lý ngôn ngữ tự nhiên và phân tích

quan điểm, phân tích quan điểm mức khía cạnh. Ngoài ra, các nghiên cứu liên quan đến hai bài toán con chính (trích rút khía cạnh, phân loại cảm xúc khía cạnh) trong phân tích quan điểm mức khía cạnh cũng được trình bày, so sánh, đánh giá chi tiết và toàn diện trong chương này.

Chương 2: Khai phá quan điểm mức khía cạnh trên các bài đánh giá sản phẩm trực tuyến. Trong chương này, một hệ thống giải quyết tổng thể ba bài toán con của bài toán phân tích quan điểm mức khía cạnh trên các bài đánh giá sản phẩm trực tuyến được trình bày. Bài toán trích rút khía cạnh được thực hiện với phương pháp học bán giám sát dựa trên xác suất có điều kiện kết hợp giải thuật bootstrapping. Bài toán phân loại cảm xúc khía cạnh được thực hiện bởi thuật toán học có giám sát Naive Bayes. Bài toán ước lượng trọng số khía cạnh được giải quyết nhờ phương pháp tiếp cận không giám sát dựa trên nội dung bài đăng và sự xuất hiện của các từ khía cạnh liên quan. Cũng trong chương 2, các bài toán của hệ thống được thử nghiệm và đánh giá trên ba bộ dữ liệu đã được công nhận bởi cộng đồng nghiên cứu quốc tế. Đồng thời các kết quả thử nghiệm cũng được so sánh và đánh giá với các phương pháp hiện đại khác.

Chương 3: Trích rút khía cạnh dựa trên Wordtovec kết hợp mô hình ngôn ngữ. Từ mô hình biểu diễn từ dạng vector, nghiên cứu sinh đề xuất một phương pháp học bán giám sát kết hợp vector từ và độ đo hỗ trợ để tính độ hỗ trợ của từ, câu đối với từng khía cạnh, từ đó thực hiện trích rút khía cạnh. Các kết quả thử nghiệm và đánh giá phương pháp đề xuất của luận án cũng được trình bày.

Chương 4: Phân lớp cảm xúc bằng cách kết hợp các bộ phân loại cơ sở. Trong chương này nghiên cứu sinh đã trình bày phương pháp phân loại cảm xúc đa lớp dựa trên Support Vector Machine và OR Gate Bayesian Network với các kỹ thuật trích chọn đặc trưng thông qua chỉ số độ lợi thông tin và thông tin tương hỗ. Phần tiếp theo là một đề xuất mô hình kết hợp các bộ phân loại cơ sở dựa trên lý thuyết Dempster-Shafer để tạo ra một bộ phân loại tổng hợp mạnh mẽ cho nhiệm vụ phân loại đa lớp. Các kết quả thực nghiệm và các phân tích đối sánh được trình bày.

Phần kết luận và hướng phát triển: Trình bày một số kết luận về ý nghĩa của những kết quả đã đạt được của luận án và một số hướng nghiên cứu tiếp theo.

Hà Nội, ngày 5 tháng 7 năm 2023

Nghiên cứu sinh

CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM VÀ PHÂN TÍCH QUAN ĐIỂM MỨC KHÍA CẠNH

1.1 Tổng quan về phân tích quan điểm

Ngày nay, truyền thông trực tuyến và truyền thông xã hội đang nhanh chóng thay thế phương tiện ngoại tuyến. Việc sử dụng Internet và các hoạt động trực tuyến (như trò chuyện, hội nghị, đặt vé, giao dịch trực tuyến, thương mại điện tử, truyền thông xã hội, viết blog và vi blog, nhấp chuột, v.v) ngày càng tăng. Phương tiện trực tuyến cung cấp các biện pháp tốt hơn để trả lời và phản hồi nhanh chóng về các vấn đề toàn cầu khác nhau trong dạng bài viết văn bản đăng tải, tin tức, ảnh, và video. Nhiều diễn đàn, blog, mạng xã hội, các website thương mại điện tử, các trang tin tức tài chính và các tài nguyên web khác đóng vai trò là các nền tảng để bày tỏ, chia sẻ rộng rãi quan điểm của người dùng. Do đó, chúng có thể được sử dụng để hiểu các quan điểm của công chúng và người tiêu dùng đối với các sự kiện xã hội, chính trị, chiến lược của các doanh nghiệp, chiến dịch tiếp thị, sản phẩm ưu tiên, giám sát [42, 43], ngoài ra, các nguồn tài nguyên này cũng có thể được sử dụng để học hành vi của người tiêu dùng, thị trường mẫu, và dự đoán xu hướng của xã hội [44, 45].

Để tạo ra các ứng dụng thực tiễn hiệu quả, cộng đồng nghiên cứu và các nhà phát triển ứng dụng đang làm việc nghiêm túc trong lĩnh vực phân tích quan điểm suốt hai mươi năm qua. Phân tích quan điểm là một nghiên cứu đo lường về các quan điểm, tình cảm, cảm xúc, và thái độ đã thể hiện trong các văn bản đối với một thực thể [4]. Phân tích quan điểm là nhiệm vụ phát hiện, trích rút và phân loại các quan điểm, tình cảm, thái độ liên quan đến các chủ đề khác nhau được thể hiện trong văn bản đầu vào [3]. Phân tích quan điểm giúp các nhà quản lý, các chuyên gia hoạch định chiến lược doanh nghiệp thấy được các kết quả khác nhau như quan sát tâm trạng cộng đồng về sự kiện chính trị, trí tuệ thị trường [45], đo lường sự hài lòng của khách hàng, dự đoán doanh thu phim [43] và nhiều hơn nữa. Phân tích quan điểm cũng giúp người tiêu dùng trở nên thông minh hơn trong các quyết định tiêu dùng của họ.

Phân tích quan điểm là lĩnh vực nghiên cứu sử dụng các kỹ thuật khác nhau trong các lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), tra cứu thông tin (Information Retrieval - IR), trích rút thông tin (Information Extraction - IE), khai phá dữ liệu (Data Mining - DM) có cấu trúc và không có cấu trúc. Phần lớn dữ liệu có sẵn trong thế giới thực là không có cấu trúc (như văn bản, tiếng nói, âm thanh, video, v.v..). Điều này đặt ra những thách thức nghiên cứu quan trọng. Để giải quyết với dữ liệu văn bản không cấu trúc như vậy, nhiều nỗ lực nghiên cứu đã được đề xuất trong những năm gần đây, và các nghiên cứu phân tích quan điểm tự động là một sự mở rộng nghiên cứu trong lĩnh vực NLP [3]. Phân tích quan điểm không phải

là vấn đề đơn lẻ, thay vào đó nó là một vấn đề đa diện. Nhiều vấn đề khác nhau cần được thực hiện để khai thác quan điểm từ văn bản đưa ra. Các công việc thu thập dữ liệu, tiền xử lý dữ liệu, biểu diễn đặc trưng, trích rút và lựa chọn đặc trưng là các tác vụ phổ biến nhất đòi hỏi phải có trong khai phá quan điểm [3].

1.1.1 Các khái niệm cơ bản

Thuật ngữ khai phá quan điểm (Opinion Minning-OM) xuất hiện khoảng từ những năm đầu của thế kỉ 21. Đến năm 2005, với nghiên cứu của Bing Liu [46] thì thuật ngữ phân tích quan điểm mới được đưa ra rõ ràng hơn. Theo tác giả, các quan điểm có thể đề cập về bất cứ chủ đề nào đó, ví dụ như một sản phẩm, một tổ chức, một cá nhân, một chủ đề chính trị hoặc xã hội... Tác giả coi các đối tượng được nhận xét là các *thực thể* (entity). Thực thể này là một tập hợp các *thành phần* (component). Và như thế, các đối tượng có thể được phân ra theo các thành phần của mối quan hệ, tức là mỗi thành phần cũng có thể có các thành phần con của nó.

Posted by: John Smith Date: September 10, 2011

"(1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) However, my wife thinks it is too heavy for her."

Dịch:

Đăng bởi: John Smith Ngày: 10-09-2011

(1) Tôi đã mua một cái máy ảnh Canon G12 sáu tháng trước đây. (2) Tôi rất thích nó. (3) Chất lượng ảnh thật là tuyệt vời. (4) Thời lượng của pin cũng lâu. (5) Tuy nhiên vợ tôi nghĩ nó khá nặng so với cô ấy.

Hình 1.1: Ví dụ bài đánh giá sản phẩm máy ảnh kỹ thuật số

Ví dụ 1.1: [3](Hình 1.1)
Từ bài nhận xét này chúng ta thấy một số điểm như sau:

- Bài nhận xét có 5 câu, trong đó câu (1) đề cập đến đối tượng được miêu tả là máy ảnh Canon G12. Câu (2) thể hiện một cảm xúc tổng thể về máy ảnh Canon G12 là tích cực. Câu (3) thể hiện một cảm xúc tích cực về chất lượng ảnh của chiếc máy ảnh này. Câu (4) thể hiện một cảm xúc tích cực về thời lượng của pin. Và cuối cùng câu (5) là một cảm xúc tiêu cực về trọng lượng của máy ảnh.
- Bài nhận xét này có quan điểm từ 2 người, điều này được gọi là nguồn quan điểm (opinion sources) hoặc chủ sở hữu quan điểm (opinion holders). Người

sở hữu quan điểm trong các câu 2,3,4 là ông John Smith và người sở hữu quan điểm trong câu 5 lại là vợ của ông John Smith.

- Thời gian đăng của bài nhận xét là ngày 10 tháng 9 năm 2011. Thời gian đăng bài là quan trọng đối với người đọc bởi vì họ luôn muốn biết các quan điểm đó thay đổi như thế nào trên dòng thời gian và khuynh hướng của các quan điểm này.

Kết luận từ quan sát:

- Một quan điểm bao gồm 2 thành phần chính: một *mục tiêu g* và một *cảm xúc s* trên mục tiêu: (g,s). Trong đó g có thể là thực thể hoặc thành phần của thực thể (thành phần này chính là các thuộc tính của thực thể) trong quan điểm đã được thể hiện. Cảm xúc s là trạng thái tình cảm mang tính tích cực, tiêu cực hoặc trung lập. Đôi khi cảm xúc này còn được thể hiện bằng một định lượng khác là điểm số hoặc sao (thang điểm 1-10 hoặc 1-5 sao). Các cảm xúc này được gọi là khuynh hướng hoặc phân cực cảm xúc. Ví dụ, trong câu (2), mục tiêu của quan điểm là máy ảnh Canon G12 và trong câu (3), mục tiêu của quan điểm là chất lượng ảnh của máy ảnh Canon G12.
- Mỗi quan điểm đều có *chủ sở hữu* là h.
- Mỗi quan điểm đều có *thời gian thể hiện t* là xác định, rõ ràng.

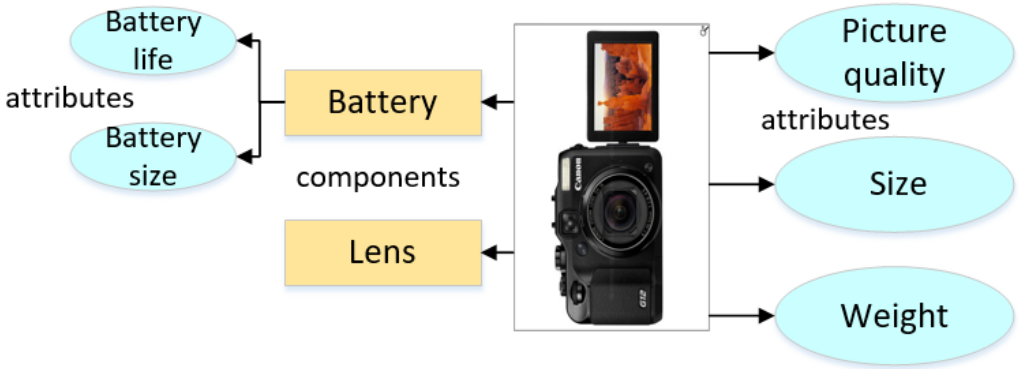
Định nghĩa 1.1 Thực thể (Entity) [47]: Thực thể e là một sản phẩm, dịch vụ, chủ đề, vấn đề, con người, tổ chức, hoặc sự kiện. Nó được mô tả với một cặp, e: (T,W), trong đó, T là một cấu trúc phân cấp của các thành phần, W là tập các thuộc tính của e. Mỗi thành phần có thể có các thành phần con và thuộc tính của nó.

Loại thực thể và biểu diễn thực thể [3]: *Loại thực thể* (Entity category) đại diện cho một thực thể duy nhất, *biểu diễn của thực thể* (Entity expression) là một từ hoặc cụm từ cụ thể mà chúng xuất hiện trong văn bản để biểu thị cho một loại thực thể.

Để nghiên cứu hiệu quả văn bản ở mức độ chi tiết tùy ý như trong *Định nghĩa 1.1* là rất khó. Hơn nữa, đối với người dùng thì cách mô tả biểu diễn phân cấp trên là khá phức tạp và khó sử dụng. Do vậy, chúng ta nên đơn giản hóa phân cấp cây thành 2 cấp độ và sử dụng thuật ngữ *khía cạnh* (Aspect) để biểu diễn cả hai là thành phần con và thuộc tính. Cây được đơn giản hóa, nút gốc là thực thể và nút lá là khía cạnh khác nhau của thực thể (Hình 1.2).

Định nghĩa 1.2 Khía cạnh (Aspect) [47]: Khía cạnh a là một thành phần con hoặc một thuộc tính của thực thể e. Ví dụ "picture quality", "battery life", "weight" là các khía cạnh của thực thể "Canon G12 camera".

Tên khía cạnh và biểu diễn khía cạnh: *Tên khía cạnh* là tên của một khía cạnh được cung cấp bởi người dùng, trong khi *biểu diễn khía cạnh* là một từ hoặc cụm từ thực tế đã xuất hiện trong văn bản mà nó chỉ ra một khía cạnh [3].



Hình 1.2: Ví dụ thực thể điện thoại iPhone gồm các thành phần và thuộc tính của nó

Biểu diễn khía cạnh rõ ràng: Các thể hiện khía cạnh trong một câu là các danh từ, cụm danh từ được gọi là biểu diễn khía cạnh rõ ràng (Explicit aspect expressions). Ví dụ, "picture quality" trong "The picture quality of this camera is great" là một biểu diễn khía cạnh rõ ràng.

Biểu diễn khía cạnh ẩn: Các dạng khác của biểu diễn khía cạnh là biểu diễn khía cạnh ẩn (Implicit aspect expressions). Ví dụ, "heavy" trong "However, my wife thinks it is too heavy for her" là một biểu diễn khía cạnh ẩn.

Định nghĩa 1.3 Cảm xúc (Sentiment) [47]: Cảm xúc s về một mục tiêu (thực thể hoặc khía cạnh) đơn giản là một quan điểm *tích cực* (positive), hoặc *tiêu cực* (negative) hoặc *trung lập* (neutral), đôi khi nó còn được thể hiện qua *điểm đánh giá* (rating).

Ví dụ 1.2: Trong ví dụ 1.1, cảm xúc đối với khía cạnh "picture quality" là tích cực thể hiện qua "amazing", nhưng cảm xúc trên khía cạnh "wieght" (khía cạnh có thể hiện ẩn) là tiêu cực thể hiện qua "too heavy".

Định nghĩa 1.4 Người sở hữu quan điểm (Opinion holder) [47]: *Người sở hữu quan điểm h* là người hay tổ chức cụ thể trực tiếp đưa ra các quan điểm về một thực thể hay một khía cạnh của thực thể.

Định nghĩa 1.5 Thời gian thể hiện quan điểm (Time) [47]: Thời gian t là thời điểm mà quan điểm về một thực thể hay một khía cạnh của thực thể xuất hiện.

Định nghĩa 1.6 Quan điểm (Opinion) [47]: Quan điểm là một bộ gồm 5 thành phần $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$. Trong đó e_i là tên của thực thể, a_{ij} là một khía cạnh j của thực thể e_i, còn s_{ijkl} là quan điểm (sự thể hiện thái độ, tình cảm, cảm xúc) trên khía cạnh a_{ij} của thực thể e_i được phát biểu bởi h_k tại thời điểm t_l, h_k là chủ sở hữu quan điểm, và t_l là thời gian khi quan điểm được thể hiện bởi h_k.

Ví dụ 1.3: Quan sát ví dụ 1.1 và *định nghĩa 1.6* ta có thể xác định các quan điểm cụ thể như sau:

(Canon G12, general, positive, Jonh Smith, September 10, 2011)

(Canon G12, picture-quality, positive, Jonh Smith, September 10, 2011)

(Canon G12, bettery-life, positive, Jonh Smith, September 10, 2011)

(Canon G12, weight, negative, Jonh Smith's wife, September 10, 2011)

Định nghĩa 1.6 đã đưa ra một cách nhìn khái quát và đầy đủ về quan điểm và các thành phần của quan điểm.

1.1.2 Các nhiệm vụ trong phân tích quan điểm

Đối tượng nghiên cứu của phân tích quan điểm

Đưa ra một văn bản chứa quan điểm d, khám phá tất cả năm thành phần quan điểm $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ trong d, nhiệm vụ chính được bắt nguồn từ bộ năm thành phần của phân tích quan điểm. Từ những thảo luận trong phần (1.1.1) *mô hình thực thể* (model of entity) và *mô hình văn bản quan điểm* (model of opinion document) [47] được xác định.

Định nghĩa 1.7 Mô hình thực thể (Model of entity): Một thực thể e_i được đại diện bởi chính nó như là một tổng thể và một tập hữu hạn các khía cạnh $A_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$. e_i có thể được biểu diễn bởi bất kỳ một tập hữu hạn của các biểu diễn thực thể của nó $\{ee_{i1}, ee_{i2}, ..., ee_{is}\}$. Mỗi khía cạnh $a_{ij} \in A_i$ của thực thể e_i có thể được biểu diễn với bất kỳ một tập hữu hạn của các biểu diễn khía cạnh $\{ae_{ij1}, ae_{ij2}, ..., ae_{ijm}\}$.

Định nghĩa 1.8 Mô hình văn bản quan điểm (Model of opinion document): Một văn bản quan điểm d chứa các quan điểm trên một tập các thực thể $\{e_1, e_2, ..., e_r\}$ và một tập con các khía cạnh của chúng từ một tập các chủ sở hữu quan điểm $\{h_1, h_2, ..., h_p\}$ ở một số điểm thời gian cụ thể.

Các bài toán chính của phân tích quan điểm

Kết hợp *Định nghĩa 1.6* với mô hình thực thể và mô hình văn bản quan điểm, các bài toán chính trong phân tích quan điểm từ một văn bản quan điểm d được tóm tắt như sau:

Bài toán 1 (trích rút và phân loại thực thể): Trích rút tất cả các biểu diễn thực thể có trong d, thực hiện gom nhóm các biểu diễn tương đồng thành các cụm, mỗi cụm là một thực thể riêng lẻ e_i.

Bài toán 2 (trích rút và phân loại khía cạnh): Trích rút tất cả các thể hiện khía cạnh của các thực thể, và phân loại các thể hiện khía cạnh vào các cụm. Mỗi một cụm thể hiện khía cạnh đơn nhất a_{ij} của e_i.

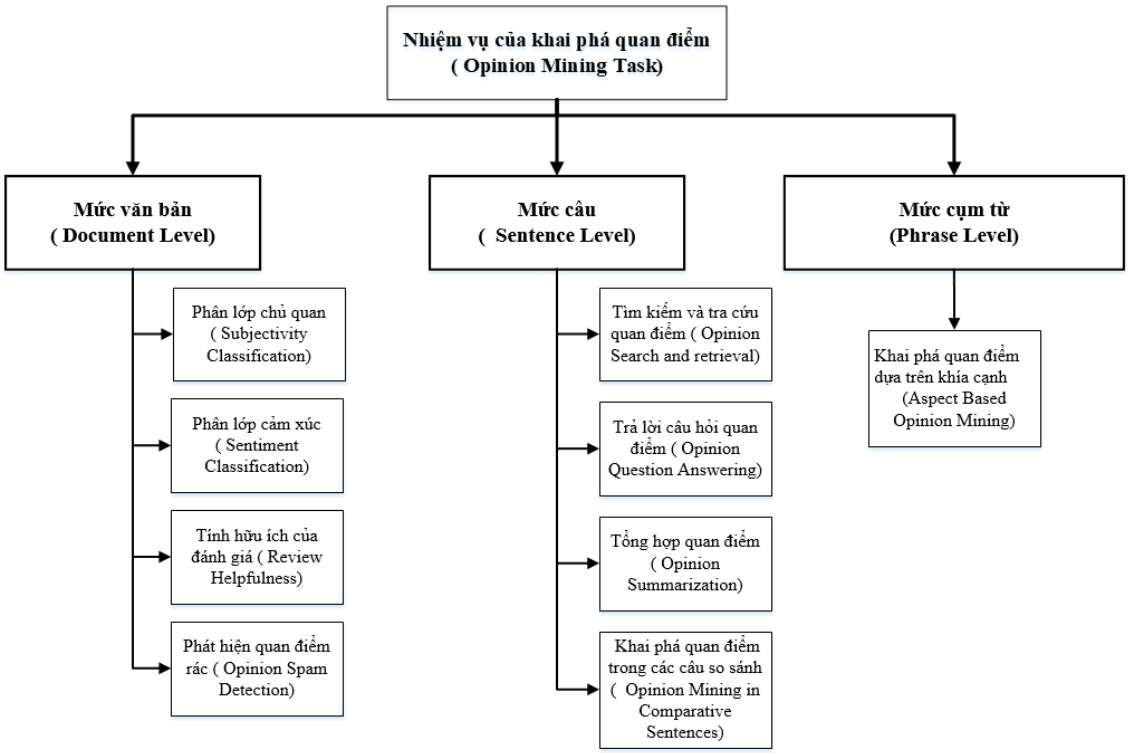
Bài toán 3 (trích rút và phân loại chủ sở hữu quan điểm): Trích rút các chủ sở hữu quan điểm của các quan điểm thể hiện trong văn bản hoặc dữ liệu có cấu trúc và sau đó phân loại chúng.

Bài toán 4 (trích rút và chuẩn hóa thời gian): Trích rút thời gian khi các quan điểm được đưa ra và chuẩn hóa chúng.

Bài toán 5 (phân lớp cảm xúc quan điểm): Xác định một quan điểm trên một khía cạnh a_{ij} là tích cực, tiêu cực hoặc trung lập, hoặc gán nhãn điểm đánh giá ngữ nghĩa đối với khía cạnh.

Bài toán 6 (tổng hợp và sinh bộ năm của quan điểm): Tạo ra tất cả bộ năm của quan điểm $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ thể hiện trong văn bản d dựa trên kết quả của các nhiệm vụ nêu trên.

1.1.3 Các mức độ phân tích quan điểm



Hình 1.3: Phân loại nhiệm vụ khai phá quan điểm theo các mức độ khác nhau

Các quan điểm thường được tìm thấy từ các văn bản người dùng. Văn bản có thể là một từ, cụm từ, câu hoặc một bài viết. Xem xét một cách chi tiết của các nghiên cứu hiện tại, và cách thức tiếp cận trong phân tích quan điểm ở góc nhìn của xử lý ngôn ngữ tự nhiên, ở mức tổng quát hoá, phân tích quan điểm được tập trung chủ yếu ở 03 cấp (mức độ-level) chính: mức văn bản (document-level), mức câu (sentence-level), và mức cụm từ (phrase-level) [3]. Mỗi mức có có các nhiệm vụ con của riêng nó [48].

Các nhiệm vụ ở mỗi mức có thể cũng được áp dụng cho các mức khác [48] (xem Hình 1.3).

Mức độ văn bản: là một hình thức phân loại đơn giản. Trong đó, toàn bộ tài liệu của văn bản đã cho được coi như một đơn vị thông tin cơ bản. Nhiệm vụ chính là xem xét toàn bộ văn bản đầu vào và phân loại xem nó có biểu diễn bất kỳ một cảm xúc tổng thể nào hay không. Cảm xúc trong một tài liệu hoặc đoạn văn hoàn chỉnh có được bằng cách giả định rằng tài liệu đang xem xét có quan điểm về một đối tượng duy nhất. Vì vậy, sẽ không phù hợp nếu tài liệu có quan điểm về các đối tượng khác nhau. Một số nghiên cứu [8–10], đã cố gắng để thực hiện các nhiệm vụ khác nhau trong phân tích quan điểm mức văn bản.

Mức độ câu: là một phân tích chi tiết của mức văn bản, trong đó xác định tính phân cực cho mỗi câu và mỗi câu có thể có quan điểm khác nhau. Ở cấp độ câu có hai bước chính liên quan: một là phân loại một câu là câu chủ quan hay câu khách quan; hai là phân loại phân cực của câu chủ quan. Các nghiên cứu [11–13, 49], đã giải quyết các vấn đề của mỗi nhiệm vụ con của mức này.

Mức độ cụm từ: ở mức độ này, phân lớp được thực hiện theo cách xử lý trơn mịn hơn. Ở đây, các thuộc tính hoặc các khía cạnh của các thực thể được quan tâm chủ yếu và phân cực được tính toán cho từng khía cạnh riêng lẻ.

Cả hai việc phân tích ở mức độ văn bản và mức độ câu không khám phá được chính xác cái gì làm cho người ta thích và không thích. Ví dụ câu "although the service is not that great, I still love this restaurant.". Rõ ràng, câu có một tinh thần tích cực cho toàn bộ nhà hàng, nhưng không thể nói rằng câu này là hoàn toàn tích cực. *Khai phá quan điểm mức khía cạnh* tới dưới nhiệm vụ khai phá quan điểm mức cụm từ. Trong nhiều ứng dụng, mục tiêu quan điểm được miêu tả bằng thực thể hoặc khía cạnh khác nhau của chúng. Phân tích quan điểm dựa trên mức độ này, việc tổng hợp hệ thống của các quan điểm về các thực thể và các khía cạnh của chúng có thể được tạo ra. Điều này biến văn bản phi cấu trúc thành dữ liệu có cấu trúc, và có thể sử dụng cho tất cả các loại phân tích định tính và phân tích định lượng. Một số phương pháp cho khai phá quan điểm dựa trên khía cạnh [16, 19, 36, 38] đã được đề xuất để trích rút các thuộc tính hoặc các khía cạnh từ các nhận xét của người dùng.

1.1.4 Vấn đề đặc trưng trong phân tích quan điểm

Các loại đặc trưng trong phân tích quan điểm

- ***Các đặc trưng dựa trên sự xuất hiện và dựa trên tần số:***

Cách phổ biến nhất để mô tả một đoạn văn bản là biểu diễn dưới dạng vector nhị phân, trong đó mỗi phần tử tương ứng với một thuật ngữ trong từ điển (dạng one-hot-

vector). Phần tử ở chỉ mục thứ i^{th} trong vectơ biểu diễn văn bản được đặt giá trị 1 nếu thuật ngữ đó xuất trong văn bản, ngược lại khi không xuất hiện thì chúng có giá trị là 0. Tương tự như vậy, ứng với mỗi phần tử thứ i^{th} trong vectơ biểu diễn là giá trị nguyên mang thông tin số lần xuất hiện của chúng trong văn bản.

- ***Các đặc trưng Unigram và N-Gram:***

Một Unigram đề cập đến một từ đơn trong văn bản và n-gram đại diện cho một chuỗi (có thứ tự) các từ liền kề nhau trong một câu. Mặc dù, đặc trưng n-gram có nhiều thông tin hơn đặc trưng unigram do chúng là một nhóm từ có tính đến vị trí trong câu, nhưng chúng có hiệu quả hơn trong việc tăng hiệu suất của mô hình phân tích hay không vẫn là một vấn đề tranh luận [50].

- ***Đặc trưng gán nhãn từ loại (Part of Speech – POS):***

Trong xử lý ngôn ngữ tự nhiên và đặc biệt trong lĩnh vực phân tích quan điểm, biểu diễn từ loại là một dạng đặc trưng được sử dụng khá phổ biến. Một số loại từ có nhiều khả năng mang thông tin về cực của một câu hoặc một tài liệu, do đó, POS có thể là một công cụ phân biệt tốt để phát hiện những từ đó. Nhiều báo cáo đã chỉ ra rằng tính từ rất quan trọng trong việc xác định phân cực cảm xúc của một tài liệu [3, 44]. Một số khác lại chỉ ra danh từ rất có ý nghĩa trong việc xác định chủ đề hoặc đối tượng mục tiêu của một tài liệu quan điểm [11].

- ***Đặc trưng cú pháp (Syntax):***

Một số nhà nghiên cứu đã thực hiện trên đặc trưng cú pháp bằng cách sử dụng cây cú pháp [47]. Có những kết quả tốt trong công bố [51], tuy nhiên, theo Ng và cộng sự [52] việc bổ sung các đặc trưng cú pháp phụ thuộc không mang lại bất kỳ cải tiến nào so với phân loại dựa trên n-gram.

- ***Đặc trưng phủ định (Negations):***

Việc sử dụng các từ phủ định trong câu có thể làm đảo lộn hoàn toàn thái cực cảm xúc của câu. Bỏ qua “not” trong câu “He does not like the color blue.” dẫn đến kết quả tích cực là giả. Việc đảo cực của một phủ định dựa trên giả định ngây thơ rằng mỗi từ phủ định đảo ngược cực của các từ theo sau đó (trong một cửa sổ) đúng trong một số trường hợp. Tuy nhiên, đó không phải là một quy tắc chung. Bên cạnh những từ mang nghĩa phủ định rõ ràng, có những thuật ngữ khác cũng mang tính phủ định (ví dụ một số động từ). Tuy vậy, các từ này còn phụ thuộc ngữ cảnh của câu hoặc đoạn văn. Do đó, dạng này rất khó xác định trong phân tích phân cực của quan điểm.

- ***Các đặc trưng theo hướng chủ đề:***

Cảm xúc của một câu đưa ra có thể theo một chủ đề cụ thể. Từ “fast” trong ngữ cảnh các bài đánh giá về xe hơi được coi là tích cực trong khi nó có thể bị coi là tiêu cực trong các bài đánh giá về phim.

Các phương pháp lựa chọn đặc trưng trong phân tích quan điểm

Đặc trưng mô tả đặc tính của dữ liệu. Một đặc trưng có thể không liên quan, có liên quan và dư thừa. Để loại bỏ các đặc trưng không liên quan hoặc dư thừa, các phương pháp *lựa chọn đặc trưng* (Feature Selection - FS) khác nhau được sử dụng. FS là một quá trình xác định và loại bỏ các đặc trưng để giảm số chiều trong không gian đặc trưng giúp cải thiện độ chính xác trong các nhiệm vụ phân tích quan điểm [50]

Các phương pháp FS có thể kể đến các *phương pháp dựa trên từ vựng* (lexicon-based methods) và *phương pháp thống kê* (statistical methods) [53]. Trong các phương pháp dựa trên từ vựng, các đặc trưng được tạo ra bởi con người. Ưu điểm của cách tiếp cận này là tính hiệu quả vì đặc trưng được xử lý cẩn thận. Tuy nhiên, việc lựa chọn các đặc trưng thủ công là một quá trình lâu dài và khó khăn. SentiWordNet8 [54] là điển hình của phương pháp này. Các phương pháp thống kê hoàn toàn tự động và được sử dụng nhiều nhất để lựa chọn đặc trưng, nhưng chúng thường không tách biệt được các đặc trưng mang tính cảm xúc khỏi các đặc trưng không mang tính cảm xúc [55].

- ***Cách tiếp cận bộ lọc***

Đây là phương pháp lựa chọn đặc trưng phổ biến [56]. Các phương pháp này lựa chọn các đặc trưng dựa trên các đặc điểm chung của dữ liệu huấn luyện mà không sử dụng bất kỳ thuật toán học máy nào. Các đặc trưng được xếp hạng dựa trên một số biện pháp thống kê và sau đó các đặc trưng xếp hạng cao nhất được chọn. Các phương pháp lọc ít tốn kém hơn về mặt tính toán và phù hợp với các tập dữ liệu có số lượng lớn các đặc trưng [55]. Một số phương pháp lựa chọn đặc trưng điển hình trong cách tiếp cận này là: độ lợi thông tin, chỉ số Chi-square (CHI), tần suất tài liệu, thông tin tương hỗ [7].

- ***Cách tiếp cận bao bọc***

Cách tiếp cận này phụ thuộc vào các thuật toán học máy vì nó đánh giá một tập hợp con các đặc trưng dựa trên kết quả hiệu suất của thuật toán học máy được sử dụng. Sự phụ thuộc này làm cho các phương pháp trình bao bọc thường lặp đi lặp lại và chuyên sâu về tính toán, nhưng chúng có thể xác định các đặc trưng hoạt động tốt nhất ứng với mô hình cụ thể đó [57]. Phương pháp trình bao bọc là sự kết hợp của các thuật toán học (ví dụ NB, SVM, ...) và các chiến lược tạo tập con các đặc trưng (ví dụ lựa chọn tiến hoặc lùi) [7].

- ***Cách tiếp cận nhúng***

Cách tiếp cận này kết hợp quá trình lựa chọn đặc trưng trong quá trình thực thi thuật toán của mô hình. Chúng sử dụng các thuật toán phân loại có khả năng lựa chọn các đối tượng đặc trưng được tập hợp sẵn của riêng nó [58]. Do đó, nó hiệu quả về mặt tính toán so với cách tiếp cận trình bao bọc. Tuy nhiên, cách tiếp cận này là trường

hợp dành riêng cho thuật toán học ứng dụng [55, 57]. Cách tiếp cận nhúng phổ biến là dựa trên các thuật toán cây quyết định khác nhau (ví dụ CART, C4.5, và ID3) cùng với các thuật toán khác [7]

- ***Cách tiếp cận lai***

Cách tiếp cận này là sự kết hợp của các cách tiếp cận bộ lọc và trình bao bọc, nhưng nói chung là các phương pháp kết hợp các cách tiếp cận khác nhau để có được tập đặc trưng tốt nhất có thể [59].

1.2 Phân tích quan điểm mức khía cạnh

Trong các ứng dụng thực tế, phân tích quan điểm ở mức độ văn bản hoặc mức độ câu thường không đủ cho việc phát triển các nghiên cứu hay hệ thống ứng dụng [3]. Bởi vì, người dùng không chỉ bày tỏ quan điểm của mình trên các văn bản và câu văn, mà còn ở các thuộc tính/khía cạnh và thực thể. Thông tin thu được ở mức độ tài liệu hoặc mức độ câu không đủ để người dùng đưa ra một quyết định đúng đắn. Do vậy, việc xem xét sâu hơn vào các khía cạnh và thực thể đã đưa ra một hướng nghiên cứu mới được gọi là *khai phá quan điểm mức khía cạnh* [11]. Ví dụ, một người cần mua một chiếc điện thoại di động với chất lượng camera tuyệt vời sẽ chỉ quan tâm tìm kiếm các bài đánh giá về khía cạnh "chất lượng ảnh" hơn là các nhận xét tổng quát về điện thoại di động.

Trong thực tế, mọi người có thể bày tỏ các quan điểm khác nhau về nhiều khía cạnh đồng thời trong cùng một bài đánh giá và thậm chí trong cùng một câu đánh giá về sản phẩm đó. Ví dụ: "This book had a good storyline, but the paper quality is bad", người nhận xét đưa ra một quan điểm tích cực đề cập trên khía cạnh cốt truyện và một quan điểm tiêu cực đề cập trên khía cạnh chất lượng giấy của quyển sách. Việc phân chia câu đa khía cạnh thành nhiều câu hoặc cụm từ có các khía cạnh đơn lẻ là thách thức trong khai phá quan điểm mức khía cạnh [3, 4, 7].

1.2.1 Quy trình phân tích quan điểm mức khía cạnh

Mục tiêu chính của phân tích quan điểm mức khía cạnh là khám phá ra tất cả các cảm xúc (xác định thái độ của người nói hoặc người viết) tồn tại trong các tài liệu về các khía cạnh khác nhau của một thực thể [3, 7]. Mô hình quy trình phân tích quan điểm mức khía cạnh được mô tả như trong Hình 1.4 [6], trong đó mỗi phần có nhiệm vụ như sau:

Thu thập dữ liệu (Data collection): Dữ liệu được thu thập từ các tài nguyên web như blog, mạng xã hội, web đánh giá (Twitter, Facebook, Amazon, Tripadvisor, ...). Người dùng sử dụng các công cụ, kỹ thuật quét web để thu được dữ liệu thích hợp,

ngoài ra một số bộ dữ liệu phổ biến có thể được cung cấp bởi các nghiên cứu trước đó.

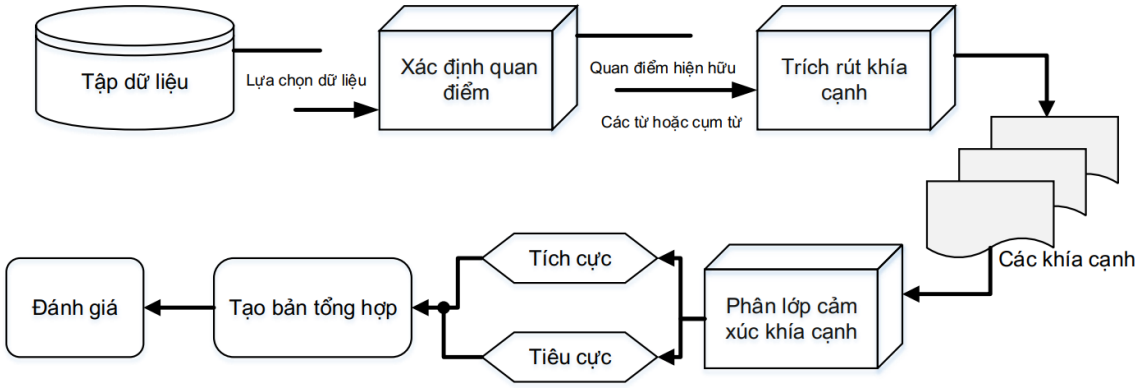
Nhận diện quan điểm (Opinion identification): Tất cả các văn bản chứa quan điểm sẽ được xác định, sau đó chúng cần được xử lý để loại bỏ những bình luận không phù hợp và giả mạo.

Trích rút khía cạnh (Aspect extraction): Trong pha này, tất cả các thể hiện khía cạnh xuất hiện trong văn bản được xác định và trích rút theo các quy trình.

Phân lớp cảm xúc (Sentiment classification): Kết quả của pha trích rút khía cạnh có thể coi là đầu vào (tiền xử lý) của giai đoạn phân lớp cảm xúc. Trong bước này, các quan điểm trên từng khía cạnh được phân lớp bằng các kỹ thuật học máy khác nhau từ giám sát, bán giám sát đến không giám sát.

Tạo ra bản tổng hợp (Production summary): Dựa trên kết quả của các bước trước đó, trong bước tạo ra bản tổng hợp, một bản tóm tắt của các kết quả quan điểm được đưa ra dưới các dạng thức khác nhau như văn bản hoặc biểu đồ.

Đánh giá (Evaluation): Việc thực hiện phân lớp quan điểm có thể được đánh giá thông qua việc sử dụng các chỉ số đánh giá như độ chính xác (accuracy), độ chụm (precision), độ đo triệu hồi (recall), và độ đo f1(f-score).



Hình 1.4: Quy trình phân tích quan điểm dựa trên khía cạnh

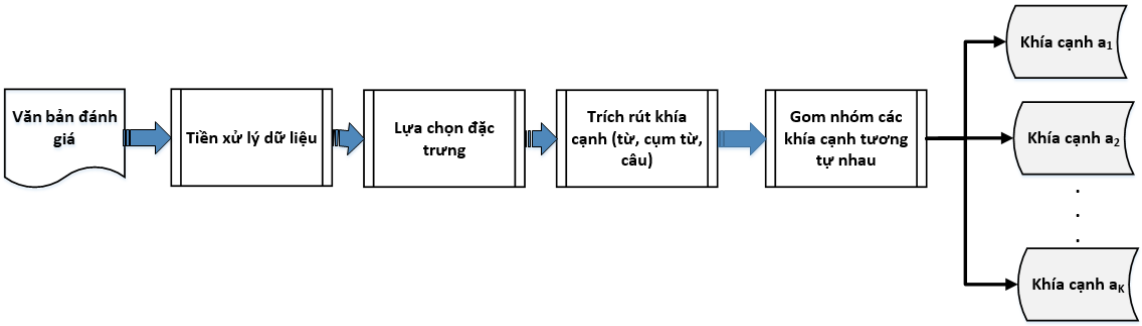
1.2.2 Các bài toán trong phân tích quan điểm mức khía cạnh

Có ba bài toán chính trong khai phá quan điểm dựa trên khía cạnh. Các bài toán đó là: (1) *trích rút khía cạnh*; (2) *phân lớp cảm xúc khía cạnh*; (3) *tổng hợp quan điểm* [3, 5, 11].

Bài toán trích rút khía cạnh

Quá trình xác định các từ cảm xúc (thể hiện quan điểm) và các từ khía cạnh của thực thể từ một câu đã cho, sau đó gán nhãn khía cạnh, trích rút chúng, gom nhóm các khía cạnh tương tự nhau trong một danh mục được gọi là trích rút khía cạnh.

Bài toán trích rút khía cạnh có thể được chia thành 2 bài toán con là *trích rút thuật ngữ khía cạnh* (Aspect Term Extraction – ATE) và gom nhóm các thuật ngữ khía cạnh này thành *danh mục khía cạnh* (Aspect Category Detection - ACD). Quy trình trích rút khía cạnh được thể hiện như trong Hình 1.5.

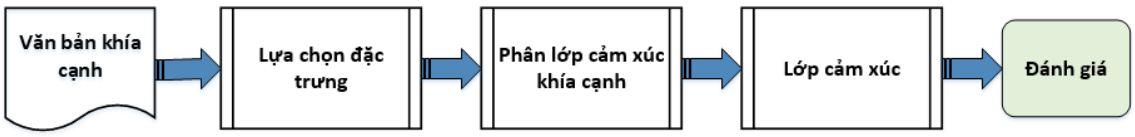


Hình 1.5: Quy trình trích rút khía cạnh

Bài toán phân lớp cảm xúc dựa trên khía cạnh

Vấn đề phân lớp cảm xúc là việc thực hiện phân lớp cảm xúc các khía cạnh đã được trích rút (từ bài toán trích rút khía cạnh) thành các cực cảm xúc như tích cực, tiêu cực, và trung lập hoặc xác định điểm đánh giá theo các thang điểm nhất định (ví dụ từ 1-10 điểm hoặc 1-5 sao). Quy trình phân loại cảm xúc khía cạnh được thể hiện như trong hình 1.6.

Phân lớp cảm xúc khía cạnh có hai dạng chính [3]. *Phân cực cảm xúc* (Sentiment polarity) là xác định xem cảm xúc của người dùng về một khía cạnh là tích cực, tiêu cực hay trung lập. *Đánh giá điểm cảm xúc* (Sentiment rating) gán một điểm số cảm xúc cho khía cạnh đó.



Hình 1.6: Quy trình phân loại cảm xúc khía cạnh

Bài toán tổng hợp quan điểm dựa trên khía cạnh

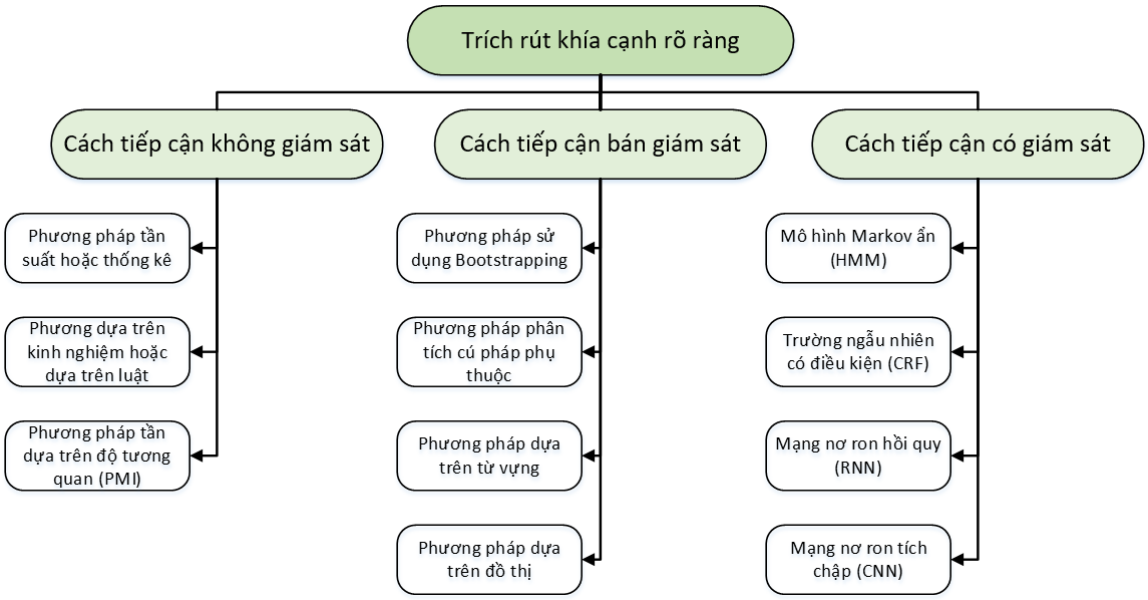
Nhiệm vụ này thực hiện cấu trúc lại toàn bộ các quan điểm người dùng về một thực thể (dựa trên kết quả của các bài toán trước đó) dưới dạng tóm tắt các thành phần chính, các đánh giá định tính hoặc định lượng cụ thể trên từng thành phần.

1.2.3 Các cách tiếp cận trích rút khía cạnh

Qua các cuộc khảo sát [3–7] cho thấy, rất nhiều công bố khác nhau đã được viết về các lĩnh vực khác nhau của phân tích quan điểm. Cũng theo thống kê cho thấy phân tích quan điểm dựa trên khía cạnh là một trong những lĩnh vực thu hút nhiều nhà nghiên cứu quan tâm. Trong ba vấn đề chính của phân tích quan điểm dựa trên khía cạnh, bài toán trích rút khía cạnh là bài toán quan trọng và nhiều thách thức nhất [5].

1.2.3.1 Các phương pháp trích rút khía cạnh rõ ràng

Việc trích rút các khía cạnh mà chúng có các thể hiện khía cạnh trong câu là các danh từ và cụm danh từ được gọi là trích rút khía cạnh rõ ràng. Trích rút khía cạnh rõ ràng có thể phân thành ba loại theo cách tiếp cận học tập: không giám sát, bán giám sát và có giám sát (xem Hình 1.7).



Hình 1.7: Phân loại các phương pháp trích rút khía cạnh rõ ràng

- ***Trích rút khía cạnh rõ ràng theo cách tiếp cận học không giám sát:***

Cách tiếp cận này được thực hiện với dữ liệu không có nhãn, và đây cũng là phương pháp phổ biến được các nhà nghiên cứu áp dụng cho các miền, các ngôn ngữ, và các bộ dữ liệu khác nhau [5]. Các phương pháp không giám sát thường dựa trên các thông tin thống kê (tần suất xuất hiện của các danh từ hoặc cụm danh từ), các thông tin về cấu trúc ngôn ngữ và kinh nghiệm (phân tích cú pháp, mẫu cú pháp), các thông tin tương quan giữa các thành phần trong ngôn ngữ (thông tin tương hỗ, thông tin giữa từ và văn bản) để thực hiện phát hiện các thể hiện khía cạnh. Sau đó, các thể hiện khía cạnh này được gom nhóm và trích rút cùng các quan điểm trên đó. Các phương pháp

này bao gồm các phương pháp dựa trên tần suất và thống kê (frequency or statistical) [11, 44, 46, 60], phương pháp dựa trên kinh nghiệm hoặc dựa trên luật (heuristic- or rule-based) [16, 61, 62], và phương pháp dựa trên điểm thông tin tương hỗ (Pointwise Mutual Information - PMI) [44].

- ***Trích rút khía cạnh rõ ràng theo cách tiếp cận học bán giám sát:***

Những kỹ thuật này phụ thuộc một phần vào đầu vào của người dùng và cần một số yếu tố hạt giống ban đầu để bắt đầu thuật toán. Các phương pháp theo cách tiếp cận này thường sử dụng các thuật toán mở rộng như bootstraping [17, 36], trình phân tích cú pháp phụ thuộc để tìm ra mối quan hệ giữa các từ quan điểm và các khía cạnh [18, 63], hoặc sử dụng từ điển đồng nghĩa kết hợp với một số mô hình như phân bố Dirichlet ẩn, thuật toán PageRank để tìm ra các khía cạnh [64, 65].

- ***Trích rút khía cạnh rõ ràng theo cách tiếp cận học giám sát:***

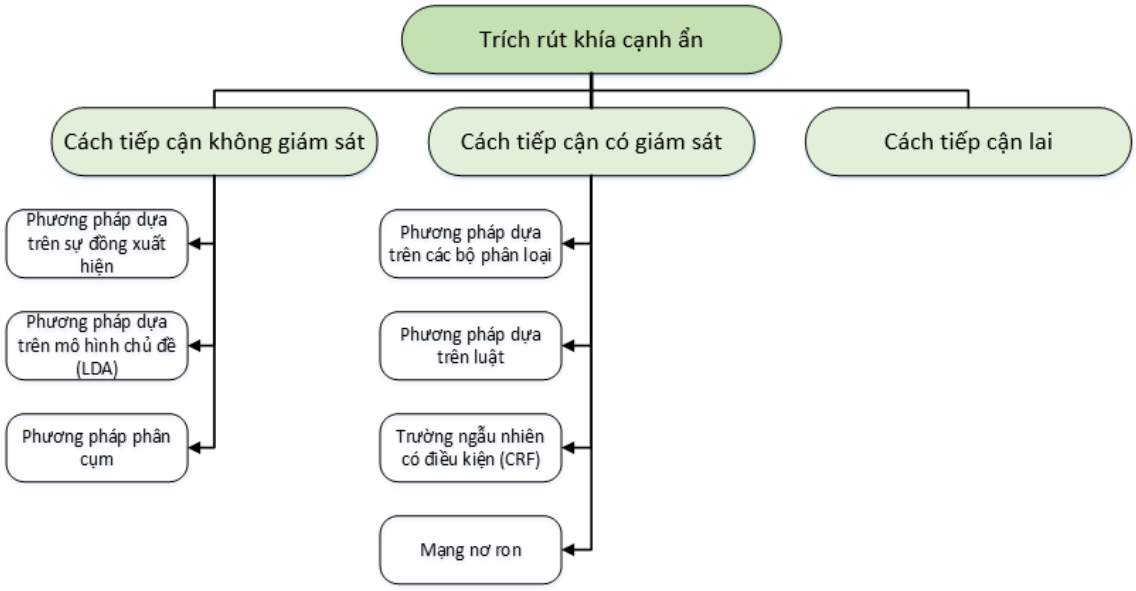
Học có giám sát đề cập đến việc tạo ra một mô hình dựa trên một tập dữ liệu huấn luyện có gán nhãn, và sau đó áp dụng nó vào một tập dữ liệu thử nghiệm chưa được gán nhãn. Vì vậy bài toán xác định khía cạnh và quan điểm là bài toán về chú thích. Trong cách tiếp cận có giám sát, các mô hình học thường được sử dụng là Mô hình Markov ẩn (Hidden Markov Model- HMM) và trường ngẫu nhiên có điều kiện (Conditional Random Field-CRF) [19–21]. Gần đây một số nghiên cứu trích rút khía cạnh dựa trên mô hình mạng nơ ron tích chập (Convolutional Neural Network-CNN) đã được đề xuất. Mô hình này sử dụng cấu trúc mạng thần kinh nhiều lớp, trong đó có một lớp tích chập với nhiều bộ lọc. CNN có thể trích rút các mẫu n-gram vì chúng có khả năng học theo ngữ cảnh [22, 23]. Mạng nơ ron hồi quy (Recurrent Neural Network - RNN) cũng là một mô hình khác của học sâu sử dụng thông tin tuần tự để dự đoán một từ trong một chuỗi từ. RNN có số lớp cố định, độ dài đầu vào có thể thay đổi nên cần xử lý đệ quy. Các nghiên cứu điển hình có thể kể đến là [24–26]. Gần đây hơn, một cơ chế chú ý trong mạng học sâu đã phát triển thành Transformer encoder-decoder [66], điều này mang lại lợi ích cho việc song song hóa của CNN và vấn đề kiểm soát sự phụ thuộc trong chuỗi từ dài của RNN. Kiến trúc Transformer sau đó phát triển thành mô hình BERT (Bidirectional Encoder Representations from Transformers) [67]. Đã có nhiều nghiên cứu [68–70] sử dụng mô hình BERT trong các mô hình học khác nhau để thực thi nhiệm vụ trích rút khía cạnh.

1.2.3.2 Các phương pháp trích rút khía cạnh ẩn

Trích rút khía cạnh ẩn là việc phát hiện các khía cạnh từ các câu không tường minh, các câu có các thể hiện khía cạnh không phải là các danh từ hay cụm danh từ. Các phương pháp trích rút khía cạnh ẩn cũng có thể được chia thành ba loại chính: không giám sát, có giám sát và cách tiếp cận lai (xem Hình 1.8).

- ***Trích rút khía cạnh ẩn theo cách tiếp cận học không giám sát:***

Các thông tin hữu ích được trích rút từ kho dữ liệu và được áp dụng để khám phá các khía cạnh tiềm ẩn. Ngoài ra, các kiến thức ngôn ngữ hoặc miền lĩnh vực có sẵn trên Internet cũng được sử dụng. Các phương pháp không giám sát thường dựa trên sự đồng xuất hiện (của các thuật ngữ trong kho ngữ liệu hoặc cơ sở kiến thức), dựa trên các mô hình thống kê (mô hình chủ đề) [71–73], và dựa trên phân cụm để phát hiện khía cạnh ngầm định [74].



Hình 1.8: Phân loại các phương pháp trích rút khía cạnh ẩn

- ***Trích rút khía cạnh ẩn theo cách tiếp cận học có giám sát:***

Cách tiếp cận này, dữ liệu gán nhãn được sử dụng để huấn luyện thuật toán và sau đó thuật toán được dùng để dự đoán khía cạnh ẩn cho một câu mới. Các phương pháp học giám sát chia thành các phương pháp dựa trên phân lớp (NB, SVM, cây quyết định, rừng ngẫu nhiên) [75, 76], các phương pháp dựa trên luật [62, 77], các phương pháp dựa trên nhãn tuần tự (CRF) [78]. Cùng với sự phát triển của mạng học sâu và mô hình BERT, các nghiên cứu trích rút khía cạnh ẩn cũng ngày càng nhiều và hiệu quả hơn [79–81].

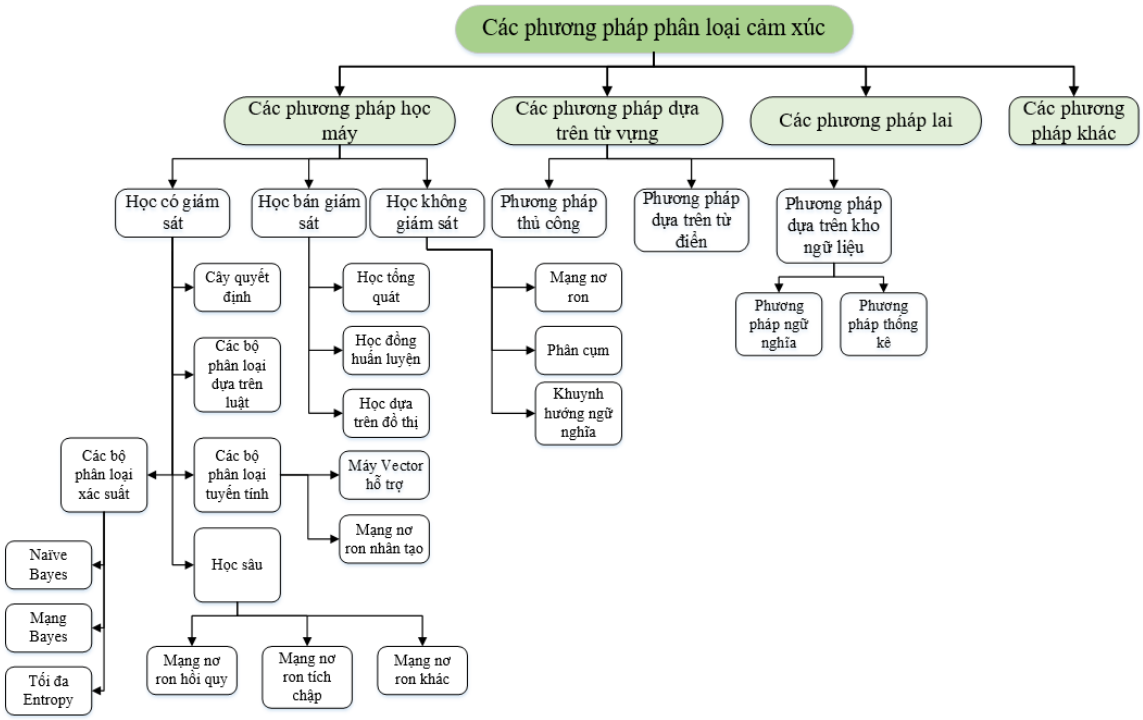
- ***Trích rút khía cạnh ẩn theo cách tiếp cận lai:***

Đây là cách kết hợp nhiều phương pháp khác nhau để xác định khía cạnh ngầm định. Chiến thuật kết hợp có thể tiến hành nối tiếp hoặc song song [82–84].

1.2.4 Các phương pháp phân loại cảm xúc khía cạnh

Các cách tiếp cận hiện nay cho bài toán phân lớp cảm xúc có thể được phân loại dựa trên nhiều quan điểm khác nhau (ví dụ: cách nhìn của toàn văn bản, mức độ chi

tiết của phân tích văn bản) [7]. Tuy nhiên hầu hết các tài liệu thường chia các cách tiếp cận phân lớp cảm xúc thành ba loại: cách tiếp cận học máy, cách tiếp cận dựa trên từ vựng, và các phương pháp lai [4, 7] (xem Hình 1.9).



Hình 1.9: Phân loại các phương pháp phân loại cảm xúc khía cạnh

- ***Các phương pháp phân lớp cảm xúc dựa trên học máy:***

Học máy là một trong những kỹ thuật thú vị nhất và được sử dụng rộng rãi do khả năng thích ứng và độ chính xác của nó trong lĩnh lực phân tích quan điểm [4, 7]. Cách tiếp cận này sử dụng đặc trưng ngôn ngữ bằng cách sử dụng các cơ chế học có giám sát, bán giám sát, không giám sát, học tăng cường, và học sâu [4, 7]. Cách tiếp cận có giám sát [27–29, 31–33, 85, 86] được áp dụng khi nhiệm vụ phân loại có một tập các lớp cụ thể đã được gán nhãn. Khi không có dữ liệu được gán nhãn thì phương pháp không giám sát [34, 87, 88] có thể là chìa khóa trong tình huống này. Mặt khác, phương pháp bán giám sát [36, 89, 90] có thể được sử dụng cho các tập dữ liệu chưa được gán nhãn mà trong đó nó bao gồm một số ví dụ đã được gán nhãn. Các thuật toán học tăng cường [91] sử dụng các cơ chế thử và sai để giúp tác nhân tương tác với môi trường xung quanh từ đó thu được phần tính toán tích lũy tối đa. Trong khi đó, các phương pháp học sâu [85, 92–95] là các phương pháp hiện đại sử dụng mạng nơ ron nhiều tầng để tạo ra các đặc trưng nổi bật và hiệu, quả đồng thời thu được hiệu suất cao trong việc giải quyết bài toán phân lớp cảm xúc [7]. Cùng với sự xuất hiện của mô hình BERT [67], phân tích quan điểm dựa trên khía cạnh với việc thực thi đa nhiệm vụ cũng được giải quyết một cách dễ dàng hơn, hiệu quả hơn. Các nghiên cứu

điển hình như trong [69, 96–98]. Các phương pháp cụ thể trong cách tiếp cận học máy được chỉ ra trong Hình 1.9.

- ***Các phương pháp phân lớp cảm xúc dựa trên từ vựng:***

Các phương pháp tiếp cận dựa trên từ vựng (lexicon-based) còn được gọi là cách tiếp cận dựa trên tri thức. Phương pháp này dự đoán điểm cảm xúc dựa trên từ vựng cảm xúc và dữ liệu không được gán nhãn. Phương pháp này yêu cầu một nguồn từ vựng có tên là từ vựng cảm xúc (một danh sách các từ được xác định trước) liên kết với định hướng ngữ nghĩa của chúng dưới dạng các từ tiêu cực hoặc tích cực được thể hiện bằng các điểm số (chẳng hạn như $+1$, -1 hoặc 0 tương ứng cho các từ tích cực, tiêu cực hoặc trung lập hoặc một giá trị phản ánh sức mạnh hoặc cường độ cảm xúc) [3, 4]. Định hướng phân cực cảm xúc cuối cùng của một tài liệu quan điểm có được bằng cách tính toán các giá trị định hướng ngữ nghĩa của các từ tạo nên nó. Có ba kỹ thuật chính để tạo các từ điển chú thích là phương pháp thủ công [99], phương pháp dựa trên từ điển [54] và phương pháp dựa trên kho ngữ liệu [100–102].

- ***Các phương pháp phân lớp cảm xúc dựa trên phương pháp lai:***

Các phương pháp lai là phương pháp kết hợp cả cách tiếp cận từ vựng và học máy. Nó kết hợp tính thông lượng của phân tích từ vựng và tính linh hoạt của cách tiếp cận học máy để đối phó với sự mơ hồ và tích hợp ngữ cảnh của các từ cảm xúc [103]. Lý do chính đằng sau phương pháp kết hợp là thừa hưởng độ chính xác cao từ học máy và tính ổn định từ phương pháp dựa trên từ vựng. Cách tiếp cận lai giúp khắc phục các yếu điểm và tận dụng các lợi thế của các phương pháp cơ sở. Điểm phân cực cảm xúc của các từ trong từ điển là đầu vào cho bộ phân loại cảm xúc. Hầu hết các nghiên cứu theo cách tiếp cận này đã sử dụng các phương pháp dựa trên từ vựng để gắn nhãn phân cực của từ, sau đó tiếp tục được sử dụng trong bộ phân loại cảm xúc. Một số nghiên cứu điển hình như [104, 105].

1.3 Một số kiến thức học máy liên quan được sử dụng trong luận án cho phân tích quan điểm mức khía cạnh

1.3.1 Thuật toán bootstrap

Phương pháp Bootstrapping [106] do nhà thống kê học Bradley Efron thuộc đại học Stanford phát triển từ cuối thập niên 1979. Phương pháp Bootstrapping được xem là phương pháp chuẩn trong phân tích thống kê và đã làm nên một cuộc cách mạng trong thống kê vì nó giải quyết được nhiều vấn đề mà trước đây khó giải quyết được.

Bootstrapping là phương pháp lấy mẫu dựa trên nguyên lý chọn mẫu có hoàn lại. Từ mẫu ban đầu lấy lại các mẫu ngẫu nhiên cùng cỡ với mẫu gốc bằng phương pháp

lấy mẫu có hoàn lại, gọi là mẫu bootstrap. Với mỗi mẫu lấy lại, người ta tính được giá trị tham số thống kê quan tâm gọi là tham số bootstrap. Sự phân bố của các tham số thống kê mẫu bootstrap là phân phối bootstap.

Xem xét một mẫu ngẫu nhiên cỡ n được quan sát từ một phân bố xác suất hoàn toàn chưa biết F. $X_i = x_i; X_i \sim_{ind} F; i = 1,...,n$. Giả sử rằng $\mathbf{X} = (X_1, X_2, ..., X_n)$ và $\mathbf{x} = (x_1, x_2, ..., x_n)$ lần lượt ký hiệu là các mẫu ngẫu nhiên và các quan sát thực tế của chúng. Cho một biến ngẫu nhiên được chỉ định $R(\mathbf{X}, F)$ phụ thuộc vào biến $\mathbf{X}$ và cả phân bố chưa biết F, ước lượng phân bố mẫu của R trên cơ sở các dữ liệu quan sát $\mathbf{x}$. Phương pháp bootstrap cho một mẫu đơn giản như sau:

1. Xây dựng phân phối xác suất mẫu $\widehat{F}$, đặt một lượng $\frac{1}{n}$ cho mỗi điểm $x_1, x_2, ..., x_n$
2. Với $\widehat{F}$ cố định, vẽ một mẫu ngẫu nhiên của kích thước n từ $\widehat{F}$ nói rằng:

   $X_i^* = x_i^*, X_i^* \sim_{ind} F, i = 1,...,n.$

   Gọi đây là mẫu bootstrap, $\mathbf{X}^* = (X_1^*, X_2^*, ..., X_n^*), \mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$

   Lưu ý rằng không có phân phối hoán vị vì các giá trị của $\mathbf{X}^*$ được lựa chọn với sự thay thế từ tập $\{x_1, x_2, \cdots, x_n\}$
3. Xấp xỉ hàm phân bố mẫu $R(\mathbf{X}, F)$ bằng phân bố xác định bởi bootstrap $R^* = R(\mathbf{X}^*, \widehat{F})$ nghĩa là phân bố R^* được xác định bởi cấu trúc ngẫu nhiên (ở bước 2) tạo ra với F được giữ cố định ở giá trị quan sát của nó.
4. Lặp lại các bước trên nhiều lần cho đến khi có được tham số mong muốn.

1.3.2 Cơ sở lý thuyết biểu diễn từ Word to Vector

1.3.2.1 Một số khái niệm trong biểu diễn từ Word to Vector

Khái niệm 1.1 Biểu diễn phân phối (Distributed representations) [107]: Biểu diễn phân phối là mỗi thực thể được biểu diễn dưới dạng một vectơ giá trị ("một mẫu của các hoạt động "), ý nghĩa của thực thể và mối quan hệ của nó với các thực thể khác được ghi lại bởi các hoạt động trong vectơ và sự tương đồng giữa các vectơ khác nhau.

Khái niệm 1.2 Phân bố ngữ nghĩa (Distributional semantics) [40]: Trong phân bố ngữ nghĩa, ý nghĩa của một từ có thể được bắt nguồn từ sự phân bố của nó trong một kho ngữ liệu, tức là một từ được tổng hợp từ các ngữ cảnh mà nó đang được sử dụng. Các từ có xu hướng xuất hiện trong các ngữ cảnh tương tự sẽ có ngữ nghĩa tương tự.

Giả thuyết phân phối (Distributional hypothesis) về ngôn ngữ và ý nghĩa của từ nói rằng các từ xuất hiện trong cùng một ngữ cảnh có xu hướng có ý nghĩa giống nhau. Khi gặp một câu có từ không xác định (không biết ý nghĩa thực sự của nó), có thể suy ra nghĩa của từ đó dựa trên ngữ cảnh mà nó xuất hiện.

Ma trận từ-ngữ cảnh (Word-context Matrices): Trong đó, mỗi hàng i đại diện cho một từ, mỗi cột j đại diện cho ngữ cảnh ngôn ngữ mà ở đó các từ có thể xuất hiện.

Một điểm trong ma trận $M_{[i,j]}$ định lượng độ mạnh liên kết giữa một từ và ngữ cảnh trong một kho ngữ liệu lớn. Nói cách khác, mỗi từ được biểu diễn dưới dạng một vectơ thưa thớt trong không gian có số chiều lớn, mã hóa một túi ngữ cảnh có trọng số mà nó xuất hiện.

Định nghĩa 1.9 Ma trận từ-ngữ cảnh: Ký hiệu V_W là tập hợp các từ (các từ vựng) và V_C là tập hợp các ngữ cảnh có thể xuất hiện của tất cả các từ. Giả sử rằng mỗi từ và mỗi ngữ cảnh đều được lập chỉ mục, sao cho w_i là từ thứ i trong từ điển V_W và c_j là từ ngữ cảnh thứ j trong từ điển ngữ cảnh V_C. Ma trận $\mathbf{M}^f \in \mathbb{R}^{|V_W| \times |V_C|}$ là ma trận từ-ngữ cảnh, được xác định như là $\mathbf{M}^f_{[i,j]} = f(w_i, c_j)$, trong đó, trong đó f là một phép đo liên kết về độ mạnh giữa một từ và ngữ cảnh.

Khái niệm 1.3 Độ đo tương tự: Độ đo tương tự giữa các từ được đo lường bởi các độ đo để thấy sự tương đồng về mặt ngữ nghĩa hoặc nội dung của chúng.

Các hàm khoảng cách khác nhau có thể được sử dụng để đo độ tương tự giữa các vectơ từ, chúng được sử dụng để biểu thị khoảng cách ngữ nghĩa giữa các từ liên kết. Khi các từ được biểu diễn dưới dạng vectơ, có thể tính toán sự tương đồng giữa các từ bằng cách tính toán sự tương đồng giữa các vectơ tương ứng. Một số các độ đo tương tự phổ biến và hiệu quả là tính độ tương tự cosin, Jacaard tổng quát, điểm thông tin tương hỗ (PMI).

1.3.2.2 Thuật toán nhúng từ W2V

Thuật toán Word2vec được phát triển bởi Tomáš Mikolov và các đồng nghiệp [108]. Word2Vec không phải là một thuật toán đơn lẻ, nó là một gói phần mềm thực hiện hai biểu diễn ngữ cảnh khác nhau (CBOW và Skip-Gram) và hai mục tiêu tối ưu hóa khác nhau (lấy mẫu âm (Negative Sampling) và toán tử Softmax phân cấp (Hierarchical Softmax)).

Mô hình CBOW (Continuous bag of words)

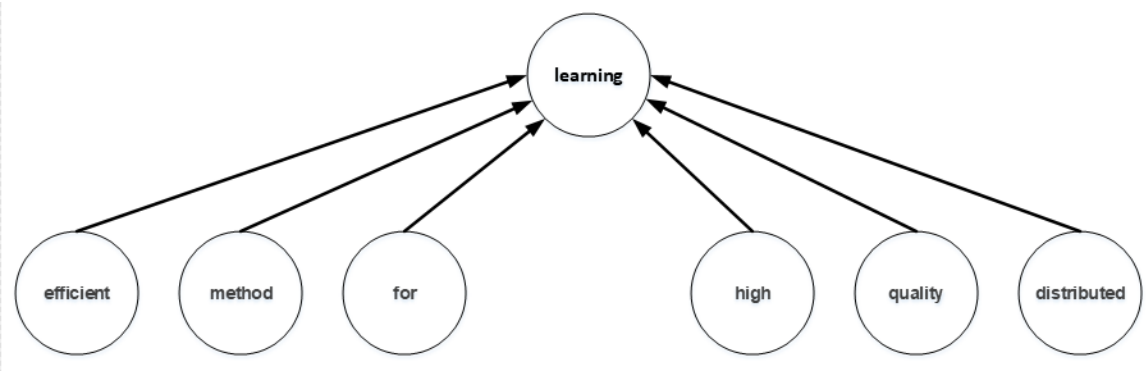
Mô hình túi từ liên tục giả sử rằng từ đích trung tâm được tạo ra dựa trên các từ ngữ cảnh phía trước và phía sau nó trong một chuỗi văn bản. Với một chuỗi văn bản "an efficient method for learning high quality distributed vector", xét từ đích trung tâm là "learning" với cửa sổ bằng 3, mô hình CBOW quan tâm đến xác suất có điều kiện để sinh ra từ đích "learning" dựa trên các từ ngữ cảnh là "efficient", "method", "for", "high", "quality", "distributed" được mô tả như Hình 1.10.

Đối với một từ đa ngữ cảnh $c_{1:k}$ mô hình CBOW của Word2vec xác định vectơ ngữ cảnh $\mathbf{c}$ là tổng của các vectơ nhúng của các thành phần ngữ cảnh $\mathbf{c} = \sum_{i=1}^{k} \mathbf{c}_i$, sau đó

xác định điểm số $s(w,c) = \mathbf{w} \cdot \mathbf{c}$ kết quả là:

$$P(D = 1|w, c_{1:k}) = \frac{1}{1 + e^{-(\mathbf{w}\cdot\mathbf{c}_1 + \mathbf{w}\cdot\mathbf{c}_2 + ... + \mathbf{w}\cdot\mathbf{c}_k)}} \tag{1.1}$$

Mô hình CBOW làm mất thông tin thứ tự giữa các phần tử ngữ cảnh. Nhưng đổi lại, nó cho phép sử dụng các ngữ cảnh có độ dài thay đổi. Tuy vậy, đối với các ngữ cảnh có độ dài cố định, CBOW vẫn có thể giữ lại thông tin thứ tự bằng cách bao gồm vị trí tương đối giữa chúng như một phần của phần tử nội dung, nghĩa là bằng cách gán nhãn vectơ nhúng khác nhau cho các phần tử ngữ cảnh ở các vị trí tương đối khác nhau.



Hình 1.10: Mô hình CBOW quan tâm đến xác suất có điều kiện tạo ra từ đích trung tâm dựa trên các từ ngữ cảnh cho trước

Mô hình Skip-Gram

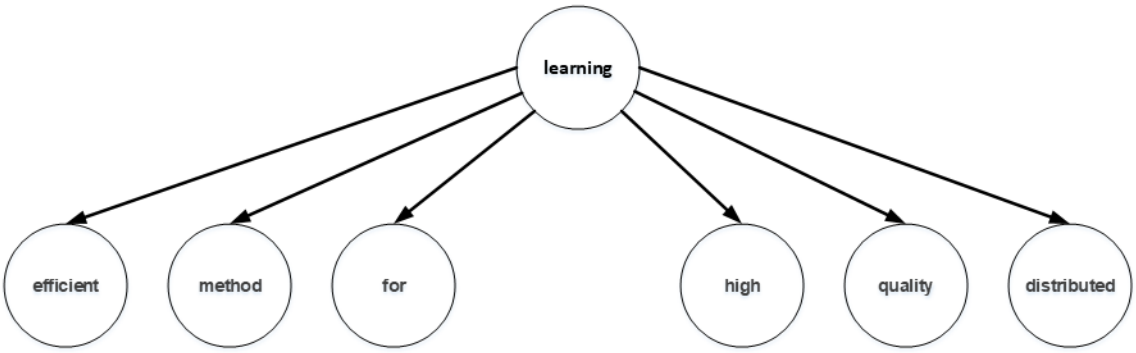
Mô hình Skip-gram giả định rằng một từ có thể được sử dụng để sinh ra các từ xung quanh nó trong một chuỗi văn bản. Ví dụ ta có chuỗi văn bản "an efficient method for learning high quality distributed vector". Ta sử dụng từ "learning" làm từ đích trung tâm và cửa sổ ngữ cảnh bằng 3, mô hình Skip-gram quan tâm đến xác suất có điều kiện sinh ra các từ ngữ cảnh với một từ đích trong khung cửa sổ P(efficient, method, for, high, quality, distributed | learning), xem mô tả Hình 1.11.

Mô hình Skip-gram của điểm số Word2vec tách rời sự phụ thuộc giữa các yếu tố ngữ cảnh. Đối với k thành phần ngữ cảnh $c_{\{1:k\}}$, biến thể Skip-gram giả định rằng các phần tử c_i trong ngữ cảnh là độc lập với nhau, có thể coi chúng như k ngữ cảnh khác nhau, nghĩa là mỗi cặp từ-ngữ cảnh $(w, c_{\{i:k\}})$ sẽ được biểu diễn trong D dưới dạng k ngữ cảnh khác nhau: $(w,c_1), \ldots, (w,c_k)$. Hàm điểm $s(w,c)$ được xác định như trong

CBOW, nhưng bây giờ mỗi một ngữ cảnh là một vector nhúng đơn lẻ:

$$\begin{gathered} P(D=1|w,c_i) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{c_i}}} \\ P(D=1|w,c_{1:k}) = \prod_{i=1}^{k} P(D=1|w,c_i) = \prod_{i=1}^{k} \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{c_i}}} \\ \log P(D=1|w,c_{1:k}) = \sum_{i=1}^{k} \log \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{c_i}}} \end{gathered} \tag{1.2}$$

Trong khi đưa ra các giả định về tính độc lập mạnh mẽ giữa các thành phần của ngữ cảnh, mô hình Skip-gram rất hiệu quả trong thực tế và được sử dụng rất phổ biến.



Hình 1.11: Mô hình Skip-gram quan tâm đến xác suất có điều kiện tạo ra các từ ngữ cảnh với một từ đích trung tâm cho trước

1.3.3 Phân loại hai lớp máyvec tơ hỗ trợ

Máy vector hỗ trợ (Support Vector Machines-SVM) được phát triển bởi Cortes và Vapnik [109], ban đầu chỉ áp dụng cho phân loại nhị phân và là một phương pháp hiệu quả trong nhiều ứng dụng. SVM dựa trên ý tưởng của bộ phân loại siêu phẳng, hoặc khả năng phân tách tuyến tính và tối đa hóa lợi nhuận. Đây là một hệ thống học tập sử dụng không gian giả thuyết của các hàm tuyến tính trong không gian đặc trưng với số chiều lớn.

Giả sử, các điểm dữ liệu được coi là các bộ giá trị $(\mathbf{X}, C)$ có thể tách rời tuyến tính. Trong đó $x_i \in \mathbb{R}^V$ là các giá trị đặc trưng, $c \in \{\pm 1\}$ là các lớp (+1 là lớp tích cực, -1 là lớp tiêu cực). Một siêu phẳng đối với D ví dụ (văn bản cảm xúc)

$$(\mathbf{x}_1, c_1), ..., (\mathbf{x}_D, c_D); c_i \in \{-1, 1\} \tag{1.3}$$

được cho là có thể phân tách tuyến tính nếu tồn tại một vector $\mathbf{w}$ và một giá trị vô hướng b sao cho thỏa mãn bất đẳng thức (1.4) đối với tất cả các phần tử của tập huấn

luyện (1.3).

$$\begin{aligned}(\mathbf{w}.\mathbf{x}_i)+b \geq 1 \text{ if } c_i = 1;\\ (\mathbf{w}.\mathbf{x}_i)+b \leq -1 \text{ if } c_i = -1\end{aligned} \tag{1.4}$$

Viết lại (1.4) dưới dạng (1.5), khi đó siêu phẳng tối ưu (1.6) là mặt phẳng duy nhất phân tách dữ liệu huấn luyện với một lề lớn nhất. Điều này có nghĩa là hai siêu phẳng trong (1.4) có cùng khoảng cách đến (1.6) và không có phần tử nào của tập mẫu nằm giữa hai siêu phẳng đó.

$$c_i(\mathbf{w}.\mathbf{x}_i+b) \geq 1; i = \{1,...,D\} \tag{1.5}$$

$$\mathbf{w}_0.\mathbf{x}+b_0 = 0 \tag{1.6}$$

Sau khi tìm được siêu phẳng phân tách, việc phân loại một mẫu mới chỉ là việc kiểm tra hàm dấu của mẫu mới.

$$c(\mathbf{x}) = sign(\mathbf{w}\mathbf{x}+b) \tag{1.7}$$

Trong bài toán phân loại văn bản dạng nhị phân, quá trình huấn luyện là quá trình tìm ra một siêu phẳng phân tách vectơ văn bản trong một lớp này với một lớp còn lại sao cho khoảng cách giữa hai siêu phẳng càng lớn càng tốt.

1.3.4 Phân loại đa lớp Naive Bayes

Đây là bộ phân loại dựa trên lý thuyết Bayes. Trong phân loại văn bản, phương pháp này phụ thuộc nhiều vào các kỹ thuật trích rút đặc trưng dạng BOW, do đó vị trí của các từ trong văn bản bị bỏ qua, đồng thời sự xuất hiện của các từ được xem là độc lập với nhau. Cho một văn bản d và một tập các nhãn C, xác suất để văn bản d thuộc về một nhãn $c_i \in C$ được xác định bởi luật Bayes là:

$$P(c_i|d) = \frac{p(c_i)p(d|c_i)}{p(d)} \tag{1.8}$$

trong đó $p(c_i)$ là xác suất tiên nghiệm của lớp c_i, $p(d|c_i)$ là xác suất tiên nghiệm của văn bản d được gán nhãn lớp c_i, và $p(d)$ là xác suất tiên nghiệm của văn bản d.

Dựa trên giả thuyết với điều kiện các đặc trưng là độc lập, Naive Bayes xác định xác suất hậu nghiệm của một lớp khi đưa ra một văn bản với việc sử dụng phân bố của các từ (các đặc trưng) công thức (1.8) được viết lại thành:

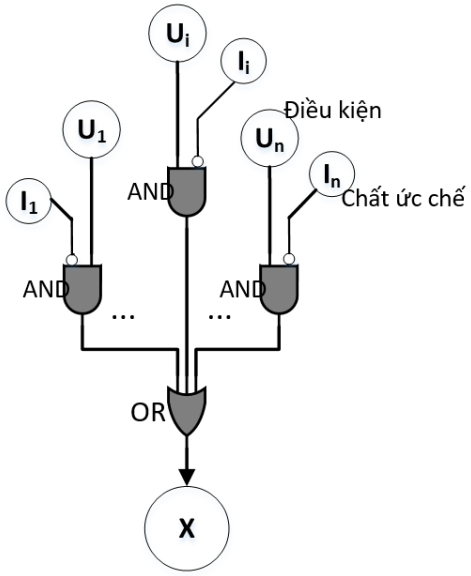
$$P(c_i|d) = \frac{p(c_i)p(w_1|c_i)p(w_2|c_i)...p(w_q|c_i)}{p(d)} \tag{1.9}$$

Hàm phân lớp sẽ gán nhãn cho lớp có giá trị xác suất hậu nghiệm lớn nhất như sau:

$$\widehat{c} = \underset{c_i \in C}{argmax}(P(c_i|d)) \tag{1.10}$$

1.3.5 Tương tác không kết hợp (Nhiễu cổng OR - Noisy OR-gate)

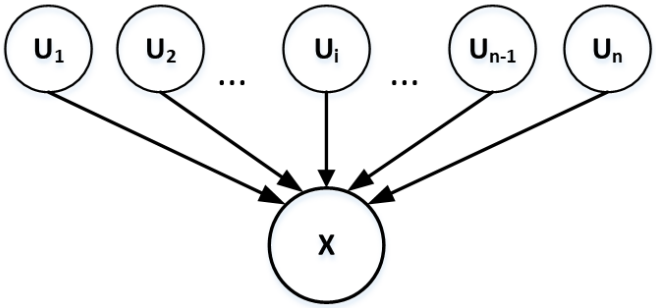
Một ngoại lệ đáng chú ý được gọi là *tương tác không kết hợp* (Disjunctive interaction) xảy ra khi bất kỳ thành viên nào của tập các nguyên nhân (các cha hoặc mẹ) có khả năng gây ra một sự kiện nhất định (biến quan tâm) và khả năng này không giảm đi khi một số điều kiện xảy ra đồng thời [110]. Nghĩa là, nếu một sự kiện X là hệ quả của một trong hai điều kiện nhân quả U_1 và U_2, thì cơ chế có thể ức chế sự xuất hiện của X bởi tác động U_1 là độc lập với cơ chế có thể ức chế sự xuất hiện của X dưới sự tác động của U_2. Mỗi cơ chế ức chế hoạt động như một biến độc lập. Ý tưởng này có thể được mô tả như trong Hình 1.12 [110].



Hình 1.12: Mô hình chuẩn về các tương tác không kết hợp giữa nhiều nguyên nhân $U_1,...,U_n$ dự đoán cùng một hệ quả X

Trong Hình 1.12, X biểu diễn cho một dự đoán hoặc một hệ quả và được xem như là đầu ra của cổng logic OR, mỗi đầu vào của cổng OR là đầu ra của một cổng AND thể hiện cho sự kết hợp của nguyên nhân giải thích cho X là U_i và một phủ định của cơ chế ức chế của U_i là I_i. Các đầu vào $\mathbf{U} = \{U_1,...,U_n\}$ là các cha hoặc mẹ của X trong mạng Bayes, chúng thường đại diện cho các giả thuyết, giải thích, phỏng đoán,

các yếu tố nhân quả hoặc các điều kiện để giải thích cho sự xuất hiện của X. Các chất ức chế $I_1,...,I_n$ biểu diễn cho các trường hợp ngoại lệ hoặc bất thường cản trở mối quan hệ bình thường giữa U và X. Chúng thường không được biểu diễn bằng các nút trong mạng Bayes, nhưng được tóm tắt ngầm định trong xác suất kết nối giữa U và X là $P(X|U_1,U_2,...,U_n)$ [110].



Hình 1.13: Mô hình mạng Bayes cổng OR nguyên nhân $U_1,...,U_n$ và hệ quả X

Chúng ta hãy xem xét một đồ thị mạng Bayes có các nút tương ứng 1-1 với biến nhị phân X là hệ quả của tập n biến nguyên nhân $\mathbf{U} = \{U_1,...,U_n\}$ (xem Hình 1.13). Theo nguyên lý mạng Bayes, xác suất hậu nghiệm của X khi biết n phân bố U, $P(X|U_1,U_2,...,U_n)$ bao gồm 2^n tham số độc lập. Tuy nhiên dựa trên nguyên tắc độc lập nhân quả do Pearl đề xuất (Noisy OR-gate) các tham số này giảm xuống còn n. Việc giải thích hàm xác suất có điều kiện của mạng Noisy OR được giải thích bởi Fabio G.Cozman trong [111] và được viết lại bởi Kuang Zhou trong [112] như sau: Ý tưởng bắt đầu với xác suất p_i, đây là xác suất mà $\{X = True\}$ với điều kiện $\{U_i = True\}$ và $\{U_j = False\}$ đối với $j \neq i$, nghĩa là:

$$p_i = p\{X = T|U_i = T; \{U_j = F\}_{j=1,i\neq j}^{n}\} \tag{1.11}$$

Xác suất p_i thường được gọi là "xác suất liên kết" và nó mô tả thực tế rằng nguyên nhân phụ thuộc giữa U_i và X có thể bị ức chế. Nếu trạng thái của biến U_i là "True" thì sẽ có một khả năng $(1-p_i)$ mà khả năng này làm cho nó chuyển trạng thái thành "False". Nếu U_i là "False" thì nó vẫn ở trạng thái "False". Ký hiệu kết quả của việc lật trạng thái của U_i là $\xi_i, i = 1,2,...,n$ khi đó

$$p(X = \alpha|U_1,U_2,...,U_n) = \sum_{\alpha_1 \vee \alpha_2 \vee ... \vee \alpha_n} p(\xi_1 = \alpha_1|U_1)...p(\xi_n = \alpha_n|U_n) \tag{1.12}$$

trong đó giá trị của α, α_i là một trong hai giá trị "True" hoặc "False". Do đó hoàn Noisy OR là sự kết hợp của các "noisy" của U_i. Gọi $\mathbf{U}_T$ là tập của các U_i có trạng thái "True" và $\mathbf{U}_F$ là tập của các U_i có trạng thái "False", phân bố của biến X điều kiện

trên các biến $U_1, U_2, ..., U_n$ là:

$$p(X = T|U_1, U_2, ..., U_n) = 1 - \prod_{i:U_i \in \mathbf{U}_T} (1 - p_i) \tag{1.13}$$

1.4 Các phương pháp đánh giá kết quả phân tích quan điểm

Để đánh giá hiệu suất trong các nhiệm vụ phân tích quan điểm dựa trên khía cạnh các tiêu chí được lấy từ nhiệm vụ tra cứu thông tin [113, 114] bao gồm các độ đo accuracy (AC), precision (P), recall (R) và f1-score (F1) (kết quả nằm trong khoảng từ $0-1$) được xác định bởi các công thức (1.14 - 1.17). Trong phân lớp nhị phân người ta thường phân chia thành hai lớp gọi là lớp dương (Positive-P) và lớp âm (Negative-N). TP là số các mẫu dương được chuẩn đoán chính xác là dương, FN là số các mẫu dương bị chuẩn đoán sai thành âm, FP là số mẫu âm bị chuẩn đoán sai thành dương, TN là số mẫu âm được chuẩn đoán chính xác là âm.

- **Tiêu chí accuracy**: là tỉ lệ của tất cả các kết quả phân lớp đúng trên toàn bộ các mẫu kiểm tra.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.14}$$

- **Tiêu chí precision**: là tỉ lệ của các mẫu dương được phân lớp đúng với tổng số các mẫu dương đầu ra của hệ thống. Tiêu chí này cho biết tỉ lệ chuẩn đoán đúng của hệ thống đối với mẫu dương.

$$P = \frac{TP}{TP + FP} \tag{1.15}$$

- **Tiêu chí recall**: là tỉ lệ các kết quả đúng mà hệ thống đã gán so với tổng số các mẫu dương có trong cơ sở dữ liệu kiểm tra. Tiêu chí này cho biết tỉ lệ bỏ sót các mẫu dương của hệ thống.

$$R = \frac{TP}{TP + FN} \tag{1.16}$$

- **Tiêu chí F1-score**: là tiêu chí đến từ sự kết hợp của precision và recall. F1 là trung bình hài hòa (harmonic mean) của precision và recall. Tiêu chí này cho biết mức độ chính xác trung bình hợp lý của hệ thống trên hai tiêu chí precision và recall. Điều này có nghĩa là một hệ thống có P=0.5 và R=0.5 thì tốt hơn một hệ thống có P=0.3 và R=0.8.

$$F1 = 2 * \frac{P * R}{P + R} \tag{1.17}$$

Precision, Recall và F1score cũng có thể được sử dụng để đánh giá hiệu suất khai phá khía cạnh sản phẩm (ví dụ trích rút khía cạnh) và tính độ đo precision và recall theo công thức (1.18) và (1.19).

$$P = \frac{|extractedaspects \cap goldaspects|}{|extractedaspects|} \tag{1.18}$$

$$R = \frac{|extractedaspects \cap goldaspects|}{|goldaspects|} \tag{1.19}$$

Hơn nữa một số các chỉ số khác cũng được sử dụng trong khai phá quan điểm dựa trên khía cạnh như sai số tuyệt đối trung bình vĩ mô (Macro-averaged Mean Absolute Error - MAE^M), trung bình bình phương sai số (Mean Square Error - MSE), trung bình bình phương sai số theo khía cạnh (Δ^2_{aspect}), độ tương quan khía cạnh (ρ_{aspect}), và độ tương quan giữa các khía cạnh (ρ_{review}). Trong các chỉ số đánh giá này, nhãn của quan điểm hoặc cảm xúc được giả định là một biến số nguyên.

MAE^M [115] được tính bằng công thức (1.20) thích hợp để giải quyết tập dữ liệu mất cân bằng cao.

$$MAE^M(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|y_j|}\sum_{y_i \in \mathbf{y}_j}|y_i - \widehat{y}_i| \tag{1.20}$$

trong đó k là số nhãn quan điểm hoặc cảm xúc (ví dụ $k = 2$ đối với phân lớp nhị phân), y là vector nhãn chính xác, $\widehat{y}$ là vector nhãn dự đoán, và y_j là tập hợp con của kho dữ liệu được tạo bởi các bài đánh giá có nhãn chính xác là j.

MSE một thước đo được sử dụng rộng rãi trong bài toán hồi quy, được sử dụng để đánh giá hiệu suất của phân lớp quan điểm/cảm xúc [37]. Chỉ số này được tính theo công thức (1.21), trong đó, n là số lượng bài đánh giá có trong kho ngữ liệu, y_j và $\widehat{y}_i$ là nhãn chính xác và nhãn dự đoán của bài đánh giá thứ i.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2 \tag{1.21}$$

Δ^2_{aspect} là một thước đo đánh giá mức độ sai số bình phương của các dự đoán phân lớp đối với giá trị thực sự của các quan điểm hoặc cảm xúc. Thước đo này được tính theo công thức (1.22).

$$\Delta^2_{aspect} = \frac{\sum_{i=1}^{Q}\sum_{j=1}^{K}(r_{i_j} - \widehat{r}_{i_j})^2}{Q \times K} \tag{1.22}$$

trong đó K là số khía cạnh của sản phẩm, Q là số bài đánh giá có trong kho ngữ liệu về sản phẩm đó, $\widehat{r}_{i_j}$ là nhãn dự đoán và r_{i_j} là nhãn chính xác của điểm đánh giá quan

điểm/cảm xúc của khía cạnh j trong bài đánh giá thứ i về sản phẩm.

ρ_{aspect} là thước đo để đánh giá xem điểm đánh giá khía cạnh được dự đoán có thể duy trì thứ tự tương đối của chúng tốt như thế nào trong một bài đánh giá với điểm đánh giá thực. ρ_{aspect} được xác định như sau:

$$\rho_{aspect} = \frac{\sum_{i=1}^{Q} \rho_{\mathbf{r}_i, \widehat{\mathbf{r}}_i}}{Q} \tag{1.23}$$

trong đó Q là số bài đánh giá về sản phẩm, $\rho_{\mathbf{r}_i, \widehat{\mathbf{r}}_i}$ là độ tương quan Pearson giữa hai vector dự đoán $\widehat{\mathbf{r}}_i$ và vector mang giá trị đúng $\mathbf{r}_i$ của điểm đánh giá quan điểm hoặc cảm xúc trên tất cả các khía cạnh trong bài đánh giá thứ i. Độ tương quan Pearson $\rho_{\mathbf{r}_i, \widehat{\mathbf{r}}_i}$ được xác định theo công thức (1.24).

$$\rho_{r_i, \widehat{r}_i} = \frac{\sum_{j=1}^{K} (r_{i_j} - \overline{r_{i_j}}) * (\widehat{r_{i_j}} - \overline{\widehat{r_{i_j}}})}{\sqrt{\sum_{j=1}^{K} (r_{i_j} - \overline{r_{i_j}})^2} * \sqrt{\sum_{j=1}^{K} (\widehat{r_{i_j}} - \overline{\widehat{r_{i_j}}})^2}} \tag{1.24}$$

trong đó K là số khía cạnh của sản phẩm, $\overline{r_{i_j}} = \frac{\sum_{j=1}^{K} r_{i_j}}{K}$ là giá trị điểm trung bình của bài đánh giá thứ i qua tất cả các khía cạnh của sản phẩm, $\overline{\widehat{r_{i_j}}} = \frac{\sum_{j=1}^{K} \widehat{r_{i_j}}}{K}$ là giá trị trung bình dự đoán của bài đánh giá thứ i qua tất cả các khía cạnh của sản phẩm.

Hai thước đo Δ^2_{aspect} và ρ_{aspect} là để đánh giá kết quả cho mỗi bài đánh giá. Kết quả trên toàn bộ tập dữ liệu được đánh giá bằng cách sử dụng thước đo đánh giá tương quan giữa các khía cạnh $\rho_{preview}$:

$$\rho_{preview} = \frac{\sum_{j=1}^{K} \rho(\overrightarrow{r_j}, \overrightarrow{\widehat{r}_j})}{K} \tag{1.25}$$

trong đó $\rho(\overrightarrow{r_j}, \overrightarrow{\widehat{r}_j})$ là độ tương quan Pearson giữa hai vector $\overrightarrow{r_j}$ và $\overrightarrow{\widehat{r}_j}$. $\rho(\overrightarrow{r_j}, \overrightarrow{\widehat{r}_j})$ được tính theo công thức (1.26)

$$\rho(\overrightarrow{r_j}, \overrightarrow{\widehat{r}_j}) = \frac{\sum_{i=1}^{Q} (r_{i_j} - \overline{r_{i_j}}) * (\widehat{r_{i_j}} - \overline{\widehat{r_{i_j}}})}{\sqrt{\sum_{i=1}^{Q} (r_{i_j} - \overline{r_{i_j}})^2} * \sqrt{\sum_{i=1}^{Q} (\widehat{r_{i_j}} - \overline{\widehat{r_{i_j}}})^2}} \tag{1.26}$$

trong đó Q là số bài đánh giá của sản phẩm, $\overline{r_{i_j}} = \frac{\sum_{i=1}^{Q} r_{i_j}}{Q}$ là giá trị điểm trung bình của khía cạnh thứ j qua tất cả các bài đánh giá của sản phẩm, $\overline{\widehat{r_{i_j}}} = \frac{\sum_{i=1}^{Q} \widehat{r_{i_j}}}{Q}$ là giá trị điểm trung bình dự đoán của khía cạnh thứ j qua tất cả các bài đánh giá của sản phẩm.

Nhiệm vụ phân lớp quan điểm hoặc cảm xúc khía cạnh còn là nhiệm vụ phân loại nhiều nhãn, dó đó các độ đo Precision, Recall và F1 nhằm đánh giá vào một lớp nên mang tính cục bộ. Vì vậy các độ đo trung bình vĩ mô (macro-average) và trung bình

vi mô (micro-average) tương ứng của các chỉ số P,R, F1 cần được xem xét thêm.

Các thước đo trung bình vĩ mô xem xét đánh giá từng lớp. Độ đo macro precision và macro recall được tính như sau:

$$MacroPrecision = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FN_i} = \frac{1}{|C|}\sum_{i=1}^{|C|}P_i \tag{1.27}$$

$$MacroRecall = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FP_i} = \frac{1}{|C|}\sum_{i=1}^{|C|}R_i \tag{1.28}$$

trong đó $|C|$ là số lượng các lớp quan điểm/cảm xúc, P_i và R_i là các độ đo Precision và Recall trương ứng của lớp i. Độ đo F1 trung bình vĩ mô tương ứng được sử dụng trong [116, 117] được tính như sau:

$$MacroF1 = 2\frac{MacroPrecision \times MacroRecall}{MacroPrecision + MacroRecall} \tag{1.29}$$

Khác với trung bình vĩ mô, trung bình vi mô tập trung vào từng mẫu của tập dữ liệu. Các độ đo Micro precision và Micro recall được tính như sau:

$$MicroPrecision = \frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}(TP_i + FN_i)} \tag{1.30}$$

$$MicroRecall = \frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}(TP_i + FP_i)} \tag{1.31}$$

Phép đo F1 trung bình vi mô tương ứng sẽ là:

$$MicroF1 = 2\frac{MicroPrecision \times MicroRecall}{MicroPrecision + MicroRecall} \tag{1.32}$$

1.5 Kết luận chương 1

Trong những năm gần đây, cùng với sự phát triển bùng nổ của các phương tiện truyền thông trực tuyến, thương mại điện tử cũng phát triển không ngừng và trở thành một xu thế toàn cầu. Điều này làm thúc đẩy sự quan tâm nghiên cứu nhiều ứng dụng khác nhau như: ứng dụng phân tích dữ liệu người tiêu dùng trong các chiến dịch quảng cáo và tiếp thị; ứng dụng hệ tư vấn và hỗ trợ ra quyết định mua hàng; ứng dụng phân tích sản phẩm ưu tiên; ứng dụng chăn sóc khách hàng và giám sát danh tiếng; ứng dụng phân tích hành vi người tiêu dùng để dự đoán xu hướng tiêu dùng của xã hội... Trong các ứng dụng đó, bài toán phân tích quan điểm dựa trên khía cạnh đối với bài đánh giá sản phẩm đóng vai trò là đầu vào quan trọng. Chương 1 của luận án trình

bày các khái niệm cơ bản trong lĩnh vực nghiên cứu phân tích quan điểm nói chung và phân tích quan điểm mức khía cạnh nói riêng. Ba vấn đề chính được quan tâm trong bài toán phân tích quan điểm dựa trên khía cạnh bao gồm các nhiệm vụ chính trong phân tích quan điểm dựa trên khía cạnh, các phương pháp trích rút khía cạnh, các phương pháp phân lớp cảm xúc khía cạnh. Các nhiệm vụ này cũng chính là các nghiên cứu trọng tâm của luận án.

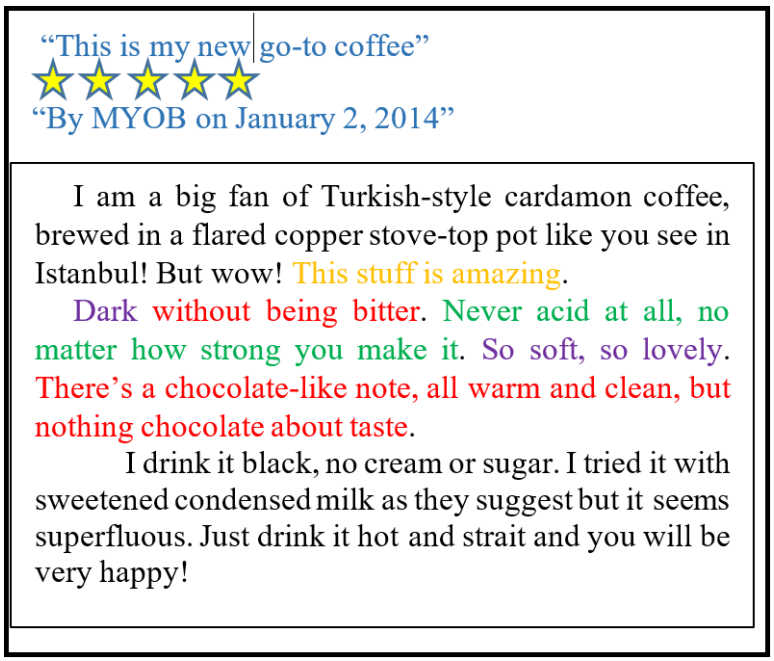
CHƯƠNG 2: PHÂN TÍCH QUAN ĐIỂM MỨC KHÍA CẠNH TRÊN CÁC BÀI ĐÁNH GIÁ SẢN PHẨM TRỰC TUYẾN

2.1 Đặt vấn đề

Trong những năm gần đây, dưới sự phát triển của công nghệ thông tin và các ứng dụng trong nhiều lĩnh vực khác nhau, đặc biệt trong thương mại điện tử, các tương tác xã hội trên trang thương mại điện tử hỗ trợ cho nhiều người dùng một phương thức bày tỏ quan điểm mới thông qua việc đánh giá sản phẩm. Những quan điểm hoặc đánh giá này thường đóng vai trò quan trọng trong việc cải thiện chất lượng sản phẩm, dịch vụ hoặc giúp các nhà quản lý hoạch định chiến lược phát triển của công ty, doanh nghiệp. Tuy nhiên, khi số lượng đánh giá được tạo ra bởi khách hàng quá lớn, không thể xử lý dễ dàng bằng các phương pháp thủ công, điều này đặt ra một thách thức về cách khai thác thông tin hiệu quả về sản phẩm, dịch vụ thông qua các công cụ, công nghệ hiện đại hơn bao gồm: tổng hợp quan điểm của người dùng [35–37], trích rút thông tin từ các bài đánh giá [16, 19, 38, 71, 73, 118], phân tích cảm xúc của người dùng [27–29, 32, 85, 94, 95], v.v. Trong chương này, luận án tập trung vào giải quyết vấn đề phân tích quan điểm của người dùng trên các bài đánh giá sản phẩm trực tuyến mức khía cạnh. Cụ thể hơn, chương này trình bày một hệ thống đầu cuối để giải quyết ba bài toán: trích rút khía cạnh được đề cập trong các bài đánh giá về một sản phẩm, suy ra điểm đánh giá của người dùng cho từng khía cạnh đã xác định và ước lượng trọng số đặt ra trên mỗi khía cạnh của người dùng.

Đánh giá của người dùng thường đề cập đến các khía cạnh khác nhau, đó là các thuộc tính hoặc thành phần của sản phẩm. Đối với mỗi một khía cạnh, người dùng thường đưa ra các quan điểm của họ thông qua việc thể hiện thái độ tích cực hoặc tiêu cực về khía cạnh đó. Giả sử có một đánh giá được tạo bởi người dùng như trong Hình 2.1, người dùng thích sản phẩm cà phê, được thể hiện bằng một điểm đánh giá tổng thể là năm sao. Tuy nhiên, những quan điểm tích cực về *thể hiện bên ngoài* (body), *vị* (taste), *mùi hương* (aroma) và *độ chua* (acidity) của cà phê cũng được đưa ra. Làm thế nào để hiểu nội dung bài đánh giá và các vấn đề mà người dùng đề cập? Nếu quan điểm của người dùng chỉ được xem xét trên tổng thể bài đánh giá thì chỉ thấy được một quan điểm tích cực đối với sản phẩm cà phê. Nếu quan điểm của người dùng được phân tích chi tiết trong từng câu của bài đánh giá thì quan điểm được nhìn thấy sẽ là quan điểm tích cực hoặc tiêu cực trên mỗi câu, mà không thấy được nhiều thông tin hơn trong toàn bộ bài đánh giá. Phân tích quan điểm dựa trên khía cạnh giải quyết vấn đề phân tích chi tiết trên những khía cạnh (*thể hiện bên ngoài, vị, hương thơm, độ chua*) của sản phẩm mà người dùng đã đề cập đến trong bài đánh giá của họ. Mức độ chi tiết là người dùng đã đề cập đến những khía cạnh nào trong bài đánh giá của họ,

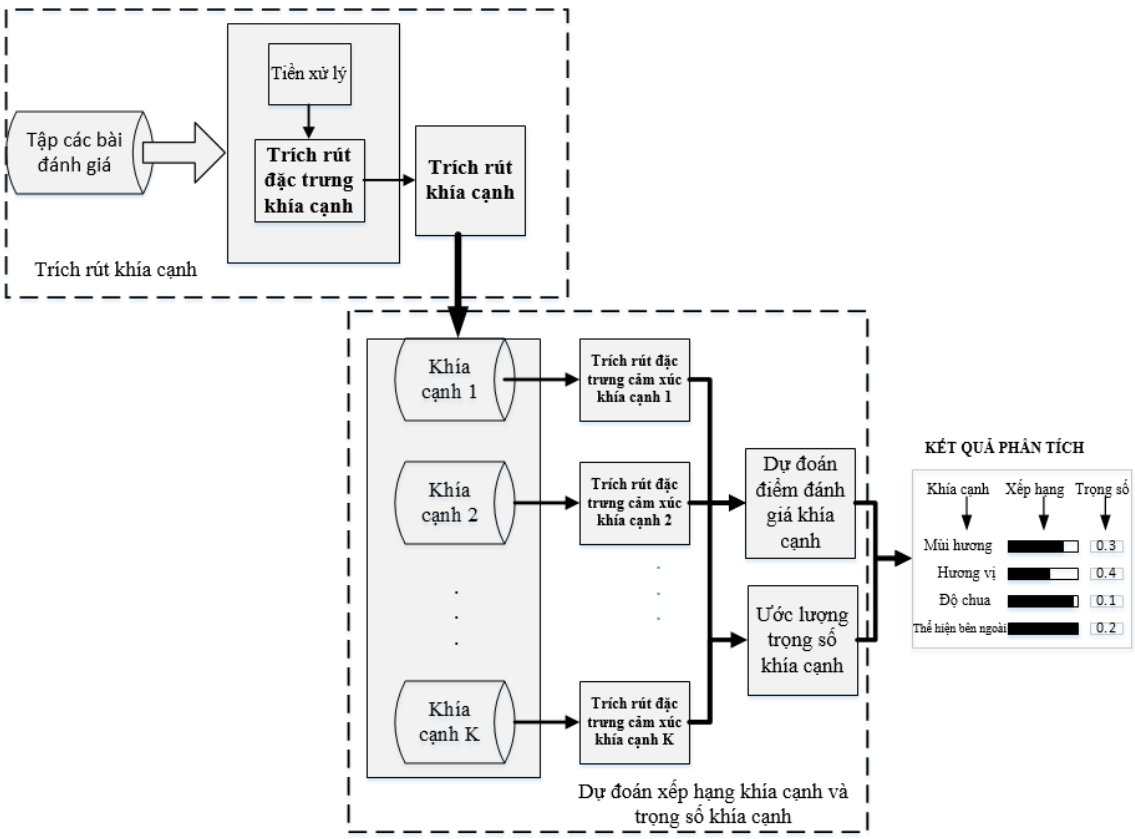
độ hài lòng hoặc quan điểm của khách hàng đối với mỗi khía cạnh đó, và sau cùng là mức độ quan tâm của mỗi khách hàng trên mỗi khía cạnh.



Hình 2.1: Một bài đánh giá về sản phẩm cà phê Trung Nguyên trên trang Amazone

Bài toán phân tích quan điểm mức khía cạnh bao gồm ba bài toán con là: (1) *Bài toán trích rút khía cạnh* tạo ra các phần (như từ trong câu hoặc câu trong bài đánh giá) đề cập đến một khía cạnh cụ thể của sản phẩm; (2) *Bài toán phân lớp cảm xúc khía cạnh* là thông qua đo lường biểu thị cảm xúc tích cực - tiêu cực hoặc dựa trên điểm đánh giá của người dùng đối với từng khía cạnh đã được trích rút trong bài toán (1); (3) *Bài toán xác định trọng số khía cạnh* là việc đánh giá mức độ quan tâm của người dùng đối với từng khía cạnh sản phẩm.

Luận án đề xuất một mô hình hệ thống nối tiếp gồm 2 modul (Hình 2.2), modul trích rút khía cạnh, modul dự đoán điểm đánh giá khía cạnh và ước lượng trọng số khía cạnh. Đối với modul trích rút khía cạnh, tập dữ liệu gồm các bài đánh giá được tiền xử lý, sau đó, các đặc trưng khía cạnh được lựa chọn. Một phương pháp trích rút khía cạnh sẽ xác định các thể hiện khía cạnh rõ ràng và thể hiện khía cạnh ẩn, từ đó câu sẽ được gán nhãn khía cạnh tương ứng. Đầu ra của modul này là tập các phần riêng lẻ của các câu đã được gán cùng nhãn khía cạnh. Đồng thời, đầu ra của modul trích rút khía cạnh là đầu vào của modul dự đoán điểm đánh giá khía cạnh và ước lượng trọng số khía cạnh. Hai vấn đề này được giải quyết một cách độc lập. Đầu ra của modul thứ hai là kết quả của hệ thống và có thể được sử dụng cho bài toán tổng hợp quan điểm mức khía cạnh.

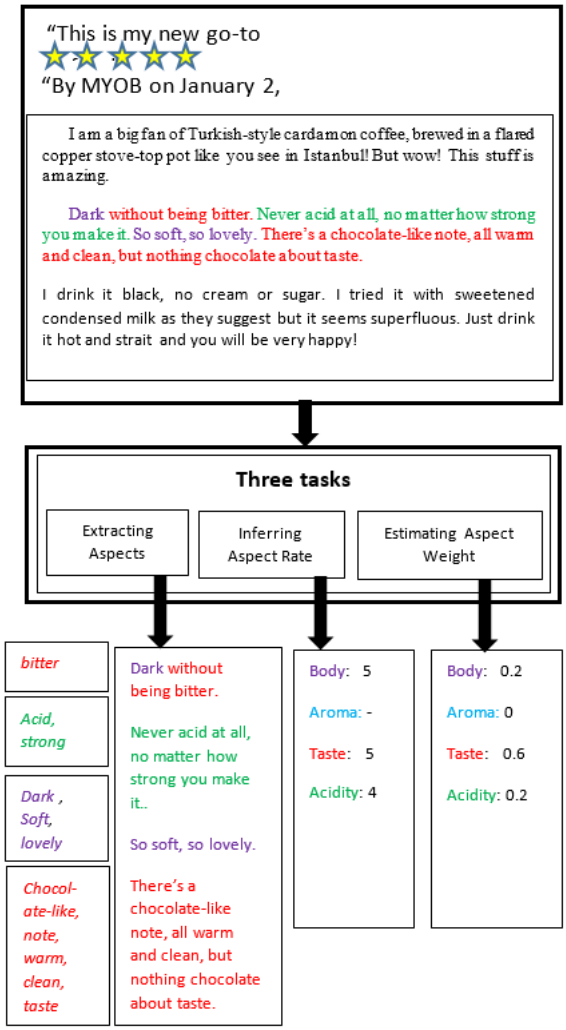


Hình 2.2: Mô hình hệ thống phân tích quan điểm mức khía cạnh các bài đánh giá sản phẩm trực tuyến

Bài toán trích rút khía cạnh là xác định tất cả các khía cạnh xuất hiện trong bài đánh giá. Một thách thức là, một số khía cạnh được đề cập rõ ràng và một số thì không. Ví dụ, trong bài đánh giá được đưa ra ở Hình 2.1, vị và độ chua của cà phê được đề cập rõ ràng (thông qua các danh từ "taste" và "acid"), nhưng thể hiện bên ngoài và hương thơm thì không được chỉ định rõ ràng. Một số các nghiên cứu trước đây chỉ đề cập đến việc xác định các khía cạnh rõ ràng [16, 19, 23, 26]. Phương pháp trích rút khía cạnh được đề xuất trong luận án sẽ xác định cả hai dạng khía cạnh rõ ràng và khía cạnh ẩn. Một khó khăn khác của trích rút khía cạnh là có thể tạo ra nhiều nhiễu (các phi khía cạnh). Làm thế nào để giảm thiểu nhiễu trong khi vẫn có thể xác định được các khía cạnh hiếm (tần suất xuất hiện thấp) và khía cạnh quan trọng cũng là một mối quan tâm khác của phương pháp đề xuất. Ngoài ra luận án cũng quan tâm đến vấn đề ngữ nghĩa của từ trong các ngữ cảnh khác nhau của chúng để cải thiện hơn nữa độ chính xác trích rút khía cạnh. Hướng tới giải quyết những vấn đề trên, trong modul trích rút khía cạnh (Hình 2.2) luận án đề xuất hai phương pháp trích rút khía cạnh là "trích rút khía cạnh sử dụng xác suất có điều kiện kết hợp kỹ thuật bootstraping" (Chương 2) và "trích rút khía cạnh dựa trên biểu diễn từ Word2vec và độ đo hỗ trợ" (Chương 3).

Sau khi các khía cạnh được xác định, việc suy ra điểm đánh giá của người dùng đối với các khía cạnh này sẽ cung cấp cái nhìn rõ hơn về mức độ hài lòng của người dùng đối với chúng. Trong nhiều ứng dụng thực tế, đặc biệt là trong các bài đánh giá sản phẩm trực tuyến, điểm đánh giá rời rạc (thang từ 1 đến 10 hoặc từ 1 đến 5) thường được sử dụng để thể hiện mức độ hài lòng của người dùng. Do vậy, bài toán phân lớp cảm xúc khía cạnh có thể được xem như bài toán phân loại đa lớp. Những vấn đề khó khăn của phân loại đa lớp như phân loại chính xác các lớp liền kề hoặc lân cận nhau, hay vấn đề mất cân bằng giữa các lớp trong dữ liệu đa lớp đều được luận án quan tâm và giải quyết. Trong modul dự đoán xếp hạng khía cạnh và trọng số khía cạnh (Hình 2.2), bộ phân lớp Naive Bayes (Chương 2) được sử dụng để giải quyết vấn đề phân lớp cảm xúc khía cạnh sau khi đã được trích rút. Mặc dù phương pháp này khá đơn giản, tuy vậy, nó đã được chứng minh là khá hiệu quả qua nhiều nghiên cứu và các hệ thống ứng dụng được công khai [3, 4, 7]. Tuy vậy, để nâng cao hơn nữa độ chính xác trong việc xác định điểm đánh giá khía cạnh, luận án cũng đã đề xuất một phương pháp phân lớp cảm xúc bằng cách kết hợp các bộ phân loại cơ sở (SVM và mạng Bayesian cổng OR) dựa trên luật kêt hợp Dempster cho bài toán phân loại đa lớp (Chương 4).

Các khía cạnh có chứa cảm xúc cũng có thể được người dùng kết hợp với các trọng số khác nhau để thể hiện tầm quan trọng tương đối của các khía cạnh. Người dùng thường đưa ra đánh giá tổng thể về ấn tượng chung đối với một sản phẩm. Đánh giá tổng thể không phải lúc nào cũng đủ thông tin. Tuy nhiên có thể giả định rằng điểm đánh giá tổng thể trên một sản phẩm là tổng trọng số của điểm đánh giá mà người dùng đưa ra trên nhiều khía cạnh của sản phẩm, trong đó, trọng số về cơ bản đo lường mức độ quan trọng của các khía cạnh. Một số nghiên cứu trước đây [35, 36] thực hiện suy ra điểm đánh giá khía cạnh và trọng số khía cạnh đồng thời dựa trên phương pháp hồi quy bằng cách sử dụng cả nội dung bài đánh giá và điểm đánh giá tổng thể. Luận án đề xuất một cách tiếp cận khác (Chương 2), cụ thể, trọng số của khía cạnh được tính bằng cách sử dụng tần suất của từ khía cạnh trong bài đánh giá và tính nhất quán của khía cạnh trên tất cả các bài đánh giá. Cách tiếp cận này không yêu cầu bất cứ thông tin nào về điểm đánh giá tổng thể hay điểm đánh giá khía cạnh từ người dùng. Hình 2.3 mô tả chi tiết ba bài toán nhỏ của bài toán phân tích quan điểm mức khía cạnh đối với các bài nhận xét sản phẩm trực tuyến.



Hình 2.3: Các bài toán con của bài toán phân tích quan điểm dựa trên khía cạnh

2.2 Các nghiên cứu liên quan

2.2.1 Trích rút khía cạnh

Trích rút khía cạnh từ các bài đánh giá trực tuyến là nhiệm vụ chính trong phân tích quan điểm mức khía cạnh. Các kỹ thuật sớm nhất được đề xuất cho trích rút khía cạnh là các kỹ thuật không giám sát như trong [11, 44, 46, 60, 62, 119]. Cách tiếp cận này đã được các nhà nghiên cứu sử dụng rộng rãi để trích rút các khía cạnh từ các bài đánh giá trực tuyến. Các phương pháp dựa trên tần suất [11, 44, 46, 60] xem xét các danh từ hoặc cụm danh từ có tần xuất cao là các ứng viên. Tuy nhiên, các phương pháp tiếp cận dựa trên tần suất có thể bỏ qua các khía cạnh có tần suất thấp. Một số cách tiếp cận dựa trên kinh nghiệm, luật, bộ lọc phức tạp được sử dụng để khắc phục vấn đề này, song kết quả không được như mong đợi vì một số khía cạnh vẫn bị bỏ sót [60–62]. Hơn nữa, các phương pháp này gặp khó khăn trong việc xác định các khía

cạnh ẩn. Điểm nổi bật của cách tiếp cận không giám sát là không yêu cầu bất kỳ một sự gán nhãn nào từ dữ liệu, đơn giản, ít chi phí. Tuy nhiên, độ chính xác trích rút chưa thực sự cao [4].

Các kỹ thuật có giám sát như trường ngẫu nhiên có điều kiện (Conditional Random Field - CRF) [19–21], bộ nhớ ngắn hạn dài hạn (long short term memory-LSTM) [120, 121] được nghiên cứu để nâng cao hiệu quả trong vấn đề trích rút khía cạnh rõ ràng. Mặc dù vậy, các kỹ thuật này lại đòi hỏi cần có một tập dữ liệu được gán nhãn thủ công để huấn luyện mô hình và do đó có thể tốn kém chi phí. Mặt khác, các kỹ thuật học sâu (LSTM) có thể bị giảm hiệu suất khi khám phá dữ liệu có nội dung ẩn hoặc câu dài do thiếu bộ nhớ [122].

Để có thể giảm chi phí do việc gán nhãn dữ liệu cho các mô hình học giám sát, đồng thời tăng hiệu quả trích rút khía cạnh, các kỹ thuật học bán giám sát cũng đã được nghiên cứu rộng rãi. Các kỹ thuật phổ biến trong cách tiếp cận này là mạng nơ ron hồi quy (Recurrent Neural Network - RNN) [24, 25], dựa trên ngữ nghĩa (semantic based) [18], dựa trên từ vựng (lexicon-based) [64]. Đặc biệt, phương pháp sử dụng đặc trưng ngôn ngữ kết hợp giải thuật bootstrap cũng đã có nhiều đề xuất [17, 119, 123, 124].

Nghiên cứu của Li và các cộng sự [119] đã đề xuất một mô hình hai bước để trích rút khía cạnh và các từ quan điểm liên quan. Ban đầu các cụm từ mô tả khía cạnh sản phẩm và cảm xúc liên quan được trích rút thông qua kỹ thuật bootstrap. Một tiêu chí tổng hợp dựa trên hai phép đo là mức độ phổ biến và độ tin cậy được sử dụng để đánh giá các mẫu và các đặc trưng. Sau đó các đặc trưng được nhóm lại thành các khía cạnh dựa trên độ tương đồng trong mạng WordNet. Qiyun Zhao và các cộng sự [17] đã đưa ra một phương pháp dựa trên cây cú pháp phụ thuộc kết hợp từ lõi quan điểm và giải thuật bootstrap để khám phá mối quan hệ giữa các từ quan điểm và khía cạnh. Đầu tiên, từ lõi quan điểm, cú pháp phụ thuộc và văn bản đánh giá là các đầu vào. Sau đó trình phân tích cú pháp giúp nắm bắt cấu trúc phụ thuộc bên trong mỗi câu. Hai nhóm từ quan điểm và mục tiêu ứng viên của quan điểm (khía cạnh) được tạo bằng cách sử dụng các quy tắc xác định trước. Sau đó, họ trích rút lặp đi lặp lại các từ quan điểm và mục tiêu dựa trên các quy tắc được xác định và tập kết quả hiện có. Trong quá trình trích rút, một đồ thị liên kết giữa các từ quan điểm và từ mục tiêu quan điểm được xây dựng để xác định trọng số của các mối liên quan này (thể hiện bằng các cạnh của đồ thị). Từ đồ thị quan hệ giữa từ quan điểm và mục tiêu quan điểm, các quy tắc cắt tỉa được áp dụng để có được các kết quả tốt nhất.

2.2.2 Phân lớp cảm xúc

Hầu hết các nghiên cứu trước đây liên quan đến phân loại cảm xúc đều tập trung vào phân loại nhị phân (tích cực và tiêu cực) [9, 27, 28, 32, 33, 85, 94] và ba lớp (tích

cực, tiêu cực, trung lập) [29, 86, 125]. Một số nghiên cứu đã cố gắng chia nhỏ các bài đánh giá trực tuyến thành các lớp nhỏ hơn (ví dụ: rất tích cực, tích cực, trung lập, tiêu cực, rất tiêu cực) để mô tả cường độ của cảm xúc [36, 126]. Một số nghiên cứu khác giải quyết nhiệm vụ phân loại nhiều lớp như là bài toán phân loại các loại cảm xúc khác nhau (ví dụ: vui, tức giận, buồn, ngạc nhiên) [41, 127, 128]. Mặc dù những công việc này cũng xử lý nhiều lớp, nhưng chúng khác với nhiệm vụ được đề cập trong luận án này như nghiên cứu sinh đã miêu tả trước đó.

Các phương pháp phân loại truyền thống để phân loại cảm xúc bao gồm các phương pháp tuyến tính [27, 28, 30], các phương pháp dựa trên xác suất [31, 32, 86], các phương pháp dựa trên luật [9], và các phương pháp cây quyết định [33, 129]. Một cuộc khảo sát của Hemmatian [6] và Ravi [4] cho thấy NB và SVM đã thu hút được sự chú ý của các nhà nghiên cứu do có nhiều ưu điểm so với các phương pháp khác. Ngoài NB và SVM, mạng nơ-ron cũng là cách tiếp cận máy học phổ biến để phân loại cảm xúc [4]. Mô hình học sâu sử dụng mạng nơ-ron hồi quy (RNNs) và mạng nơ-ron tích chập (CNNs) đã được áp dụng thành công trong bài toán phân lớp cảm xúc [85, 94, 95]. Mặc dù các thuật toán học sâu có độ chính xác vượt trội so với các thuật toán học máy truyền thống, nhưng chúng yêu cầu lượng dữ liệu đủ lớn để hoạt động tốt. Ngoài ra, học sâu cần sử dụng CPU nhiều hơn khoảng 10 lần so với học máy truyền thống [130].

2.2.3 Trọng số khía cạnh

Đánh giá khía cạnh là sự đo lường mức độ hài lòng của người dùng về các khía cạnh của sản phẩm. Trong khi đó trọng số của khía cạnh đo lường mức độ quan trọng của khía cạnh do người dùng đặt ra. Cho đến nay chỉ có một số nghiên cứu được thực hiện để khám phá các trọng số mà người đánh giá đặt ra trên các khía cạnh [35]. Trong [131], mô hình hồi quy hedonic được sử dụng để xác định trọng số của từng khía cạnh thông qua việc sử dụng nhu cầu sản phẩm như một hàm mục tiêu. Nhưng trọng số thu được là chung cho tất cả các bài đánh giá mà không tính đến sở thích cá nhân của từng người dùng. Trong [63], các tác giả sử dụng mô hình hồi quy xác suất (Probabilistic Regression Model - PRM) để ước lượng trọng số khía cạnh. Nghiên cứu này giả định ước lượng tổng thể được rút ra từ phân bố Gaussian với giá trị trung bình là tích số của điểm đánh giá khía cạnh và trọng số khía cạnh. Đối với mỗi bài đánh giá, với các điểm đánh giá khía cạnh đã biết, trọng số khía cạnh với xác suất hậu nghiệm có khả năng xảy ra nhất được suy ra với tần suất xuất hiện là kiến thức tiên nghiệm. Trong [36, 37], PRM cũng được sử dụng để ước tính trọng số khía cạnh. Trong nghiên cứu này, một mô hình đồ thị xác suất được giới thiệu để ước tính đồng thời điểm đánh giá khía cạnh và trọng số của mỗi khía cạnh. Wang và các cộng sự [35] đề xuất mô hình

hồi quy lớp ẩn (Latent Class Regression Model - LCRM) trong một mô hình đồ thị xác suất để đồng thời thực hiện cả hai nhiệm vụ dự đoán điểm đánh giá khía cạnh và ước lượng trọng số khía cạnh.

2.3 Các khái niệm cơ bản trong bài toán phân tích quan điểm mức khía cạnh

Bài đánh giá của người dùng i về một sản phẩm được ký hiệu di. Thông thường, mỗi bài đánh giá có thể chứa nhiều câu. Trong đó, mỗi câu chứa nhiều từ w_j trong tập hợp phổ quát của tất cả các từ có thể có.

Định nghĩa 2.1 Tập các bài đánh giá (Review Text Documents): Tập các bài đánh giá được ký hiệu là $\mathcal{D} = \{d_1, d_2, ..., d_D\}$ là một tập các bài viết nhận xét về một loại sản phẩm, D là số bài đánh giá trong toàn bộ dữ liệu.

Định nghĩa 2.2 Từ điển (Vocabulary): Giả sử rằng có V các từ được tách ra từ tập các bài đánh giá $\mathcal{D}$. Tập các từ này được gọi là từ điển $\mathcal{V} = \{w_j | j = \overline{1,V}\}$.

Định nghĩa 2.3 Khía cạnh (Aspect): Khía cạnh là một đặc điểm (một thuộc tính hoặc một thành phần) của sản phẩm. Ví dụ, "*taste*", "*aroma*" và "*body*" là một số khía cạnh có thể có của sản phẩm "*coffee*". Giả định rằng có K khía cạnh được đề cập trong tất cả các bài đánh giá, được ký hiệu là $\mathcal{A} = \{a_k | k = \overline{1,K}\}$. Một khía cạnh a_k được biểu diễn bằng một tập hợp các từ và được ký hiệu là $a_k = \{w | w \in \mathcal{V}, A(w) = a_k\}$, trong đó a_k là tên của khía cạnh, w là một từ trong tập hợp $\mathcal{V}$ và $A(.)$ là một toán tử ánh xạ một từ tới một khía cạnh.

Ví dụ: những từ như "*taste*", "*aftertaste*" và "*mouth feel*" có thể đặc trưng cho khía cạnh "*taste*" của sản phẩm "*coffee*".

Định nghĩa 2.4 Từ lõi khía cạnh (Aspect Core Words): Cho một khía cạnh a_k, một tập hợp rất ít các từ có trong từ điển miêu tả rất rõ ràng khía cạnh a_k được gọi là từ lõi khía cạnh, ký hiệu là $\mathcal{C}_k = \{w_{kj} \in \mathcal{V} | w_{kj} \longrightarrow a_k, j = \overline{1,N}\}$, trong đó w_{kj} là từ mô tả khía cạnh a_k, N là số từ lõi của khía cạnh a_k. Tập từ lõi khía cạnh này không giao thoa sang tập từ lõi khía cạnh khác. Các từ lõi có thể được cung cấp bởi người dùng hoặc bởi một số chuyên gia lĩnh vực.

Định nghĩa 2.5 Từ khía cạnh (Aspect Words): Tập tất cả các từ có trong từ điển $\mathcal{V}$ mà chúng có thể mô tả về khía cạnh a_k (các từ này khác với các từ lõi khía cạnh) được gọi là các từ khía cạnh, ký hiệu là $\mathcal{T}_k = \{w_{kj} \in \mathcal{V}, w_{kj} \notin \mathcal{C}_k | w_{kj} \longrightarrow a_k, j = \overline{1,M}\}$. Các từ khía cạnh không thuộc tập các từ lõi khía cạnh. M là số từ khía cạnh của khía cạnh a_k.

Định nghĩa 2.6 Điểm đánh giá khía cạnh (Aspect Rating): Cho một văn bản đánh giá của người dùng d_i, một vector K chiều $\mathbf{r}_i \in \mathbb{R}^K$ được sử dụng để biểu diễn điểm đánh giá của K khía cạnh trong văn bản đánh giá d_i, ký hiệu là $\mathbf{r}_i = \{r_{i_1}, r_{i_2}, ..., r_{i_K}\}$,

trong đó r_{i_k} là một giá trị số cho biết đánh giá của người dùng về khía cạnh a_k, và $r_{i_k} \in [r_{min}, r_{max}]$ (ví dụ dải phạm vi của r_{i_k} có thể thuộc từ 1 đến 5).

Định nghĩa 2.7 Trọng số khía cạnh (Aspect Weight): Trọng số khía cạnh biểu hiện sự quan tâm của người dùng đối với một hoặc một vài khía cạnh cụ thể của sản phẩm. Cho một văn bản đánh giá của người dùng d_i, một vector K chiều $\alpha_i \in \mathbb{R}^K$ được sử dụng để biểu diễn mức độ quan tâm của người dùng đối với K khía cạnh trong văn bản đánh giá d_i, ký hiệu là $\alpha_i = \{\alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_K}\}$, trong đó α_{i_k} là một giá trị số cho biết trọng số đánh giá của người dùng về khía cạnh a_k, và $\alpha_{i_k} \in [0,1]$, và $\sum_{k=1}^{K} \alpha_{i_k} = 1$. Trọng số một khía cạnh cao hơn đồng nghĩa với việc người dùng nhấn mạnh nhiều hơn vào khía cạnh tương ứng.

Định nghĩa 2.8 Điểm đánh giá tổng thể của bài đánh giá (Review overall Rating): Cho một văn bản đánh giá của người dùng d_i, một giá trị số $y_i \in \mathbb{R}^+$ biểu diễn điểm đánh giá tổng thể của người dùng về một sản phẩm trên tất cả các khía cạnh của sản phẩm. Giá trị điểm đánh giá tổng thể này tương tự như điểm đánh giá khía cạnh.

Bài toán trích rút khía cạnh: Mục tiêu của nhiệm vụ này là trích rút các khía cạnh được đề cập trong một bài đánh giá. Giả định rằng mỗi khía cạnh là một phân phối xác suất trên tất cả các từ. Đồng thời cũng giả định rằng mỗi câu trong văn bản của bài đánh giá có thể đề cập đến nhiều khía cạnh. Do đó, phương pháp đề xuất trong chương này trích rút các khía cạnh dựa trên xác suất có điều kiện của các từ sao cho mỗi câu có thể gán nhiều nhãn.

Bài toán dự đoán điểm đánh giá khía cạnh: Nhiệm vụ này là suy ra vector $\mathbf{r}_i$ của điểm xếp hạng khía cạnh (được xác định trong định nghĩa 2.6) cho một bài đánh giá d_i. Điểm đánh giá của một khía cạnh phản ánh cảm xúc của người dùng về khía cạnh đó thường được thể hiện bằng các từ cảm xúc (tích cực hoặc tiêu cực). Người dùng càng sử dụng nhiều từ ngữ mang tính tích cực, thì điểm đánh giá họ muốn đặt ra trên khía cạnh đó càng cao. Trong luận án, áp dụng phương pháp học có giám sát, phương pháp Naive Bayes, để học điểm đánh giá khía cạnh trong đó các từ cảm xúc được xem xét như là các đặc trưng đầu vào.

Bài toán ước lượng trọng số khía cạnh: Nhiệm vụ này là ước tính trọng số không âm α_i mà người dùng đặ trên khía cạnh a_{ik} của văn bản d_i (được xác định trong Định nghĩa 2.7). Về cơ bản, trọng số của một khía cạnh đo lường mức độ quan trọng được đưa ra bởi người dùng đối với khía cạnh đó. Người ta quan sát thấy rằng, mọi người thường nói nhiều hơn về các khía cạnh mà họ quan tâm trong cùng một bài đánh giá. Bên cạnh đó, có quan điểm cho rằng một khía cạnh là quan trọng thường được nhiều người khác chia sẻ. Dựa trên những quan sát này, một công thức được đưa ra để tính toán trọng số khía cạnh. Công thức tính đến số lần xuất hiện của các từ thảo luận về khía cạnh trong một bài đánh giá và tần suất của các câu văn bản thảo luận về cùng

một khía cạnh trên tất cả các bài đánh giá.

Các ký hiệu chính được sử dụng trong hệ thống phân tích quan điểm mức khía cạnh của luận án được trình bày trong Bảng 2.1.

Bảng 2.1: Các ký hiệu sử dụng trong phân tích quan điểm mức khía cạnh

Ký hiệu	**Mô tả**
$\mathcal{D} = \{d_i \mid i = \overline{1,D}\}$	Tập các văn bản đánh giá của một sản phẩm, D là số bài đánh giá
$\mathcal{V} = \{w_j \mid j = \overline{1,V}\}$	Tập từ điển của một sản phẩm, V là số từ có trong từ điển
$\mathcal{A} = \{a_k \mid k = \overline{1,K}\}$	Tập các khía cạnh của một sản phẩm, K là số các khía cạnh
$\mathcal{C}_k = \{w_{kj} \in \mathcal{V} \mid w_{kj} \longrightarrow a_k,$ $j = \overline{1,N}\}$	Tập từ lõi khía cạnh của một sản phẩm, N là số từ lõi của một khía cạnh a_k
$\mathcal{T}_k = \{w_{kj} \in \mathcal{V}, w_{kj} \notin \mathcal{C}_k \mid$ $w_{kj} \longrightarrow a_k, j = \overline{1,M}\}.$	Tập từ khía cạnh của một sản phẩm, M là số từ có trong tập từ khía cạnh a_k
$\mathbf{r}_i \in \mathbb{R}^K$	Điểm đánh giá khía cạnh của người dùng thông qua bài đánh giá d_i, $\mathbf{r}_i = \{r_{i_1}, r_{i_2}, ..., r_{i_K}\}$
$\alpha_i \in \mathbb{R}^K$	Trọng số khía cạnh, $\alpha_i = \{\alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_K}\}$
$y_i \in \mathbb{R}^+$	Điểm đánh giá tổng thể của sản phẩm mà người dùng thể hiện thông qua bài đánh giá d_i
r_{i_k}	Điểm đánh giá khía cạnh thứ k của bài đánh giá d_i, $r_{i_k} \in [1,5]$
α_{i_k}	Trọng số khía cạnh thứ k của bài đánh giá d_i, $\alpha_{i_k} \in [0,1]$

2.4 Hệ thống phân tích quan điểm mức khía cạnh các bài đánh giá sản phẩm trực tuyến

2.4.1 Trích rút khía cạnh sử dụng xác suất có điều kiện kết hợp kỹ thuật Bootstraping

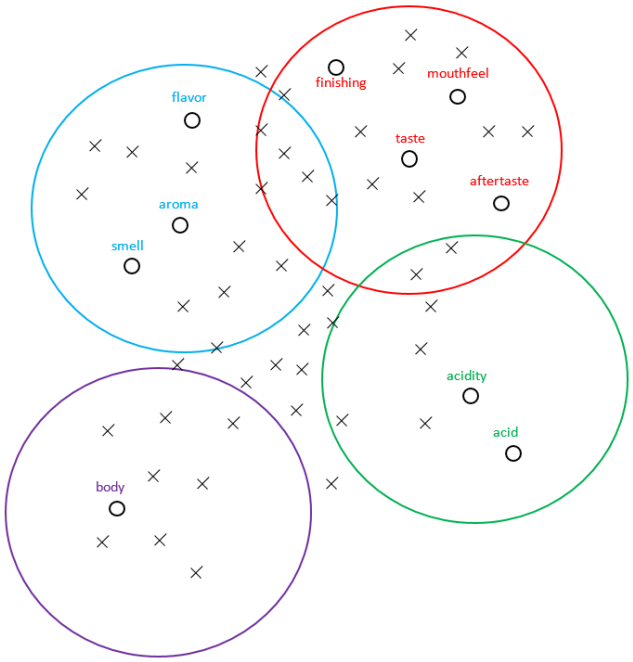
Hầu hết dữ liệu thực tế (các bài đánh giá sản phẩm trực tuyến) đều không được gán nhãn khía cạnh, do vậy luận án đã lựa chọn hướng tiếp cận của bài toán là các phương pháp dựa trên học bán giám sát. Cụ thể, luận án đã sử dụng giả định rằng, một tập hợp phổ quát của tất cả các khía cạnh có thể có cho mỗi sản phẩm đều biết trước cùng với

các từ khía cạnh được gọi là *từ lõi khía cạnh* (thuật ngữ mô tả chính xác khía cạnh). Giả định này là thực tế và chi phí thấp. Tập dữ liệu không cần phải gán nhãn đầy đủ. Việc xác định sẵn tập các khía cạnh và một số từ lõi điển hình liên quan đến các khía cạnh khá đơn giản vì có thể bổ sung thêm thông tin từ nhà cung cấp sản phẩm hoặc các chuyên gia miền lĩnh vực [5]. Sau đó, bài toán trích rút khía cạnh là ánh xạ các câu hoặc cụm từ trong bài đánh giá thành các tập hợp con tương ứng với từng khía cạnh. Thách thức chính ở đây là trong nhiều bài đánh giá, các câu không chứa đủ các từ lõi khía cạnh hoặc thậm chí không có bất kỳ từ lõi khía cạnh nào. Điều này dẫn đến việc gán nhãn cho các câu không chính xác. Do đó, cần phải mở rộng các từ lõi khía cạnh thành tập các *từ khía cạnh* phong phú hơn dựa trên dữ liệu đã biết (các bài đánh giá).

Trong một số phương pháp bán giám sát, tập hợp các từ khía cạnh được xây dựng dựa trên tần suất, bộ lọc theo cú pháp phụ thuộc, mô hình Bayes hoặc mô hình Markov ẩn. Các phương pháp này thường gây ra các hiện tượng bỏ qua các khía cạnh có tần xuất thấp hoặc tạo ra các nhiễu (phi khía cạnh) và gặp khó khăn trong việc xác định các khía cạnh ẩn. Để vượt qua các vấn đề này, luận án đề xuất một phương pháp sử dụng mô hình xác suất có điều kiện kết hợp kỹ thuật Bootstrap để trích rút các từ khía cạnh. Việc sử dụng xác suất có điều kiện giúp phát hiện các khía cạnh có tần suất thấp. Kỹ thuật Bootstrap giúp việc mở rộng các từ lõi khía cạnh và từ khía cạnh trở nên dễ dàng hơn với chi phí thấp. Thêm vào đó, để loại bỏ các nhiễu (các từ phi khía cạnh), luận án đã sử dụng đặc trưng TF-IDF hỗ trợ lựa chọn ra các từ có trọng số quan trọng đối với các bài nhận xét và trong toàn tập dữ liệu. Đặc trưng POS lựa chọn ra không chỉ các từ là danh từ và cụm danh từ mà còn lựa chọn các từ là tính từ, trạng từ, động từ giúp cho các khía cạnh ẩn không bị bỏ qua.

Hình 2.4 minh họa bốn khía cạnh của một sản phẩm cà phê được thể hiện bằng các từ khía cạnh tương ứng của chúng, trong đó ký hiệu O thể hiện các từ lõi khía cạnh, biểu tượng X đại diện cho các từ xuất hiện trong dữ liệu. Đối với sản phẩm cà phê này, bốn khía cạnh "*body*", "*taste*", "*aroma*" và "*acidity*" đã được biết đến. Tập hợp các từ lõi khía cạnh tương ứng với các khía cạnh này lần lượt là {body}, {taste, aftertaste, finishing, mouthfeel}, {aroma, smell, flavor} và {acid, acidity}. Các từ lõi khía cạnh sau đó được mở rộng bằng cách bổ sung các từ có xác suất xuất hiện cao trong cùng một câu mà các từ lõi đó xuất hiện. Tập hợp các từ khía cạnh được đại diện bởi bốn vòng tròn. Các vòng tròn này có thể chồng lên nhau, điều này cho biết rằng một số từ khía cạnh có thể thuộc về các khía cạnh khác nhau.

Giả sử rằng $\mathcal{A} = \{a_1, a_2, ..., a_K\}$ là tập K khía cạnh của sản phẩm. a_k là tập các từ đại diện cho khía cạnh a_k và tần suất xuất hiện của chúng luôn lớn hơn ngưỡng θ (θ được xác định trong quá trình thực nghiệm). Mục tiêu là thu thập tập các từ mà chúng



Hình 2.4: Từ lõi với các khía cạnh

xuất hiện trong các câu của toàn bộ ngữ liệu thuộc về khía cạnh a_k. Tập hợp các từ của hai khía cạnh có thể trùng nhau, do đó một số thuật ngữ có thể thuộc về nhiều khía cạnh. Đầu tiên, các câu có chứa ít nhất một từ trong tập từ lõi khía cạnh $\mathcal{C}_k$ ban đầu của khía cạnh được định vị (gán nhãn). Sau đó, tất cả các từ bao gồm danh từ, cụm danh từ, tính từ, trạng từ và động từ xuất hiện trong những câu này được tìm kiếm. Các từ xuất hiện nhiều hơn một ngưỡng θ nhất định được bổ sung vào tập hợp các từ khía cạnh. Các từ có số lần xuất hiện lớn nhất trong tập hợp các từ khía cạnh mới tìm được sẽ được thêm vào tập hợp các từ lõi khía cạnh. Tập hợp các từ khía cạnh và các từ lõi khía cạnh được cập nhật, các từ này được sử dụng để tìm các câu được gán nhãn tiếp theo. Quá trình nói trên được lặp lại cho đến khi không tìm thấy thêm từ mới.

Thuật toán trích rút khía cạnh được đề xuất hoạt động như sau: Trước hết, văn bản của bài đánh giá được chia thành các câu (bước 2). Sau đó, các nhãn khía cạnh từ tập $\mathcal{A}$ của tất cả các nhãn được gán cho mọi câu của tập văn bản bài đánh giá $\mathcal{D}$ dựa trên các từ lõi $\mathcal{C}_k$ ban đầu về khía cạnh (bước 3). Dựa trên các câu được nhãn khía cạnh ban đầu này, tập các từ lõi khía cạnh $\mathcal{C}_k$ và tập từ khía cạnh $\mathcal{T}_k$ cho mọi khía cạnh được cập nhật. Các từ trong tập $\mathcal{T}_k$ có số lần xuất hiện lớn nhất sẽ được đưa vào tập $\mathcal{C}_k$, các từ trong $\mathcal{V}$ mà chúng không thuộc $\mathcal{C}_k$ có xác suất điều kiện khi biết $\mathcal{S}_k$ lớn hơn ngưỡng θ được cập nhật vào tập $\mathcal{T}_k$ (bước 4). Các nhãn cho tất cả các câu được cập nhật bằng cách sử dụng tập từ lõi khía cạnh và tập từ khía cạnh mới (bước 5). Bước 4 và bước 5 được lặp lại cho đến khi không tìm thấy tập từ khía cạnh mới nào nữa hoặc số lần lặp lại vượt quá ngưỡng nhất định.

Algorithm 1 Thuật toán trích rút khía cạnh

Input:

$\mathcal{D}$: tập văn bản các bài nhận xét $\mathcal{D} = \{d_1, d_2, ..., d_D\}$;

$\mathcal{A}$: tập các khía cạnh $\mathcal{A} = \{a_1, a_2, ..., a_K\}$;

$\mathcal{C}$: tập các từ lõi khía cạnh $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$, $\mathcal{C}_k$ là tập các từ lõi của khía cạnh a_k;

$\mathcal{V}$: tập từ điển (kho tất cả các từ trong toàn bộ ngữ liệu) $\mathcal{V} = \{w_1, w_2, ..., w_P\}$;

$Loop$: số lần lặp;

θ: ngưỡng xác suất có điều kiện.

Output:

$\mathcal{S}_k$: tập các câu được gán nhãn khía cạnh a_k, với $k = \overline{1,K}$.

▷ **Bước 1**: Khởi tạo các biến

$\mathcal{S} = \emptyset$; ▷ $\mathcal{S}$ là tập tất cả các câu trong kho ngữ liệu

for $k = 1$ to K **do** $\mathcal{S}_k = \emptyset$;

end for ▷ $\mathcal{S}_k$ là tập tất cả các câu gán nhãn a_k với $k = \overline{1,K}$

for $k = 1$ to K **do** $\mathcal{T}_k = \emptyset$;

end for ▷ $\mathcal{T}_k$ là tập các từ khía cạnh của khía cạnh a_k, với $k = \overline{1,K}$

$Loop_{Execution} = 0$; ▷ Số lần lặp thực hiện khởi tạo ban đầu là 0

▷ **Bước 2**: Tách câu

for mỗi văn bản $d_i \in \mathcal{D}$ **do** $\mathcal{S} \longleftarrow Segment(d_i)$

end for ▷ Tách mỗi văn bản trong tập dữ liệu thành các câu lưu trong $\mathcal{S}$

▷ **Bước 3**: Gán tất cả các nhãn khía cạnh a_k tới câu s_h nếu câu chứa ít nhất một từ lõi trong $\mathcal{C}_k$ của $\mathcal{C}$

for mỗi câu $s_h \in \mathcal{S}$ **do**

  for $k = 1$ to K **do**

    for mỗi từ $w_{hj} \in s_h$ **do** // w_{hj} là từ thứ j trong câu s_h

      if $w_{hj} \in \mathcal{C}_k$ **then** $\{s_h \longleftarrow label(a_k); \mathcal{S}_k \longleftarrow s_h\}$

      end if

    end for

  end for

end for

repeat

▷ **Bước 4**: Cập nhật tập từ lõi khía cạnh và tập từ khía cạnh

  for $k = 1$ to K **do Cập_nhật** $(\mathcal{C}_k, \mathcal{T}_k)$;

  end for

  $Loop_{Execution}++$;

Algorithm 1 Thuật toán trích rút khía cạnh (continued)

▷ **Bước 5**: Gán lại nhãn khía cạnh a_k cho câu

for mỗi câu $s_h \in \mathcal{S}$ **do**
  for $k = 1$ to K **do**
    $m(k) = 0$; ▷ $m(k)$: số từ khía cạnh của khía cạnh a_k có trong câu s_h
    for mỗi từ $w_{hj} \in s_h$ **do**
      if $w_{hj} \in \mathcal{C}_k$ **then** $\{s_h \longleftarrow label(a_k); \mathcal{S}_k \longleftarrow s_h\}$
      else
        if $w_{hj} \in \mathcal{T}_k$ **then** $m(k) = m(k) + 1$;
        end if
      end if
    end for
    if $m(k) = argmax(m(k))$ **then** $\{s_h \longleftarrow label(a_k); \mathcal{S}_k \longleftarrow s_h\}$
    end if
  end for
end for
until $(\mathcal{C}_k, \mathcal{T}_k)$ không thay đổi $\| \; Loop_{Execution} >= Loop$

Algorithm 2 Thủ tục cập nhật từ lõi khía cạnh và từ khía cạnh

Input:

$\mathcal{V}$: tập từ điển (kho tất cả các từ trong toàn bộ ngữ liệu) $\mathcal{V} = \{w_1, w_2, ..., w_P\}$;

$\mathcal{C}_k$: tập các từ lõi được gán nhãn khía cạnh a_k

$\mathcal{T}_k$: tập các từ khía cạnh được gán nhãn khía cạnh a_k.

$\mathcal{S}_k$: tập các câu chứa khía cạnh a_k

θ: ngưỡng xác suất có điều kiện để mở rộng tập từ khía cạnh $0 \leq \theta \leq 1$.

Output:

$\mathcal{C}_k$: tập các từ lõi được gán nhãn khía cạnh a_k được cập nhật mới.

$\mathcal{T}_k$: tập các từ khía cạnh được gán nhãn khía cạnh a_k được cập nhật mới.

procedure CẬP NHẬT$(\mathcal{C}_k, \mathcal{T}_k)$ ▷ Cập nhật tập từ lõi khía cạnh, tập từ khía cạnh
  for mỗi từ $w_j \in \mathcal{S}_k$ **do**
    $Count(j) = 0$; ▷ Khởi tạo số đếm của từ
  end for
  $Count_{max} = 0$; ▷ Khởi tạo số lần xuất hiện lớn nhất của tất cả các từ có trong khía cạnh a_k
  for mỗi từ $w_j \in \mathcal{S}_k$ **do**
    if $w_j \in \mathcal{V}$ **then** $Count(j) = Count(j) + 1$;
    end if
    $Count_{max} = max(Count(j))$;
  end for

Algorithm 2 Thủ tục cập nhật từ lõi khía cạnh và từ khía cạnh (continued)

for mỗi từ $w_j \in \mathcal{S}_k$ **do**

▷ Cập nhật tập từ lõi khía cạnh a_k

if $w_j \in \mathcal{V}\&Count(j) = Count_{max}\&w_j \notin \mathcal{C}_k$ **then** $\mathcal{C}_k \longleftarrow w_j$

end if

▷ Cập nhật tập từ khía cạnh a_kvới từ mới w_j mà xác suất có điều kiện của nó khi biết tập $\mathcal{S}_k$ vượt quá ngưỡng θ. $Pr(w_j|\mathcal{S}_k) = p(w_j).p(\mathcal{S}_k|(w_j)/p(\mathcal{S}_k)$

if $w_j \in \mathcal{V}$ & $w_j \notin \mathcal{C}_k$ & $Pr(w_j|\mathcal{S}_k) \geq \theta$ **then** $\mathcal{T}_k \longleftarrow w_j$

end if

end for

end procedure

Đánh giá độ phức tạp của các thuật toán

Mệnh đề 2.1 [Độ phức tạp của thủ tục cập nhật từ lõi khía cạnh]: Độ phức tạp của thủ tục cập nhật từ lõi khía cạnh **CẬP NHẬT** $(\mathcal{C}_k, \mathcal{T}_k)$ là O(T_k) với T_k là số từ có trong tất cả các câu được gán nhãn khía cạnh a_k.

Chứng minh:

Thủ tục này gồm: 2 câu lệnh gán (dòng 2,3) nên đều có thời gian chạy là O(1); Vòng lặp **for** (dòng 4-8) có thân gồm lệnh **if** (dòng 5-6) và lệnh gán (dòng 7). Vòng lặp **for** có số lần lặp là T_k (số từ có trong tất cả các câu được gán nhãn khía cạnh a_k). Thân của lệnh **if** là một lệnh gán nên có thời gian chạy là O(1). Lệnh gán (dòng 7) cũng có thời gian chạy là O(1). Vì vậy, thời gian chạy của vòng lặp **for** (dòng 4-8) là O(T_k); Vòng lặp **for** (dòng 9-14) có thân gồm 2 lệnh **if** (dòng 10-11 và dòng 12-13). Vòng lặp **for** có số lần lặp là T_k. Hai lệnh **if** đều có thân là các lệnh gán nên thời gian chạy của mỗi lệnh đều là O(1). Do đó, thời gian chạy của vòng lặp **for** (dòng 9-14) là O(T_k). Như vậy, độ phức tạp của thủ tục **CẬP NHẬT** $(\mathcal{C}_k, \mathcal{T}_k)$ là O(T_k). Mệnh đề đã được chứng minh.

Mệnh đề 2.2 [Độ phức tạp của thuật toán trích rút khía cạnh]: Độ phức tạp của thuật toán trích rút khía cạnh là O(LKT) với L là số lần cập nhật $(\mathcal{C}_k, \mathcal{T}_k)$, K là số khía cạnh của sản phẩm và T số từ được tách ra từ tất cả các văn bản có trong kho dữ liệu.

Chứng minh:

Thuật toán trích rút khía cạnh gồm 5 bước. Bước 1 gồm các dòng lệnh 1-6, bước 2 gồm các dòng lệnh 7-8, bước 3 gồm các dòng lệnh 9-16, bước 4 gồm các dòng lệnh 18-20, bước 5 gồm các dòng lệnh 21-34.

Bước 1 gồm 2 lệnh gán (dòng 1,6) và 2 vòng lặp **for** (dòng 2-3 và 4-5). Lệnh gán có thời gian thực hiện là O(1). Hai vòng lặp **for** đều có số lần lặp là K (số khía cạnh của sản phẩm) và thân của chúng là các lệnh gán có thời gian thực hiện là O(1). Như

vậy, thời gian thực hiện của bước 1 là O(K).

Bước 2 là một vòng lặp **for** với số lần lặp là D (số văn bản có trong kho dữ liệu). Thân của vòng lặp là một lệnh gán có thời gian thực hiện là O(1). Vậy bước 2 của thuật toán có thời gian thực hiện là O(D).

Bước 3 gồm 3 vòng lặp **for** lồng nhau. Lệnh **for** thứ nhất (dòng 9-16) có số lần lặp là S (số câu được tách ra từ D văn bản). Thân của nó là vòng lặp **for** thứ hai (dòng 10-15) có số lần lặp là K. Thân của vòng lặp thứ 2 là vòng lặp **for** thứ 3 có số lần lặp là N (số từ có trong câu). Vòng lặp thứ 3 có thân là lệnh **if** (dòng 12-13) thực hiện một lệnh gán có thời gian chạy là O(1). Do đó, thời gian chạy của bước 3 là O(SKN). Tuy nhiên SN=T (số từ được tách ra từ D văn bản có trong kho dữ liệu). Vì vậy, bước 3 của thuật toán có thời gian thực hiện là O(KT).

Bước 4 gồm một vòng lặp **for** (dòng 18-19) và một lệnh gán (dòng 20). Lệnh gán có thời gian thực hiện là O(1). Vòng lặp **for** có số lần lặp là K và thân của nó là lời gọi thủ tục **CẬP NHẬT** ($\mathcal{C}_k, \mathcal{T}_k$). Thủ tục này có thời gian thực hiện là O(T_k), vì thế vòng lặp **for** có thời gian thực hiện là O(KT_k). Mặt khác $KT_k = T$, vì thế bước 4 của thuật toán có thời gian thực hiện là O(T).

Bước 5 gồm 3 vòng lặp **for** lồng nhau. Lệnh **for** thứ nhất (dòng 21-34) có số lần lặp là S. Thân của nó là vòng lặp **for** thứ hai (dòng 22-33) có số lần lặp là K. Thân của vòng lặp thứ 2 gồm lệnh gán (dòng 23) có thời gian thực hiện là O(1), lệnh **if** (dòng 31-32) thực hiện một lệnh gán có thời gian thực hiện là O(1) và vòng lặp **for** thứ 3 có số lần lặp là N. Vòng lặp thứ 3 có thân là lệnh **if** (dòng 25-29) thực hiện một lệnh gán có thời gian chạy là O(1). Do đó, thời gian chạy của bước 5 là O(SKN). Ta đã biết SN=T, Vì vậy, bước 5 của thuật toán có thời gian thực hiện là O(KT).

Bước 4 và bước 5 được lặp lại trong vòng lặp repeat-until (dòng 17-35) với số lần lặp là L(Loop). Như vậy thời gian thực hiện tổng hợp của thuật toán trích rút khía cạnh là O(LKT), trong đó L là số lần cập nhật $\mathcal{C}_k, \mathcal{T}_k$, K là số khía cạnh của sản phẩm, T là số từ được tách ra từ D văn bản có trong kho dữ liệu huấn luyện.

Từ thời gian thực hiện của bước 1 là O(K), bước 2 là O(D), bước 3 là O(KT), bước 4 và bước 5 là O(LKT), ta suy ra độ phức tạp của thuật toán trích rút khía cạnh là O(LKT). Mệnh đề được chứng minh.

2.4.2 Dự đoán điểm đánh giá khía cạnh dựa trên phân lớp Naive Bayes

Vấn đề dự đoán điểm đánh giá khía cạnh có thể được coi là vấn đề phân loại đa lớp, trong đó điểm đánh giá (từ 1 đến 5) được coi là các nhãn và các *từ cảm xúc* (sentiment word) được xem xét như là các đặc trưng. Trong hầu hết các công việc liên quan đến phân tích quan điểm, tính từ và trạng từ được sử dụng như những từ ứng viên chỉ cảm xúc. Tính từ và trạng từ được phát hiện dựa trên kỹ thuật gán nhãn từ loại (POS).

Người ta nhận ra rằng một số cụm từ cũng có thể được sử dụng để thể hiện tình cảm tùy thuộc vào các ngữ cảnh khác nhau. Ví dụ trong hai câu sau: "*we have big problem with staff*", và "*we have a big room*", hai cụm danh từ "big problem" và "big room" truyền tải những cảm xúc trái ngược nhau, trong khi cả hai cụm từ đều chứa cùng một tính từ cảm xúc "big". Để nắm bắt được ngữ cảnh dạng này, luận án sử dụng trích rút các đặc trưng bi-gram theo một số mẫu cú pháp cố định được đề xuất trong [8]. Chỉ những mẫu cố định của hai từ liên tiếp, trong đó một từ là tính từ hoặc trạng từ và từ còn lại cung cấp ngữ cảnh mới được xem xét.

Điểm đánh giá khía cạnh hoặc điểm đánh giá chung của người dùng về một sản phẩm thường có dạng dữ liệu số (số thực/số nguyên). Đối với cách tiếp cận hồi quy có thể dự đoán các điểm này ở dạng số thực. Tuy nhiên với cách tiếp cận phân lớp, các giá trị điểm được coi là các giá trị số nguyên. Trong bài toán của luận án, các giá trị điểm đánh giá ở dạng số thực sẽ được quy đổi về số nguyên theo quy tắc làm tròn. Điều này có nghĩa là một điểm số là 3.3 được quy về điểm số 3, điểm số 3.7 được quy về điểm số 4.

Cho một văn bản đánh giá d_i, điểm đánh giá của khía cạnh a_k với Q đặc trưng (ký hiệu là $(f_1, f_2, ...f_q, ..., f_Q)$) được trích rút xác định dựa trên xác suất điểm r_{i_k} thuộc về lớp $c \in C_{class} = \{1,2,3,4,5\}$. Xác suất là:

$$p(r_{i_k} \in c | f_1, f_2, ..., f_q) = \frac{p(f_1, f_2, ..., f_q | r_{i_k} \in c) * p(r_{i_k} \in c)}{p(f_1, f_2, ..., f_q)} \tag{2.1}$$

Giả định rằng các đặc trưng là độc lập, sau đó (2.1) được chuyển thành:

$$p(r_{i_k} \in c | f_1, f_2, ..., f_Q) = \frac{(\prod_{q=1}^{Q} p(f_q | r_{i_k} \in c)) * p(r_{i_k} \in c)}{\sum_{q=1}^{Q} p(f_q)} \tag{2.2}$$

trong đó:

$p(f_q | r_{i_k} \in c) = \frac{n_{a_k}(f_q, c)}{n_{a_k}(c)}$ là xác suất mà đặc trưng f_q thuộc về lớp c, $n_{a_k}(f_q, c)$ là số câu được gán nhãn điểm c của khía cạnh a_k mà chúng có chứa đặc trưng f_q, $n_{a_k}(c)$ là tất cả số câu có nhãn khía cạnh là a_k và có nhãn điểm khía cạnh là c;

$p(r_{i_k} \in c) = \frac{n_{a_k}(c)}{n_{a_k}}$ là xác suất mà điểm đánh giá khía cạnh r_{i_k} thuộc về lớp c, n_{a_k} là tất cả số câu có nhãn khía cạnh là a_k

$p(f_q) = \frac{n_{a_k}(f_q)}{n_{a_k}}$ là xác suất của đặc trưng f_q, $n_{a_k}(f_q)$ là số câu có nhãn khía cạnh a_k mà chúng chứa đặc trưng f_q.

Để làm mịn công thức (2.2) phép biến đổi Laplace được sử dụng, $p(f_q | r_{i_k} \in c)$ trở thành:

$$p(f_q | r_{i_k} \in c) = \frac{n_{a_k}(f_q, c) + 1}{n_{a_k}(c) + |V_{a_k}| + 1} \tag{2.3}$$

trong đó: $\left|V_{a_k}\right|$ là tổng số đặc trưng có trong khía cạnh a_k.

Theo công thức (2.2), mẫu số là tương đương với tất cả các nhãn lớp c, do đó điểm đánh giá khía cạnh r_{i_k} được gán nhãn c khi xác suất $p(r_{i_k} \in c|f_1, f_2, ..., f_Q)$ là lớn nhất.

$$\widehat{c} = \arg \max_{c \in C_{class}} (\prod_{q=1}^{Q} p(f_q|r_{i_k} \in c) * p(r_{i_k} \in c)) \tag{2.4}$$

2.4.3 Ước lượng trọng số khía cạnh dựa trên tần suất khía cạnh trong bài đánh giá và trong toàn bộ kho ngữ liệu

Thông qua việc nghiên cứu các bài đánh giá sản phẩm, nghiên cứu sinh nhận thấy rằng nếu người dùng quan tâm nhiều hơn đến một khía cạnh (cho biết khía cạnh đó quan trọng đối với người dùng), họ sẽ đề cập nhiều hơn về nó trong bài đánh giá. Hơn nữa, có ý tưởng cho rằng một khía cạnh quan trọng thường được nhiều người dùng chia sẻ. Hay nói một cách khác, mỗi quan điểm của người dùng này cũng sẽ có ảnh hưởng tới người dùng khác. Sau quan sát này, nghiên cứu sinh đề xuất ước tính trọng số của khía cạnh bằng cách tính toán thông qua hai thành phần: Thành phần trọng số của khía cạnh a_k trong văn bản d_i được ký hiệu là ED_{ik}, và thành phần trọng số của khía cạnh thông qua tất cả các văn bản (toàn bộ kho dữ liệu) được ký hiệu là EC_k. Lưu ý rằng, theo cách này, điểm phân cực của từ cảm xúc không được sử dụng như trong một số cách tiếp cận khác. Thay vào đó các phép đo xác suất của từ và câu liên quan đến một khía cạnh trong bài đánh giá và trong kho dữ liệu được xem xét. Ý tưởng này tương tự như ý tưởng sử dụng TF-IDF để đo mức độ quan trọng của từ ở một mức độ nào đó.

Cho một bài đánh giá d_i và một khía cạnh a_k, thành phần trọng số ED_{ik} của khía cạnh a_k được xác định dựa trên xác suất xuất hiện của khía cạnh a_k trong văn bản d_i như sau:

$$ED_{ik} = \frac{\sum_{j=1}^{N_i} w_{ikj}}{N_i} \tag{2.5}$$

trong đó w_{ikj} là từ thứ j trong các từ khía cạnh của khía cạnh a_k xuất hiện trong văn bản d_i, và N_i là số từ khía cạnh xuất hiện trong văn bản d_i của tất cả các khía cạnh.

Thành phần trọng số EC_k được tính dựa trên số câu khía cạnh a_k xuất hiện trong toàn bộ kho dữ liệu về sản phẩm theo công thức (2.6).

$$EC_k = \frac{\sum_{h=1}^{M} s_{kh}}{M} \tag{2.6}$$

trong đó s_{kh}, là câu thứ h trong kho ngữ liệu được gán nhãn khía cạnh a_k, và M là tổng

số câu có trong kho ngữ liệu.

Cuối cùng, trọng số α_{ik} cho khía cạnh a_k của bài đánh giá d_i được tính như sau:

$$\alpha_{ik} = \frac{ED_{ik} * EC_k}{\sum_{k=1}^{K} ED_{ik} * EC_k} \tag{2.7}$$

trong đó K là số khía cạnh được xác định của sản phẩm, mẫu số $\sum_{k=1}^{K} ED_{ik} * EC_k$ chuẩn hóa giá trị của α_{ik} trong đoạn $[0,1]$.

2.5 Kết quả thực nghiệm

2.5.1 Dữ liệu và môi trường thử nghiệm

Các thực nghiệm được thực hiện bằng cách sử dụng ba bộ dữ liệu khác nhau bao gồm: bộ dữ liệu đánh giá khách sạn được thu thập từ Tripadvisor.com được sử dụng trong [36], bộ dữ liệu đánh giá bia được thu thập từ Beeradvocate.com sử dụng trong [132] và bộ dữ liệu đánh giá cà phê Trung Nguyên được thu thập từ trang web Amazon.com.

Tập dữ liệu Khách sạn bao gồm bảy khía cạnh khác nhau đó là value, room, location, cleanliness, check-in/front desk, service và business services. Tập dữ liệu này bao gồm gần 194,000 bài viết đánh giá về 1,759 khách sạn. Tập dữ liệu về bia có năm khía cạnh riêng biệt là aroma (or smell), palate (or feel), taste, appearance (or look), and overall. Bộ dữ liệu này khá lớn với hàng triệu lượt đánh giá. Một tập hợp con gồm 50.000 đánh giá về bia được sử dụng trong thử nghiệm. Tập dữ liệu về cà phê với bốn khía cạnh là aroma, taste, acidity, và body gồm 1200 bài đánh giá thuộc 17 loại cà phê khác nhau. Thông tin thống kê về ba bộ dữ liệu được trình bày trong Bảng 2.2.

Bảng 2.2: Thống kê ba bộ dữ liệu Khách sạn, Bia, Cà phê

	Tập dữ liệu Khách sạn	**Tập dữ liệu Bia**	**Tập dữ liệu Cà phê**
Số bài đánh giá	193,661	50,000	1200
Tổng số câu	1,790,880	509,320	5289
Trung bình số câu trên một bài đánh giá	9.25	10.19	4.41

Trong bài toán trích rút khía cạnh, tập từ lõi (Bảng 2.3) của từng khía cạnh đối với từng sản phẩm được sử dụng làm đầu vào cho thuật toán trích rút khía cạnh được mô tả trong phần 2.4.1, trong đó ngưỡng θ (ngưỡng xác suất có điều kiện) được thử

nghiệm từ 0.05 đến 0.5, giới hạn bước lặp L=10. Các bộ kiểm tra 2500, 2000 và 500 câu được chọn ngẫu nhiên từ bộ dữ liệu khách sạn, bộ dữ liệu bia và bộ dữ liệu cà phê tương ứng để kiểm tra độ chính xác của bài toán trích rút khía cạnh. Đối với bài toán dự đoán điểm đánh giá khía cạnh, tập dữ liệu sau khi thực hiện bài toán (1) được chia thành 5 phần. Sau đó kiểm tra chéo với 5_$fold$ được thực hiện cho bài toán (2), (3).

Bảng 2.3: Thống kê khía cạnh và từ lõi khía cạnh của ba bộ dữ liệu Khách sạn, Bia, Cà phê

Tập dữ liệu	Khía cạnh	Từ lõi khía cạnh
Hotel	Value	Value, price, worth
	Room	Room, rooms
	Location	Location
	Cleanliness	Dirty, smelled, clean
	Check in/front desk	Staff
	Service	Service, breakfast, food
	Business service	Internet, wifi
Bia	Appearance	Appearance, color, colors, coloring, head, foam
	Aroma	Aroma, aromas, smell, smelling
	Palate	Palate, mouth, feel, mouth feel
	Taste	Taste, tastes, aftertaste, in the end, finish, finishing
	Overall	Overall
Cà phê	Aroma	Aroma, aromas, smell, smelling, flavor, flavors
	Taste	Taste, tastes, aftertaste, finish, finishing, mouth feel
	Acidity	Acid, acidity
	Body	Body, aged, vintage

2.5.2 Tiền xử lý và trích chọn đặc trưng

Tiền xử lý dữ liệu: Các kỹ thuật xử lý ngôn ngữ tự nhiên khác nhau được sử trong các bước tiền xử lý văn bản, trích rút đặc trưng và lựa chọn đặc trưng.

Mã hóa tách câu: Văn bản được chia tách thành các phần nhỏ hơn là câu và từ. Tại bước này văn bản đầu vào được chia ra thành các phần nhỏ dựa trên các dấu hiệu nhận biết của câu là dấu “.”, “!”, “?” và dấu “,”.

Xóa dấu câu và các ký tự đặc biệt: Sau bước tách câu các dấu câu bị lỗi như "...",

"!!!", "???", và các ký tự đặc biệt như biểu tượng cảm xúc được loại bỏ khỏi các phần văn bản (câu hoặc cụm từ).

Gán nhãn từ loại: Các từ được phân loại thành các loại từ khác nhau như danh từ, tính từ, trạng từ, động từ nhờ các công cụ gán nhãn POS.

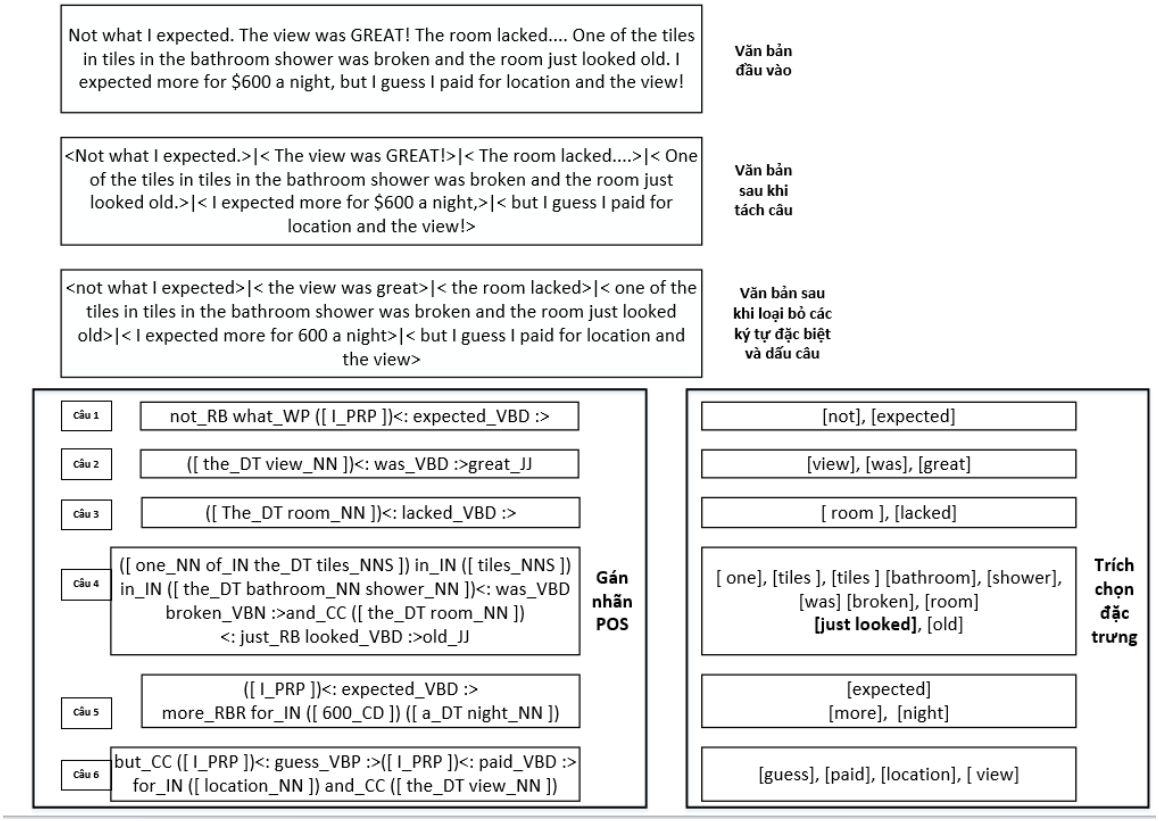
Chuyển đổi chữ hoa: Tất cả các chữ cái được chuyển sang chữ cái thường.

Loại bỏ từ dừng: Từ dừng bao gồm các mạo từ, liên từ, và giới từ đươc loại khỏi dữ liệu văn bản.

Trích rút và lựa chọn đặc trưng: Nhiệm vụ này rất quan trọng vì chất lượng của các đặc trưng ảnh hưởng trực tiếp đến hiệu quả của mô hình khai phá khía cạnh. Luận án sử dụng đặc trưng mức từ dạng Uni-gram và Bi-gram để làm đặc trưng cơ bản.

Lựa chọn đặc trưng là lọc và loại bỏ các đặc trưng ít liên quan hoặc rất ít ảnh hưởng trong vấn đề trích rút khía cạnh. Nhiều nghiên cứu chỉ ra rằng danh từ và cụm danh từ là những chỉ số quan trọng của khía cạnh [3]. Tuy nhiên, trong các nghiên cứu về trích rút khía cạnh ẩn [36] các dạng từ như tính từ, trạng từ, động từ cũng mang ngữ nghĩa ẩn ám chỉ đến khía cạnh nào đó. Trong câu *"This camera is expensive."*, từ "expensive" ám chỉ đến khía cạnh "price" của thực thể "camera". Thêm vào đó, phương pháp của luận án còn sử dụng đặc trưng tần suất từ - tần suất nghịch đảo văn bản (TF-IDF) để lọc ra các đặc trưng có trọng số quan trọng đối với toàn tập dữ liệu. Tuy nhiên, để phù hợp với thực tế, trong công thức TF-IDF nghiên cứu sinh hiệu chỉnh lại tính theo số câu có chứa thuật ngữ đang xem xét, nghĩa là $\frac{\#_d(w)}{\sum_{w' \in d}(w')} \log \frac{|\mathcal{S}|}{|s \in \mathcal{S} : w \in s|}$. Trong đó $\#_d(w)$ là số lần xuất hiện của từ w trong văn bản d, $\sum_{w' \in d}(w')$ là tổng số lần xuất hiện của tất cả các từ có trong d, $|\mathcal{S}|$ là tổng số câu tách ra từ tập dữ liệu D, $|s \in \mathcal{S} : w \in s|$ là tổng số câu chứa từ w.

Đặc trưng Uni-gram được trích rút là các danh từ, tính từ, động từ, trạng từ. Đặc trưng Bi-gram được trích rút theo mẫu cú pháp được đề xuất trong [8]. Chỉ những mẫu cố định của hai từ liên tiếp, trong đó một từ là tính từ hoặc trạng từ và từ còn lại cung cấp ngữ cảnh mới được xem xét. Hai từ liên tiếp được trích rút nếu thẻ POS của chúng tuân theo bất kỳ quy tắc nào trong Bảng 2.4, trong đó JJ là thẻ tính từ, NN là thẻ danh từ, RB là thẻ trạng từ, VB là thẻ động từ. Ví dụ, với quy tắc số 2 trong bảng này có nghĩa là hai từ liên tiếp được tách ra nếu từ đầu tiên là trạng từ (hoặc các trạng từ so sánh), từ thứ hai là tính từ, và từ thứ ba (từ không được trích rút) là một từ không phải danh từ. Trong câu "Quite dry, with a good grassy note" hai từ "quite dry" và "good grassy" được trích rút vì chúng thỏa mãn quy tắc thứ 2 và quy tắc thứ 3 tương ứng. Một ví dụ minh họa về quá trình tiền xử lý và trích chọn đặc trưng được thể hiện trong Hình 2.5



Hình 2.5: Ví dụ mô tả quá trình tiền xử lý và trích chọn đặc trưng

Bảng 2.4: Các luật trích rút đặc trưng bi-gram dựa trên POS

	Từ đầu tiên	Từ thứ hai	Từ thứ ba (không trích rút)
1.	JJ	NN or NNS	Bất kỳ từ nào
2.	RB, RBR, or RBS	JJ	Not NN nor NNS
3.	JJ	JJ	Not NN nor NNS
4.	NN or NNS	JJ	Not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	Bất kỳ từ nào

2.5.3 Kết quả và đánh giá

Trích rút khía cạnh

Để đánh giá hiệu quả của phương pháp đề xuất, độ đo precision được sử dụng. Bảng 2.5 cho thấy hiệu suất của phương pháp này đối với ba bộ dữ liệu cho vấn đề trích rút khía cạnh. Độ chính xác trung bình tương ứng là 0,786, 0,803 và 0,653 lần lượt cho bộ dữ liệu khách sạn, bộ dữ liệu bia và bộ dữ liệu cà phê. Phương pháp đề

xuất đạt được hiệu suất tốt trên bộ dữ liệu khách sạn và bia. Tuy nhiên, đối với bộ dữ liệu cà phê, kết quả không tốt như mong đợi. Điều này là do trong bộ dữ liệu cà phê, người dùng thường chỉ đưa ra cái nhìn chung về sản phẩm. Hơn nữa, tập dữ liệu này chủ yếu chứa các bài đánh giá ngắn, với số lượng câu trung bình là 4.5, so với 10 và 9 của tập dữ liệu khách sạn và tập dữ liệu bia.

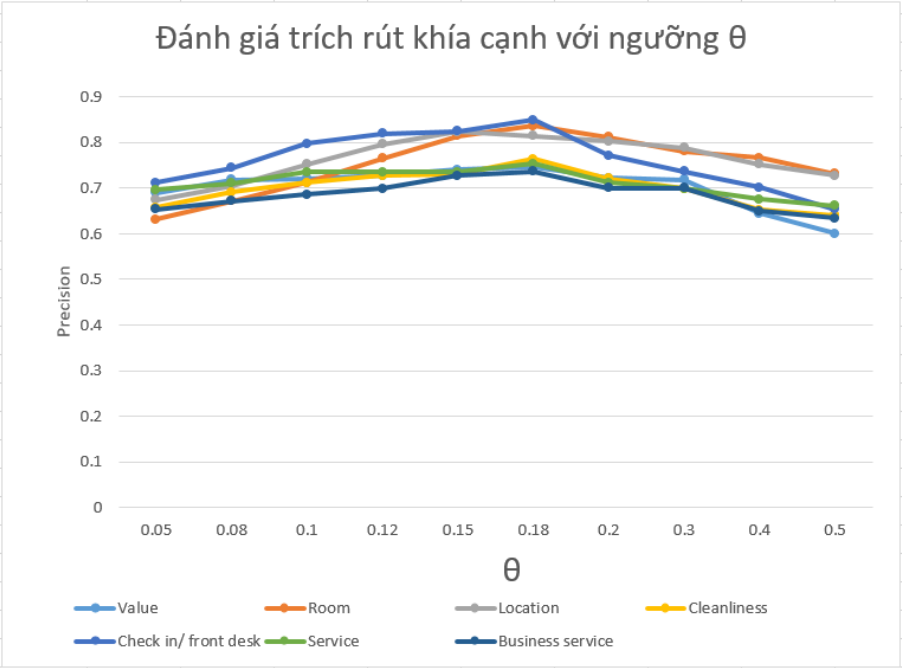
Bảng 2.5: Kết quả trích rút khía cạnh trên ba bộ dữ liệu Khách sạn, Bia, Cà phê

Dữ liêu Khách sạn		**Dữ liệu Bia**		**Dữ liệu Cà phê**	
Khía cạnh	$P_{recision}$	**Khía cạnh**	$P_{recision}$	**Khía cạnh**	$P_{recision}$
Value	0.747	Appearance	0.750	Aroma	0.667
Room	0.837	Aroma	0.857	Taste	0.677
Location	0.814	Palate	0.857	Acidity	0.667
Cleanliness	0.764	Taste	0.848	Body	0.600
Check in/ front desk	0.850	Overall	0.704		
Service	0.754				
Business service	0.737				
Trung bình	0.786	**Trung bình**	0.803	**Trung bình**	0.653

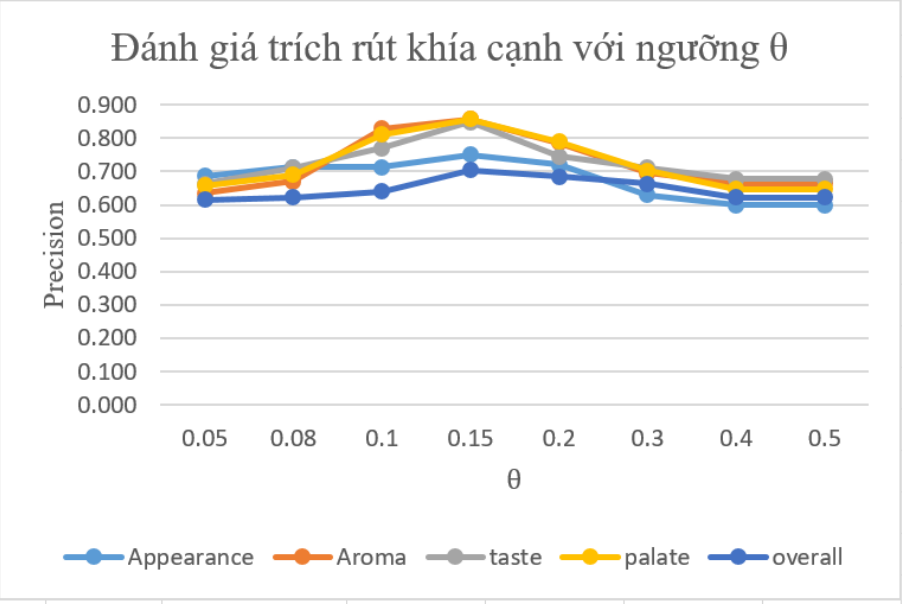
Để phân tích kỹ hơn về tác động của các tham số thuật toán đối với khả năng trích rút đúng khía cạnh (sử dụng độ đo Precision) trong từng bộ dữ liệu, luận án xem xét ngưỡng θ là ngưỡng xác suất để lấy mở rộng các tập từ khía cạnh và phân tích ảnh hưởng của tham số này đến hiệu quả của phương pháp. Bằng thực nghiệm, ngưỡng θ được thay đổi nhằm xác định giá trị mà tại đó phương pháp đề xuất hoạt động tốt nhất.

Kết quả cho thấy, ngưỡng θ cần tìm lần lượt có giá trị 0.18 (Hình 2.6), 0.15 (Hình 2.7) và 0.08 (Hình 2.8) tương ứng với các bộ dữ liệu Khách sạn, Bia, và Cà phê. Hình 2.6, 2.7 chỉ ra rằng với các bộ dữ liệu lớn (Khách sạn, Bia) độ chính xác của phương pháp đề xuất thuộc khoảng từ 0.6 đến 0.85. Mặt khác đường cong (Precision) của các khía cạnh khác nhau trong hai bộ dữ liệu này là khá tương đồng nhau, điều này thể hiện tính tương quan giữa các khía cạnh là cao. Tuy nghiên, theo Hình 2.8 thì điểm ngưỡng θ giữa các khía cạnh trong bộ dữ liệu Cà phê là không đồng nhất. Với khía cạnh Aroma và Taste ngưỡng θ ở điểm 0.08, với khía cạnh Acidity ngưỡng θ ở điểm 0.09 và khía cạnh Body ngưỡng θ ở điểm 0.1. Thêm vào đó, kết quả dự đoán khía cạnh giảm rất thấp (khoảng 0.2) khi ngưỡng θ cao. Ngoài ra, theo số liệu trong quá trình thực nghiệm cho thấy số lần lặp cập nhật từ lõi khía cạnh và từ khía cạnh của

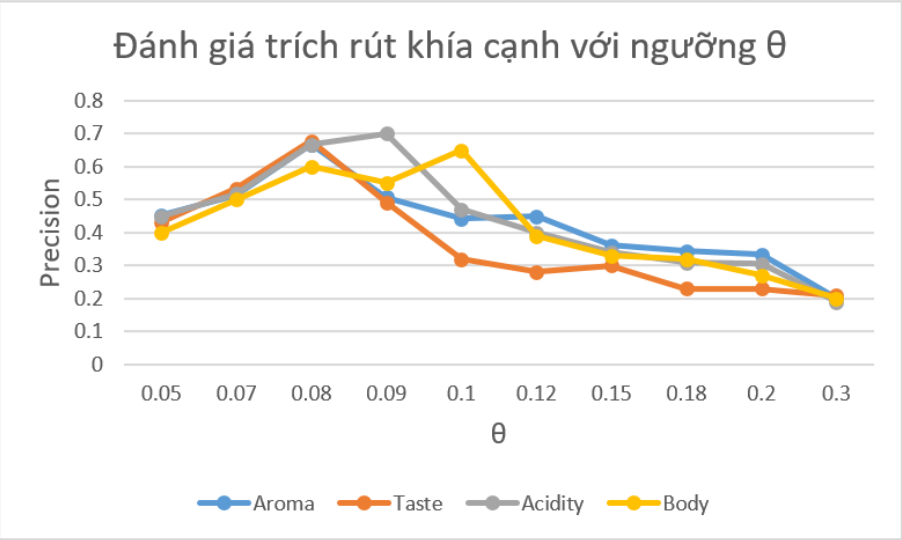
sản phẩm Cà phê chỉ là 4-6 lần tùy thuộc giá trị ngưỡng θ. Như vậy phương pháp đề xuất phù hợp với các bộ dữ liệu lớn với mật độ câu và từ cao, không phù hợp với tập dữ liệu nhỏ và dữ liệu thưa. Phương pháp đề xuất có thể cài đặt đơn giản xong mất chi phí khi phải dò ngưỡng θ trong quá trình thực nghiệm.



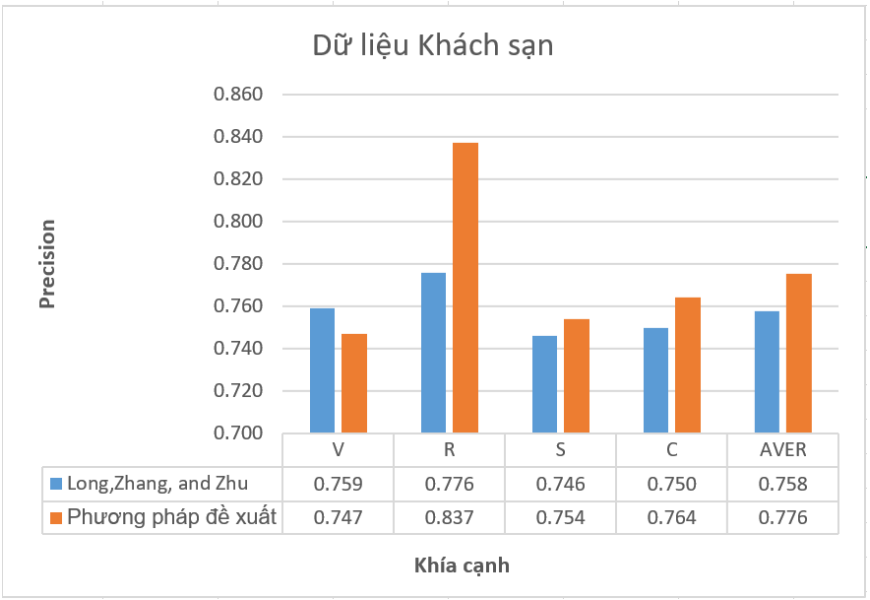
Hình 2.6: Hiệu quả của phương pháp đề xuất ứng với các ngưỡng θ khác nhau đối với bộ dữ liệu Khách sạn



Hình 2.7: Hiệu quả của phương pháp đề xuất ứng với các ngưỡng θ khác nhau đối với bộ dữ liệu Bia



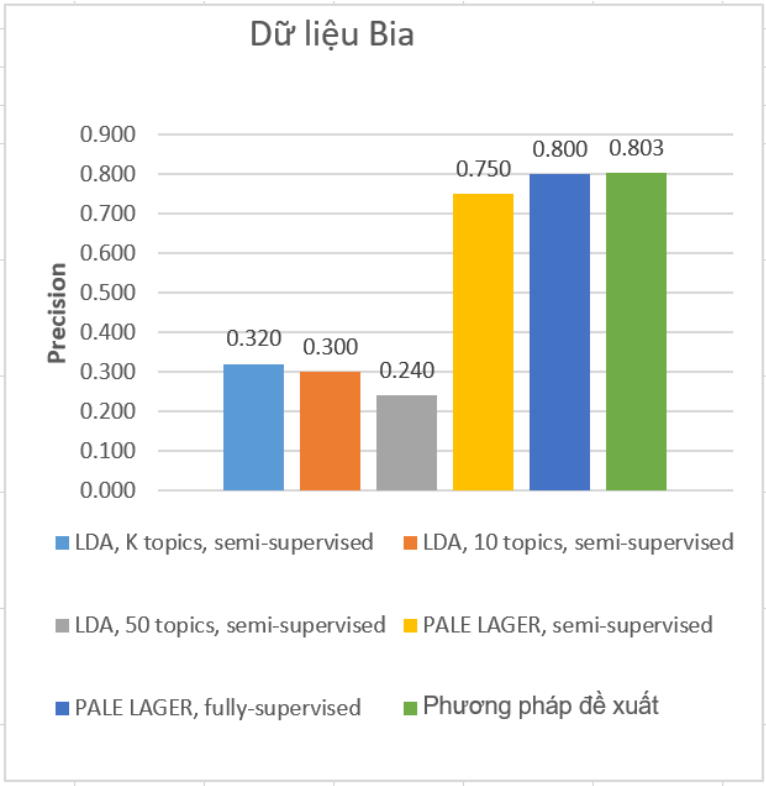
Hình 2.8: Hiệu quả của phương pháp đề xuất ứng với các ngưỡng θ khác nhau đối với bộ dữ liệu Cà phê



Hình 2.9: Kết quả so sánh phương pháp đề xuất với phương pháp của Long và các cộng sự

Phương pháp đề xuất của luận án được so sánh với phương pháp dựa trên tần suất của Long và các cộng sự [60] trên tập dữ liệu khách sạn. Phương pháp này sử dụng một tập nhỏ các từ lõi khía cạnh, sau đó các từ lõi khía cạnh này được sử dụng để tạo các từ mở rộng thông qua trình cú pháp phụ thuộc (tìm ra các mối liên quan đến sự xuất hiện của các từ lõi khía cạnh và từ mở rộng trong một văn bản). Đối với mỗi cặp thực thể - khía cạnh, các từ lõi khía cạnh, từ mở rộng và các từ liên quan được tạo thành một nhóm. Cuối cùng, một độ đo khoảng cách thông tin được sử dụng để lựa

chọn các ứng viên toàn diện nhất về một khía cạnh nào đó. Hình 2.9 cho thấy rằng phương pháp đề xuất tốt hơn so với phương pháp của Long ở các khía cạnh room (R), service (S) và Clean (C). Nhưng phương pháp của Long tốt hơn trong việc phát hiện khía cạnh Value (V).



Hình 2.10: Kết quả phương pháp đề xuất so sánh với LDA và PALE LAGER

Phương pháp đề xuất cũng được so sánh với hai phương pháp dựa trên mô hình chủ đề trong [132] và trong [133] trên tập dữ liệu bia.

Phương thức trong [133] là phương pháp bán giám sát, được gọi là LDA. Trong [132], trích rút khía cạnh và dự đoán điểm đánh giá khía cạnh được thực hiện trong cùng một khuôn khổ và được gọi là PALE LAGER (Preference and Attribute Learning from Labeled Groundtruth and Explicit Ratings). Các tác giả coi mỗi câu chỉ thảo luận về một khía cạnh duy nhất. Một phân bố xác suất mà một câu s thảo luận về một khía cạnh k với điểm đánh giá khía cạnh r được thiết lập. Việc học các phân bố này được tách thành hai vectơ tham số liên quan đến các từ mô tả khía cạnh và các từ mô tả cảm xúc đối với các khía cạnh. Từ hàm xác suất được xây dựng trên các vectơ tham số, các mô hình học không giám sát, có giám sát và bán giám sát trên tập dữ liệu được thực hiện. Do phương pháp đề xuất trong luận án có thể được coi là một phương pháp bán giám sát, vì vậy nó được so sánh với PALE LAGER(một phương pháp bán giám sát) và với PALE LAGER (một phương pháp giám sát hoàn toàn). Kết quả trong Hình 2.10 cho thấy rằng phương pháp đề xuất của luận án tốt hơn LDA với tỷ lệ khoảng

cách lớn và hơi nhỉnh hơn PALE LAGER (một phương pháp bán giám sát) và PALE LAGER (một phương pháp có giám sát).

Dữ liệu thống kê từ kía cạnh được mở rộng được trình bày trong các Bảng 2.6, 2.7, 2.8.

Bảng 2.6: Tập từ khía cạnh của dữ liệu Cà phê

Khía cạnh	**Từ khía cạnh**
Aroma	bran, brew, butter, charr, chocolate, citrus, fruit, honey, love, lover, organic, press, quality, smooth, lemon, smoke, stuff...
Taste	bitter, bitterness, chocolate, honey, salt, freshness, brew, love, lover, mild, organic, press, quality, roaster, smooth, soft, sour, stuff, sweet, sweetness, syrup, . . .
Acidity	acidic, acidness, sourness, . . .
Body	love, press, smooth, richness, thick, thin, soft, . . .

Bảng 2.7: Tập từ khía cạnh của dữ liệu Khách sạn

Khía cạnh	**Từ khía cạnh**
Value	hotel, charge, cost, discount, dollars
Room	bathroom, bathrooms, bed, beds, bath, floor, floors, chair, chairs, balcony, shower, lobby, noise, pool, queen, couple, sheraton, coffee, desk, hotel, suite, tv, view, water, window, carpet, closet, doors, furniture, king, pillows, sink, toilet, tub, toiletries, . . .
Location	airport, area, center, downtown, hotel, market, place, places, restaurant, shop, shops, shopping, show review, street, view, views, neighborhood, square, waterfront, . . .
Cleanliness	hotel, floor, shelf, desk, chair, bag, door, lobby, stairs, . . .
Check in/ front desk	desk, clerk, lounge, luggage, reception, checkout, . . .
Service	bar, bars, coffee, concierge, food, park, parking, restaurant, wine, buffet, . . .
Business service	Tv, television, wireless, hotwire, cable, computer, connection, free, freeway, . . .

Bảng 2.8: Tập từ khía cạnh của dữ liệu Bia

Khía cạnh	**Từ khía cạnh**
Appearance /Look	black, body, brown, bubble, bud, copper, lace, lacing, dots, drip, dust, back, finger, fuzzy, fluff, golden, half, laye orange, straw, surface, top, white, yellow...
Aroma /Smell	bacon, banana, basil, caramel, cheese, cream, dry citrus, fruitiness, honey, light, malt, malts, meat, mint, nose, pear, perfume, pill, pine, roast, sandalwood, smoke, smoky, spice, sweet, sweetness, yeast...
Palate/Feel	alcohol, body, carbonation, cream, drinkability, dry, hoppy, light, round, spring, summer, ...
Taste	alcohol, avalanche, balance, bitterness, body, bread, burn, caramel, carbonation, cheese, chocolate, clear, cocoa, coffee, complexity, flavors, flavors, fruit, fruitiness, ginger, grains, malt, matinee, meat, Medium-dry, oats, pear, roast, smoke, smoothness, spring, subtleties, summer, sweet, throat, toffee, tongue, vanilla, wood, ...
Overall	beer, beers, bottle, drink, beverage, style, glass, sense, quality, brewery, level, lace, brewpub ...

Dự đoán điểm đánh giá khía cạnh

Không giống như đánh giá của nhiệm vụ trích rút khía cạnh được thực hiện ở mức độ câu (gán nhãn khía cạnh cho câu), trong nhiệm vụ dự đoán điểm đánh giá khía cạnh kết quả dựa trên việc xem xét ở cấp độ văn bản (mỗi khía cạnh gồm một tập hợp các câu được gán nhãn cùng khía cạnh). Để đánh giá hiệu suất của phương pháp đề xuất trong nhiệm vụ này, luận án thực hiện đánh giá trên ba độ đo: độ đo sai số bình phương trung bình theo khía cạnh (Δ^2_{aspect}), độ đo tương quan khía cạnh (ρ_{aspect}), và độ đo tương quan khía cạnh qua tất cả các bài đánh giá (ρ_{review}). Hai độ đo đầu là đánh giá kết quả cho mỗi bài đánh giá, còn độ đo cuối là để đánh giá trong trong toàn bộ kho dữ liệu.

Kết quả của phương pháp đề xuất được so sánh với hai phương pháp của Long [60] và của Wang [36] trên bộ dữ liệu khách sạn. Long đề xuất hai phương pháp dự đoán điểm khía cạnh dựa trên SVM và BN. Phương pháp của Wang cho dự đoán điểm đánh giá khía cạnh dựa trên hồi quy ẩn (Latent Rating Regression - LRR), phương pháp này thực hiện hai nhiệm vụ dự đoán điểm đánh giá khía cạnh và ước lượng trọng số khía cạnh trong cùng một quy trình. Kết quả so sánh được chỉ ra trong Bảng 2.9. Kết

Bảng 2.9: So sánh kết quả phương pháp đề xuất với một số phương pháp về nhiệm vụ dự đoán điểm đánh giá khía cạnh

Phương pháp	Δ^2_{aspect}	ρ_{aspect}	ρ_{review}
Long và cộng sự với SVM	0.286	0.557	0.708
Long và cộng sự với BN	0.441	0.429	0.591
LRR	0.896	0.464	0.618
Phương pháp đề xuất	0.101	0.583	0.757

quả cho thấy phương pháp đề xuất hoạt động tốt hơn các phương pháp của Long và Wang trên cả ba tham số đánh giá.

Ước lượng trọng số khía cạnh

Thông thường, trọng số khía cạnh không được đề cập trực tiếp trong bài đánh giá của người dùng. Để đánh giá tính đúng đắn của ước lượng trọng số khía cạnh, luận án sử dụng điểm đánh giá tổng thể trên sản phẩm. Khi đưa ra điểm đánh giá tổng thể của một bài đánh giá đối với sản phẩm, giả định rằng, điểm đánh giá tổng thể là tổng trọng số của các điểm đánh giá trên nhiều khía cạnh của sản phẩm [63]. Theo giả định này, điểm đánh giá tổng thể của sản phẩm thông qua bài đánh giá được xác định như trong (2.8). Các giá trị ước lượng trọng số khía cạnh kết hợp với các điểm đánh giá khía cạnh của bài viết được sử dụng để tính điểm đánh giá tổng thể. Điểm đánh giá tổng thể này sẽ được so sánh với điểm đánh giá tổng thể thực tế do người dùng đưa ra để đánh giá hiệu quả của phương pháp đề xuất.

$$y_i = \sum_{k=1}^{K} r_{ik}\alpha_{ik} \tag{2.8}$$

trong đó r_{ik} và α_{ik} lần lượt là điểm đánh giá và trọng số của khía cạnh thứ k trong bài đánh giá d_i.

Phương pháp đề xuất được so sánh với phương pháp của Wang [36] dựa trên độ đo lỗi bình phương trung bình của điểm đánh giá tổng thể ($\Delta^2_{overallrating}$) cho ba tập dữ liệu. Kết quả trong bảng 2.10 cho thấy phương pháp đề xuất có thể so sánh được với phương pháp của Wang.

Bảng 2.10: MSE của điểm đánh giá tổng thể

Phương pháp	**Bộ dữ liệu sản phẩm**		
	Khách sạn	**Bia**	**Cà phê**
LRR	0.905	0.856	1.234
Phương pháp đề xuất	0.1456	0.1423	0.1904

2.6 Kết luận chương 2

Trong chương 2, nghiên cứu sinh trình bày một mô hình hệ thống nối tiếp gồm 2 modul giải quyết ba bài toán con của bài toán phân tích quan điểm mức khía cạnh. Bài toán trích rút khía cạnh tạo ra các phần của bài đánh giá đề cập đến một khía cạnh cụ thể của sản phẩm. Bài toán phân lớp cảm xúc khía cạnh xác định tính phân cực hoặc điểm đánh giá độ hài lòng của người dùng đối với từng khía cạnh đã được trích rút. Bài toán xác định trọng số khía cạnh là việc đánh giá mức độ quan tâm của người dùng đối với từng khía cạnh sản phẩm.

Để trích rút các khía cạnh được đề cập đến trong bài đánh giá về một sản phẩm, nghiên cứu sinh đã đề xuất một phương pháp sử dụng xác suất có điều kiện của các từ kết hợp với giải thuật Boostraping. Phương pháp đề xuất trích rút cả hai dạng khía cạnh rõ ràng và khía cạnh ẩn. Bên cạnh đó, các khía cạnh quan trọng nhưng có tần suất thấp cũng được phát hiện và trích rút một cách hiệu quả. Kết quả thử nghiệm cho thấy phương pháp đề xuất phù hợp với các bộ dữ liệu lớn, đối với các bộ dữ liệu nhỏ chưa thực sự hiệu quả. Ngoài ra cách tiếp cận biểu diễn từ theo dạng "Bag of Word" trong phương pháp đề xuất chưa xem xét từ trong các ngữ cảnh khác nhau của chúng, do đó vấn đề này tiếp tục được nghiên cứu trong Chương 3 của luận án.

Bài toán suy ra điểm đánh giá của người dùng cho từng khía cạnh được thực hiện dựa trên bộ phân loại Naive Bayes. Cách tiếp cận này khá đơn giản và cũng đã đem lại một số kết quả khả quan. Tuy nhiên, để cải thiện hơn nữa hiệu quả của phân lớp cảm xúc khía cạnh, nghiên cứu sinh đã đề xuất một phương pháp tiếp cận dựa trên sự kết hợp của các bộ phân loại cơ sở với nền tảng kết hợp là lý thuyết Dempster được trình bày trong Chương 4.

Bài toán ước lượng trọng số mà người dùng đặt trên mỗi khía cạnh của sản phẩm được giải quyết bằng cách sử dụng số lần xuất hiện của các từ thảo luận về khía cạnh trong một bài đánh giá và tần suất của các câu văn thảo luận về cùng một khía cạnh trên tất cả các bài đánh giá. Với cách tiếp cận này, việc ướng lượng trọng số khía cạnh không phụ thuộc vào các giá trị biết trước của điểm đánh giá khía cạnh cũng như điểm đánh giá tổng thể của người dùng về sản phẩm trong mỗi bài đánh giá.

CHƯƠNG 3: TRÍCH RÚT KHÍA CẠNH DỰA TRÊN BIỂU DIỄN TỪ WORD2VEC VÀ ĐỘ ĐO HỖ TRỢ

3.1 Đặt vấn đề

Hiện nay, có nhiều cách tiếp cận để giải quyết nhiệm vụ trích rút khía cạnh, tuy nhiên những tiếp cận này có một vài hạn chế. Điều đó giới hạn việc ứng dụng chúng trong thực tiễn. Một số phương pháp dựa trên tần suất [38, 44] tuy đơn giản song nó tạo ra nhiều phi khía cạnh và bỏ qua những khía cạnh có tần suất thấp. Ngoài ra, cách tiếp cận này cần điều chỉnh thủ công các tham số gây khó khăn cho việc chuyển sang tập dữ liệu khác. Các phương pháp trích rút dựa trên luật [16, 61] cố gắng vượt qua được nhược điểm của phương pháp tần số, song lại tạo ra các phi khía cạnh mà chúng khớp với mẫu xác định trước. Một cách tiếp cận cũng rất được các nhà nghiên cứu quan tâm là mô hình chủ đề [71–73, 133] để trích rút khía cạnh ẩn. Mô hình chủ đề thực hiện đồng thời hai nhiệm vụ trích rút khía cạnh và gom nhóm khía cạnh theo phương thức không giám sát. Mặc dù mô hình chủ đề có thể phát hiện các khía cạnh ẩn trong các bài nhận xét, tuy nhiên, phương pháp này có một số giới hạn là nó cần một lượng lớn dữ liệu và cần có những điều chỉnh đáng kể để đạt được kết quả hợp lý. Một cách tiếp cận khác có thể khắc phục các hạn chế của các phương pháp không giám sát là học có giám sát [20, 118, 134]. Phương pháp học có giám sát có thể tự động học các tham số mô hình từ dữ liệu huấn luyện. Mặc dù vậy, phương pháp học giám sát đòi hỏi một tập các thực thể được gán nhãn để chuẩn bị trước cho việc xác định các khía cạnh từ các bài nhận xét. Việc gán nhãn trước cho dữ liệu dẫn đến khó có thể áp dụng cùng một mô hình cho các lĩnh vực sản phẩm khác nhau. Đặc biệt, gần đây với sự phát triển mạnh mẽ của các mô hình học sâu, nhiều nghiên cứu trích rút khía cạnh dựa trên các mạng học sâu đã được xuất bản [120, 121]. Các mô hình học sâu giúp cho việc học đặc trưng ở mức cao và học các đặc trưng phức tạp được dễ dàng hơn. Việc biểu diễn dữ liệu văn bản dạng vectơ từ hoặc vectơ văn bản đã trở nên thuận tiện hơn và đem lại nhiều cải thiện đáng kể về độ chính xác cho bài toán trích rút khía cạnh.

Trước khi xuất hiện các mô hình nhúng từ thì hầu hết các nghiên cứu đều rất khó khăn trong vấn đề xem xét các ngữ cảnh khác nhau của từ. Mặt khác các thách thức cơ bản mà nhiệm vụ trích rút khía cạnh phải đối mặt như xác định khía cạnh ẩn, xác định tính đa khía cạnh và tính thích ứng miền vẫn chưa được giải quyết triệt để. Để vượt qua các thách thức này luận án đề xuất một phương pháp bán giám sát dựa trên biểu diễn từ dạng W2V kết hợp mô hình ngôn ngữ để trích rút khía cạnh. Với cách biểu diễn từ dạng W2V phương pháp có thể giải quyết vấn đề ngữ cảnh của từ khía cạnh và từ quan điểm giúp xác định tốt khía cạnh ẩn và khía cạnh có tần suất thấp.

Một độ đo hỗ trợ đơn giản giúp giải quyết vấn đề tương quan trong trích rút khía cạnh. Mặt khác phương pháp đề xuất sử dụng tập các từ lõi khía cạnh giúp tăng độ chính xác hơn so với phương pháp mô hình chủ đề. Tính thích ứng của mô hình cũng được đáp ứng trong quá trình học biểu diễn vectơ từ.

3.2 Các nghiên cứu liên quan

Các nghiên cứu đầu tiên trích rút các khía cạnh là dựa trên mô hình không giám sát [3], mô hình thống kê từ/thuật ngữ được sử dụng nhiều trong cách tiếp cận này. Các phương pháp này không yêu cầu dữ liệu huấn luyện được gán nhãn và có chi phí thấp. Các kỹ thuật không giám sát *thống kê* (statistical) [38, 44], kỹ thuật dựa trên *bộ lọc* (rule based) [16, 61] cho trích rút khía cạnh rõ ràng. Kỹ thuật dựa trên *mô hình chủ đề* (topic modeling) để trích rút khía cạnh ẩn [71–73]. Các kỹ thuật này dễ dàng tìm được các khía cạnh chung, song khó có thể tìm thấy các khía cạnh cục bộ, thường bỏ qua các khía cạnh có tần suất thấp, và gặp khó khăn trong việc xác định khía cạnh ẩn [3, 4].

Để nâng cao độ chính xác trong nhiệm vụ trích xuất khía cạnh, các kỹ thuật học máy có giám sát được sử dụng. Kỹ thuật này sử dụng dữ liệu được gán nhãn để trích rút cả khía cạnh rõ ràng và khía cạnh ẩn. Nói cách khác, các kỹ thuật học giám sát sử dụng các thuật toán yêu cầu huấn luyện từ dữ liệu có nhãn. Các kỹ thuật có giám sát là *trường ngẫu nhiên có điều kiện* (Conditional Random Field - CRF) [19–21] cho trích rút khía cạnh rõ ràng. *Mô hình hệ thống phân cấp* (Hierarchy) [118, 134] cho khía cạnh ngầm định. Các kỹ thuật *Bộ nhớ ngắn hạn dài hạn* (long short term memory-LSTM) [120, 121] được sử dụng để khai phá cả hai loại khía cạnh rõ ràng và khía cạnh ẩn. Các phương pháp học giám sát có ưu điểm độ chính xác cao. Tuy nhiên, khó khăn của phương pháp này là nó đòi hỏi dữ liệu phải được gán nhãn trước. Trong thực tế, việc gán nhãn cho dữ liệu trích rút khía cạnh là rất khó khăn và tốn nhiều chi phí.

Các kỹ thuật bán giám sát sử dụng cả dữ liệu có nhãn và dữ liệu không có nhãn để trích rút cả khía cạnh rõ ràng và khía cạnh ẩn. Các kỹ thuật bán giám sát sử dụng các thuật toán yêu cầu huấn luyện trong một ngữ cảnh hạn chế nhất định. Một số ví dụ về nghiên cứu theo cách tiếp cận này là *Mạng Nơ ron hồi quy* (Recurrent Neural Network - RNN) [24, 25] cho trích rút các khía cạnh rõ ràng. Cách tiếp cận *Dựa trên ngữ nghĩa* (semantic based) [18] cho khía cạnh ẩn và *Dựa trên từ vựng* (lexicon-based) [64] cho sự kết hợp trích rút cả khía cạnh rõ ràng lẫn khía cạnh ẩn.

3.3 Một số khái niệm cơ bản trong mô hình trích rút khía cạnh dựa trên biểu diễn từ Word2vec

Trong bài toán trích rút khía cạnh của Chương 3 có sử dụng lại một số khái niệm được đề cập trong Chương 2, cụ thể là: *Định nghĩa 2.1 Tập các bài đánh giá* ($\mathcal{D} = \{d_1, d_2, ..., d_D\}$); *Định nghĩa 2.2 Từ điển* ($\mathcal{V} = \{w_j | j = \overline{1,V}\}$); *Định nghĩa 2.3 Khía cạnh* ($\mathcal{A} = \{a_k | k = \overline{1,K}\}$); *Định nghĩa 2.4 Từ lõi khía cạnh* ($\mathcal{C}_k = \{w_{kj} \in \mathcal{V} | w_{kj} \rightarrow a_k, j = \overline{1,N}\}$);

Định nghĩa 3.1 Vectơ từ (Word vector): Cho một từ w_j, một vectơ P chiều $\mathbf{x}_{w_j} \in \mathbb{R}^P$ được sử dụng để biểu diễn cho P ngữ cảnh khác nhau của từ w_j trong toàn bộ không gian ngữ cảnh của kho ngữ liệu. Ký hiệu $\mathbf{x}_{w_j} = \{x_{1w_j}, x_{2w_j}, ..., x_{Pw_j}\}$, trong đó $x_{pw_j} (p = \overline{1,P})$ là một giá trị số thực có được nhờ quá trình huấn luyện Word2vec.

Định nghĩa 3.2 Vectơ từ lõi khía cạnh (Aspect core word vector): Mỗi từ lõi của khía cạnh a_k, $w_{kj} \in \mathcal{C}_k$ được ánh xạ tương ứng tới một vectơ trong tập vectơ từ được gọi là Vectơ từ lõi khía cạnh ký hiệu $\mathbf{x}_{core_{a_k}}$.

Định nghĩa 3.3 Độ hỗ trợ của từ đối với khía cạnh ($supp(w_j \rightarrow a_k)$): Độ hỗ trợ của từ w_j đối với khía cạnh a_k là một giá trị biểu diễn cho khả năng từ w_j có thể mô tả về khía cạnh a_k. Hay nói một cách khác, độ hỗ trợ thể hiện mối quan hệ tương quan giữa từ w_j và khía cạnh a_k. Nếu độ hỗ trợ càng lớn thì mối liên hệ giữa chúng càng mật thiết. Độ hỗ trợ được tính toán dựa trên sự cải tiến của độ đo Euclidean như trong công thức (3.1).

$$supp(w_j \longrightarrow a_k) = \frac{1}{N} \sum_{t=1}^{N} \frac{1}{\sum_{p=1}^{P} (x_{pw_j} - x_{pcore_{ta_k}})^2} \tag{3.1}$$

trong đó: $supp(w_j \longrightarrow a_k)$ là độ hỗ trợ của từ w_j đối với khía cạnh a_k; N là số từ lõi của khía cạnh a_k; P là số chiều của vectơ từ; x_{pw_i} là giá trị của chiều thứ p (trong biểu diễn vectơ từ) của từ w_j; $x_{pcore_{ta_k}}$ là giá trị của chiều thứ p (trong biểu diễn vectơ từ) của từ lõi thứ t thuộc về khía cạnh a_k.

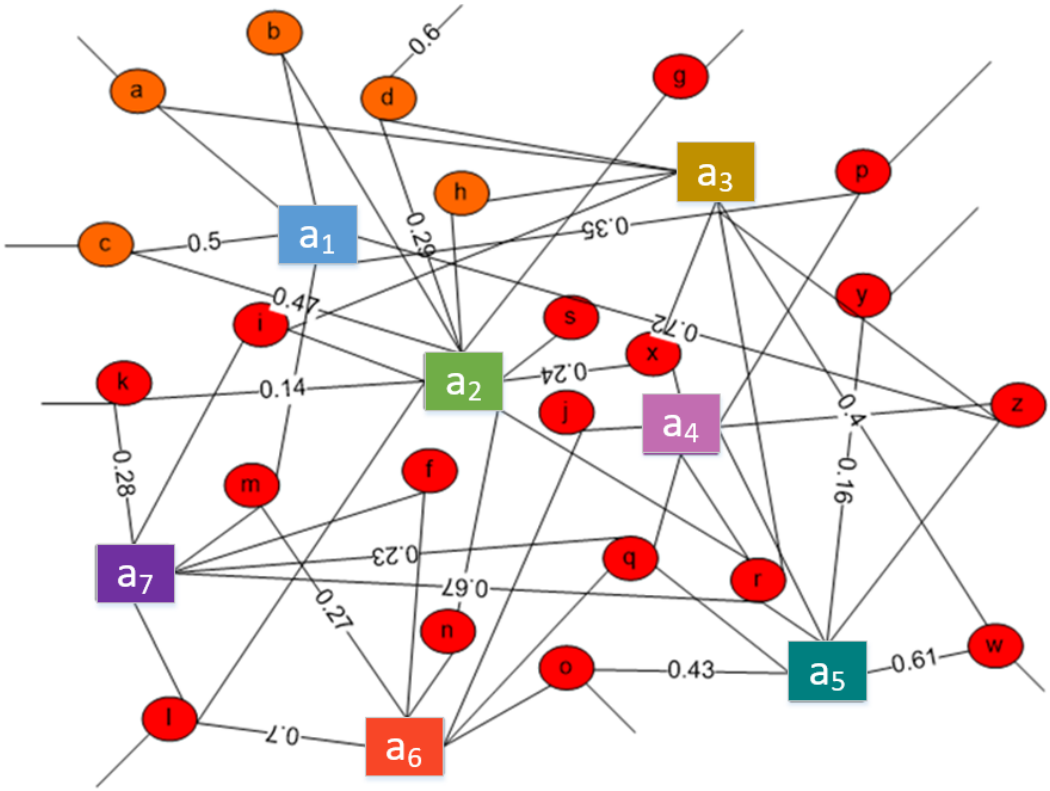
Ví dụ, Hình 3.1 mô tả tập các từ được xây dựng dựa trên độ đo hỗ trợ của từ đó đối với các khía cạnh, trong đó a_k (k=1,2,...,7) là các khía cạnh; các ký hiệu: a,b,c,..., w là các từ trong từ điển $\mathcal{V}$; các cạnh nối từ các a_k đến các từ (có trọng số) thể hiện độ hỗ trợ của từ đối với khía cạnh a_k (các giá trị này là khác nhau). Độ hỗ trợ của từ đối với khía cạnh là khác nhau thể hiện mối quan hệ của mỗi khía cạnh với mỗi từ là khác nhau.

Định nghĩa 3.4 Độ hỗ trợ của câu đối với khía cạnh ($supp(S \rightarrow a_k)$): Độ hỗ trợ của một câu S đối với khía cạnh a_k là một giá trị biểu diễn cho khả năng câu S có thể

mô tả về khía cạnh a_k. Nếu độ hỗ trợ càng lớn thì khả năng câu S miêu tả chính xác về khía cạnh a_k càng cao. Độ hỗ trợ của câu S đối với khía cạnh a_k được tính toán dựa trên trung bình độ hỗ trợ của tất cả các từ w_j có trong câu S đối với khía cạnh a_k theo công thức (3.2).

$$supp(S \longrightarrow a_k) = \frac{1}{Q} \sum_{j=1}^{Q} supp(w_j \longrightarrow a_k) \tag{3.2}$$

trong đó: Q là số từ có trong câu S.



Hình 3.1: Độ hỗ trợ của từ đối với khía cạnh

3.4 Trích rút khía cạnh dựa trên biểu diễn từ Word2vec và độ đo hỗ trợ

Một sản phẩm có thể có nhiều khía cạnh khác nhau, ví dụ sản phẩm khách sạn có thể là *phòng* (rooms), *vị trí* (location), hay *dịch vụ* (service). Mỗi khía cạnh này được người dùng sử dụng một tập hợp các từ để mô tả về chúng trong bài viết. Luận án giả sử rằng, mỗi khía cạnh a_k được thể hiện bởi một tập các từ. Tập các từ này có thể mô tả trực tiếp hoặc gián tiếp khía cạnh và được gọi là *tập từ khía cạnh* (aspect words). Giả sử rằng, ban đầu có một tập rất ít các từ có trong từ điển miêu tả rất rõ ràng khía cạnh a_k, các từ này được gọi là *từ lõi khía cạnh* (aspect core words) a_k. Các từ này

đồng thời có thể được xem là tâm của cụm phân bố khía cạnh a_k (trong Hình 3.1 được thể hiện bằng các hình vuông). Mỗi một từ trong từ điển $\mathcal{V}$ sẽ có một mối tương quan đối với từng khía cạnh của sản phẩm, hay còn gọi là độ hỗ trợ của từ đối với khía cạnh. Trong Hình 3.1 các từ được thể hiện bằng các hình tròn, độ hỗ trợ của từ đối với mỗi khía cạnh được biểu diễn bằng cạnh liên kết giữa hình tròn và hình vuông. Giá trị của độ hỗ trợ này là các số thực. Độ hỗ trợ này có thể được biểu diễn dựa trên độ đo khoảng cách (độ tương đồng) giữa các từ trong không gian biểu diễn từ. Một câu có chứa nhiều từ. Câu được gán nhãn khía cạnh a_k nếu trong câu đó có nhiều từ thuộc cụm phân bố của a_k, hoặc chứa các từ rất gần với tâm của cụm a_k.

Một vấn đề rõ ràng là, các từ trong một câu không phải là các từ rời rạc mà giữa chúng có mối quan hệ ngữ cảnh phụ thuộc nhau, thậm chí giữa các từ trong các câu khác nhau cũng có thể có mối quan hệ này. Làm thế nào để biểu diễn được mối tương quan này? Các kỹ thuật biểu diễn văn bản trong NLP như BOW hay one-hot-vector khó có thể biểu diễn được mối quan hệ này. Tuy nhiên, với sự phát triển của các kỹ thuật học sâu, các mô hình biểu diễn word2vec hay Golove là các mô hình biểu diễn rất hiệu quả trong nắm bắt ngữ nghĩa của từ theo ngữ cảnh. Khi huấn luyện word2vec các từ tương quan với nhau sẽ được nhúng trong các ngữ cảnh giữa chúng và được mã hóa bằng các chiều của vectơ từ. Ví dụ: khi chạy word2vec bằng ngôn ngữ python với câu lệnh model.similarity('room','door') => kết quả: 0.9998. Từ ví dụ này thấy rằng 'room' và 'door' có độ tương đồng về ngữ cảnh là 0.9998 tức là khả năng "room" và "door" thuộc cùng chủ đề là 0.9998. Nhờ đó ta có thể dự đoán được rằng nếu trong câu có xuất hiện từ "door" tức là khả năng cao câu đó thuộc khía cạnh "room". Luận án đề xuất phương pháp sử dụng biểu diễn đặc trưng word2vec và dựa trên các độ đo khoảng cách để gán nhãn khía cạnh cho câu/cụm từ trong bài đánh giá sản phẩm trực tuyến.

Nhiều nghiên cứu đã chỉ ra các danh từ, tính từ, trạng từ hoặc động từ trong câu thường mang nhiều thông tin liên quan đến khía cạnh của sản phẩm. Luận án cũng sử dụng cách tiếp cận này để nâng cao hiệu quả trong bài toán xác định và trích rút khía cạnh. Các từ này được lựa chọn và tính toán trong pha kiểm tra của mô hình đề xuất. Mô hình trích rút khía cạnh dựa trên học không giám sát sử dụng biểu diễn Word2vec được mô tả như Hình 3.2.

Pha huấn luyện: Tập bài đánh giá về sản phẩm được huấn luyện Word2vec để thu được tập các vectơ từ (word-vectors). Tập các từ lõi khía cạnh (aspect core word) được so khớp với các vectơ từ để thu được các vectơ từ lõi khía cạnh (aspect core word vector). Sau đó, các vectơ từ lõi khía cạnh và các vectơ từ được dùng để tính độ hỗ trợ của từ đối với từng khía cạnh.

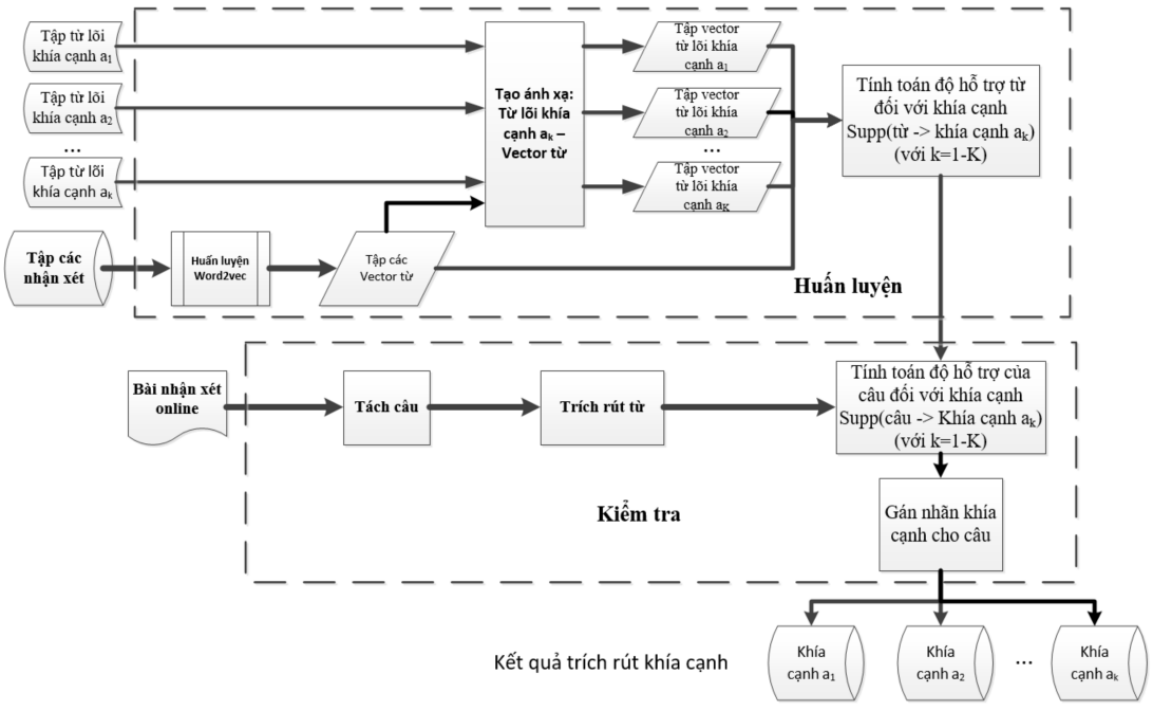
Bước 1 (chuẩn bị dữ liệu): Tập các nhận xét online của người dùng về sản phẩm

khác nhau được thực hiện thao tác tách câu, chuẩn hóa câu (chuyển về các ký tự thường, loại bỏ các ký tự thừa, các biểu tượng). Một tập rất ít các từ miêu tả chính xác một khía cạnh cụ thể được xác định trước (các từ này thường được thống kê từ dữ liệu hoặc được xác định từ các chuyên gia) được gọi là từ lõi khía cạnh.

Bước 2 (Huấn luyện word2vec): Đầu vào sẽ là tập các câu đã được chuẩn hóa từ bước 1. Tại bước này sử dụng công cụ word2vec chạy trên ngôn ngữ python để vectơ hóa các từ. Kết quả huấn luyện word2vec thu được tập các vector từ

Bước 3 (Tạo tập các vectơ từ lõi khía cạnh): Tập các từ lõi khía cạnh được ánh xạ tương ứng thành các vectơ và được gọi là các vectơ từ lõi khía cạnh.

Bước 4 (Tính supp(Từ -> Khía cạnh)): Từ tập các vectơ từ, tính độ hỗ trợ của từng từ đối với từng khía cạnh. Độ hỗ trợ của từ w_j đối với khía cạnh a_k được tính theo công thức (3.1).



Hình 3.2: Gán nhãn khía cạnh của câu dựa trên word2vec và độ đo hỗ trợ

Pha kiểm tra: Từ một bài nhận xét online, các câu được chuẩn hóa, được gán POS. Tiếp theo, tách các từ trong câu và trích rút từ (danh từ, tính từ, trạng từ, động từ) trong từng câu. Ánh xạ các từ này thành các vector từ tương ứng. Sử dụng độ hỗ trợ của các từ (đã được tính toán tại pha huấn luyện) để tính độ hỗ trợ của câu đối với từng khía cạnh. So sánh độ hỗ trợ này với ngưỡng hoặc chọn độ hỗ trợ lớn nhất để gán khía cạnh cho câu.

Bước 1 (Tách câu): Tách câu, chuẩn hóa câu.

Bước 2 (Trích rút từ): Câu được thực hiện gán nhãn POS (thông qua các công cụ

gán nhãn POS) Trích rút các từ (danh từ, tính từ, trạng từ, động từ) trong câu. Sau đó, so khớp các từ được trích rút với các từ đã được huấn luyện để xác định độ hỗ trợ của từ với từng khía cạnh.

Bước 3 (supp(Câu -> Khía cạnh)): Dựa vào độ hỗ trợ của từ với từng khía cạnh tính độ hỗ trợ của từng câu đối với từng khía cạnh theo công thức (3.2).

Bước 4 (Gán nhãn khía cạnh cho câu): Từ độ hỗ trợ của câu đối với khía cạnh, so sánh với ngưỡng hoặc lấy giá trị lớn nhất để xác định nhãn khía cạnh cho câu.

Phương pháp đề xuất được thực hiện một cách khá đơn giản. Sau khi huấn luyện Word2vec các từ sẽ được tính toán độ hỗ trợ đối với từng từ lõi khía cạnh. Các giá trị này được lưu vào trong một bảng cơ sở dữ liệu có số cột là số các từ lõi của tất cả các khía cạnh, số hàng là số từ có trong từ điển của kho dữ liệu. Trong pha trích rút (kiểm tra), câu được tách thành các từ, sau đó sử dụng các truy vấn cơ sở dữ liệu để lấy các giá trị độ hỗ trợ của từ đối với khía cạnh, tính toán độ hỗ trợ của câu đối với khía cạnh và xác định khía cạnh cuối cùng. Câu sẽ được gán nhãn khía cạnh khi độ hỗ trợ lớn hơn ngưỡng xác định. Ưu điểm của phương pháp đề xuất là tính toán đơn giản, có thể trích rút cả khía cạnh rõ ràng và khía cạnh ẩn, không phụ thuộc vào tần suất xuất hiện của từ. Tuy vậy, phương pháp này có nhược điểm là độ chính xác trích rút phụ thuộc vào ngưỡng để gán nhãn. Mặt khác, với biểu diễn word2vec, ngữ cảnh của từ mới chỉ được xem xét ở mức tổng thể trong kho dữ liệu mà chưa xem xét ở mức cục bộ (mức câu hoặc về câu). Điều này là khá quan trọng trong phân tích quan điểm khía cạnh đối với các câu ghép có nhiều khía cạnh khác nhau và mức độ phân cực cảm xúc khía cạnh khác nhau. Trong tương lai, nghiên cứu sinh sẽ hướng tới giải quyết vấn đề này để nâng cao hơn nữa hiệu quả của bài toán trích rút khía cạnh.

3.5 Kết quả thực nghiệm

3.5.1 Tiền xử lý dữ liệu

Tập dữ liệu được tiền xử lý trước. Từ các bài nhận xét các câu được tách nhờ các kỹ thuật token. Các từ trong câu được chuẩn hóa (mỗi từ được phân cách bằng dấu cách), coi mỗi từ là một token. Bước tiếp theo bộ từ vựng được xây dựng, trong đó các từ xuất hiện dưới 5 lần sẽ xem như token “<unk>” (unknow) đại diện cho các từ hiếm gặp

Để loại bỏ các từ có tần suất khá phổ biến như “the”, “a”, “an” xuất hiện trong các ngữ cảnh với các từ bình thường (vì chúng thường không có nhiều thông tin, ví dụ: đối với từ “cat” thì các sự kiện như “a cat” và “the cat” sẽ không nhận được điểm số cao hơn so với “cute cat” và “small cat”). Do đó, khi huấn luyện mô hình embedding từ, ta có thể thực hiện lấy mẫu con [135] trên các từ. Cụ thể, mỗi từ w_i được gán chỉ

số trong tập dữ liệu sẽ bị loại bỏ với một xác suất nhất định. Xác suất loại bỏ được tính như sau:

$$P(w_i) = \max(1 - \sqrt{\frac{t}{f(w_i)}}, 0) \tag{3.3}$$

trong đó, $f(w_i)$ là tần suất xuất hiện của từ w_i trong tập dữ liệu huấn luyện, nghĩa là $f(w_i) = \frac{\#w_i}{\sum_{w' \in V_w} \#w'}$, và hằng số t là một siêu tham số (luận án hiệu chỉnh từ $10^{-5} \div 10^{-4}$). Như vậy, từ w_i chỉ có thể bị loại trong lúc lấy mẫu con khi $f(w_i) > t$, tần suất của từ càng cao thì xác suất loại bỏ càng lớn.

Bộ từ vựng thu được có kích thước là: 27,721 từ đối với tập dữ liệu khách sạn; 10,583 từ đối với tập dữ liệu bia; 1,751 từ đối với tập dữ liệu cà phê. Mỗi từ được biểu diễn bởi 50 chiều.

3.5.2 Huấn luyện Word2vec

Với dữ liệu sau khi thực hiện tiền xử lý, từ đích trung tâm và từ ngữ cảnh được thực hiện trích xuất với kích thước cửa sổ tối đa (max_window_size) là 5 (thông thường là các giá trị 2,5,10). Các từ ngữ cảnh được trích rút với khoảng cách của nó tới từ đích trung tâm không quá kích thước cửa sổ tối đa. Khoảng cách này được lấy ngẫu nhiên từ 1 đến max_window_size theo phân phối đều.

Tiếp theo, để huấn luyện gần đúng các mẫu âm liên tục được thực hiện lấy mẫu. Với mỗi cặp từ đích trung tâm và ngữ cảnh, lấy ngẫu nhiên k từ nhiễu theo khuyến nghị của Word2vec [135]. Các từ phủ định w có thể được lấy mẫu theo tần suất dựa trên kho ngữ liệu của chúng $\frac{\#w_i^{0.75}}{\sum_{w' \in V_w} \#(w')^{0.75}}$. Cuối cùng, thực hiện huấn luyện lại Word2vec với biến thể skip-gram bằng cách sử dụng các tầng embedding và phép nhân minibatch. Trước khi đưa vào tính toán độ tương tự của từ và của câu, các vectơ từ tương ứng với các từ trong từ điển $\mathcal{V}$ được đưa vào cơ sở dữ liệu vectơ từ để lưu trữ. Thông tin chi tiết dữ liệu huấn luyện được trình bày trong Bảng 3.1.

Bảng 3.1: Thống kê dữ liệu huấn luyện Word2vec

	Tập dữ liệu Khách sạn	**Tập dữ liệu Bia**	**Tập dữ liệu Cà phê**
Số bài đánh giá	193,661	50,000	1,200
Tổng số câu	1,790,880	509,320	5,289
Từ vựng (từ)	27,721	10,583	1,751
Số chiều	50	50	50
Thời gian huấn luyện	3p26.891s	1p2.495s	4.695s

3.5.3 *Tạo cơ sở dữ liệu và lựa chọn đặc trưng tính toán*

Bảng dữ liệu vectơ từ: Mỗi từ trong $\mathcal{V}$ sau khi được huấn luyện Word2vec sẽ chuyển thành một vectơ được lưu trong ma trận $\mathcal{VV}ec$.

Bảng dữ liệu vectơ từ lõi khía cạnh: Từ tập $\mathcal{C}_k$, ánh xạ mỗi từ có trong $\mathcal{C}_k$ thành một vectơ từ lõi khía cạnh $\mathbf{x}_{corea_k}$ chứa trong $\mathcal{CV}ec_k$.

Bảng dữ liệu độ hỗ trợ của từ đối với khía cạnh: Với mỗi vectơ từ $\mathbf{x}_j \in \mathcal{VV}ec$ tính độ hỗ trợ đối với từng khía cạnh a_k và lưu trong ma trận độ hỗ trợ của từ với khía cạnh. Ma trận này có V hàng và K cột. Mỗi hàng là một từ trong từ điển $\mathcal{V}$, mỗi cột tương ứng với một khía cạnh. Giá trị ở vị trí hàng j cột k thể hiện độ hỗ trợ của từ w_j đối với khía cạnh a_k. Bảng 3.2 cho biết các thông tin thống kê liên quan đến quá trình huấn luyện tính toán độ hỗ trợ của từ đối với các khía cạnh trong ba bộ dữ liệu thử nghiệm.

Bảng 3.2: Thống kê dữ liệu huấn luyện độ hỗ trợ của từ đối với khía cạnh

	Tập dữ liệu Khách sạn	**Tập dữ liệu Bia**	**Tập dữ liệu Cà phê**
Số khía cạnh	7	5	4
Tổng số từ lõi khía cạnh	15	21	17
Độ hỗ trợ max - min	1000 - 0.161	1000 - 2.007	1000 - 4.406
Độ hỗ trợ trung bình	9.031	8.104	16.951
Thời gian huấn luyện	1h4p21.193s	8p37.170s	6.647s

Lựa chọn đặc trưng tính toán: Để nâng cao hiệu quả tính toán khi xác định độ hỗ trợ của câu S đối với từng khía cạnh, luận án sử dụng các đặc trưng dựa trên POS như danh từ, tính từ, trạng từ, động từ.

3.5.4 *Kết quả thực nghiệm*

Các thử nghiệm được thực hiện bằng cách sử dụng các bộ dữ liệu như trong miêu tả Mục 2.5.1. Thời gian gán nhãn trung bình đối với một văn bản kiểm tra: 0.3856s đối với dữ liệu khách sạn, 0.1059s đối với dữ liệu bia và 0.039s đối với dữ liệu cà phê.

Để đánh giá hiệu quả của phương pháp đề xuất, trong phần này luận án sử dụng hai độ đo là precision, recall và F1-score. Kết quả thử nghiệm trên ba bộ dữ liệu được thể hiện trong Bảng 3.3, 3.4, 3.5. Từ kết quả ba bảng dữ liệu cho thấy, hiệu xuất trên dữ liệu Bia cho kết quả tốt nhất vớiđiểm F1 trung bình là 0.831, các giá trị điểm F1 tương ứng với các khía cạnh dao động từ 0.798 (khía cạnh Appearance) đến 0.888 (khía cạnh Aroma). Tiếp theo sau là hiệu suất của bộ dữ liệu Khách sạn với giá trị trung bình F1 là 0.778, giá trị F1 thấp nhất là 0.709 (khía cạnh Service), giá trị F1 cao

nhất là 0.842 (khía cạnh Business service). Cuối cùng, bộ dữ liệu Cà phê thu được hiệu suất trung bình là 0.675, điểm F1 thấp nhất là 0.623 (khía cạnh Acidity), điểm F1 cao nhất là 0.716 (khía cạnh Body). Với bộ dữ liệu Cà phê thu được hiệu xuất chưa cao có thể do một số nguyên nhân như các bài viết ngắn, các quan điểm trong bài đánh giá thường ít chi tiết, sự xuất hiện các đặc trưng thường thưa thớt (khả năng bắt ngữ cảnh của từ giảm).

Bảng 3.3: Kết quả trích rút khía cạnh đối với bộ dữ liệu Khách sạn

Khía cạnh	**Precision**	**Recall**	**F1-score**
Value	0.774	0.753	0.763
Room	0.788	0.751	0.769
Location	0.823	0.794	0.808
Cleanliness	0.767	0.728	0.747
Check in/ front desk	0.804	0.800	0.802
Service	0.736	0.684	0.709
Business service	0.850	0.835	0.842
Trung bình	**0.792**	**0.764**	**0.778**

Bảng 3.4: Kết quả trích rút khía cạnh đối với bộ dữ liệu Bia

Khía cạnh	**Precision**	**Recall**	**F1-score**
Appearance	0.795	0.800	0.798
Aroma	0.875	0.901	0.888
Palate	0.862	0.792	0.826
Taste	0.843	0.826	0.834
Overall	0.821	0.803	0.812
Trung bình	**0.839**	**0.824**	**0.831**

Bảng 3.5: Kết quả trích rút khía cạnh đối với bộ dữ liệu Cà phê

Khía cạnh	**Precision**	**Recall**	**F1-score**
Aroma	0.702	0.684	0.693
Taste	0.666	0.659	0.663
Acidity	0.654	0.600	0.623
Body	0.712	0.720	0.716
Trung bình	**0.684**	**0.666**	**0.675**

Để đánh giá đầy đủ hơn hiệu suất của phương pháp mà luận án đề xuất, phương pháp đề xuất được tiến hành thử nghiệm và so sánh kết quả với hai phương pháp cơ sở là LDA [133] và của Long và các cộng sự [60] trên cùng bộ dữ liệu sử dụng độ đo precision. Kết quả cho thấy hiệu suất của phương pháp đề xuất khả quan hơn so với cách tiếp cận của Long, Zhang and Zhu và mô hình LDA đơn thuần.

Bảng 3.6: So sánh kết quả phương pháp đề xuất với phương pháp LDA và Long et al. trên tập dữ liệu Khách sạn với độ đo precision

Khía cạnh	**PP LDA [133]**	**PP Long et al. [60]**	**PP đề xuất**
Value	0.65	0.76	0.77
Room	0.47	0.78	0.79
Location	0.56	0.75	0.82
Cleanliness	0.60	0.75	0.77
Check in/front desk	0.65	0.74	0.80
Service	0.59	0.75	0.74
Business service	0.60	0.75	0.85
Trung bình	**0.589**	**0.754**	**0.791**

3.6 Kết luận chương 3

Trong chương này, nghiên cứu sinh đã đề xuất một mô hình trích rút khía cạnh của các bài nhận xét sản phẩm trực tuyến. Trong đó, phương pháp đề xuất đã khai thác hiệu quả biểu diễn đặc trưng từ dạng vectơ. Một từ điển (dạng vectơ từ) được phát triển từ Word2Vec và chúng được sử dụng để tính toán trọng số của thuật ngữ cốt lõi bằng thước đo hỗ trợ. Kết quả thử nghiệm đã chỉ ra rằng, phương pháp này hoạt động tốt trên các bộ dữ liệu của thế giới thực so với một số phương pháp khác và nó có thể được áp dụng cho một số lĩnh vực khác nhau.

CHƯƠNG 4: PHÂN LỚP CẢM XÚC BẰNG CÁCH KẾT HỢP CÁC BỘ PHÂN LOẠI CƠ SỞ

4.1 Đặt vấn đề

Như đã đề cập trong Chương 2, bài toán phân lớp quan điểm hoặc cảm xúc mà luận án xem xét là nhiệm vụ phân loại nhiều lớp của các đánh giá sản phẩm trực tuyến. Cụ thể hơn, với đánh giá của người viết về sản phẩm hoặc dịch vụ cần dự đoán quan điểm của người đó thuộc một trong năm nhóm sau: tích cực cảm xúc (emotional positive), tích cực lý trí (rational positive), trung lập (neutral), tiêu cực lý trí (rational negative), và tiêu cực cảm xúc (emotional negative). Có nhiều ứng dụng thực tế của bài toán này như phát triển sản phẩm, hiểu phản hồi và đáp ứng nguyện vọng của khách hàng, xác định chiến lược tiếp thị hiệu quả.

Làm thế nào để phân loại chính xác một bài đánh giá vào các lớp lân cận do sự khác biệt tương đối nhỏ giữa các lớp này là thách thức chính trong nhiệm vụ dự đoán điểm đánh giá khía cạnh. Ví dụ, "5 sao" gần với "4 sao" hơn là "1 sao". Do những đặc điểm này, lỗi phân loại nhầm lẫn bởi độ không chắc chắn hoặc độ không rõ ràng cao thường xảy ra. Đối với bài toán phân loại văn bản, khi các vector đặc trưng không chứa đầy đủ thông tin hoặc khi một số lớp có điểm xác xuất tương đồng nhau thì tính mơ hồ càng cao và khả năng phân lớp chính xác càng thấp.

Vấn đề dữ liệu mất cân bằng cũng là một thách thức khác của bài toán phân loại quan điểm hoặc cảm xúc nhiều lớp. Mật độ phân phối của các thể hiện (ví dụ) trên các nhãn lớp là không cân bằng, thậm chí có thể chênh lệch rất lớn dẫn đến độ chính xác khi phân loại trên các lớp thiểu số sẽ thấp. Các bộ phân loại truyền thống thường được thiết kế để giải quyết bài toán phân loại với dữ liệu cân bằng nên khó có thể thực hiện tốt trên dữ liệu mất cân bằng (do thường phân loại sai ở các lớp thiểu số). Mặc dù, đã có nhiều nghiên cứu giải quyết vấn đề mất cân bằng dữ liệu tuy nhiên hầu hết chỉ tập trung vào vấn đề phân loại nhị phân. Các công cụ và kỹ thuật hỗ trợ trực tiếp hoặc giải quyết vấn đề phân loại không cân bằng cho bài toán phân loại đa lớp vẫn còn khan hiếm.

Khó khăn thứ ba liên quan đến cấu trúc văn bản của các bài đánh giá sản phẩm trực tuyến. Thông thường các văn bản này là các văn bản ngắn, có cấu trúc lỏng lẻo, đặc trưng thưa thớt, phụ thuộc nhiều vào ngữ cảnh và thường không cung cấp đủ thông tin cơ bản để có hàm phân biệt tốt giữa các văn bản khác nhau. Hơn nữa, trong các văn bản ngắn thường có nhiều nhiễu, lộn xộn, lỗi ngữ pháp, lỗi chính tả, từ viết tắt, tiếng lóng, thuật ngữ bất thường. Do đó, để nâng cao hiệu quả phân loại văn bản ngắn, việc tiền xử lý dữ liệu và lựa chọn đặc trưng trở nên đặc biệt quan trọng.

Hướng tới việc khắc phục những khó khăn nêu trên, trong chương này của luận án,

nghiên cứu sinh đề xuất việc phát triển một phương pháp phân loại hiệu quả dựa trên ý tưởng kết hợp các bộ phân loại khác nhau, bổ sung và khắc phục yếu điểm của mỗi bộ phân loại riêng lẻ. Quan trọng hơn, bằng cách kết hợp các bộ phân loại khác nhau cung cấp nhiều loại bằng chứng khác nhau, có thể cải thiện độ chính xác của việc phân loại, đặc biệt là trong trường hợp có độ không chắc chắn và mơ hồ cao. Trong cộng đồng học máy, nhiều phương pháp kết hợp bộ phân loại đã được phát triển như bỏ phiếu, tổng trọng số, tích và hợp nhất trung bình có trọng số. Sự kết hợp của các bộ phân loại vẫn tự nó là một bộ phân loại, tuy nhiên, không có gì đảm bảo cho một bộ phân loại tối ưu toàn cục bằng cách kết hợp. Để có được một kết quả mạnh mẽ, chúng ta phải xem xét cẩn thận các khía cạnh sau: bản chất của thông tin được kết hợp, cấu trúc của sự kết hợp và việc lựa chọn các bộ phân loại cơ sở phù hợp.

Phương pháp dựa trên lý thuyết Dempster-Shafer (DS) được sử dụng nhằm mục đích kết hợp nhiều bộ phân loại với nhau để tạo nên một bộ phân loại mạnh mẽ và hiệu quả hơn. Trong các tình huống có sự không chắc chắn, phương pháp kết hợp dựa trên lý thuyết DS đã được chứng minh là hiệu quả và do đó rất phù hợp với vấn đề của luận án. Lý thuyết DS đã được áp dụng thành công trong các bài toán phân loại khác nhau, tuy nhiên, các nghiên cứu trước đây chủ yếu tập trung vào các phân lớp nhị phân và có rất ít các nghiên cứu xử lý các vấn đề phân loại cảm xúc đa lớp.

Vấn đề tiếp theo là lựa chọn số lượng và các bộ phân loại cơ sở phù hợp trong bộ kết hợp. Mặc dù số lượng các bộ phân loại cơ sở có thể kết hợp với nhau trong lý thuyết DS là không giới hạn, nhưng hiệu quả của phương pháp kết hợp sẽ giảm khi kết hợp quá nhiều bộ phân loại với nhau. Mục tiêu của nghiên cứu là làm thế nào để có được bộ phân loại cuối cùng mạnh mẽ và chính xác trong khi chỉ cần sử dụng tối thiểu các bộ phân loại. Máy vectơ hỗ trợ được chọn làm thuật toán phân loại thành phần đầu tiên vì những lý do sau: SVM có khả năng tổng quát hóa tốt trong không gian đặc trưng nhiều chiều; SVM rất thích hợp cho các vấn đề với các đặc trưng dày đặc và các thể hiện thưa thớt. Mặc dù vậy, với dữ liệu không cân bằng hoặc khi dữ liệu không chắc chắn thì SVM không thực sự tốt, vì SVM chuẩn chỉ được thiết kế cho phân loại nhị phân. Vấn đề này của SVM sẽ được giải quyết trong thành phần thứ hai của phương pháp kết hợp. Hướng đến mục tiêu này, luận án đề xuất một thuật toán phân loại cảm xúc nhiều lớp dựa trên mô hình tương tác loại trừ đa nguyên nhân (được gọi là nhiễu cổng OR) được đề xuất bởi Pearl [110] và được phát triển thành mạng Bayesian cổng OR (OR Gate Bayesian Network - OGBN) áp dụng cho bài toán phân loại văn bản bởi Campos và cộng sự [136].

4.2 Các nghiên cứu liên quan

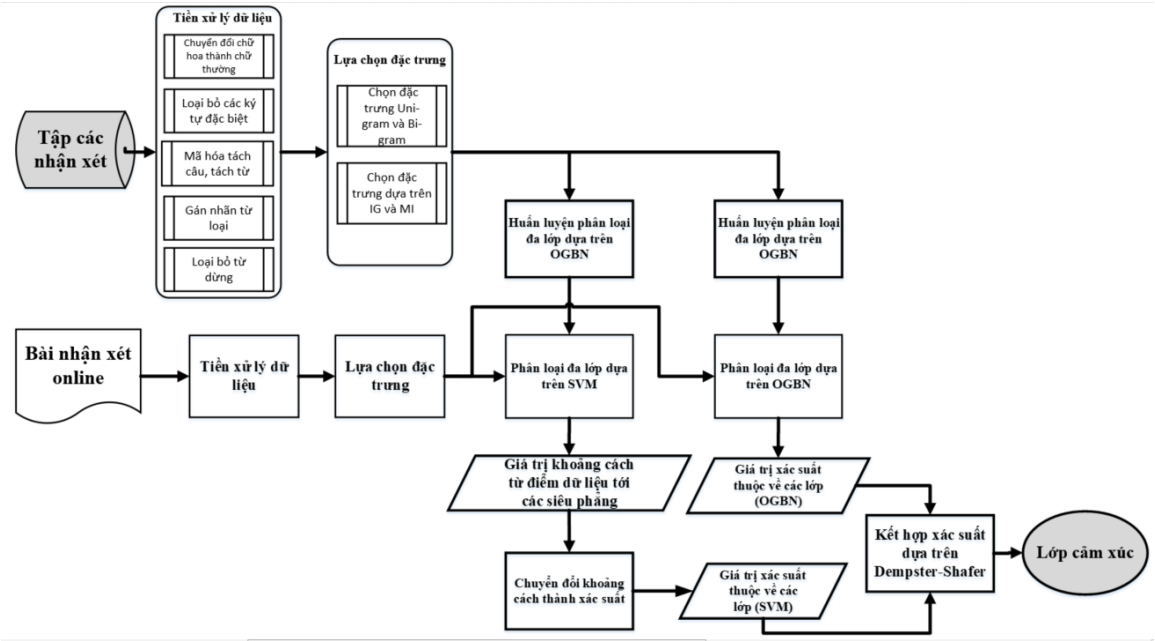
Các nghiên cứu về phân loại đa lớp cảm xúc còn ít, nhưng vấn đề này cũng đã nhận được một số sự chú ý trong những năm gần đây. Các tác phẩm tiên tiến hơn đã được trình bày, chúng đã đi sâu hơn vào việc đánh giá mức độ mạnh mẽ của cảm xúc (ví dụ: rất tiêu cực, tiêu cực, ít tiêu cực, trung lập, ít tích cực, tích cực, và rất tích cực, v.v.) [137, 138]. Một số nghiên cứu [41, 127] đề xuất cách tiếp cận mới cho phân loại cảm xúc đa lớp, trong đó văn bản được phân loại thành nhiều lớp cảm xúc khác nhau (ví dụ: tình yêu, niềm vui, sự ngạc nhiên, tức giận, buồn bã, sợ hãi, v.v.). Các nghiên cứu này cũng sử dụng các phương pháp tương tự như đối với phân loại nhị phân. Các nghiên cứu dựa trên học máy truyền thống như [41, 127, 138], một số khác sử dụng cách tiếp cận học sâu như là [92, 93]. Các phương pháp đa phân lớp cảm xúc thường là một phần mở rộng của phân lớp cảm xúc nhị phân sử dụng các chiến lược cải tiến khác nhau [137].

Gần đây, các nhà nghiên cứu đã đề xuất sử dụng các bộ phân loại dung hợp (kết hợp nhiều bộ phân loại) để tạo ra một bộ phân loại mạnh mẽ với độ chính xác cao hơn cho việc phân loại cảm xúc. Để tạo ra bộ phân loại hợp nhất hiệu quả, có hai cách tiếp cận: cách tiếp cận khuôn khổ phụ thuộc và khuôn khổ độc lập [139]. Nghiên cứu điển hình cho khung phụ thuộc là mô hình Boosting [29, 140]. Các thuật toán Boosting có thể cải thiện độ chính xác của các bộ phân loại cơ sở yếu. Tuy nhiên, cơ sở dữ liệu khổng lồ đặt ra một số thách thức đối với các thuật toán Boosting như độ phức tạp tính toán và các vấn đề lưu trữ. Các phương pháp cho khung độc lập bao gồm các phương pháp trọng số và phương pháp học tổng hợp (meta-learning). Các nghiên cứu sử dụng cách tiếp cận học trọng số khác nhau có thể kể đến [86, 141]. Các phương pháp trọng số thường được sử dụng khi các bộ phân loại cơ sở thực hiện cùng một nhiệm vụ và có hiệu suất tương đương. Các nghiên cứu theo cách tiếp cận học tổng hợp [142] cho phép kết hợp nhiều thuật toán phân loại cơ sở trên nhiều phương thức kết hợp khác nhau, từ đó tìm ra một kết hợp tối ưu nhất có thể. Mặc dù đã có một số nghiên cứu tạo ra các bộ phân loại kết hợp, chúng chủ yếu giải quyết bài toán phân lớp cảm xúc dạng nhị phân. Hiện có rất ít nghiên cứu về cách tiếp cận này cho vấn đề phân loại cảm xúc đa lớp.

Trong các vấn đề phân lớp cảm xúc, các phương pháp kết hợp đã cho thấy khả năng cải thiện độ chính xác [126, 139]. Lý thuyết kết hợp các xác suất DS là một trong những phương pháp kết hợp nổi trội trong lĩnh vực này. Cách tiếp cận này thường có hiệu suất tốt hơn các phương pháp tính trọng số khác (ví dụ: trung bình và tổng trọng số) [143]. Một số nghiên cứu trước đây đã áp dụng và phát triển lý thuyết DS cho vấn đề phân loại cảm xúc như [126, 144]. Tuy nhiên, những nghiên cứu này sử dụng

một thuật toán phân lớp để thực hiện phân lớp trên từng câu của bài đánh giá. Sau đó mỗi câu đã được phân lớp này được xem xét như một bằng chứng kết hợp. Cách tiếp cận này là sự kết hợp các yếu tố thành phần (câu) trong một văn bản, kết quả phân loại cuối cùng của văn bản là điểm đánh giá có trọng số (dựa trên lý thuyết DS) của các câu. Cách tiếp cận của nghiên cứu sinh hoàn toàn khác. Nghiên cứu sinh sử dụng nhiều bộ phân loại độc lập khác nhau để xác định điểm số cảm xúc của văn bản đánh giá của người dùng. Từ các kết quả này, dựa trên phương pháp tổng hợp theo lý thuyết DS, kết quả điểm đánh giá cảm xúc cuối cùng của văn bản sẽ được đưa ra.

4.3 Phân loại cảm xúc đa lớp bằng cách kết hợp các bộ phân loại cơ sở



Hình 4.1: Mô hình phân loại cảm xúc đa lớp bằng cách kết hợp SVM và OGBN dựa trên luật DS

Như đã giới thiệu ở Mục 4.1, nghiên cứu sinh đề xuất một phương pháp kết hợp giải quyết vấn đề phân loại đa lớp dựa trên sự kết hợp các giá trị đầu ra xác suất từ hai bộ phân lớp cơ sở (OGBN và SVM) theo luật kết hợp DS với dữ liệu phi cấu trúc. Lớp dự đoán cuối cùng dựa trên kết quả tổng hợp từ các thuật toán cơ sở. Hình 4.1 trình bày các bước thực hiện của phương pháp đề xuất.

Quá trình tiền xử lý văn bản (xem Mục 2.5.2) được thực hiện, các đặc trưng uni-gram, bi-gram, IG và MI được lựa chọn. Tập dữ liệu sau khi được trích chọn đặc trưng được chia làm hai phần là tập dữ liệu huấn luyện (80%) và tập dữ liệu kiểm tra (20%). Sau khi huấn luyện SVM và OGBN ta thu được các mô hình phân loại đa lớp dựa trên SVM và mô hình phân loại đa lớp dựa trên OGBN. Dữ liệu kiểm tra (là các bài đánh

giá sau khi đã được tiền xử lý và trích chọn đặc trưng) sẽ được đưa vào các mô hình phân loại đa lớp SVM và OGBN. Mô hình phân loại đa lớp dựa trên SVM có đầu ra là giá trị khoảng cách từ điểm dữ liệu đến các siêu phẳng tương ứng với các lớp, đây không phải là một giá trị xác suất. Do đó, các điểm khoảng cách này cần được biến đổi thành các giá trị xác suất tương ứng thông qua một hàm chuyển đổi. Đầu ra của mô hình phân loại đa lớp dựa trên OGBN là các giá trị xác suất tương ứng với mỗi lớp. Hia giá trị xác suất này (xác suất chuyển đổi từ SVM và xác suất từ OGBN) trở thành các đầu vào của luật kết hợp DS. Lớp có giá trị xác suất lớn nhất theo luật kết hợp DS sẽ được gán nhãn điểm cho bài đánh giá.

4.3.1 Phân loại cảm xúc đa lớp dựa trên SVM

Trong trường hợp tập dữ liệu đa lớp, có hai chiến lược chính để mở rộng SVM làm việc với nhiều lớp: *Một với một* (One-vs-one-OVO) và *một với tất cả* (One-vs-all-OVA).

Đối với chiến lược OVO, nếu có C lớp, vấn đề phân loại được biến đổi thành $C(C-1)/2$ bài toán phân lớp nhị phân. Việc phân tích kết quả $C(C-1)/2$ bộ phân lớp nhị phân với cách thức biểu quyết hoặc biểu quyết có trọng số sẽ xác định được nhãn lớp của mẫu kiểm tra x. Chiến lược OVO có độ phức tạp là $O(C^2)$.

Chiến lược OVA tạo ra một bộ phân loại duy nhất cho mỗi lớp. Trong mỗi bộ phân loại này, các mẫu dương là các mẫu có nhãn của lớp đó, các mẫu âm là các mẫu có nhãn của các lớp còn lại. Nếu có C lớp, chiến lược OVA cần C bộ phân loại nhị phân để huấn luyện. độ phức tạp của chiến lược OVA là $O(C)$. Một mẫu x mới được gán cho lớp mà đầu ra bộ phân loại của nó theo (4.1) xuất ra giá trị dương lớn nhất (nghĩa là cực đại lề) như trong (4.2).

$$y(x) = \mathbf{w}\mathbf{x} + b = \sum_{i=1}^{D} \alpha_i c_i (\mathbf{x} x_i + b) \tag{4.1}$$

$$c = arg \max_{1 \le c \le C} y_c(x) \tag{4.2}$$

Trong (4.2), C là số lớp, $x = \{x_1, ..., x_D\}$ là vector đặc trưng cho văn bản hoặc câu, y_c là bộ phân lớp tách lớp c (có điểm đánh giá cảm xúc là c) ra khỏi các lớp còn lại. Quyết định dự đoán trong (4.2) xác định bộ phân loại y_c có khoảng cách xa nhất đối với mẫu kiểm tra x. Điều này lý giải y_c có sức mạnh phân loại tốt nhất trong số tất cả các bộ phân loại, để tách x ra khỏi các lớp còn lại.

Vì độ phức tạp của chiến lược OVO là $O(C^2)$ và độ phức tạp của chiến lược OVA là $O(C)$, nên có thể thấy, chiến lược OVO cần số lượng bộ phân loại nhị phân lớn hơn nhiều so với chiến lược OVA. Hơn nữa, chiến lược OVA có kích thước tập dữ liệu huấn

luyện lớn hơn phương án OVO do nó sử dụng toàn bộ toàn bộ tập dữ liệu huấn luyện ban đầu để huấn luyện. Chiến lược OVA được sử dụng trong mô hình đề xuất của luận án vì mục đích hiệu quả.

4.3.2 Biến đổi đầu ra của SVM thành xác suất

SVM tạo ra một giá trị chưa được hiệu chỉnh trong (4.1) và (4.2), đây không phải là một giá trị xác suất. Vì nghiên cứu sinh sử dụng lý thuyết Dempster-Shafer để kết hợp các bộ phân loại, cho nên cần chuyển đổi giá trị khoảng cách của bộ phân loại SVM nhiều lớp để xuất ra các giá trị xác suất hậu nghiệm.

Platt [145] đề xuất một phương pháp để ước lượng SVM hậu nghiệm bằng cách sử dụng một hàm sigmoid và điểm số SVM. Thay vì ước tính mật độ có điều kiện của lớp $p(y|c)$, ông ấy sử dụng mô hình tham số để điều chỉnh trực tiếp xác suất hậu nghiệm $P(c=1|y)$. Giả sử siêu phẳng tối ưu đã học là $wx+b=0$, khi đó điểm SVM của x^* là $y(x^*)=wx^*+b$. Theo Platt, xác suất hậu nghiệm của SVM có thể được ước tính như sau:

$$p(c=1|y(x^*))=\frac{1}{1+exp(Ay(x^*)+B)} \tag{4.3}$$

trong đó $y(x)$ được xác định trong (4.1).

Mô hình sigmoid này tương đương với việc giả định rằng đầu ra của SVM tỷ lệ với log odds của một ví dụ dương. Các tham số A và B của (4.3) được hợp lý hóa bằng cách sử dụng ước tính khả năng xảy ra tối đa từ tập huấn luyện $(y_i;c_i)$. Đầu tiên, chúng ta hãy xác định một tập huấn luyện mới $(y_i;t_i)$; trong đó t_i là xác suất mục tiêu được xác định như sau:

$$t_i=\frac{c_i+1}{2} \tag{4.4}$$

Các tham số A và B được tìm thấy bằng cách tối thiểu hóa khả năng lỗi của dữ liệu huấn luyện, đây là một hàm lỗi chéo entropy:

$$min-\sum_i t_i log(p_i)+(1-t_i)log(1-t_i) \tag{4.5}$$

trong đó:

$$p_i=\frac{1}{1+exp(Ay_i+B)} \tag{4.6}$$

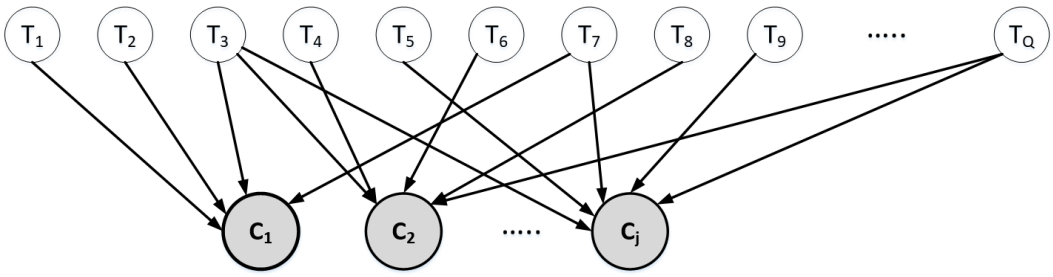
Sự tối thiểu hóa trong (4.5) là sự tối thiểu hóa hai tham số. Do đó, nó có thể được thực hiện bằng bất kỳ trong số thuật toán tối ưu nào. Luận án sử dụng mã giả trong [145] để ước tính các tham số A và B của (4.3) phù hợp theo hàm sigmoid.

4.3.3 *Phân loại cảm xúc đa lớp dựa trên mạng Bayesian cổng Noisy-OR*

Tuy SVM có thể giải quyết tốt các bài toán phân lớp có số chiều đặc trưng lớn, nhưng khi gặp phải dữ liệu không chắc chắn hoặc mất cân bằng cao SVM sẽ bị giảm hiệu quả. Ngoài ra độ chính xác của SVM cũng sẽ giảm trong bài toán phân loại đa lớp vì cơ sở lý thuyết ban đầu của SVM được thiết kế trên phân lớp nhị phân [137]. Thêm nữa, việc phân tách các lớp quá gần nhau hoặc phủ lấp lên nhau, SVM cũng gặp khó khăn.

Để vượt qua một số yếu điểm đã nêu trên của phương pháp SVM đa lớp, luận án đề xuất phương pháp kết hợp SVM với một bộ phân loại đa lớp hiệu quả khác (bộ phân loại này có thể bổ sung cho SVM) từ đó tạo nên một bộ phân loại tổng hợp chính xác hơn, mạnh mẽ hơn. Bộ phân loại kết hợp với SVM dựa trên tương tác loại trừ đa nguyên nhân (Multi-cause Disjunctive Interaction - MCDI), phương pháp này dựa trên cách tiếp cận mạng Bayesian cổng Noisy-OR được đề xuất trong [110] và được áp dụng cho phân lớp văn bản trong [136].

Cách tiếp cận theo mô hình mạng Bayes cổng OR là một biến thể của mạng Bayes nên nó hiển nhiên là một bộ phân loại đa lớp, mang các ưu điểm của mạng Bayes như hoạt động tốt trên dữ liệu có độ mất cân bằng cao [136]. Thêm vào đó, mạng Bayes cổng OR có độ phức tạp là $O(n)$ thấp hơn so với độ phức tạp của mạng Bayes là $O(2^n)$ [110–112].



Hình 4.2: Bộ phân lớp mạng Bayes Noisy OR-gate

Mô hình phân lớp cảm xúc của một bài đánh giá từ khách hàng như sau: Tập tác đặc trưng $\{f_q\}$ thu được nhờ các bước tiền xử lý và lựa chọn đặc trưng. Mỗi đặc trưng $\{f_q\}$ được xem như là một nút T_q còn được gọi là *nút nguyên nhân*. Các nút này có giá trị tương ứng 0 hoặc 1 thể hiện đặc trưng $\{f_q\}$ không xuất hiện hoặc xuất hiện trong văn bản đánh giá d_i. Mỗi lớp trong C_{class} lớp là một nút *nút kết quả* ký hiệu là C_j. Trong bài toán phân loại đa lớp mà luận án giải quyết $C_{class} = \{1,2,3,4,5\}$. Cấu trúc mạng là cố định. Nếu đặc trưng $\{f_q\}$ xuất hiện trong dữ liệu huấn luyện của lớp c_j thì trong mô hình cấu trúc mạng xuất hiện một cung có hướng đi từ nút nguyên nhân T_q đến nút kết quả C_j. Tập các nút T_q có cung kết nối với nút C_j được gọi là các

cha của C_j và được ký hiệu là $Pa(c_j)$ (xem Hình 4.2). Theo công thức (1.13), xác suất xuất hiện của lớp c_j ($c_j = true$) khi biết văn bản d_i được xác định như sau:

$$p(c_j|d_i) = 1 - \prod_{T_q \in Pa(c_j) \cap d_i} (1 - p(f_q)) \tag{4.7}$$

trong đó, $p(f_q)$ là xác suất xuất hiện của chỉ nguyên nhân T_q làm cho nút kết quả mang giá trị "True" (nghĩa là buộc lớp c_j xuất hiện). Xác suất này có thể được ước lượng trực tiếp $\widehat{p}(c_j|f_q)$ sử dụng xấp xỉ Laplace:

$$\widehat{p}(c_j|f_q) = \frac{N_{jq} + 1}{N_{\bullet q} + 2} \tag{4.8}$$

trong đó N_{jq} là số lần mà đặc trưng f_q xuất hiện trong các văn bản của lớp c_j; $N_{\bullet q}$ là số lần mà đặc trưng f_q xuất hiện trong tất cả các văn bản của kho dữ liệu, tức là $N_{\bullet q} = \sum_{c_j \in C_{class}} N_{jq}$.

Hàm phân lớp của văn bản d_i sẽ được xác định dựa vào các giá trị của $p(c_j|d_i)$ theo công thức (4.7). Khi đó hàm phân lớp sẽ là:

$$c = arg \max_{c \in \{c_1, c_2, ..., c_5\}} (1 - \prod_{T_q \in Pa(C_j) \cap d_i} (1 - p(f_q)) \tag{4.9}$$

4.3.4 Mô hình kết hợp sử dụng lý thuyết Dempster-Shafer

Dempster-Shafer là một lý thuyết toán học về bằng chứng dựa trên hàm niềm tin và lý luận hợp lý [146]. Lý thuyết DS có khả năng biểu diễn cho sự không chắc chắn và thiếu hiểu biết, và do đó, nó có thể được sử dụng để kết hợp các bộ phân loại nhằm cải thiện độ chính xác.

Lý thuyết DS bắt đầu bằng cách giả định *một khung phân biệt* là một tập hữu hạn các giả thuyết loại trừ lẫn nhau về một miền lĩnh vực nào đó. Trong bối cảnh vấn đề của luận án, khung phân biệt được đưa ra bởi một tập $C = \{c_1, ..., c_n\}$, trong đó n là số lớp, và c_j được gọi là nhãn lớp. Một siêu tập hợp $\mathbf{P}(C)$ là tập của tất cả các tập con có thể có của các lớp $\mathbf{P}(C) = \{\emptyset, \{c_1\}, ..., \{c_n\}, \{c_1, c_2\}, ..., \{c_1, ..., c_n\}\}$. Ví dụ với $n = 5$ thì siêu tập hợp $\mathbf{P}(C)$ sẽ có $2^5 = 32$ tập hợp con. Lý thuyết DS gán *hàm giá trị* (mass value) m trong khoảng từ 0 đến 1 cho mỗi tập con $A \in \mathbf{P}(C)$ của siêu tập hợp và thỏa mãn những điều kiện sau:

$$m(\phi) = 0; \sum_{A \in \mathbf{P}(C)} m(A) = 1 \tag{4.10}$$

Đưa ra hai bằng chứng được biểu thị bằng hai phép toán gán xác suất cơ bản m_1 và

m_2, quy tắc kết hợp của Dempster (còn được gọi là *hàm tổng trực giao khối lượng* (orthogonal sum mass function) [147] và ký hiệu bởi $m = m_1 \oplus m_2$) được định nghĩa như sau:

$$m_{1,2}(A) = m_1 \oplus m_2(A) = \frac{\sum_{X,Y \in \mathbf{P}(C); X \cap Y = A} m_1(X) m_2(Y)}{1 - \sum_{X,Y \in \mathbf{P}(C); X \cap Y = \emptyset} m_1(X) m_2(Y)} \tag{4.11}$$

với $A \in \mathbf{P}(C)$ gọi là các giả thuyết.

Có bốn quy tắc được mở rộng từ quy tắc kết hợp Dempster là quy tắc hợp nhất dạng tuyển của Dubois và Prade, quy tắc kết hợp trung tâm của Zhang, quy tắc kết hợp thống nhất của Inagaki và quy tắc của Yager [147]. Tất cả các quy tắc kết hợp đều liên quan đến bốn tính chất đại số: tính giao hoán $A \oplus B = B \oplus A$; tính đồng nhất $A \oplus A = A$; tính xấp xỉ $A \oplus B \approx A' \oplus B$, trong đó $A' \approx A$ (A' là rất gần A); và tính kết hợp $A \oplus (B \oplus C) = (A \oplus B) \oplus C$, trong đó $\oplus$ biểu thị phép toán kết hợp. Với phương pháp đề xuất, luận án sử dụng quy tắc kết hợp ban đầu của Dempster. Hai nguồn bằng chứng được kết hợp theo phương trình (4.11), nhưng nó có thể mở rộng để kết hợp một số nguồn tùy ý khi cần. Lưu ý rằng mẫu số của (4.11) là giống nhau đối với tất cả các giả thuyết, trong thực tế, việc tính toán chỉ cần tính tử số.

Tiếp theo, việc cần làm là xác định các hàm khối lượng sẽ được sử dụng trong bài toán. Trong lý thuyết DS, hàm khối lượng có thể được xây dựng đơn giản hay phức tạp tùy thuộc vào mục đích sử dụng và các điều kiện của bài toán. Các hàm khối lượng được sử dụng trong đề xuất của luận án chỉ định một hệ số từ 0 đến 1 cho một giả thuyết (tập hợp con của các lớp $A \in \mathbf{P}(C)$) theo đầu ra xác suất của các bộ phân loại khác nhau (nói cách khác là các nguồn bằng chứng). Tuy nhiên, ngay cả với định nghĩa hàm khối lượng khá đơn giản này, đối với mỗi thuật toán phân loại được sử dụng, nhiều bộ phân loại phải được xây dựng. Điều này hoàn toàn không thực tế khi số lượng lớp lớn. Để giải quyết vấn đề này, luận án đề xuất một phương pháp gần đúng để tính toán công thức (4.11) dựa trên tính gần đúng của quy tắc Dempster như sau:

Cho $\Theta = \mathbf{P}(C) \setminus C$, điều này nghĩa là Θ tính cho tất cả các tập con của $\mathbf{P}(C)$ có lực lượng lớn hơn 1, với mỗi giả thuyết tương ứng với một lớp riêng biệt c_j, khi đó:

$$\sum_{X,Y \in \mathbf{P}(C); X \cap Y = c_j} m_1(X) m_2(Y) \approx m_1(c_j) m_2(c_j) + m_1(c_j) m_2(\Theta) + m_1(\Theta) m_2(c_j) \tag{4.12}$$

Theo công thức (4.10), $m(\Theta)$ được xác định bởi:

$$m(\Theta) = 1 - \sum_{c_j \in C} m(c_j) \tag{4.13}$$

Lưu ý rằng $m(\Theta)$ trong công thức (4.13) vẫn chiếm các tập con của $\mathbf{P}(C)$ mà chúng không phải là siêu tập con của c_j. Để giảm tác động của những tập con như vậy luận án sử dụng xấp xỉ $\tilde{m}$ như sau:

$$\tilde{m}(\Theta) = \frac{|\{A \in \Theta; A \ni c_j\}|}{|\Theta|}(1 - \sum_{c_j \in C} m(c_j)) \tag{4.14}$$

Cho một mẫu văn bản đánh giá d_i, đối với mỗi lớp c_j, hàm khối lượng được cấu trúc dựa vào ma trận nhầm lẫn (CM_φ) và giá trị xác suất của mỗi lớp c_j ($p(c_j|d_i)$) được xác định bởi bộ phân lớp φ. Ma trận nhầm lẫn có hai chiều, một chiều được lập chỉ mục bởi lớp thực tế của một đối tượng và chiều kia được lập chỉ mục bởi lớp mà bộ phân loại dự đoán. Trong ma trận nhầm lẫn (Bảng 4.1) N_{jk} biểu diễn số mẫu thuộc về lớp c_j nhưng được phân loại vào lớp c_k. Hàm khối lượng cho lớp c_j khi biết ma trận nhầm lẫn của trình phân loại φ được xác định như sau:

$$m_\varphi(c_j) = \frac{2P_\varphi(c_j)R_\varphi(c_j)}{P_\varphi(c_j) + R_\varphi(c_j)} \cdot \frac{p_\varphi(c_j|d_i)}{\sum_{j=1}^{n} p_\varphi(c_j|d_i)} \tag{4.15}$$

trong đó $P_\varphi(c_j) = n_{jj}/\sum_{k=1}^{n} n_{kj}$ and $R_\varphi(c_j) = n_{jj}/\sum_{k=1}^{n} n_{jk}$. Phương trình (4.15) đảm bảo rằng các điều kiện trong (4.10) được thỏa mãn cho các hàm khối lượng được sử dụng trong luận án này.

Bảng 4.1: Ma trận nhầm lẫn

CM_φ		Lớp dự đoán		
		c_1	... c_k...	c_n
Lớp thực tế	c_1	N_{11}	N_{1k}	N_{1n}
	...		...	
	c_j	N_{j1}	...N_{ik}...	N_{jn}
	...		...	
	c_n	N_{n1}	N_{nk}	N_{nn}

Xem xét ví dụ 3.1 để hiểu rõ hơn về phương pháp kết hợp hai bộ đa phân lớp dựa trên SVM và mạng Bayes Noisy OR-gate. Hình 4.3 cho biết kết quả khả năng (xác suất) mà văn bản đánh giá "725721" thuộc về các lớp khác nhau qua hai bộ phân lớp dựa trên SVM và mạng Bayes Noisy OR-gate. Đồng thời, bảng 4.2 biểu diễn ma trận nhầm lẫn của bộ phân lớp dựa trên SVM và ma trận nhầm lẫn của các bộ phân lớp dựa trên mạng Bayes Noisy OR-gate. Kết quả gán các giá trị hàm khối lượng cho từng lớp $m_\varphi(c_j)$ và cho $m_\varphi(\Theta)$ được trình bày trong bảng 4.3 Xét với lớp c_1, giá trị $m_{SVM}(c_1)$ được tính như sau:

Result 1 (SVM)

<DOCUMENT name="hotel-725721" category="c5">
<CATEGORY bpa = "0.553">{c1}</ CATEGORY >
<CATEGORY bpa = "0.725">{c2}</ CATEGORY >
<CATEGORY bpa = "0.847">{c3}</ CATEGORY >
<CATEGORY bpa = "0.877">{c4}</ CATEGORY >
<CATEGORY bpa = "0.813">{c5}</ CATEGORY >
</ DOCUMENT >

Result 1 (Noisy OR-gate BN)

<DOCUMENT name=" hotel-725721" category="c5">
<CATEGORY bpa = "0.769">{c1}</ CATEGORY >
<CATEGORY bpa = "0.724">{c2}</ CATEGORY >
<CATEGORY bpa = "0.669">{c3}</ CATEGORY >
<CATEGORY bpa = "0.930">{c4}</ CATEGORY >
<CATEGORY bpa = "0.999">{c5}</ CATEGORY >
</ DOCUMENT >

Hình 4.3: Ví dụ kết quả đầu ra từ hai bộ phân lớp dựa trên SVM và mạng Bayes Noisy OR-gate

Bảng 4.2: Ma trận nhầm lẫn từ hai bộ phân lớp dựa trên SVM và mạng Bayes noisy OR-gate

CM_{SVM}		**Lớp dự đoán**				
		c_1	c_2	c_3	c_4	c_5
Lớp thực tế	c_1	**1007**	114	43	29	7
	c_2	63	**1194**	27	14	2
	c_3	27	101	**2116**	104	52
	c_4	0	15	136	**4608**	241
	c_5	0	114	232	312	**6442**

CM_{OGBN}		**Lớp dự đoán**				
		c_1	c_2	c_3	c_4	c_5
Lớp thực tế	c_1	**1031**	100	41	28	0
	c_2	56	**1200**	25	19	0
	c_3	24	95	**2128**	104	49
	c_4	0	12	129	**4627**	232
	c_5	0	110	224	262	**6504**

$$m_{SVM}(c_1) = \frac{2*\frac{1007}{(1007+63+27+0+0)}*\frac{1007}{1007+114+43+29+7}}{\frac{1007}{(1007+63+27+0+0)}+\frac{1007}{(1007+114+43+29+7)}} * \frac{0.553}{(0.553+0.725+0.847+0.877+0.813)} = 0.0635$$

Bảng 4.3: Kết quả các hàm khối lượng cho ví dụ 3.1

Hàm khối lượng	$A = \{c_j\}$					
	c_1	c_2	c_3	c_4	c_5	Θ
m_{SVM}	0.0635	0.0800	0.0948	0.1052	0.0992	0.3096
m_{OGBN}	0.0839	0.0754	0.0703	0.1048	0.1144	0.3063
$m_{(SVM,OGBN)}$	0.0508	0.0539	0.0575	0.0757	0.0771	

4.4 Kết quả thực nghiệm

4.4.1 Bộ dữ liệu thực nghiệm

Bảng 4.4 trình bày một số thông tin về các tập dữ liệu được sử dụng trong thực nghiệm với mô hình phân loại cảm xúc đa lớp bằng cách kết hợp các bộ phân lớp cơ bản. Các bộ dữ liệu này bao gồm các văn bản đánh giá của người dùng đối với các sản phẩm khác nhau và đã được sử dụng trong thực nghiệm với các phương pháp đề xuất của Chương 2 trong luận án này. Tuy nhiên với bộ dữ liệu cà phê ban đầu chỉ có 1200 bài đánh giá (số lượng tương đối nhỏ so với hai bộ còn lại), do vậy nghiên cứu sinh đã tiến hành thu thập thêm dữ liệu trên trang amazon.com. Trong bộ dữ liệu khách sạn, các khía cạnh khác nhau của khách sạn như phòng, vị trí, các dịch vụ...được người dùng quan tâm và đánh giá. Trong bộ dữ liệu bia và cà phê, người đánh giá sử dụng các cảm nhận khác nhau để mô tả đánh giá của họ về các khía cạnh như mùi thơm, hương vị, thể hiện bề ngoài của đồ uống. Tuy nhiên bia và cà phê là hai loại đồ uống khác nhau do đó các từ cảm xúc được sử dụng để mô tả đánh giá cũng khác nhau.

Bảng 4.4: Thông tin tổng hợp các bộ dữ liệu

	Bộ dữ liệu Khách sạn	**Bộ dữ liệu Bia**	**Bộ dữ liệu Cà phê**
Bài đánh giá	193,661	50,000	12,000
Số câu	1,790,880	509,320	62,896
Trung bình số câu trên bài đánh giá	9.25	10.19	5.24
Tập dữ liệu huấn luyện (bài đánh giá)	170,000	45,000	10,800
Tập dữ liệu kiểm tra (bài đánh giá)	17,000	5,000	1,200

Vấn đề mất cân bằng giữa các lớp khác nhau của dữ liệu đều xuất hiện trong các bộ dữ liệu thử nghiệm. Đối với tập dữ liệu về bia, lớp có số bài đánh giá ít nhất kém hơn 100 lần so với lớp có số bài nhiều nhất. Vấn đề này trở nên trầm trọng hơn khi sử dụng phân loại OVA cho phân loại cảm xúc đa lớp dựa trên SVM. Bảng 4.5 cho biết phân phối của các lớp trong ba bộ dữ liệu, các ký hiệu: lớp tiêu cực cảm xúc c_1; lớp tiêu cực lý trí c_2; lớp trung lập c_3; lớp tích cực lý trí c_4; lớp tích cực cảm xúc c_5.

Bảng 4.5: Phân bố của các lớp cảm xúc trong các bộ dữ liệu

Bộ dữ liệu		**Lớp** c_1	**Lớp** c_2	**Lớp** c_3	**Lớp** c_4	**Lớp** c_5	**Tổng**
Bài đánh giá	Khách sạn	12,565	13,415	24,892	61,254	81,535	193,661
	Bia	230	1,245	5,785	27,224	15,516	50,000
	Cà phê	654	857	1,413	4,142	4934	12,000

4.4.2 Tiền xử lý và lựa chọn đặc trưng

Quá trình tiền xử lý được thực hiện theo các bước đã được trình bày trong Chương 2 của luận án này. Tuy nhiên đối với phần lựa chọn đặc trưng trong phân loại cảm xúc, nghiên cứu sinh tiến hành hai thử nghiệm khác nhau để tìm ra bộ đặc trưng giúp nâng cao hiệu suất của phương pháp đề xuất. Trước tiên, bộ đặc trưng cơ sở được tạo ra như trong phương pháp đề xuất tại Chương 2 của luận án (uni-gram, bi-gram). Tiếp theo bộ đặc trưng rút gọn được tạo ra dựa trên bộ đặc trưng cơ sở thông qua các phép lọc đặc trưng với *độ lợi thông tin* (Information Gain - IG) và thông tin tương hỗ (Mutual Information - MI) giữa đặc trưng và lớp cảm xúc.

Đối với phân loại SVM, cần có biểu diễn dưới dạng vector đặc trưng. Với mục đích này, văn bản được chuyển thành biểu diễn số thông qua việc sử dụng kỹ thuật túi từ. Mỗi văn bản được mã hóa thành một vector có độ dài cố định, trong đó độ dài là số đặc trưng trong từ điển. Giá trị của mỗi vị trí trong vector thể hiện số lần xuất hiện của mỗi từ trong văn bản. Luận án cũng thử nghiệm với phương pháp TF-IDF cho mục đích này, tuy nhiên cả hai phương pháp đều có tác dụng giống nhau do cách lựa chọn đặc trưng cảm xúc.

Lựa chọn đặc trưng: Nhiệm vụ này là loại bỏ các đặc trưng không liên quan và dư thừa. Nhiều nghiên cứu đã chỉ ra rằng các tính từ và trạng từ là các chỉ báo quan trọng của cảm xúc [3]. Các đặc trưng Uni-gram là các tính từ và trạng từ trong tập hợp các

từ của bộ dữ liệu được lựa chọn bằng cách sử dụng thẻ POS thu được trong bước tiền xử lý. Các đặc trưng bi-gram được tạo ra bằng cách trích rút theo các mẫu cú pháp cố định được miêu tả trong Bảng 2.4 [8]. Bằng cách chỉ sử dụng tính từ và trạng từ, nghiên cứu sinh thu được tập các đặc trưng cơ sở ban đầu bao gồm 69,314 đặc trưng cho bộ dữ liệu khách sạn, 55,231 đặc trưng cho bộ dữ liệu bia, và cuối cùng là 19,099 đặc trưng cho bộ dữ liệu cà phê. Bộ đặc trưng này được gọi là đặc trưng "Uni+Bi".

Để giảm số đặc trưng hơn nữa, việc sử dụng phương pháp lựa chọn đặc trưng dựa trên kỹ thuật IG và MI được áp dụng. IG đo lường lượng thông tin thu được để dự đoán lớp bằng cách biết sự hiện diện hay vắng mặt của một đặc trưng trong một tài liệu. Độ lợi thông tin trong bài toán phân loại là thước đo mức độ phổ biến của một đặc trưng trong một lớp cụ thể so với mức độ phổ biến của nó trong tất cả các lớp khác. Một đặc trưng xuất hiện chủ yếu trong các bài đánh giá tích cực hiếm khi xuất hiện trong các bài đánh giá tiêu cực thường chứa độ lợi thông tin cao. Ví dụ sự hiện diện của từ "wonderful" trong bài đánh giá khách sạn là một chỉ báo mạnh mẽ cho thấy bài đánh giá đó là tích cực. Điều này có nghĩa là từ "wonderful" có độ lợi thông tin cao. Độ lợi thông tin của đặc trưng f được xác định theo công thức (4.16).

$$\begin{aligned} I(f) = &-\sum_{\mathrm{j}=1}^{\mathrm{n}} P(c_j) log(P(c_j)) \\ &+ P(f) \sum_{j=1}^{n} P(c_j|w) log(P(c_j|f)) \\ &+ (1-P(f)) \sum_{j=1}^{n} (1-P(c_j|f)) log(1-P(c_j|f)) \end{aligned} \tag{4.16}$$

trong đó:

$I(f)$: là độ lợi thông tin của đặc trưng f;

$P(c_j) = D_{c_j}/D$: là xác suất của lớp cảm xúc thứ j, D_{c_j} là số văn bản đánh giá thuộc về lớp cảm xúc là c_j;

$P(f) = D(f)/D$: là xác suất mà đặc trưng f xuất hiện trong toàn bộ kho dữ liệu của một loại sản phẩm, $D(f)$ là số văn bản cảm xúc mà chúng có chứa đặc trưng f;

$P(c_j|f) = D_{c_j}(f)/D(f)$: là xác suất có điều kiện của lớp cảm xúc c_j khi một đặc trưng f xuất hiện trong văn bản đánh giá, $D_{c_j}(f)$ là số văn bản thuộc về lớp cảm xúc c_j mà chúng có chưa đặc trưng f.

Thông tin tương hỗ $M_{c_j}(f)$ giữa đặc trưng f và lớp cảm xúc c_j được xác định dựa trên mức đồng xuất hiện giữa đặc trưng f và lớp cảm xúc c_j, $M_{c_j}(f)$ được định nghĩa

như sau:

$$M_{c_j}(f) = log\frac{P_{c_j}(f)}{P(f).P(c_j)} = log\frac{D_{c_j}(f).D}{D(f).D_{c_j}} \tag{4.17}$$

và lượng thông tin tương hỗ trung bình của một đặc trưng f qua tất cả các lớp cảm xúc được xác định như sau:

$$M_{avg}(f) = \sum_{j=1}^{n} P(c_j)M_{c_j}(f) \tag{4.18}$$

Bộ đặc trưng thứ hai sau khi thu gọn được gọi là "Uni+bi+IG+MI", bộ đặc trưng này có 6,000 đặc trưng đối với dữ liệu khách sạn, 5,000 đặc trưng đối với dữ liệu bia và 2,000 đặc trưng đối với dữ liệu cà phê. Thông tin chi tiết của hai bộ đặc trưng đối với các bộ dữ liệu được trình bày trong Bảng 4.6.

Bảng 4.6: Số chiều của hai tập đặc trưng trong ba bộ dữ liệu

Bộ dữ liệu	**Số chiều của đặc trưng Uni+Bi**	**Số chiều của đặc trưng Uni+Bi+IG+MI**
Khách sạn	69,314	6,000
Bia	55,231	5,000
Cà phê	19.099	2,000

4.4.3 Kết quả và thảo luận

Phương pháp đề xuất được thực hiện bằng ngôn ngữ lập trình C#. Ba bộ dữ liệu được mô tả trong Mục 2.5.1 được sử dụng. Các độ đo để đánh giá hiệu quả của phương pháp đề xuất bao gồm độ Accuracy (ACC), Precision (P), Recall (R), và độ đo f1-score (F1).

Thư viện LibSVM được luận án sử dụng để tiến hành cài đặt các thử nghiệm cho bộ phân loại đa lớp SVM. Nhân *SigmoidKernel* được lựa chọn, và các tham số khác là $C = 1$, $gamma = 10^{-5}$, $degree = Math.Log((prio0 + 1)/(prio1 + 1))$, số mẫu thuộc lớp tích cực là $prio1$ và số mẫu thuộc lớp không tích cự là $prio0$. Đối với các thử nghiệm trên mạng Bayes cổng nhiễu OR không có tham số.

Để đánh giá hiệu quả của phương pháp đề xuất, hai thí nghiệm đã được triển khai. Trong thí nghiệm thứ nhất, luận án so sánh độ chính xác của bộ phân loại đa lớp dựa trên SVM và bộ phân loại đa lớp dựa trên mạng Bayes Noisy OR-gate với hai bộ đặc trưng đầu vào khác nhau. Thử nghiệm tiếp theo để đánh giá hiệu quả của phương pháp kết hợp hai bộ phân loại cơ sở dựa trên lý thuyết DS. Luận án đánh giá sự cải thiện

tổng thể của mô hình kết hợp. Đồng thời các khía cạnh khác liên quan đến vấn đề dữ liệu mất cân bằng, vấn đề phân loại sai các lớp lân cận cũng được thảo luận.

Bảng 4.7: So sánh hai bộ phân lớp cơ sở trên ba bộ dữ liệu

Bộ dữ liệu	**Độ đo**					
	Bộ phân loại	**đặc trưng**	P(%)	R(%)	F1(%)	Acc (%)
Bia	SVM-based	Uni+Bi	74.37	79.42	76.81	89.54
		Uni+Bi +IG+MI	78.13	83.44	80.70	91.36
	OGBN-based	Uni+Bi	82.29	92.18	86.95	93.96
		Uni+Bi +IG+MI	83.11	91.35	87.03	93.54
Khách sạn	SVM-based	Uni+Bi	86.43	86.45	86.44	86.43
		Uni+Bi +IG+MI	87.75	89.36	88.55	90.39
	OGBN-based	Uni+Bi	89.06	90.80	89.92	91.45
		Uni+Bi +IG+MI	88.62	90.21	89.41	91.12
Cà phê	SVM-based	Uni+Bi	81.40	81.82	81.61	82.83
		Uni+Bi +IG+MI	89.33	89.41	89.37	90.08
	OGBN-based	Uni+Bi	94.41	93.42	93.91	94.08
		Uni+Bi +IG+MI	93.77	92.95	93.36	93.67

Bảng 4.7 cho thấy độ chính xác dự đoán của phương pháp dựa trên SVM và OGBN trên ba bộ dữ liệu, mỗi bộ sử dụng hai tập đặc trưng đầu vào. Bộ đặc trưng "Uni+Bi" chứa các tính từ, trạng từ đơn hoặc cụm 2 từ có chứa tính từ, trạng từ. Thông qua chỉ số IG và MI cao của các đặc trưng "Uni+Bi", Tập đặc trưng "Uni+Bi+IG+MI" được lựa chọn. Trong kết quả của cả ba bộ dữ liệu đều cho thấy Bộ phân loại dựa trên OGBN hoạt động tốt hơn bộ phân loại dựa trên SVM. Kết quả này xác nhận phân tích trước đây của nghiên cứu sinh rằng SVM hoạt động tốt với phân loại văn bản nhị phân, nhưng gặp khó khăn khi xử lý với đa phân loại đa lớp. Một quan sát thú vị ở đây là, cùng một bộ phân lớp, với tập đặc trưng có số chiều lớn ("Uni+Bi") thì bộ phân lớp dựa trên OGBN hoạt động tốt hơn, trong khi với tập đặc trưng thu gọn (""Uni+Bi+IG+MI"") thì bộ phân lớp SVM hoạt động tốt hơn.

Bảng 4.8: So sánh phương pháp kết hợp với hai bộ phân loại cơ sở

Bộ dữ liệu	**Độ đo**				
	Bộ phân loại	P(%)	R(%)	F1(%)	Accuracy(%)
Bia	SVM-based	78.13	83.44	80.70	91.36
	OGBN-based	83.11	91.35	87.03	93.54
	DS based integration	**88.17**	**94.69**	**91.32**	**95.36**
Khách sạn	SVM-based	87.75	89.36	88.55	90.39
	OGBN-based	88.62	90.21	89.41	91.12
	DS based integration	**91.89**	**92.76**	**92.32**	**93.66**
Cà phê	SVM-based	89.33	89.41	89.37	90.08
	OGBN-based	93.77	92.95	93.36	93.67
	DS based integration	**95.81**	**95.63**	**95.72**	**95.83**

Bảng 4.8 cho biết độ chính xác của hai bộ phân loại cơ sở và phương pháp kết hợp dựa trên lý thuyết DS. Như chúng ta có thể thấy trong bảng phương pháp kết hợp hoạt động tốt hơn cả hai phương pháp dựa trên SVM và dựa trên OGBN đối với cả ba bộ dữ liệu. Lưu ý rằng trong thử nghiệm này cả ba phương pháp đều sử dụng tập đặc trưng ”Uni+Bi+IG+MI”. Mặc dù phương pháp kết hợp có thể sử dụng các tập đặc trưng đầu vào khác nhau cho các bộ phân loại cơ sở khác nhau, nhưng để hạn chế chi phí tính toán trong các mô hình (do số chiều đặc trưng lớn) nên nghiên cứu sinh đã lựa chọn một tập đặc trưng thu gọn cho cả hai bộ phân loại cơ sở. Kết quả cho thấy phương pháp kết hợp có hiệu suất tốt hơn so với bộ phân loại dựa trên SVM (ACC từ 3.27% đến 5.75%) và so với bộ phân loại dựa trên OGBN (ACC từ 1.82% đến 2.54%). Tuy nhiên kết quả đã được bao phủ bởi các lớp chiếm đa số.Phân tích sau đây sẽ cho thấy rằng phương pháp được đề xuất cải thiện đáng kể hiệu suất phân loại của các lớp thiểu số và khắc phục vấn đề dữ liệu mất cân bằng mà phương pháp dựa trên SVM phải đối mặt.

Bảng 4.9, 4.10, 4.11 cho thấy số các mẫu bị phân loại nhầm lẫn giữa hai lớp kề cận nhau trong ba phương pháp. Chúng ta dễ dàng nhận thấy rằng, Với phương pháp kết hợp DS, số lượng mẫu giữa các lớp kề cận bị phân loại nhầm lẫn giảm đáng kể, nhất là đối với các lớp khó phân biệt (giữa c_1 và c_2, giữa c_4 và c_5). Trong bộ dữ liệu Khách sạn, tỉ lệ phân loại nhầm lẫn giữa các lớp c_1 và c_2 của phương pháp kết hợp DS giảm so với phương pháp dựa trên SVM và dựa trên OGBN lần lượt là 38.4% và 30.1%. Tỉ lệ giảm giữa các lớp c_4 và c_4 của phương pháp kết hợp DS so với phương pháp dựa trên SVM và dựa trên OGBN lần lượt là 38.0% và 30.6%.

Bảng 4.9: Các mẫu đã bị phân loại sai của các lớp kề của ba phương pháp trên tập dữ liệu Bia

	Bộ phân loại	SVM based	OGBN-based	DS based integration
số mẫu bị phân loại sai	$c_1 \to c_2$	6	2	2
	$c_2 \to c_1$	10	0	3
	Tổng	16	2	5
	$c_2 \to c_3$	10	7	2
	$c_3 \to c_2$	36	34	19
	Tổng	46	41	21
	$c_3 \to c_4$	18	14	7
	$c_4 \to c_3$	56	29	29
	Tổng	74	43	36
	$c_4 \to c_5$	132	78	78
	$c_5 \to c_4$	51	45	37
	Tổng	183	123	115

Bảng 4.10: Các mẫu đã bị phân loại sai của các lớp kề của ba phương pháp trên tập dữ liệu Khách sạn.

	Bộ phân loại	SVM based	OGBN-based	DS based integration
Số mẫu bị phân loại sai	$c_1 \to c_2$	114	100	56
	$c_2 \to c_1$	63	56	53
	Tổng	177	156	109
	$c_2 \to c_3$	27	25	25
	$c_3 \to c_2$	101	95	68
	Tổng	128	120	93
	$c_3 \to c_4$	104	104	100
	$c_4 \to c_3$	136	129	122
	Tổng	240	233	222
	$c_4 \to c_5$	241	232	180
	$c_5 \to c_4$	312	262	163
	Tổng	553	494	343

Bảng 4.11: Các mẫu đã bị phân loại sai của các lớp kề của ba phương pháp trên tập dữ liệu Cà phê.

	Bộ phân loại	SVM based	OGBN-based	DS based integration
Số mẫu bị phân loại sai	$c_1 \rightarrow c_2$	18	16	6
	$c_2 \rightarrow c_1$	10	4	4
	Tổng	28	20	10
	$c_2 \rightarrow c_3$	8	8	8
	$c_3 \rightarrow c_2$	7	4	4
	Tổng	15	12	12
	$c_3 \rightarrow c_4$	4	4	4
	$c_4 \rightarrow c_3$	7	4	4
	Tổng	11	8	8
	$c_4 \rightarrow c_5$	23	12	8
	$c_5 \rightarrow c_4$	18	16	7
	Tổng	41	28	15

Bảng 4.12: Sự cải thiện hiệu suất của phương pháp kết hợp so với phương pháp dựa trên SVM đối với các lớp thiểu số

Bộ dữ liệu	**Lớp**	**Tỉ lệ mất cân bằng**	**Mô hình SVM-based (Acc-%)**	**Mô hình DS based integration (Acc-%)**
Bia	c_1	1:216	60.00	90.00
	c_2	1:39	84.76	97.62
Khách sạn	c_1	1:14	83.92	91.25
	c_2	1:13	91.85	92.54
Cà phê	c_1	1:17	84.17	95.00
	c_2	1:13	89.00	94.00

Vấn đề cải thiện độ chính xác của phương pháp kết hợp DS so với phương pháp dựa trên SVM đối với các lớp thiểu số cũng được phân tích và đánh giá. Các tỉ lệ mất cân bằng của hai lớp c_1 và c_2 cùng với độ chính xác phân loại trên các lớp này được quan sát và so sánh. Kết quả cụ thể được chỉ ra trong bảng 4.12. Các số liệu trong bảng này cho thấy, phương pháp kết hợp DS cho kết quả tốt hơn so với phương pháp dựa trên SVM trong cả ba bộ dữ liệu. Điều này khẳng định mô hình kết hợp DS giữa được các ưu điểm, loại bỏ các nhược điểm của các mô hình (dựa trên SVM và dựa trên OGBN) như đã thảo luận ban đầu.

4.5 Kết luận chương 4

Trong chương này luận án xem xét giải quyết nhiệm vụ phân loại cảm xúc khía cạnh đa lớp (tích cực cảm xúc, tích cực lý trí, trung lập, tiêu cực lý trí và tiêu cực cảm xúc). Nghiên cứu sinh đã đề xuất một mô hình kết hợp hai bộ phân loại đa lớp cơ sở dựa trên SVM và dựa trên mạng Bayesian cổng OR sử dụng phương pháp kết hợp dựa trên lý thuyết Dempster-Shafer. Phương pháp kết hợp này có thể giải quyết một số vấn đề khó khăn trong bài toán phân loại đa lớp như dữ liệu không cân bằng, tính mơ hồ không rõ ràng, tính liền kề giữa các lớp lân cận.

Phương pháp đề xuất được tiến hành thử nghiệm trên ba bộ dữ liệu online. Kết quả cho thấy độ chính xác (ACC) của phương pháp kết hợp cao hơn so với các phương pháp cơ sở từ 2% đến 5%.

KẾT LUẬN

Những kết quả nghiên cứu của luận án

Mục tiêu của luận án là nghiên cứu bài toán phân tích quan điểm mức khía cạnh các bài viết đánh giá sản phẩm của người dùng trực tuyến. Bài đánh giá về sản phẩm/dịch vụ của người dùng trực tuyến thể hiện quan điểm, mức độ quan tâm của người dùng đối với sản phẩm/dịch vụ đó. Đồng thời, các bài đánh giá này cũng thể hiện tính cá nhân hóa của người dùng, xu hướng tiêu dùng và định hướng thị trường của sản phẩm/dịch vụ. Để phân tích chi tiết các quan điểm của người dùng qua các bài đánh giá trực tuyến, nghiên cứu sinh đã thực hiện nghiên cứu và giải quyết các vấn đề của bài toán phân tích quan điểm mức khía cạnh. Ba nhiệm vụ chính trích rút khía cạnh, phân lớp cảm xúc khía cạnh, ước lượng trọng số khía cạnh được nghiên cứu sinh tìm hiểu và đề xuất các giải pháp giải quyết vấn đề. Các đóng góp chính của luận án bao gồm:

- Đề xuất hệ thống nối tiếp thực hiện ba nhiệm vụ trích rút khía cạnh, dự đoán điểm cảm xúc khía cạnh, ước lượng trọng số khía cạnh của bài toán phân tích quan điểm dựa trên khía cạnh.

  Với nhiệm vụ trích rút khía cạnh, nghiên cứu sinh đề xuất một kỹ thuật học bán giám sát dựa trên xác suất có điều kiện kết hợp thuật toán bootstrapping để thực hiện bài toán. Đồng thời kỹ thuật bán giám sát này được kết hợp với các kỹ thuật lựa chọn đặc trưng dựa trên TF-IDF và POS để nâng cao hiệu suất của phương pháp. Phương pháp đề xuất có thể giải quyết các vấn đề về dữ liệu có gán nhãn, vấn đề phát hiện khía cạnh ẩn và các khía cạnh có tần suất thấp.

  Với nhiệm vụ dự đoán điểm cảm xúc khía cạnh, phương pháp học giám sát Naive Bayes được thực hiện. Cách tiếp cận này có khả năng giải quyết bài toán đa lớp và dữ liệu mất cân bằng.

  Với nhiệm vụ ước lượng trọng số khía cạnh, một cách tiếp cận không giám sát dựa trên nội dung bài viết của người dùng và tính phổ quát trong toàn bộ kho ngữ liệu được nghiên cứu. Phương pháp đề xuất giúp giải quyết được tính cá nhân hóa trên từng người dùng nhưng lại không yêu cầu phải biết điểm đánh giá cảm xúc từng khía cạnh cũng như điểm đánh giá tổng thể của bài viết.
- Luận án đề xuất một phương pháp bán giám sát để cải thiện hiệu suất trích rút khía cạnh dựa trên biểu diễn W2V kết hợp độ đo hỗ trợ. Phương pháp đề xuất có thể giải quyết tốt đối vớitrích rút khía cạnh ẩn và đặc biệt giải quyết được vấn đề phụ thuộc ngữ cảnh của từ trong nhiệm vụ này.
- Luận án đề xuất một phương pháp kết hợp hai bộ phân loại mạnh mẽ là Support

Vector Machine và OR Gate Bayesian Network dựa trên lý thuyết Dempster để giải quyết nhiệm vụ phân lớp cảm xúc khía cạnh.Phương pháp đề xuất có hiệu quả vượt trội so với hai phương pháp cơ sở. Đặc biệt phương pháp kết hợp có thể giải quyết vấn đề phân tách các lớp gần nhau, vấn đề dữ liệu mất cân bằng trong bài toán phân loại đa lớp.

Ý nghĩa và khả năng ứng dụng vào thực tiễn

Phân tích quan điểm mức khía cạnh các bài đánh giá sản phẩm trực tuyến là một bài toán có vai trò quan trọng trong nghiên cứu cũng như trong ứng dụng thực tiễn đối với các hoạt động của tổ chức, doanh nghiệp, đặc biệt là các doanh nghiệp kinh doanh trực tuyến.

Phân tích quan điểm người dùng là bài toán có tính ứng dụng cao, bởi vì ý kiến đánh giá của người tiêu dùng sẽ giúp ích cho những người dùng khác trong quá trình tìm hiểu và ra quyết định lựa chọn dùng/mua sản phẩm. Hầu hết người dùng trực tuyến đều rất quan tâm đến những đánh giá của người dùng khác đối với các sản phẩm/dịch vụ mà họ quan tâm. Những đánh giá tích cực/tiêu cực về sản phẩm giúp người tiêu dùng có lựa chọn chính xác cho quyết định tiêu dùng của họ. Bên cạnh đó, về phía doanh nghiệp cũng rất quan tâm đến quan điểm của người dùng. Nguồn thông tin này rất có ý nghĩa với doanh nghiệp trong việc hoạch định các chiến lược quảng bá/tiếp thị, các chiến lược bán hàng, chiến lược quản trị quan hệ khách hàng, v.v. Ngoài ra, dựa vào quan điểm của người dùng đối với/sản phẩm dịch vụ, các nhà quản lý cấp cao có thể phân tích thị trường, dự đoán xu hướng tiêu dùng, dự đoán khả năng phát triển sản phẩm từ đó đưa ra các chiến lược kinh doanh hiệu quả.

Kết quả bài toán phân tích quan điểm mức khía cạnh của người dùng có thể là đầu vào hữu ích cho các hệ thống hỗ trợ ra quyết định, các hệ thống phân tích và dự đoán thì trường, các hệ thống phân tích dự đoán xu thế tiêu dùng của xã hội, v.v. Các hệ thống này giúp cho doanh nghiệp giảm chi phí đầu vào, nâng cao chất lượng sản phẩm, nâng cao chất lượng dịch vụ chăm sóc khách hàng, nắm bắt thị trường, tối đa hóa lợi nhuận, định hướng phát triển cho doanh nghiệp hiệu quả.

Những vấn đề còn hạn chế của luận án

Ngoài những đóng góp chính của luận án, luận án vẫn còn một số vấn đề cần tiếp tục nghiên cứu và cải thiện gồm:

Thứ nhất, luận án mới chỉ tập trung nghiên cứu phân tích quan điểm với dữ liệu là các bài nhận xét sản phẩm mà chưa quan tâm đến các bài viết dạng khác như bài đăng trên mạng xã hội, blog về các vấn đề khác (lập trường chính trị, các chủ đề xã hội hiện đại), hoặc dữ liệu dạng email. Ngoài ra luận án cũng cần nghiên cứu các quan điểm người dùng trên dạng dữ liệu hình ảnh, video, .. trên phương tiện trực tuyến.

Thứ hai, một vấn đề còn để ngỏ của phân tích mức khía cạnh là tổng hợp quan

điểm. Mặc dù, nhiệm vụ tổng hợp quan điểm dựa trên kết quả của hai nhiệm vụ trích rút khía cạnh và phân lớp cảm xúc khía cạnh, song vẫn cần có một kết quả cuối cùng hoàn thiện hơn cho toàn bộ nhiệm vụ.

Thứ ba, mặc dù các đóng góp đã có tính hiệu quả nhưng đối với hệ thống ứng dụng thế giới thực vẫn cần có các kết quả tốt hơn nữa. Do vậy, hướng nghiên cứu này vẫn cần được nghiên cứu sâu hơn nữa và có kết quả tốt hơn nữa.

Hướng nghiên cứu tiếp theo

Từ những kết quả nghiên cứu đã được thực hiện và các hạn chế đã được chỉ ra, nghiên cứu sinh đề xuất một số nghiên cứu mở rộng như sau:

Thứ nhất, thực hiện các nghiên cứu tổng hợp quan điểm từ các kết quả đã công bố của luận án.

Thứ hai, mở rộng phạm vi nghiên cứu trên các dạng bài viết quan điểm khác ngoài dạng bài viết đánh giá sản phẩm trên phương tiện trực tuyến.

Thứ ba, nghiên cứu sâu hơn các phương pháp học máy để có thể kết hợp các phương pháp học khác nhau nhằm cải thiện hiệu suất tổng thể của hệ thống trong nhiệm vụ đặt ra.

CÁC CÔNG TRÌNH CÔNG BỐ

CT1. Nguyễn Thị Ngọc Tú, Nguyễn Thị Thu Hà, Nguyễn Long Giang, Nguyễn Việt Anh, Nguyễn Trần Quốc Vinh. "*Một phương pháp phân loại đa lớp hiệu quả trong phân tích quan điểm*". Hội nghị quốc gia lần thứ XV "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin", HÀ NỘI, 11/2022, ISBN: 978-604-357-119-6 p517-526

CT2. Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen. "*Language model combined with word2vec for product's aspect based extraction*". ICIC Express Letters, Volume 14, Number 11, 2020, ISSN 1881-803X P1033-1040 (SCOPUS).

CT3. Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen. "*Mining Aspects of Customer's Review on the Social Network*". Journal of Big Data, Volume6, Issue 1, 12/2019, ISSN: 2196-1115 (SCOPUS - Q1).

CT4. Nguyễn Thị Ngọc Tú, Bùi Khánh Linh, Nguyễn Thị Thu Hà, Nguyễn Việt Anh, Nguyễn Ngọc Cương. "*Trích rút khía cạnh sản phẩm dựa trên mô hình ngôn ngữ kết hợp với Word2Vec*". Hội thảo quốc gia lần thứ XXI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, Thanh Hóa, 27-28/7/2018, ISBN: 978-604-67-1104-9 P343 - 349.

CT5. Nguyễn Thị Ngọc Tú, Nguyễn Đức Long, Nguyễn Khắc Giáo, Nguyễn Thị Thu Hà, Nguyễn Việt Anh. "*Một phương pháp phân tích quan điểm đánh giá của người dùng đối với chất lượng sản phẩm dựa trên các nhận xét cá nhân*". Hội nghị quốc gia lần thứ X "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin", ĐÀ NẴNG, 8/2017, ISBN: 978-604–913-614-6 p585-594.

TÀI LIỆU THAM KHẢO

[1] M Rushdi Saleh, Maria Teresa Martín-Valdivia, Arturo Montejo-Ráez, and LA Ureña-López. Experiments with svm to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804, 2011. 1

[2] Octavian Popescu and Carlo Strapparava. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13, 2014. 1

[3] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012. 1, 2, 4, 8, 9, 10, 13, 15, 17, 18, 19, 20, 24, 40, 57, 68, 90

[4] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015. 2, 8, 17, 23, 24, 40, 42, 43, 68

[5] Toqir A Rana and Yu-N Cheah. Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, 46(4):459–483, 2016. 2, 18, 20, 47

[6] Fatemeh Hemmatian and Mohammad Karim Sohrabi. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review*, 52(3):1495–1545, 2019. 17, 43

[7] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021. 1, 3, 16, 17, 20, 23, 40

[8] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 417–424. ACL, 2002. 1, 14, 53, 57, 91

[9] Rui Xia, Feng Xu, Jianfei Yu, Yong Qi, and Erik Cambria. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1):36–45, 2016. 2, 42, 43

[10] Nana Li, Shuangfei Zhai, Zhongfei Zhang, and Boying Liu. Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3490–3496, 2017. 1, 14

[11] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004. 1, 14, 15, 17, 18, 21, 41

[12] Fangzhao Wu, Jia Zhang, Zhigang Yuan, Sixing Wu, Yongfeng Huang, and Jun Yan. Sentence-level sentiment classification with weak supervision. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 973–976, 2017.

[13] Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124, 2016. 14

[14] Xianghua Fu, Wangwang Liu, Yingying Xu, Chong Yu, and Ting Wang. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In *Asian conference on machine learning*, pages 17–32. PMLR, 2016. 1

[15] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830, 2015. 2

[16] Anh-Dung Vo, Quang-Phuoc Nguyen, and Cheol-Young Ock. Opinion–aspect relations in cognizing customer feelings via reviews. *IEEE Access*, 6:5415–5426, 2018. 2, 14, 21, 37, 39, 67, 68

[17] Qiyun Zhao, Hao Wang, Pin Lv, and Chen Zhang. A bootstrapping based refinement framework for mining opinion words and targets. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1995–1998, 2014. 21, 42

[18] Kim Schouten, Nienke De Boer, Tjian Lam, Marijtje Van Leeuwen, Ruud Van Luijk, and Flavius Frasincar. Semantics-driven implicit aspect detection in consumer reviews. In *Proceedings of the 24th International Conference on World Wide Web*, pages 109–110, 2015. 21, 42, 68

[19] Ya Lin Miao, Wen Fang Cheng, Yi Chun Ji, Shun Zhang, and Yan Long Kong. Aspect-based sentiment analysis in chinese based on mobile reviews for bilstm-crf. *Journal of Intelligent & Fuzzy Systems*, 40(5):8697–8707, 2021. 14, 21, 37, 39, 42, 68

[20] Zarmeen Nasim and Sajjad Haider. Absa toolkit: An open source tool for aspect

based sentiment analysis. *International Journal on Artificial Intelligence Tools*, 26(06):1750023, 2017. 67

[21] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*, 2016. 21, 42, 68

[22] Binxuan Huang and Kathleen M Carley. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 1091–1096, 2018. 21

[23] Avinash Kumar, Vishnu Teja Narapareddy, Veerubhotla Aditya Srikanth, Lalita Bhanu Murthy Neti, and Aruna Malapati. Aspect-based sentiment classification using interactive gated convolutional network. *IEEE Access*, 8:22445–22453, 2020. 21, 39

[24] Ying Ding, Changlong Yu, and Jing Jiang. A neural network model for semi-supervised review aspect identification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 668–680. Springer, 2017. 21, 42, 68

[25] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, 2017. 42, 68

[26] Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. Can: Constrained attention networks for multi-aspect sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4600–4609. Association for Computational Linguistics, 2019. 2, 21, 39

[27] P Vijayaragavan, R Ponnusamy, and M Aramudhan. An optimal support vector machine based classification model for sentimental analysis of online product reviews. *Future Generation Computer Systems*, 111:234–240, 2020. 2, 23, 37, 42, 43

[28] Xiaojia Pu, Gangshan Wu, and Chunfeng Yuan. Exploring overall opinions for document level sentiment classification with structural svm. *Multimedia Systems*, 25(1):21–33, 2019. 42, 43

[29] Madiha Khalid, Imran Ashraf, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, and GyuSang Choi. Gbsvm: sentiment classification from unstructured reviews using ensemble classifier. *Applied Sciences*, 10(8):2788, 2020. 23, 37, 43, 80

[30] Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. Twitter feature selection and classification using support vector machine for aspect-based sentiment analysis. In *International conference on industrial, engineering and other applications of applied intelligent systems*, pages 269–279. Springer, 2016. 43

[31] Harshali P Patil and Mohammad Atique. Cdnb: Caviar-dragonfly optimization with naive bayes for the sentiment and affect analysis in social media. *Big data*, 8(2):107–124, 2020. 23, 43

[32] Xin Xie, Songlin Ge, Fengping Hu, Mingye Xie, and Nan Jiang. An improved algorithm for sentiment analysis based on maximum entropy. *Soft Computing*, 23(2):599–611, 2019. 37, 42, 43

[33] Phu Vo Ngoc, Chau Vo Thi Ngoc, Tran Vo THi Ngoc, and Dat Nguyen Duy. A c4. 5 algorithm for english emotional classification. *Evolving Systems*, 10(3):425–451, 2019. 23, 42, 43

[34] Yue Han, Yuhong Liu, and Zhigang Jin. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications*, 32(9):5117–5129, 2020. 2, 23

[35] Feng Wang and Li Chen. Review mining for estimating users' ratings and weights for product aspects. In *Web Intelligence*, volume 13, pages 137–152. IOS Press, 2015. 2, 37, 40, 43

[36] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792, 2010. 14, 21, 23, 40, 43, 55, 57, 64, 65

[37] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626, 2011. 33, 37, 43

[38] Roman Klinger and Philipp Cimiano. Joint and pipeline probabilistic models for fine-grained sentiment analysis: Extracting aspects, subjective phrases and

their relations. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 937–944. IEEE, 2013. 2, 14, 37, 67, 68

[39] Vaishali Ganganwar and R Rajalakshmi. Implicit aspect extraction for sentiment analysis: a survey of recent approaches. *Procedia Computer Science*, 165:485–491, 2019. 2

[40] Yoav Goldberg and Graeme Hirst. *Neural network methods in natural language processing. morgan & claypool publishers (2017)*. Morgan & Claypool, 2017. 2, 3, 25

[41] Mondher Bouazizi and Tomoaki Ohtsuki. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, 2(3):181–194, 2019. 3, 43, 80

[42] Arjun Chaudhuri. *Emotion and reason in consumer behavior*. Routledge, 2006. 8

[43] Huaxia Rui, Yizao Liu, and Andrew Whinston. Whose and what chatter matters? the effect of tweets on movie sales. *Decision support systems*, 55(4):863–870, 2013. 8

[44] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007. 8, 15, 21, 41, 67, 68

[45] Yung-Ming Li and Tsung-Ying Li. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206–217, 2013. 8

[46] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005. 9, 21, 41

[47] Bing Liu. *Web data mining*, volume 1. Springer, 2011. 10, 11, 12, 15

[48] K Vivekanandan and J Soonu Aravindan. Aspect-based opinion mining: A survey. *International Journal of Computer Applications*, 106(3), 2014. 13, 14

[49] Xianghua Fu, Wangwang Liu, Yingying Xu, and Laizhong Cui. Combine hownet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. *Neurocomputing*, 241:18–27, 2017. 14

[50] Siti Rohaidah Ahmad, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. A review of feature selection techniques in sentiment analysis. *Intelligent data analysis*, 23(1):159–189, 2019. 15, 16

[51] Ling Zhao, Ying Liu, Mingyao Zhang, Tingting Guo, and Lijiao Chen. Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model. *Information Processing & Management*, 58(5):102641, 2021. 15

[52] Vincent Ng, Sajib Dasgupta, and SM Niaz Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL 2006 main conference poster sessions*, pages 611–618, 2006. 15

[53] Rajeev Kumar and Jasandeep Kaur. Random forest-based sarcastic tweet classification using multiple feature collection. In *Multimedia Big Data Computing For IoT Applications*, pages 131–160. Springer, 2020. 16

[54] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010. 16, 24

[55] Nazrul Hoque, Dhruba K Bhattacharyya, and Jugal K Kalita. Mifs-nd: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14):6371–6385, 2014. 16, 17

[56] Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E Alsaadi. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86:105836, 2019. 16

[57] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018. 16, 17

[58] Haoyue Liu, MengChu Zhou, and Qing Liu. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3):703–715, 2019. 16

[59] Siti Rohaidah Ahmad, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. Ant colony optimization for text feature selection in sentiment analysis. *Intelligent Data Analysis*, 23(1):133–158, 2019. 17

[60] Chong Long, Jie Zhang, and Xiaoyan Zhu. A review selection approach for accurate feature rating estimation. In *Coling 2010: Posters*, pages 766–774, 2010. 21, 41, 61, 64, 77

[61] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Coling 2010: posters*, pages 1462–1470, 2010. 21, 67, 68

[62] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37, 2014. 21, 22, 41

[63] Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1496–1505, 2011. 21, 43, 65

[64] Zhijun Yan, Meiming Xing, Dongsong Zhang, and Baizhang Ma. Exprs: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, 52(7):850–858, 2015. 21, 42, 68

[65] Baizhang Ma, Dongsong Zhang, Zhijun Yan, and Taeha Kim. An lda and synonym lexicon based approach to product feature extraction from online consumer product reviews. *Journal of Electronic Commerce Research*, 14(4):304, 2013. 21

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 21

[67] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 21, 23

[68] Ning Liu and Bo Shen. Aspect term extraction via information-augmented neural network. *Complex & Intelligent Systems*, 9(1):537–563, 2022. 21

[69] Guoshuai Zhao, Yiling Luo, Qiang Chen, and Xueming Qian. Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, page 110326, 2023. 24

[70] Manju Venugopalan and Deepa Gupta. An enhanced guided lda model augmented with bert based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-based systems*, 246:108668, 2022. 21

[71] Reinald Kim Amplayo, Seanie Lee, and Min Song. Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis. *Information Sciences*, 454:200–215, 2018. 22, 37, 67, 68

[72] Christina Sauper and Regina Barzilay. Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*, 46:89–127, 2013.

[73] Shehzad Khalid, Muhammad Haseeb Aslam, and Muhammad Taimoor Khan. Opinion reason mining: Implicit aspects beyond implying aspects. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5. IEEE, 2018. 22, 37, 67, 68

[74] Zhao Fang, Qiang Zhang, Xiaoan Tang, Anning Wang, and Claude Baron. An implicit opinion analysis model based on feature-based implicit opinion patterns. *Artificial Intelligence Review*, 53(6):4547–4574, 2020. 22

[75] Hen-Hsen Huang, Jun-Jie Wang, and Hsin-Hsi Chen. Implicit opinion analysis: Extraction and polarity labelling. *Journal of the Association for Information Science and Technology*, 68(9):2076–2087, 2017. 22

[76] Muhammad Afzaal, Muhammad Usman, and Alvis Fong. Tourism mobile app with aspect-based sentiment classification framework for tourist reviews. *IEEE Transactions on Consumer Electronics*, 65(2):233–242, 2019. 22

[77] Kim Schouten, Onne Van Der Weijde, Flavius Frasincar, and Rommert Dekker. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275, 2017. 22

[78] Sanjay Chatterji, Nitish Varshney, and Ranjan Kumar Rahul. Aspectframenet: a framenet extension for analysis of sentiments around product aspects. *The Journal of Supercomputing*, 73(3):961–972, 2017. 22

[79] Murtadha Ahmed, Shengfeng Pan, Jianlin Su, Xinxin Cao, Wenze Zhang, Bo Wen, and Yunfeng Liu. Bert-asc: Implicit aspect representation learning through auxiliary-sentence construction for sentiment analysis. *arXiv preprint arXiv:2203.11702*, 2022. 22

[80] Jibran Mir, Azhar Mahmood, and Shaheen Khatoon. Multi-level knowledge engineering approach for mapping implicit aspects to explicit aspects. *CMC-COMPUTERS MATERIALS & CONTINUA*, 70(2):3491–3509, 2022.

[81] Li Yu and Xuefei Bai. Implicit aspect extraction from online clothing reviews with fine-tuning bert algorithm. In *Journal of Physics: Conference Series*, volume 1995, page 012040. IOP Publishing, 2021. 22

[82] Toqir A Rana, Yu-N Cheah, and Tauseef Rana. Multi-level knowledge-based approach for implicit aspect identification. *Applied Intelligence*, 50(12):4616–4630, 2020. 22

[83] Jinzhan Feng, Shuqin Cai, and Xiaomeng Ma. Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm. *Cluster Computing*, 22(3):5839–5857, 2019.

[84] Mohammad Tubishat and Norisma Idris. Explicit and implicit aspect extraction using whale optimization algorithm and hybrid approach. In *2018 International Conference on Industrial Enterprise and System Engineering (ICoIESE 2018)*, pages 208–213. Atlantis Press, 2019. 22

[85] Oscar Araque, Ignacio Corcuera-Platas, J Fernando Sánchez-Rada, and Carlos A Iglesias. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246, 2017. 23, 37, 42, 43

[86] Yun Wan and Qigang Gao. An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1318–1325. IEEE, 2015. 23, 43, 80

[87] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6):1134–1145, 2011. 23

[88] Sumbal Riaz, Mehvish Fatima, Muhammad Kamran, and M Wasif Nisar. Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster Computing*, 22(3):7149–7164, 2019. 23

[89] Wei Gao, Shoushan Li, Yunxia Xue, Meng Wang, and Guodong Zhou. Semi-supervised sentiment classification with self-training on feature subspaces. In *Workshop on Chinese Lexical Semantics*, pages 231–239. Springer, 2014. 23

[90] Vincent Van Asch and Walter Daelemans. Predicting the effectiveness of self-training: Application to sentiment classification. *arXiv preprint arXiv:1601.03288*, 2016. 23

[91] Weifeng Liu, Lianbo Zhang, Dapeng Tao, and Jun Cheng. Reinforcement online learning for emotion prediction by using physiological signals. *Pattern Recognition Letters*, 107:123–130, 2018. 23

[92] Lei Liu, Hao Chen, and Yinghong Sun. A multi-classification sentiment analysis model of chinese short text based on gated linear units and attention mechanism. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–13, 2021. 23, 80

[93] Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. In *arXiv preprint arXiv:1804.00831*, page 141–145. Association for Computational Linguistics, 2018. 80

[94] Chunyi Yue, Hanqiang Cao, Guoping Xu, and Youli Dong. Collaborative attention neural network for multi-domain sentiment classification. *Applied Intelligence*, 51(6):3174–3188, 2020. 37, 42, 43

[95] Weijiang Li, Fang Qi, Ming Tang, and Zhengtao Yu. Bidirectional lstm with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*, 387:63–77, 2020. 23, 37, 43

[96] Bin Liang, Xiang Li, Lin Gui, Yonghao Fu, Yulan He, Min Yang, and Ruifeng Xu. Few-shot aspect category sentiment analysis via meta-learning. *ACM Transactions on Information Systems*, 41(1):1–31, 2023. 24

[97] Olha Kaminska, Chris Cornelis, and Veronique Hoste. Fuzzy rough nearest neighbour methods for aspect-based sentiment analysis. *Electronics*, 12(5):1088, 2023.

[98] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, Hua Jin, and Dacheng Tao. Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 24

[99] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013. 24

[100] Hongyu Han, Jianpei Zhang, Jing Yang, Yiran Shen, and Yongshi Zhang. Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications*, 77(16):21265–21280, 2018. 24

[101] Swati Sanagar and Deepa Gupta. Unsupervised genre-based multidomain sentiment lexicon learning using corpus-generated polarity seed words. *IEEE Access*, 8:118050–118071, 2020.

[102] Muhammad Zubair Asghar, Anum Sattar, Aurangzeb Khan, Amjad Ali, Fazal Masud Kundi, and Shakeel Ahmad. Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language. *Expert Systems*, 36(3):e12397, 2019. 24

[103] Itisha Gupta and Nisheeth Joshi. Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. *Journal of intelligent systems*, 29(1):1611–1625, 2019. 24

[104] DV Devi, Thatiparti Venkata Rajini Kanth, Kakollu Mounika, and Nambhatla Sowjanya Swathi. Assay: Hybrid approach for sentiment analysis. In *Information and Communication Technology for Intelligent Systems*, pages 309–318. Springer, 2019. 24

[105] Kariman Elshakankery and Mona F Ahmed. Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis. *Egyptian Informatics Journal*, 20(3):163–171, 2019. 24

[106] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982. 24

[107] Geoffrey E Hinton. Distributed representations. *Technical Report CMU-CS-84-157*, 1984. 25

[108] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. 26

[109] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 28

[110] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988. 30, 31, 79, 84

[111] Fabio Gagliardi Cozman. Axiomatizing noisy-or. In *ECAI*, volume 16, page 979, 2004. 31

[112] Kuang Zhou, Arnaud Martin, and Quan Pan. The belief noisy-or model applied to network reliability analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24(06):937–960, 2016. 31, 84

[113] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999. 32

[114] David L Olson and Dursun Delen. *Advanced data mining techniques*. Springer Science & Business Media, 2008. 32

[115] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Multi-facet rating of product reviews. In *European conference on information retrieval*, pages 461–472. Springer, 2009. 33

[116] Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. Aspect sentiment classification with both word-level and clause-level attention networks. In *IJCAI*, volume 2018, pages 4439–4445, 2018. 35

[117] Chuang Fan, Qinghong Gao, Jiachen Du, Lin Gui, Ruifeng Xu, and Kam-Fai Wong. Convolution-based memory network for aspect-based sentiment analysis. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pages 1161–1164, 2018. 35

[118] Dangguo Shao, Qing An, Kun Huang, Yan Xiang, Lei Ma, Junjun Guo, and Runda Yin. Aspect-level sentiment analysis for based on joint aspect and position hierarchy attention mechanism network. *Journal of Intelligent and Fuzzy Systems*, 42(Preprint):1–12, 2022. 37, 67, 68

[119] Yan Li, Hui Wang, Zhen Qin, Weiran Xu, and Jun Guo. Confidence estimation and reputation analysis in aspect extraction. In *2014 22nd international conference on pattern recognition*, pages 3612–3617. IEEE, 2014. 41, 42

[120] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4194–4200. ijcai.org, 2018. 42, 67, 68

[121] Jianfei Yu, Jing Jiang, and Rui Xia. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(1):168–177, 2019. 42, 67, 68

[122] Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375, 2020. 42

[123] Ayoub Bagheri, Mohamad Saraee, and Franciska de Jong. An unsupervised aspect detection model for sentiment analysis of reviews. In *International conference on application of natural language to information systems*, pages 140–151. Springer, 2013. 42

[124] Zhen Hai, Kuiyu Chang, and Gao Cong. One seed to find them all: mining opinion features via association. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 255–264, 2012. 42

[125] Ana Valdivia, M Victoria Luzón, and Francisco Herrera. Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, 32(4):72–77, 2017. 43

[126] Parisa Jamadi Khiabani, Mohammad Ehsan Basiri, and Hamid Rastegari. An improved evidence-based aggregation method for sentiment analysis. *Journal of Information Science*, 46(3):340–360, 2020. 43, 80

[127] Kai Gao, Hua Xu, and Jiushuo Wang. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517–4528, 2015. 43, 80

[128] Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. Emotion classification of online news articles from the reader's perspective. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226. IEEE, 2008. 43

[129] Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Omar Qawasmeh. Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2):308–319, 2018. 43

[130] Dhvani Kansara and Vinaya Sawant. Comparison of traditional machine learning and deep learning approaches for sentiment analysis. In *Advanced computing technologies and applications*, pages 365–377. Springer, 2020. 43

[131] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. Show me the money! deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65, 2007. 43

[132] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE, 2012. 55, 62

[133] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 62, 67, 77

[134] Charu Gupta, Amita Jain, and Nisheeth Joshi. A novel approach to feature hierarchy in aspect based sentiment analysis using owa operator. In *Proceedings of 2nd International Conference on Communication, Computing and Networking*, pages 661–667. Springer, 2019. 67, 68

[135] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 73, 74

[136] LM de Campos, JM Fernández-Luna, JF Huete, and AE Romero. Or gate bayesian networks for text classification: a discriminative alternative approach to multinomial naive bayes. In *Actas del XIV Congreso Espanol sobre Tecnologıas y Lógica Fuzzy*, pages 385–390, 2008. 79, 84

[137] Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan. A method for multi-class sentiment classification based on an improved one-vs-one (ovo) strategy and the support vector machine (svm) algorithm. *Information Sciences*, 394:38–52, 2017. 80, 84

[138] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012. 80

[139] Lior Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33(1):1–39, 2010. 80

[140] Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik. Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. *IEEE access*, 8:101840–101858, 2020. 80

[141] Jordan J Bird, Anikó Ekárt, Christopher D Buckingham, and Diego R Faria. High resolution sentiment analysis by ensemble classification. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 593–606. Springer, 2019. 80

[142] Naznin Sultana and Mohammad Mohaiminul Islam. Meta classifier-based ensemble learning for sentiment classification. In *Proceedings of International Joint Conference on Computational Intelligence*, pages 73–84. Springer, 2020. 80

[143] Yaxin Bi, Jiwen Guan, and David Bell. The combination of multiple classifiers using an evidential reasoning approach. *Artificial Intelligence*, 172(15):1731–1751, 2008. 80

[144] Mohammad Ehsan Basiri, Ahmad Reza Naghsh-Nilchi, and Nasser Ghasem-Aghaee. Sentiment prediction based on dempster-shafer theory of evidence. *Mathematical Problems in Engineering*, 2014, 2014. 80

[145] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 83

[146] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics.*, 38:325–339, 1967. 85

[147] Kari Sentz and Scott Ferson. Combination of evidence in dempster-shafer theory. *Sandia Report*, 2002. 86