BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ NGOC TÚ

NGHIÊN CỨU CẢI TIẾN MỘT SỐ PHƯƠNG PHÁP PHÂN TÍCH QUAN ĐIỂM MỨC KHÍA CẠNH DỰA TRÊN HỌC MÁY

Chuyên ngành: Hệ thống thông tin

Mã số: 9 48 01 04

TÓM TẮT LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

HÀ NỘI - 2023

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Người hướng dẫn khoa học: PGS.TS Nguyễn Việt Anh

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện, họp tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi … giờ …, ngày … tháng … năm 2023.

Có thể tìm hiểu luận án tại:

- Thư viện Học viện Khoa học và Công nghệ
- Thư viện Quốc gia

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Trong thời đại công nghệ thông tin phát triển hiện nay, 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 trên các trang web mua sắm hoặc các trang mạng xã hội. 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, 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.

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. Trong phân tích 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. 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.

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). Mặc dù nhiều nghiên cứu, nhiều ứng dụng đã được thực hiện trong phân tích quan điểm mức khía cạnh, nhưng lĩnh vực này vẫn còn nhiều thách thức cần vượt qua.

- ***Đối với nhiệm vụ trích rút khía cạnh:*** Khó khăn đầu tiên là thiếu dữ liệu huấn luyện có gán nhãn trong nhiệm vụ này. Thứ hai, nhiều câu đánh giá thiếu các thể hiện khía cạnh rõ ràng (danh từ) dẫn đến khó xác định khía cạnh. Ngoài ra, có nhiều cách thức ám chỉ các khía cạnh (đặc trưng ẩn) xuất hiện khiến nhiệm vụ khai phá càng phức tạp, bởi phải xác định đặc trưng ẩn nào gán với khía cạnh nào. Thứ ba, khi một từ xuất hiện cần xem xét ngữ cảnh của nó. Đối với nhiều từ cách giải thích 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. Thứ tư, một số khía cạnh quan trọng nhưng có tần suất xuất hiện thấp dễ bị bỏ qua. 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 nhiệm vụ phân lớp cảm xúc khía cạnh:*** Thứ nhất, nhiệm vụ phân loại cảm xúc đa lớp có nhiều thách thức hơn so với phân loại hai lớp. Sự hiện diện của nhiều lớp làm cho một bộ phân loại khó xác định

biên giới giữa các lớp khác nhau hơn. Thứ hai, sự gần gũi giữa các lớp cảm xúc hoặc giữa các lớp có cùng cực cảm xúc gần như là tương tự nhau và chúng rất dễ bị phân loại nhầm lẫn nhau. Thứ ba, 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, song trong ngữ cảnh nói về tốc độ xử lý của CPU thì lại mang tính tiêu cực. Cuối cùng, sự hiện diện của phủ định có thể đảo ngược cực cảm xúc của một văn bản. Tuy nhiên, không dễ để xử lý điều 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 văn bản.

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 thực tế. 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. Tác giả luận án lựa chọn đề tài "*Nghiên cứu phát triển một số thuật toán học máy trong dự báo kinh tế*".

2. Mục tiêu nghiên cứu

Mục tiêu của luận án là đề xuất một hệ thống thực hiện ba nhiệm vụ của bài toán phân tích quan điểm mức khía cạnh đánh giá sản phẩm trực tuyến. Từ đó, nghiên cứu sinh đề xuất một số thuật toán học máy bán giám sát để trích rút khía cạnh và quan điểm, đề xuất một số thuật toán học máy có giám sát để giải quyết nhiệm vụ phân lớp quan điểm đã được trích rút từ nhiệm vụ đầu thành các cực cảm xúc khác nhau, đề xuất một cách tiếp cận mới để ước lượng trọng số khía cạnh mà người dùng đặt lên mỗi khía cạnh.

3. Các nội dung nghiên cứu

Luận án nghiên cứu các vấn đề trong phân tích quan điểm và bài toán phân tích quan điểm mức khía cạnh. Luận án nghiên cứu các phương pháp học máy truyền thống và hiện đại, đề xuất 02 thuật toán bán giám sát để trích rút khía cạnh và quan điểm từ các bài đánh giá sản phẩm trực tuyến. Thuật toán thứ nhất dựa trên xác suất có điều kiện kết hợp giải thuật bootstrapping, thuật toán thứ hai dựa trên biểu diễn WordtoVector kết hợp mô hình ngôn ngữ. Nghiên cứu sinh cũng đề xuất các phương pháp học Naïve Bayes, Support Vector Machine, mạng Bayesian cổng OR, lý thuyết kết hợp Dempster-Shafer cho nhiệm vụ phân lớp cảm xúc khía cạnh. Một phương pháp học không giám sát dựa trên nội dung bài đánh giá được đề xuất cho nhiệm vụ ước lượng trọng số khía cạnh.

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

1.1.1. Các khái niệm cơ bản

Định nghĩa 1.6 Quan điểm (opinion): 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 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.

1.1.2. Các nhiệm vụ trong phân tích quan điểm

Bài toán 1 (trích rút và phân loại thực thể

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 của thực thể e_i đại diện điển hình một khía cạnh đơn nhất a_{ij}.

Bài toán 3 (trích rút và phân loại chủ sở hữu quan điểm

Bài toán 4 (trích rút và chuẩn hóa thời gian)

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. Đây là nhiệm vụ dường như rất là đơn giản nhưng trong thực tế nó rất khó khăn trong một vài trường hợp.

1.1.3. Các mức độ phân tích quan điểm

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.

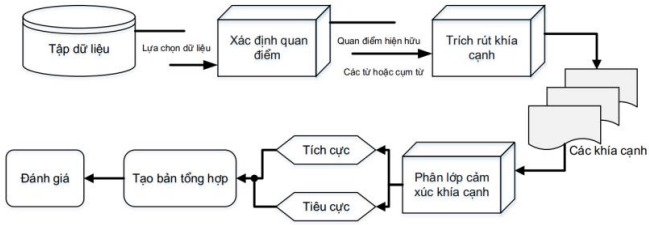
Mức độ câu: là một phân tích chi tiết của mức văn bản, trong đó xác định tính phân cực cho mỗi câu và mỗi câu có thể có quan điểm khác nhau.

Mức độ cụm từ: 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.

1.1.4. Vấn đề đặc trưng trong phân tích quan điểm

1.2 Phân tích quan điểm mức khía cạnh

1.2.1. Quy trình phân tích quan điểm mức khía cạnh



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

Bài toán trích rút khía cạnh: xác định tất cả các thuật ngữ khía cạnh có trong mỗi câu của bài đánh giá hoặc toàn bộ văn bản bài đánh giá.

Bài toán phân lớp cảm xúc dựa trên khía cạnh: cho một khía cạnh, xác định cực của từng thuật ngữ khía cạnh hoặc toàn bộ khía cạnh đó.

Bài toán xác định trọng số khía cạnh: Nhiệm vụ này xác định các khía cạnh quan trọng đánh giá tổng thể mà người dùng đưa ra.

1.2.3. Các cách tiếp cận trích rút khía cạnh

1.2.3.1 Các phương pháp trích rút khía cạnh rõ ràng

Các phương pháp 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.

- Trích rút khía cạnh rõ ràng với học không giám sát bao gồm phương pháp *dựa trên tần suất và thống kê*, phương pháp *dựa trên kinh nghiệm hoặc dựa trên luật*, và phương pháp *dựa trên điểm thông tin tương hỗ*.
- Trích rút khía cạnh rõ ràng với học bán giám sát bao gồm phương pháp sử dụng *Bootstrapping*, phương pháp *phân tích cú pháp phụ thuộc*, phương pháp *dựa trên từ điển*.
- Trích rút khía cạnh rõ ràng với học giám sát bao gồm các *mô hình Markov ẩn* (HMM), *trường ngẫu nhiên có điều kiện* (CRF), *mạng nơ ron hồi quy* (RNN), *mạng nơ ron tích chập* (CNN).

1.2.3.2 Các phương pháp trích rút khía cạnh ẩn

Các phương pháp trích rút khía cạnh ẩn có thể phân thành các phương pháp học không giám sát, có giám sát, và cách tiếp cận lai.

- Trích rút khía cạnh ẩn với học không giám sát bao gồm các phương pháp *dựa trên sự đồng xuất hiện*, phương pháp *dựa trên mô hình chủ đề*, phương pháp *dựa trên phân cụm*.
- Trích rút khía cạnh ẩn với học có giám sát bao gồm các phương pháp dựa trên phân lớp, dựa trên luật, dựa trên nhãn tuần tự.
- Trích rút khía cạnh ẩn theo cách tiếp cận lai là cách kết hợp của nhiều phương pháp khác nhau.

1.2.4. Các phương pháp phân lớp cảm xúc khía cạnh

Các cách tiếp cận hiện nay cho nhiệm vụ phân lớp cảm xúc có thể được phân loại thành cách tiếp cận học máy, cách tiếp cận dựa trên từ điển, và các phương pháp lai.

Các phương pháp phân lớp cảm xúc dựa trên học máy: gồm có học có giám sát, học không giám sát, học bán giám sát, học tăng cường, và học sâu.

- Phân lớp cảm xúc theo cách tiếp cận học có giám sát được phân thành 4 loại: *tuyến tính*, *dựa trên xác suất*, *dựa trên quy tắc*, và *cây quyết định*.
- Phân lớp cảm xúc theo cách tiếp cận học không có giám sát bao gồm các kỹ thuật *phân cụm phân cấp* và *phân cụm theo vùng*.
- Phân lớp cảm xúc theo cách tiếp cận học bán giám sát được phân thành *học tổng quát*, *học đồng huấn luyện*, *huấn luyện chọn lọc*, *học dựa trên đồ thị*, và *học đa quan điểm*.
- Phân lớp cảm xúc theo cách tiếp cận học tăng cường là phương pháp trong đó tác nhân được thưởng trong bước thời gian tiếp theo dựa trên đánh giá về hành động trước đó của nó.
- Phân lớp cảm xúc theo cách tiếp cận học sâu là dựa trên mạng ANN bao gồm các mô hình *mạng nơ ron hồi quy* (RNN), *mạng nơ ron tích chập* (CNN), và *mạng niềm tin sâu* (DBN).

Các phương pháp dựa trên từ điển: còn được gọi là cách tiếp cận dựa trên tri thức. Có ba kỹ thuật chính để tạo các từ điển chú thích là phương pháp thủ công, phương pháp dựa trên từ điển và phương pháp dựa trên kho ngữ liệu.

- Xây dựng từ vựng cảm xúc dựa trên phương pháp thủ công sử dụng từ đồng nghĩa trái nghĩa và dựa trên một từ điển có sẵn.
- Xây dựng từ vựng cảm xúc dựa trên phương pháp kho ngữ liệu, từ điển được học từ dữ liệu với cách tiếp cận thống kê và ngữ nghĩa.
- Các phương pháp lai kết hợp cả cách tiếp cận từ vựng và học máy.

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*

1.3.2. *Cơ sở lý thuyết biểu diễn từ Word to Vector*

1.3.3. *Phân loại hai lớp máyvec tơ hỗ trợ*

1.3.4. *Phân loại đa lớp Naive Bayes*

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

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

CHƯƠNG 2 KHAI PHÁ QUAN ĐIỂM MỨC KHÍA CẠNH

2.1 Đặt vấn đề

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.

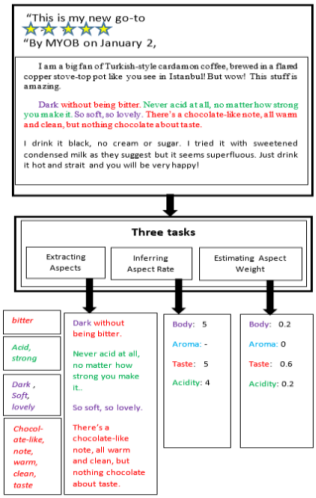
Làm thế nào để hiểu nội dung bài đánh giá và các vấn đề mà người dùng đề cập? Phân tích quan điểm dựa trên khía cạnh giải quyết vấn đề phân tích chi tiết trên những khía cạnh của sản phẩm mà người dùng đã đề cập đến trong bài đánh giá của họ. Mức độ chi tiết là người dùng đã đề cập đến những khía cạnh nào trong bài đánh giá của họ, độ hài lòng/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.

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 nhiệm vụ (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.

Bài toán trích rút khía cạnh là xác định tất cả các khía cạnh xuất hiện trong bài đánh giá. Trong nhiệm vụ này có một số thách thức như sau: Một số khía cạnh được đề cập rõ ràng và một số khác thì không. Cần trích rút được khía cạnh ẩn. Giải quyết vấn đề nhiễu (các thuật ngữ phi khía cạnh) trong khi vẫn có thể xác định các khía cạnh hiếm và quan trọng.

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 vì số lượng khía cạnh quan trọng thường nhỏ và có thể dễ dàng thu được từ các chuyên gia miền lĩnh vực. Sau đó nhiệm vụ trích rút khía cạnh trở thành xác định chính xác các khía cạnh hiện có cho các câu/phần văn bản trong bài đánh giá. Thách thức chính ở đây là trong nhiều bài đánh giá, các câu không chứa đủ các từ lõi khía cạnh, thậm chí không có bất kỳ từ lõi khía cạnh nào, và do đó có thể bị gán cho các nhãn khía cạnh sai. Vấn đề này được giải quyết bằng cách liên tục cập nhật và mở rộng các từ lõi khía cạnh thành tập các từ khía cạnh bằng cách sử dụng kỹ thuật xác suất có điều kiện kết hợp bootstrap. Bộ phân lớp Naive Bayes được sử dụng để giải quyết vấn đề phân lớp cảm xúc khía cạnh sau khi đã được trích rút. 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. Luận án đề xuất một cách tiếp cận ước lượng trọng số của khía cạnh 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. Hình 2.2 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

2.2.2 Phân lớp cảm xúc

2.2.3 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 số sản phẩm được ký hiệu d_i. có nhiều câu, mỗi câu chứa nhiều từ w_j trong tập hợp của tất cả các từ có thể có.

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

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 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. 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 ký hiệu là $a_k = \{w|w \in V, A(w) = a_k\}$, trong đó a_k là tên của khía cạnh, w là một từ thuộc $\mathcal{V}$ và $A(.)$ là một toán tử ánh xạ một từ tới một khía cạnh.

Định nghĩa 2.4 Từ lõi khía cạnh (Aspect Core Words): Cho một khía cạnh a_k, một tập rất ít các từ thuộc $\mathcal{V}$ miêu tả rất rõ ràng khía cạnh a_k được gọi là từ lõi khía cạnh, ký hiệu là $\mathcal{C}_k=\{w_{kj} \in \mathcal{V}|w_{kj} \rightarrow a_k, j = \overline{1,N}\}$, 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.

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 $\mathcal{T}_k = \{w_{kj} \in \mathcal{V}, w_{kj} \notin \mathcal{C}_k|w_{kj} \rightarrow a_k, j = \overline{1,M}\}$. 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 $\mathbf{r}_i = \{r_{i1}, r_{i2}, ..., r_{iK}\}$, trong đó r_{ik} là một giá trị số cho biết đánh giá của người dùng về khía cạnh ak, và $r_{ik} \in [r_{min}, r_{max}]$ (ví dụ r_{ik} 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 $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK}\}$, trong đó α_{ik} là một giá trị số cho biết trọng số đánh giá của người dùng về khía cạnh a_k, và $\alpha_{ik} \in [0, 1]$, và $\sum_{k=1}^{K} \alpha_{ik} = 1$.

Định nghĩa 2.8 Điểm đánh giá tổng thể của bài đánh giá (Review overall Rating): Cho văn bản đánh giá 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 sản phẩm. Giá trị điểm tổng thể này tương tự như điểm đánh giá khía cạnh.

Nhiệm vụ trích rút khía cạnh: Giả định rằng mỗi khía cạnh là một phân phối xác suất trên tất cả các từ và mỗi câu trong văn bản của bài đánh giá có thể đề cập đến nhiều khía cạnh, mục tiêu của nhiệm vụ này là trích rút các khía cạnh được đề cập trong một bài đánh giá.

Nhiệm vụ dự đoán điểm đánh giá khía cạnh: Nhiệm vụ này là suy ra vector $\mathbf{r}_i$ của điểm đánh giá khía cạnh (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 được thể hiện bằng các từ cảm xúc (tích cực hoặc tiêu cực).

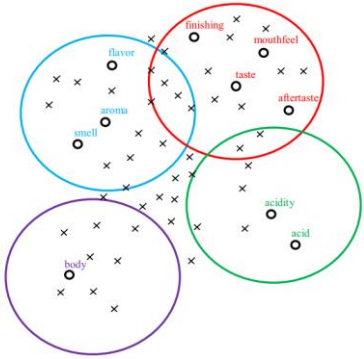
Nhiệm vụ ước lượng trọng số khía cạnh: Nhiệm vụ này là ước tính trọng số không âm α_i mà người dùng đặt trên khía cạnh a_{ik} của văn bản d_i

(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.

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

Nhãn khía cạnh được xác định dựa trên tập hợp các từ có liên quan được gọi là các *từ khía cạnh* hoặc *thuật ngữ khía cạnh* $\mathcal{T}_k$. Giả sử có một số từ khóa được chỉ định để mô tả từng khía cạnh, gọi là *từ lõi khía cạnh* $\mathcal{C}_k$. Giả định rằng tập hợp phổ quát của tất cả các khía cạnh có thể có cho mỗi sản phẩm đều biết trước. Nhiệm vụ trích xuất khía cạnh trở thành xác định chính xác các khía cạnh hiện có cho các câu trong bài đánh giá. Thách thức chính là trong nhiều bài đánh giá, các câu không chứa đủ các từ cốt lõi hoặc thậm chí không có bất kỳ từ cốt lõi nào, do đó có thể bị gán cho các khía cạnh sai. Vấn đề này được giải quyết bằng cách liên tục cập nhật và mở rộng tập các từ cốt lõi thành tập các từ khía cạnh bằng cách sử dụng kỹ thuật xác suất có điều kiện kết hợp với kỹ thuật bootstrap.



Hình 2.4 *Từ lõi với các khía cạnh*

Giả sử rằng $\mathcal{A}$= {a_1, a_2, ..., a_K} là tập của K khía cạnh của sản phẩm. a_k là tập các từ thuộc tính đại diện cho khía cạnh a_k và tần suất xuất hiện của chúng luôn lớn hơn ngưỡng θ. Mục tiêu là thu thập tập các từ mà chúng 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 (xem Hình 2.4). Đầu tiên, các câu có chứa ít nhất một từ trong tập từ cốt lõi $\mathcal{C}_k$ 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 lớn hơn ngưỡng θ được bổ sung vào tập hợp các từ khía cạnh. Các từ có số lần xuất hiện lớn

nhất trong tập từ khía cạnh mới tìm được sẽ được thêm vào tập các từ cốt lõi. Tập các từ khía cạnh và các từ cốt lõi được cập nhật, các từ này được sử dụng để gán nhãn các câu tiếp theo. Quá trình này được lặp lại cho đến khi không tìm thấy thêm từ mới.

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 được coi là các nhãn và các *từ cảm xúc* được xem xét như là các đặc trưng. Ngoài ra một số các đặc trưng bi-gram được trích rút theo các mẫu cú pháp được đề xuất trong.

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

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

Giả định rằng các đặc trưng là độc lập, điểm đánh giá khía cạnh r_{ik} được gán nhãn c khi xác suất $p(r_{ik} \in c | f_1, f_2, ..., f_Q)$ là lớn nhất.

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

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*

Đối với người dùng, nếu một khía cạnh là quan trọng, họ sẽ đề cập nhiều hơn về nó trong bài đánh giá. Hơn nữa, một ý tưởng rằng một khía cạnh quan trọng thường được nhiều người dùng chia sẻ. Số đo trọng số của khía cạnh a_k trong văn bản d_i được ký hiệu là ED_{ik}, và số đo trọng số của khía cạnh thông qua toàn bộ kho dữ liệu được ký hiệu là EC_k.

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

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

$$EC_k = \frac{\sum_{h=1}^{M} s_{kh}}{M} \quad (2.6)$$

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.

Trọng số α_{ik} cho khía cạnh a_k của bài đánh giá d_i được tính như sau:

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

2.5 Kết quả thực nghiệm

2.5.1 Dữ liệu thử nghiệm

Các thí nghiệm được thực hiện trên ba bộ dữ liệu đánh giá khách sạn được thu thập từ Tripadvisor.com, đánh giá bia được thu thập từ Beeradvocate.com và đánh giá cà phê Trung Nguyên được thu thập từ trang web Amazon.com.

2.5.2 Tiền xử lý và trích chọn đặc trưng

2.5.3 Kết quả và đánh giá

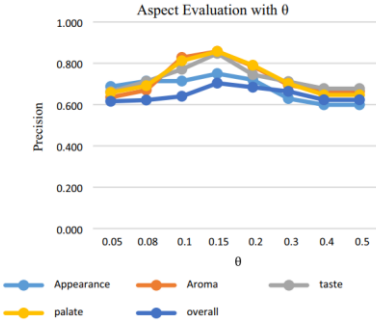
Trích rút khía cạnh

Để đánh giá hiệu quả, độ đo precision được sử dụng. Bảng 2.5 cho thấy hiệu suất của phương pháp này trong nhiệm vụ 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.

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

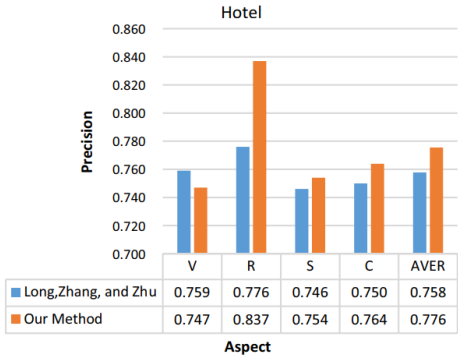
Dữ liệu Khách sạn		Dữ liệu Bia		Dữ liệu Cà phê	
Khía cạnh	$P_{recision}$	Khía cạnh	$P_{recision}$	Khía cạnh	$P_{recision}$
Value	0.747	Appearance	0.750	Aroma	0.667
Room	0.837	Aroma	0.857	Taste	0.677
Location	0.814	Palate	0.857	Acidity	0.667
Cleanliness	0.764	Taste	0.848	Body	0.600
Check in/front desk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

Trong thuật toán đề xuất, ngưỡng θ là ngưỡng xác suất để lấy mở rộng các tập từ khía cạnh. Bằng thực nghiệm, ngưỡng θ tốt nhất được thể hiện trong Hình 2.7 khoảng 0.15.



Hình 2.7 *Hiệu suất của phương pháp ứng với các ngưỡng θ khác nhau*

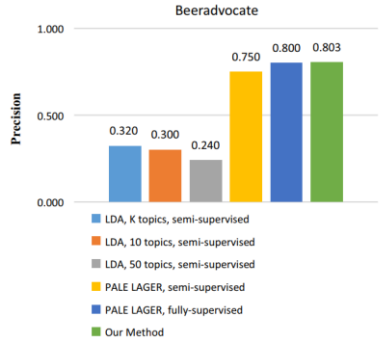
Phương pháp đề xuất của luận án được so sánh với phương pháp dựa trên tần suất trong trên tập dữ liệu khách sạn. Hình 2.9 cho thấy kết quả so sánh giữa hai phương pháp.



	V	R	S	C	AVER
Long,Zhang, and Zhu	0.759	0.776	0.746	0.750	0.758
Our Method	0.747	0.837	0.754	0.764	0.776

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

Phương pháp đề xuất cũng được so sánh với hai phương pháp dựa trên mô hình chủ đề trong (PALE LAGER bán giám sát và giám sát) và trong (LDA) trên tập dữ liệu bia. 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 (bán giám sát và có giám sát).



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

Nhiệm vụ dự đoán điểm đánh giá khía cạnh

Để đánh giá hiệu suất của phương pháp đề xuất trong nhiệm vụ này, ba độ đo: sai số bình phương trung bình theo khía cạnh ($\Delta^2{}_{aspect}$), độ tương quan khía cạnh (ρ_{aspect}), và độ tương quan khía cạnh qua tất cả các bài đánh giá (ρ_{review}) được sử dụng. Kết quả của phương pháp đề xuất được so sánh với hai phương pháp của Long và các cộng sự, Wang trên bộ dữ liệu khách sạn. Kết quả so sánh được chỉ ra trong Bảng 2.9.

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

Nhiệm vụ ước lượng trọng số khía cạnh

Phương pháp đề xuất được so sánh với phương pháp của Wang 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 được chỉ ra trong Bảng 2.10.

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 tổng thể giải quyết ba bài toán con của bài toán phân tích quan điểm mức khía cạnh: (1) trích rút các khía cạnh được đề cập đến trong bài đánh giá về một sản phẩm bằng cách sử dụng xác suất có điều kiện của các từ kết hợp với giải thuật Boostrapping; (2) suy ra điểm đánh giá của người dùng cho từng khía cạnh được xác định dựa trên bộ phân loại Naive Bayes; (3) ước lượng trọng số mà người dùng đặt trên mỗi khía cạnh bằng cách sử dụng số lần xuất hiện của các từ thảo luận về khía cạnh đó 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.

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

3.1 Đặt vấn đề

3.2 Các nghiên cứu liên quan

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

Định nghĩa 3.1 Vectơ từ (Word vector): Đưa ra một từ w_j một vectơ P chiều $\mathbf{x}_{wj} \in \mathbb{R}^P$ 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 $\mathbf{x}_{wj} = \{x_{1wj}, x_{2wj}, \ldots, x_{pwj}\}$, trong đó x_{pwj} 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, $w_{kj} \in C_k$ được ánh xạ tương ứng tới một vectơ trong tập vectơ từ được gọi là Vectơ từ lõi khía cạnh ký hiệu $\mathbf{x}_{coreak}$.

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. Độ hỗ trợ được tính toán dựa trên sự cải tiến của độ đo Euclidean như trong công thức (3.1).

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

trong đó: $supp(w_j \rightarrow a_k)$ là độ hỗ trợ của từ chủ đề w_j đối với khía cạnh a_k; N là số từ lõi của khía cạnh a_k; P là số chiều của vector từ; x_{pwi} là giá trị của chiều thứ p (trong biểu diễn vector từ) của từ wj; $x_{pcoretak}$ là giá trị của chiều thứ p (trong biểu diễn vector từ) của từ lõi thứ t thuộc về khía cạnh a_k.

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. 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).

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

trong đó: $supp(S \rightarrow a_k)$ là độ hỗ trợ câu S với khía cạnh a_k; $supp(w_j \rightarrow a_k)$ là độ hỗ trợ của từ chủ đề w_j đối với khía cạnh a_k; Q là số từ của câu S.

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

Mỗi khía cạnh a_k được thể hiện bởi một tập các từ. Từ được biểu diễn từ dưới dạng Word2vec để nắm bắt các ngữ cảnh khác nhau của từ nhằm nâng cao độ chính xác. Mô hình đề xuất được mô tả trong Hình 3.2.

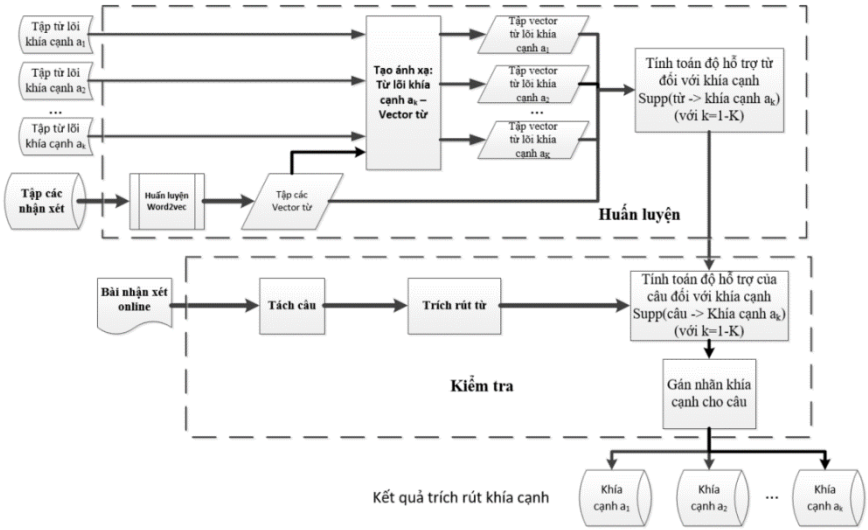
Pha huấn luyện:

Bước 1 (Dữ liệu): tách câu, chuẩn hóa câu.

Bước 2 (Huấn luyện word2vec): sử dụng công cụ word2vec chạy trên ngôn ngữ python để vector hóa các từ.

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

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:

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

Bước 2 (Trích rút từ): Trích rút các danh từ, tính từ, động từ, trạng từ trong câu. Các từ này được so khớp với các từ đã được huấn luyện để xác định độ hỗ trợ của từ đối với từng khía cạnh

Bước 3 (supp(Câu -> 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): so sánh độ hỗ trợ của câu với ngưỡng hoặc lấy giá trị lớn nhất để xác định nhãn khía cạnh cho câu.

3.5 Kết quả thực nghiệm

3.5.1 Tiền xử lý dữ liệu

3.5.2 Huấn luyện Word2vec

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

3.5.4 Kết quả thực nghiệm

Để đánh giá hiệu quả của phương pháp đề xuất, trong phần này luận án sử dụng các độ đo là precision và recall và F1. Kết quả thử nghiệm trên ba bộ dữ liệu được thể hiện trong bảng 3.3, 3.4, 3.5.

Phương pháp đề xuất được tiến hành thử nghiệm và so sánh kết quả với hai phương pháp cơ sở là LDA và của Long và các cộng sự trên bộ dữ liệu khách sạn sử dụng độ đo precision. Kết quả được chỉ ra trong bảng 3.6.

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
Appearance	0.795	0.800
Aroma	0.875	0.901
Palate	0.862	0.792
Taste	0.843	0.826
Overall	0.821	0.803
Average	**0.839**	**0.824**

Bảng 3.5 *Kết quả trích rút khía cạnh đối với bộ dữ liệu Cà Phê*

Khía cạnh	Precision	Recall
Aroma	0.702	0.684
Taste	0.666	0.659
Acidity	0.654	0.600
Body	0.712	0.720
Average	**0.684**	**0.666**

Bảng 3.6 *So sánh kết quả phương pháp đề xuất với phương pháp LDA và Long và cộng sự trên tập dữ liệu Khách sạn với độ đo precision*

Khía cạnh	PP LDA [132]	PP Long et al. [59]	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 dựa trên việc khai thác hiệu quả biểu diễn đặc trưng từ dạng vector và sử dụng chúng để tính toán trọng số của thuật ngữ cốt lõi bằng thước đo hỗ trợ. Phương pháp này hoạt động tốt trên các bộ dữ liệu của thế giới thực và nó có thể được áp dụng cho một số lĩnh vực khác nhau.

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

4.1 Đặt vấn đề

Bài đánh giá được phân thành 5 lớp dựa trên các đánh giá cảm tính và đánh giá lý tính. Thách thức chính là 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, do độ không chắc chắn, sự mơ hồ xảy ra khi vector đặc trưng không chứa đủ thông tin, do các lớp có điểm xác suất tương tự nhau. Khó khăn quan trọng khác là vấn đề dữ liệu không cân bằng. Thách thức thứ ba là tính thưa thớt của dữ liệu và phụ thuộc nhiều vào ngữ cảnh của văn bản ngắn dẫn đến khó có hàm phân biệt tốt giữa các văn bản khác nhau.

Để khắc phục những khó khăn trên, ý tưởng cơ bản là kết hợp các bộ phân loại khác nhau, có thể bổ sung cho nhau, khắc phục yếu điểm của mỗi bộ phân loại riêng lẻ, 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.

Nghiên cứu sinh đề xuất sử dụng phương pháp dựa trên lý thuyết Dempster-Shafer (DS) và sử dụng chỉ hai bộ phân loại mạnh mẽ là SVM nhiều lớp và thuật toán phân loại nhiều lớp dựa trên mô hình tương tác không kết hợp (hay OR Gate Bayesian Network - OGBN). Mục tiêu của đề xuất: sử dụng ít bộ phân loại nhất, giải quyết vấn đề dữ liệu mất cân bằng, cải thiện hiệu suất phân loại đa lớp.

Văn bản được tiền xử lý, các đặc trưng được lựa chọn là uni-gram, bi-gram, độ lợi thông tin (Information Gain - IG) và thông tin tương hỗ (Mutual Information – MI). Bài viết được phân loại dựa trên SVM và OGBN. Đầu ra của thuật toán SVM được đưa qua một hàm chuyển đổi thành giá trị xác suất. Giá trị xác suất tương ứng của SVM cùng với đầu ra xác suất OGBN trở thành đầu vào của luật kết hợp DS. Điểm đánh giá cuối cùng của bài đánh giá là lớp mà giá trị xác của nó là lớn nhất.

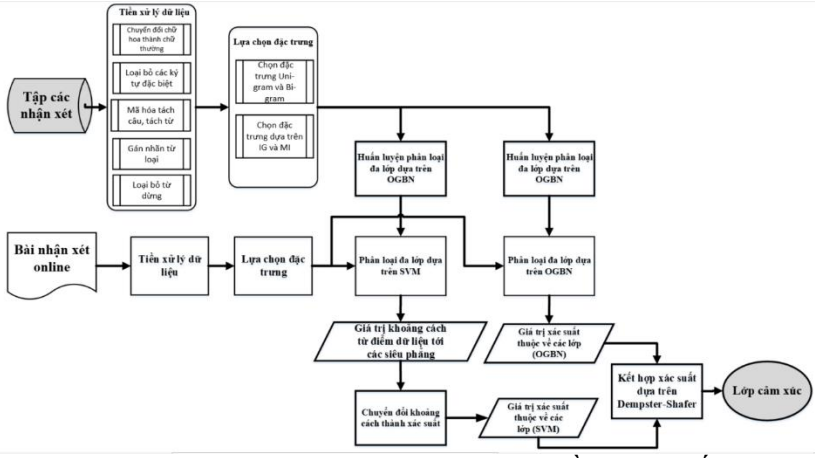
4.2 Các nghiên cứu liên quan

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

Như đã giới thiệu ở Mục 4.1, nghiên cứu sinh đề xuất một mô hình kết hợp thực hiện phân loại đa lớp bằng cách kết hợp xác suất đầu ra từ hai thuật toán phân lớp cơ sở (SVM và OGBN) dựa trên 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 mô tả quy trình trong mô hình đề xuất.

Văn bản được tiền xử lý (xem Mục 2.5.2), các đặc trưng được lựa chọn là uni-gram, bi-gram, IG và MI. Các đặc trưng biểu diễn văn bản trở thành đầu vào của các thuật toán phân loại cơ bản SVM và OGBN. Đầu ra của thuật

toán SVM là hàm khoảng cách từ điểm dữ liệu đến các siêu phẳng, đây không phải là một giá trị xác suất. Do đó, điểm khoảng cách này được đưa qua một hàm chuyển đổi thành giá trị xác suất. Giá trị xác suất tương ứng của SVM cùng với đầu ra xác suất OGBN trở thành đầu vào của luật kết hợp DS. Điểm đánh giá cuối cùng của bài đánh giá là lớp mà giá trị xác kết hợp của nó là lớn nhất.



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

4.3.1 Phân loại cảm xúc đa lớp dựa trên SVM

Trong trường hợp tập dữ liệu đa lớp, chiến lược *một với tất cả* (One-vs-all-OVA) được lựa chọn. 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).

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

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

4.3.2 Biến đổi đầu ra 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 phương pháp Dempster-Shafer được đề xuất để kết hợp các bộ phân loại, do đó cần hiệu chỉnh đầu ra bộ phân loại SVM nhiều lớp để xuất ra các giá trị xác suất hậu nghiệm.

Platt đề xuất một phương pháp để ước lượng SVM hậu nghiệm bằng cách sử dụng một hàm sigmoid và điểm số SVM như sau:

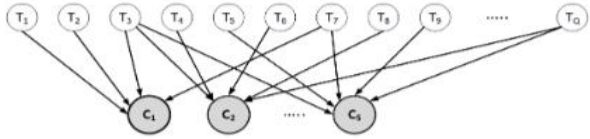
$$p(c = 1 | y(x^*)) = \frac{1}{1 + exp(Ay(x^*) + B)} \tag{4.3}$$

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

Các tham số A và B được ước lượng phù hợp với hàm sigmoid, mã giả trong đề xuất của Platt được sử dụng.

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

Mạng Bayes cổng OR kế thừa những ưu điểm của mạng Bayes. Phương pháp này hiển nhiên phù hợp với bài toán phân loại nhiều lớp và hoạt động tốt trên dữ liệu có độ mất cân bằng cao, nó cũng làm giảm độ phức tạp tính toán so với mạng Bayes từ $O(2^n)$ xuống $O(n)$.



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

Mô hình dự đoán điểm đánh giá cảm xúc của một bài đánh giá được cấu trúc như sau: tập các đặc trưng $\{f_q\}$, *mỗi* $\{f_q\}$ là một nút T_q *nút nguyên nhân*; Có C_{class} *nút kết quả* tương ứng với C_{class} các lớp. Cấu trúc mạng là cố định. Cung đi từ nút T_q đến nút kết quả C_j nếu đặc trưng $\{f_q\}$ xuất hiện trong dữ liệu huấn luyện của lớp c_j. Xác suất hậu nghiệm của mỗi lớp c_j khi biết văn bản d_i được xác định như sau:

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

Xác suất này có thể được ước lượng trực tiếp $\hat{p}(c_j|f_q)$ sử dụng xấp xỉ Laplace:

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

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 $N_{\bullet q} = \sum_{c_j} N_{jq}$. Hàm phân lớp của văn bản d_i:

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

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

Một siêu tập hợp $\mathbf{P}(C)$ là tập của tất cả các tập con có thể có của các lớp $\mathbf{P}(C) = \{\emptyset, \{c_1\}, \ldots, \{c_n\}, \{c_1, c_2\}, \ldots, \{c_1, \ldots, c_n\}\}$. 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 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) và ký hiệu bởi $m = m_1 \oplus m_2$) 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.

Cho $\Theta = \mathbf{P}(C)/C$, 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, chúng ta có:

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

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, như vậy luận án sử dụng xấp xỉ $\tilde{m}$ như sau:

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

Để cấu trúc hàm khối lượng cho mỗi lớp c_j từ một mẫu văn bản đánh giá d_i, luận án 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 với hàm khối lượng cho lớp c_j được cung cấp bởi trình phân loại φ như sau:

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

4.4 Kết quả thực nghiệm

4.4.1 Bộ dữ liệu thực nghiệm

Phân bố đánh giá của 5 lớp trong ba bộ dữ liệu được thể hiện trong Bảng 4.5, các ký hiệu: lớp tiêu cực cảm xúc $c1$; lớp tiêu cực lý trí $c2$; lớp trung lập $c3$; lớp tích cực lý trí $c4$; lớp tích cực cảm xúc $c5$.

Bảng 4.5 *Phân bố của các lớp cảm xúc trong 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

Luận án xây dựng hai bộ đặc trưng: bộ đặc trưng cơ sở (uni-gram, bi-gram); bộ đặc trưng rút gọn dựa trên bộ đặc trưng cơ sở thông qua các phép lọc đặc trưng (IG kết hợp MI). Ba thí nghiệm được thực hiện để đánh giá hiệu suất của phương pháp đề xuất.

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

Thí nghiệm đầu tiên so sánh hiệu suất của bộ phân loại đa lớp dựa trên SVM, dựa trên mạng Bayes Noisy OR-gate bằng cách sử dụng hai bộ đặc trưng đầu vào khác nhau. Bảng 4.7 cho thấy hiệu suất của phương pháp dựa trên SVM và OGBN trên ba bộ dữ liệu. Bộ phân loại dựa trên OGBN hoạt động tốt hơn bộ phân loại dựa trên SVM với OVA trong tất cả các bộ dữ liệu. Kết quả này xác nhận phân tích trước đây của chúng tôi rằng SVM hoạt động tốt với phân loại văn bản nhị phân, nhưng gặp khó khăn khi xử lý với đa phân loại văn bản. Bộ phân lớp dựa trên OGBN hoạt động tốt hơn với tập đặc trưng có số chiều lớn ("Uni+Bi"), trong khi phương pháp dựa trên SVM hoạt động tốt với tập đặc trưng được thu gọn (""Uni+Bi+IG+MI"").

Thử nghiệm thứ hai để đánh giá hiệu quả của việc kết hợp hai bộ phân loại cơ sở bằng cách sử dụng lý thuyết DS. Chúng tôi sẽ đánh giá sự cải thiện tổng thể của mô hình kết hợp, đánh giá vấn đề dữ liệu mất cân bằng và vấn đề phân loại sai giữa các lớp lân cận.

Bảng 4.8 chỉ ra phương pháp kết hợp dựa trên DS hoạt động tốt hơn cả hai phương pháp dựa trên SVM và dựa trên OGBN đối với cả ba bộ dữ liệu. Kết quả cho thấy phương pháp kết hợp vượt trội hơn một chút 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%) kết quả đã được bao phủ bởi các lớp chiếm đa số.

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 4.8 *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	**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	**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	**95.81**	**95.63**	**95.72**	**95.83**

Bảng 4.9, 4.10, 4.11 trình bày số lượng mẫu bị phân loại sai giữa hai lớp liền kề theo ba phương pháp, các lớp tiêu cực cảm xúc, tiêu cực lý trí, trung lập, tích cực lý trí, tích cực cảm xúc được ký hiệu lần lượt là c_1, c_2, c_3, c_4, c_5.

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

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

Bảng 4.10 *Các mẫu 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	Kết hợp DS
Số mẫu bị phân loại sai	$c_1 \rightarrow c_2$	114	100	56
	$c_2 \rightarrow c_1$	63	56	53
	Tổng	177	156	109
	$c_2 \rightarrow c_3$	27	25	25
	$c_3 \rightarrow c_2$	101	95	68
	Tổng	128	120	93
	$c_3 \rightarrow c_4$	104	104	100
	$c_4 \rightarrow c_3$	136	129	122
	Tổng	240	233	222
	$c_4 \rightarrow c_5$	241	232	180
	$c_5 \rightarrow c_4$	312	262	163
	Tổng	553	494	343

Bảng 4.11 *Các mẫu 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	Kết hợp DS
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

4.5 Kết luận chương 4

Trong chương này luận án xem xét giải quyết nhiệm vụ phân loại quan điểm/cảm xúc khía cạnh đa lớp. Nghiên cứu sinh đã đề xuất một mô hình kết hợp mạnh mẽ để giải quyết vấn đề trên bằng cách sử dụng phương pháp dựa trên lý thuyết Dempster-Shafer với sự lựa chọn cẩn thận các bộ phân loại cơ sở có thể bổ sung tốt nhất cho nhau. Bằng cách áp dụng phân tích điểm mạnh và điểm yếu của các phương pháp hiện có, nghiên cứu sinh đã đưa ra hai ứng cử viên của phương pháp phân loại kết hợp, đó là các phương pháp đa phân loại dựa trên SVM và đa phân loại dựa trên mạng Bayesian cổng OR. Kết quả cho thấy tính hiệu quả vượt trội của phương pháp kết hợp so với hai phương pháp cơ sở. Trong kết quả đó cũng thể hiện khả năng khắc phục những vấn đề dữ liệu không cân bằng, tính mơ hồ của dữ liệu và tính liền kề của các lớp lân cận.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Những kết quả đạt được của luận án

Phân tích quan điểm dựa trên khía cạnh các bài đánh giá sản phẩm trực tuyến được coi là một công cụ hữu ích để khám phá tính cá nhân hóa người dùng, dự đoán xu hướng tiêu dùng, định hướng thị trường sản phẩm. Nghiên

cứu trong luận án này phát triển một số thuật toán học máy để nâng cao chất lượng khai phá, phân tích quan điểm mức khía cạnh. Một số kết luận như sau:

- Đề xuất hệ thống 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***, luận án 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. 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 mô hình ngôn ngữ. 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.

2. Định hướng phát triển

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.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ CỦA TÁC GIẢ LIÊN QUAN ĐẾN ĐỀ TÀI LUẬN ÁN

1. 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
2. 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).
3. 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).
4. 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.
5. 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.