ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VŨ THỊ NHẠN

TỔNG HỢP QUAN ĐIỂM TRỰC TUYẾN CỦA NGƯỜI TIÊU DÙNG THEO TÍNH NĂNG CỦA SẢN PHẨM

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI – 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VŨ THỊ NHẠN

TỔNG HỢP QUAN ĐIỂM TRỰC TUYẾN CỦA NGƯỜI TIÊU DÙNG THEO TÍNH NĂNG CỦA SẢN PHẨM

Ngành: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Mã số: 60 48 01 04

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN VIỆT ANH

TS. BÙI QUANG HƯNG

HÀ NỘI - 2016

Lời cam đoan

Tôi xin cam đoan báo cáo luận văn này được viết bởi tôi dưới sự hướng dẫn của thầy giáo, Tiến sĩ Nguyễn Việt Anh và Tiến sĩ Bùi Quang Hưng. Tất cả các kết quả đạt được trong luận văn này là quá trình tìm hiểu, nghiên cứu của riêng tôi. Trong toàn bộ nội dung của luận văn, những điều được trình bày là của cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu khác. Các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày …… tháng ..… năm 2016

Người cam đoan

Vũ Thị Nhạn

Mục lục

Danh mục hình vẽ

Danh mục bảng biểu

Lời cảm ơn

Đầu tiên, tôi muốn gửi lời cảm ơn sâu sắc nhất đến cán bộ hướng dẫn khoa học, thầy giáo, TS. Nguyễn Việt Anh, và TS. Bùi Quang Hưng người đã đưa tôi đến lĩnh vực nghiên cứu này và đã giảng dạy trong quá trình tôi học tập tại trường Đại học Công Nghệ - Đại học Quốc Gia Hà Nội và nghiên cứu tại Viện Công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Thầy luôn truyền cho tôi nguồn cảm hứng, nhiệt huyết nghiên cứu khoa học và hết sức tận tình hướng dẫn tôi, cho tôi những lời khuyên quý báu. Mặc dù thầy rất bận với công việc giảng dạy và nghiên cứu nhưng thầy đã dành cho tôi nhiều thời gian thảo luận các ý tưởng nghiên cứu, chỉ dẫn cách nghiên cứu, giải đáp thắc mắc và động viên tôi vượt qua những vấn đề khó khăn cũng như hướng tôi tới nhiều vấn đề có giá trị khác khiến tôi muốn tìm hiểu và nghiên cứu trong tương lai.

Tôi cũng xin gửi lời cám ơn tới các Thầy, Cô giáo của Khoa Công nghệ thông tin, đã truyền dậy những kiến thức bổ ích, hiện đại về lĩnh vực Hệ thống thông tin mà tôi học tập. Tôi đã được tiếp cận một môi trường học thuật cao, hiểu được sự vất vả cũng như thành quả đạt được khi tham gia nghiên cứu khoa học.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc tới bố mẹ, anh chị và bạn bè tôi. Họ đã luôn bên cạnh tôi, ủng hộ và giúp đỡ tôi trong suốt quá trình học tập và hoàn thiện luận văn này

Học viên thực hiện luận văn

Vũ Thị Nhạn

Mở đầu

"*Người khác nghĩ gì*" luôn là một câu hỏi đặt ra cho mỗi chúng ta trong những lần ra quyết định. Khi bạn có nhu cầu mua một chiếc tivi, bạn sẽ có xu hướng tìm hiểu xem người khác nói gì về sản phẩm này. Với cùng một số tiền bỏ ra, bạn sẽ lựa chọn được những sản phẩm có những chức năng đáp ứng được yêu cầu của bạn một cách thích hợp nhất. Hay như chương trình *Ai là triệu phú* phát sóng trên truyền hình, có hai trong ba quyền trợ giúp là hỏi ý kiến của người khác.

Cùng với sự phát triển của kinh tế xã hội, Internet ngày càng phát triển. Mọi người dần biết đến các trang blog, diễn đàn hay các trang mạng xã hội khác. Đó là nơi họ cùng bày tỏ quan điểm về một vấn đề, sự kiện hay chất lượng của một sản phẩm. Đó là nguồn thông tin quan trọng đối với mọi người khi có nhu cầu tìm hiểu về vấn đề nào đó.

Đối với doanh nghiệp, khi họ đưa bất kỳ một sản phẩm nào ra thị trường, họ cần biết người tiêu dùng đánh giá như thế nào về sản phẩm của họ. Từ đó, họ có chiến lược kinh doanh cho phù hợp. Theo như các công ty lớn nhận định, ý kiến của khách hàng là một phần quan trọng trong việc hình thành quan điểm ý kiến của các khách hàng khác và sự tin tưởng vào thương hiệu, quyết định mua hàng sẽ liên quan đến các chính sách quảng bá thương hiệu của công ty họ. Với sự phong phú của các nguồn tài nguyên về quan điểm như hiện nay, cơ hội và thách thức là rất lớn trong việc sử dụng công nghệ thông tin để tìm kiếm và hiểu được ý kiến của người khác [24]

Hình 1. Khai phá quan điểm người dùng

Người tiêu dùng khi đánh giá về một sản phẩm dịch vụ nào đó, nhưng họ cũng có thể đưa ra ý kiến tổng quan nhất về một sản phẩm. Ví dụ *"Chiếc điện thoại Iphone 6s là rất tốt"*. Nhưng lại có các ý kiến đưa ra để đánh giá chất lượng của một tính năng (khía cạnh, đặc trưng) nào đó của sản phẩm. Ví dụ: "*Màn hình của chiếc Iphone 6s là đẹp"* hoặc "*camera rất nét"*. Các ý kiến phản hồi của người tiêu dùng là đa dạng và phong phú. Việc tổng hợp các ý kiến thủ công sẽ mất nhiều thời gian và sức người. Một công cụ tổng hợp ý kiến tự động của người tiêu dùng sẽ làm giảm thời gian và công sức. Chính vì vậy, tôi đã chọn hướng nghiên cứu tổng hợp quan điểm theo tính năng của sản phẩm của người tiêu dùng Việt Nam với dữ liệu chủ yếu được lấy trên các diễn đàn công nghệ. Trong luận văn của mình, tôi trình bày một phương pháp tổng hợp quan điểm, sử dụng luật lan truyền kép kết hợp với việc tách câu ghép và câu phức thành các câu đơn (mỗi một câu đơn chứa một tính năng của sản phẩm) dựa theo luật để trích xuất ra các tính năng của sản phẩm của người tiêu dùng Việt Nam. Tiếp theo, tôi sử dụng kiến thức về mẫu phổ biến để loại bỏ các dữ liệu nhiễu. Và cuối cùng, tôi sử dụng phương pháp thống kê để tổng hợp quan điểm đánh giá của người tiêu dùng về từng tính năng của sản phẩm.

Luận văn của tôi được chia thành các phần như sau:

Chương 1: Trong chương này, tôi trình bày tổng quan về khai phá quan điểm và một số khái niệm liên quan. Đồng thời, tôi trình bày những khó khăn và thách thức của khai phá quan điểm nói chung và một vài lĩnh vực ứng dụng của khai phá quan điểm được ứng dụng trên thế giới hiện nay

Chương 2: Trình bày khái quát một số pháp được các nhà nghiên cứu trên thế giới nghiên cứu và áp dụng vào việc tổng hợp ý kiến theo tính năng của sản phẩm trên thế giới cũng như ở Việt Nam hiện nay

Chương 3: Trong chương này, tôi trình bày một cách chi tiết một phương pháp tổng hợp ý kiến theo tính năng của sản phẩm được tôi nghiên cứu và thử nghiệm với dữ liệu tiếng Việt

Chương 4: Kết quả thực nghiệm được trình bày trong chương này, đồng thời tôi cũng đưa ra đánh giá về phương pháp mà tôi đã đề xuất trong chương 3

Chương 5: Kết luận

Chương 1. Tổng quan về khai phá quan điểm

1.1. Giới thiệu

Chúng ta đã biết, cùng với sự phát triển của Internet, các mạng xã hội, diễn đàn, blog như Facebook, Twitter, Zing Me,… thu hút hàng triệu người Việt Nam sử dụng. Tại đó mọi người thể hiện quan điểm của mình về rất nhiều vấn đề, rất nhiều đối tượng. Đặt tình huống chẳng hạn một người cần mua máy điện thoại mới nhưng anh ta chưa biết nên mua loại nào. Anh ta có thể hỏi ý kiến của bạn bè, nhờ sự tư vấn của người bán hàng. Một cách thông minh hơn là anh ta có thể tham khảo thông tin trên mạng, nhưng sẽ phải đọc rất nhiều bài viết. Một doanh nghiệp khi đưa một sản phẩm ra thị trường, họ rất cần biết người tiêu dùng có phản ứng như thế nào về sản phẩm của họ. Họ có thể thuê nhân viên tra cứu các thông tin trên các trang mạng xã hội – nơi mà người tiêu dùng có thể đưa ra các ý kiến về sản

Hình 1.1. Mô hình khai phá quan điểm

phẩm đó sau khi họ đã sử dụng. Tuy nhiên, việc thực hiện tổng hợp các ý kiến đánh giá đó thành một bản tổng hợp có thể nhìn trực quan nhất thì việc tổng hợp thủ công mất rất nhiều thời gian. Vì vậy, cần thiết phải có một công cụ thực hiện tổng hợp các ý kiến đó một cách tự động. Việc tự động tổng hợp ý kiến, quan điểm về một đối tượng hay vấn đề cụ thể nào đó gọi là tổng hợp quan điểm. Khi đó máy tính sẽ trợ giúp người dùng bằng cách thu thập và phân tích văn bản chứa quan điểm và đưa ra kết quả tổng hợp.

Quan điểm là ý kiến của cá nhân một người về một đối tượng nào đó trong một thời gian nhất định. Theo định nghĩa của Liu [13], một quan điểm bao gồm 5 yếu tố (e_i, a_{ij}, s_{ijkl}, h_k, t_l) trong đó e_i là tên của chủ thể, a_{ij} là đặc trưng của e_i, s_{ijkl} là quan điểm về đặc trưng a_{ij} của

e_i, h_k là người giữ quan điểm và t_l là thời điểm mà quan điểm đó được đưa ra bởi h_k. Quan điểm s_{ijkl} có thể tích cực, tiêu cực, trung lập hoặc có thể biểu diễn bởi các mức độ khác nhau.

Trong định nghĩa của Liu có một số khái niệm về đối tượng, đặc trưng, người giữ quan điểm được làm rõ như sau:

Đối tượng

Đối tượng được dùng để chỉ thực thể (người, sản phẩm, sự kiện, chủ đề,…) được đánh giá. Mỗi đối tượng có một tập các thành phần (components) hay thuộc tính (attributes) gọi chung là các đặc trưng (tính năng) (features) [12]. Mỗi thành phần hay thuộc tính lại có một tập các thành phần hay thuộc tính con. Như vậy, một đối tượng O được biểu diễn bởi một cặp [T, A] trong đó T là một cấu trúc phân cấp gồm các thành phần cha và con; A là tập các thuộc tính của đối tượng O.

Ví dụ: *Máy ảnh* có một tập thành phần như *ống kính, pin* và các thuộc tínhnhư *kích cỡ, khối lượng, chất lượng ảnh*. Thành phần *pin* có các thuộc tính con như *kích cỡ, thời gian, dung lượng*.

Các đặc trưng ẩn và hiện

Với mỗi đánh giá r bao gồm tập các câu r = {s1, s2, … sm}. Nếu đặc trưng f xuất hiện trong r, ta nói f là đặc trưng hiện (explicit feature). Ngược lại, ta nói f là đặc trưng ẩn (implicit feature) [12].

Ví dụ:

Máy ảnh này đắt quá. Đặc trưng *giá* là đặc trưng ẩn.

Màu này đẹp ghê. Đặc trưng *màu* là đặc trưng hiện..

Đoạn đánh giá

Đoạn đánh giá về một đặc trưng f của đối tượng O trong r là một tập các câu liên tiếp trong r diễn tả quan điểm tích cực hay tiêu cực về đặc trưng f. Đoạn đánh giá bao gồm tối thiểu ít nhất một câu [12].

Ví dụ:

Bộ nhớ của chiếc điện thoại này là nhỏ.

Hôm qua, tôi mua một chiếc điện thoại Iphone 5S ở siêu thị điện máy Trần Anh. Tôi rất thích nó. Kích thước của nó phù hợp với tay tôi cầm. Giá cả phải chăng mà chất lượng cũng tốt.

Hai ví dụ trên, tuy số lượng câu và độ dài là khác nhau nhưng chúng đều là các đoạn đánh giá cho sản phẩm điện thoại.

Quan điểm ẩn, hiện

Quan điểm hiện (explicit opinion) về một đặc trưng f là một câu thể hiện quan điểm mang tính chủ quan, diễn trả trực tiếp quan điểm tích cực hay tiêu cực của tác giả. Quan điểm ẩn (implicit opinion) về một đặc trưng f là câu thể hiện quan điểm tích cực hay tiêu cực một cách không tường minh [12].

VD:

Điện thoại này *đẹp* quá. Quan điểm hiện – khen chiếc điện thoại đẹp.

Máy ảnh *mới mua và đã hỏng*. Quan điểm ẩn – chê chiếc máy ảnh không tốt.

Người đánh giá

Là người hay tổ chức cụ thể đưa ra các ý kiến đánh giá của cá nhân (tổ chức). Trong trường hợp đánh giá sản phẩm, forum, blog thì người đánh giá luôn là các tác giả của đánh giá hay bài viết đó [12].

1.2. Các thách thức của khai phá quan điểm

1.2.1. Những người khác nhau có phong cách viết khác nhau

Thực tế, các bình luận hay quan điểm được đưa ra bởi những người khác nhau thì họ có cách viết khác nhau, từ cách thức sử dụng ngôn ngữ, chữ viết tắt và kiến thức của họ là một thách thức riêng của mỗi người. Mọi người đều không bày tỏ ý kiến theo cùng một cách.

1.2.2. Quan điểm thay đổi theo thời gian

Một thách thức khác cần phải xét đến là vấn đề làm thế nào để có thể theo dõi các quan điểm thay đổi theo thời gian. Một sản phẩm có thể là tốt nhất tại thời điểm này nhưng tại thời điểm 2, 3 năm sau thì nó không phải là tốt nhất nữa, người ta sẽ có nhiều sự lựa chọn hơn khi các sản phẩm mới tốt hơn về giá cả và chất lượng. Tuy nhiên, cũng có những sản phẩm ban đầu đưa ra ngoài thị trường chưa được tốt nhưng qua quá trình cải thiện chất lượng của sản phẩm hoặc dịch vụ thì lại được người tiêu dùng đánh giá cao hơn.

1.2.3. Độ mạnh của quan điểm

Xác định độ mạnh của một quan điểm là một thách thức phải đối mặt trong khai phá quan điểm. Nhiều nỗ lực đã được thực hiện để xác định các yếu tố quyết định sức mạnh của một ý kiến trong một bối cảnh nào đó. Bổ sung thêm việc phân lớp các từ thành các mức độ xu hướng quan điểm khác nhau, một số từ bổ nghĩa có thể được dùng để xác định độ mạnh của quan điểm(“*rất”, “một chút”, “hết sức”, “hơi”,...)*. Cụm từ “*rất hài lòng*” và “*hơi hài lòng*” sẽ được phân lớp thành rất tích cực và kém tích cực nếu “*rất*” và “*hơi*” được phân tích và sử dụng để xác định mức độ đối lập.

1.2.4. Quan điểm theo ngữ cảnh

Tương tự như phân loại một quan điểm là tích cực, tiêu cực hoặc trung lập có thể là một nhiệm vụ khó khăn trong khai phá quan điểm. Một từ quan điểm có thể được coi là tích cực trong một tình huống này nhưng nó lại mang tính tiêu cực trong tình huống khác.

Một quan điểm về một sản phẩm hoặc một tính năng sản phẩm cũng có thể gây hiểu lầm cho một hệ thống khai phá quan điểm để đánh giá. Ví dụ từ “*dài*” nếu được sử dụng để mô tả tuổi thọ của pin của máy tính xách tay theo một cách đó là “*tuổi thọ pin của máy tính xách tay là dài*”, nó sẽ được coi là tích cực nhưng nếu nó được sử dụng theo một cách khác đó là “*thời gian khởi động của máy tính xách tay dài*”, nó sẽ được coi là một ý kiến tiêu cực.

1.2.5. Các câu đánh giá có sự pha trộn

Một thách thức lớn đối với khai phá quan điểm xuất hiện khi mọi người thể hiện đánh giá tích cực và tiêu cực trong cùng một câu. Mọi người có nhiều ý kiến khác nhau trong cùng một câu. Những câu như vậy có thể gây khó khăn để phân tích cú pháp hoặc khai phá quan điểm.

Trong luận văn của này, tôi đã cố gắng để giải quyết vấn đề này bằng việc xây dựng công cụ tách các ý quan điểm trong các câu có đánh giá về nhiều tính năng thành các ý nhỏ. Trong đó, mỗi ý chỉ bao gồm một tính năng và một ý kiến đánh giá. (Chi tiết tôi xin trình bày ở chương 3.)

1.2.6. Quan điểm mang tính châm biếm, mỉa mai

Các quan điểm mang tính châm biếm, mỉa mai tồn tại khá nhiều trong văn bản. Trong đó một quan điểm tiêu cực nhưng lại được người nêu quan điểm thể hiện dưới dạng quan

điểm tích cực. Điều này gây khó khăn rất lớn trong quá trình phân tích quan điểm. Ví dụ "*Bộ phim hay thế này mà anh cũng rủ tôi đi xem*" khác với "*Bộ phim này rất hay*"

1.2.7. Xử lý ngôn ngữ tự nhiên trong câu quan điểm

Các ý kiến mà mọi người nêu lên trên các trang mạng xã hội thường là họ viết theo ngôn ngữ rất tự nhiên của họ. Họ có thể dùng ngôn ngữ, cách viết tắt hay các biểu tượng cảm xúc riêng. Mỗi người khác nhau sẽ có các phong cách viết khác nhau. Vì thế nên các câu đánh giá thường ở dạng bán cấu trúc. Trong khi việc cần thiết là từ những dữ liệu bán cấu trúc, được viết bằng ngôn ngữ rất tự nhiên đó, chúng ta phải đưa ra được các thông tin hữu ích. Tuy nhiên, trong đánh giá của người tiêu dùng thường, họ thường dùng các ngôn ngữ văn bản là không chính thức và không theo quy tắc ngữ pháp. Vì vậy, vấn đề xử lý ngôn ngữ tự nhiên trong việc xử lý các ý kiến đánh giá là một vấn đề cực kỳ khó khăn.

1.3. Các ứng dụng trong khai phá quan điểm

1.3.1. Nghiên cứu thị trường dành cho người mua và bán

Khi chúng ta muốn mua một sản phẩm nào, chúng ta không biết được loại sản phẩm này có phù hợp hay không, cửa hàng nào có dịch vụ khách hàng tốt, giá bán ở đâu rẻ hơn, chất lượng ở đâu tốt hơn,… thì các quan điểm về sản phẩm của những người dùng trước là một kênh thông tin quan trọng cho chúng ta.

Hay đối với những người bắt đầu kinh doanh, họ chưa biết kinh doanh mặt hàng gì, loại sản phẩm nào đang được người tiêu dùng ưa chuộng, hình thức kinh doanh nào là hợp lý, kinh doanh ở khu vực nào thu được lợi nhuận cao nhất. Khi đó, các hành vi của khách hàng sẽ hỗ trợ cho họ.

1.3.2. Cải thiện chất lượng của sản phẩm, dịch vụ

Ta xem xét một ví dụ sau: Một nhà sản xuất máy vi tính lớn đang thất vọng về doanh thu thấp bất thường của mình. Lãnh đạo công ty đưa ra câu hỏi "*Tại sao khách hàng không mua máy tính của chúng ta?*". Những thông tin cụ thể như giá thành, chất lượng sản phẩm của đối thủ cạnh tranh là mục tiêu chính để khảo sát. Ngoài ra, các đánh giá chủ quan về thiết kế, dịch vụ khách hàng,… của khách hàng cũng là các yếu tố cần được xem xét.

1.3.3. Hệ thống gợi ý

Khai phá quan điểm cũng có vài trò quan trọng như một công nghệ hỗ trợ cho các hệ thống khác. Một ứng dụng tiềm năng đó là ta có thể áp dụng khai phá quan điểm trong các

hệ thống khuyến cáo, giúp cho hệ thống đưa ra các gợi ý về các sản phẩm cho người dùng mà có khả năng người dùng quan tâm là cao nhất, tăng lợi nhuận cho doanh nghiệp.

Trong các hệ thống trực tuyến, các quảng cáo được hiển thị ở góc màn hình cần được kiểm tra xem có phù hợp với nội dung trang web hay không. Ví dụ trong một trang web có nội dung chuyên về gan mà hiển thị các quảng cáo về các sản phẩm rượu là không hợp lý, nếu hiện thị quảng cáo về các loại thuốc trị viêm gan sẽ phù hợp với nội dung hơn.

1.3.4. Hỗ trợ thông minh trong chính quyền

Thông minh trong chính quyền là một dạng ứng dụng vô cùng hữu ích đối với các chính trị gia. Chẳng hạn như khi một dự luật được đưa ra, quốc hội rất muốn lấy ý kiến của nhân dân về dự thảo luật, xem rằng nó có hợp lý hay không, nhân dân có những phản ứng như thế nào về nó. Hay đối với những cuộc bầu cử tổng thống, thủ tướng, những ý kiến đánh giá của người dân giữ một vai trò cực quan trọng đối với kết quả của cuộc bầu cử.

1.3.5. Hỗ trợ đưa ra quyết định

Khai phá quan điểm có vai trò to lớn trong việc hỗ trợ ra quyết định. Hoặc đối với những vấn đề về kinh tế xã hội khác. Đối với sự kiện chặt 6700 cây xanh ở Hà Nội, phản ứng không đồng tình của người dân đã có tác động to lớn đối với chính quyền Ủy ban nhân dân thành phố Hà Nội, Bí thư thành ủy Hà Nội phải ra quyết định xem xét và xử lý đối với sở, ban ngành liên quan[1]. Hay đối với tin tức về vụ xử phạt đối với quán café *Xin chào* tại thành phố Hồ Chí Minh đã buộc thủ tướng Nguyễn Xuân Phúc ra chỉ đạo xem xét, dừng khởi tố vụ án[2].

1.4. Các bài toán khai phá quan điểm

Khai phá quan điểm là một lĩnh vực được nghiên cứu từ những năm 90, tuy nhiên với những khó khăn và thách thức của nó mà nó vẫn được cộng đồng nghiên cứu trên thế giới và tại Việt Nam nghiên cứu. Và có thể nói nó vẫn là một chủ đề "nóng" trong cộng động nghiên cứu tại Việt Nam và trên thế giới.

Theo nghiên cứu của Liu [7], khai phá quan điểm gồm 3 bài toán chính như sau:

- Phân lớp quan điểm
- Khai phá quan điểm so sánh

[1] http://vnexpress.net/tin-tuc/thoi-su/bi-thu-ha-noi-khong-xu-ly-kieu-hoa-ca-lang-vu-chat-cay-xanh-3161498.html
[2] http://thanhnien.vn/thoi-su/thu-tuong-chi-dao-xem-xet-dung-khoi-to-vu-chu-quan-ca-phe-xin-chao-694327.html

– Tổng hợp quan điểm.

1.4.1. Phân lớp quan điểm

Với bài toán này có thể coi khai phá quan điểm như bài toán phân lớp văn bản. Bài toán phân lớp một văn bản đánh giá là tích cực hay tiêu cực. Ví dụ: với một đánh giá sản phẩm, hệ thống xác định xem nhận xét về sản phẩm ấy là tốt hay xấu. Phân lớp này thường là phân lớp ở mức tài liệu. Thông tin được phát hiện không mô tả chi tiết về những gì mọi người thích hay không thích.

Mô hình bài toán:

- Tập đánh giá D = {di}
- Hai lớp đánh giá Pos(tích cực) và Neg(Tiêu cực)
- Bộ phân lớp sẽ phân di vào một trong hai lớp Pos/Neg

Ví dụ: Chúng ta có câu đánh giá sau: *Điện thoại này đẹp quá*. Hệ thống thực hiện phân lớp câu quan điểm trên là tích cực hay tiêu cực

1.4.2. Khai phá quan điểm so sánh

Ngoài cách biểu diễn các quan điểm bằng cách trực tiếp nhận xét về đối tượng còn có một cách đánh giá là bằng cách so sánh đối tượng muốn nhận xét với một đối tượng khác. Ví dụ, khi một người nói một cái gì đó là tốt hay xấu, người ta thường yêu cầu *so với cái gì?*. Vì vậy, một trong những cách quan trọng nhất của đánh giá đối tượng là so sánh trực tiếp nó với một đối tượng tương tự khác.

Ví dụ:

"*Kiểu dáng điện thoại Samsung galaxy S4 đẹp hơn galaxy S3*" ở đây đặc trưng kiểu dáng của Samsung galaxy S4 là đối tượng được nhận xét.

1.4.3. Tổng hợp quan điểm

– Tổng hợp quan điểm dựa trên khía cạnh

Bài toán này đi chi tiết vào mức khía cạnh để làm rõ đối tượng mà người đưa ra quan điểm thích hay không thích. Đối tượng ở đây có thể là sản phẩm, dịch vụ, một chủ đề, một cá nhân hay tổ chức.

Ví dụ, trong một câu đánh giá "*pin của chiếc điện thoại Sony này là không tốt"* thì quan điểm ở đây phát biểu về tính năng *pin* của sản phẩm chiếc *điện thoại Sony*. Yêu cầu đầu ra là một bản tổng hợp chi tiết các chiều hướng quan điểm đến các tính năng của sản phẩm.

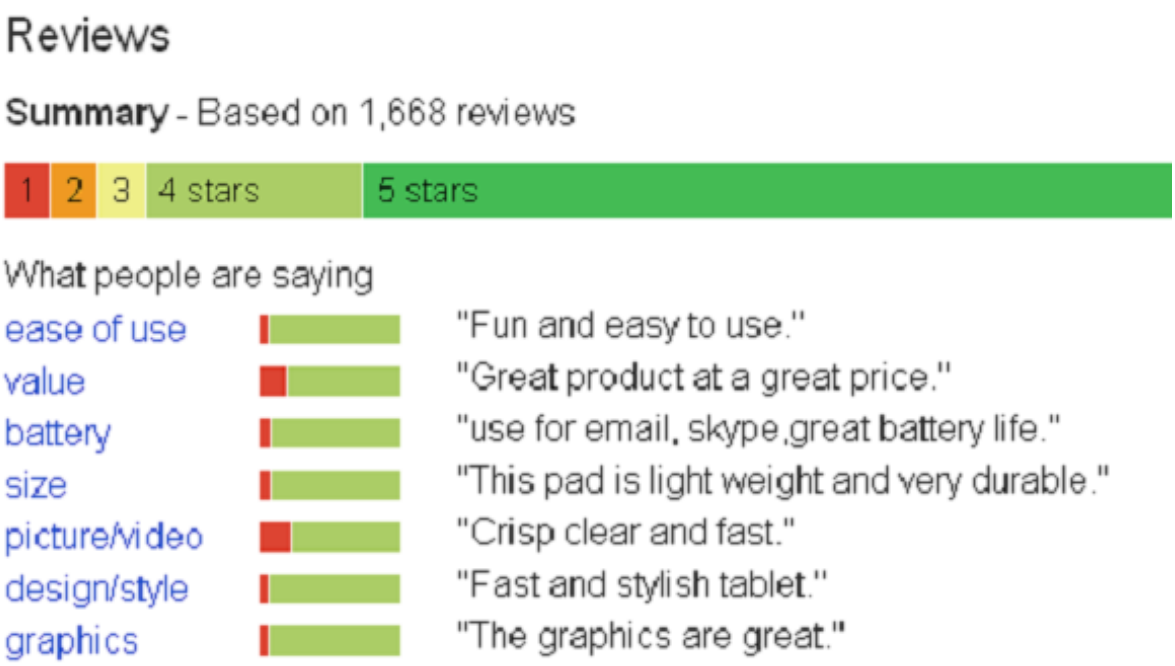
– Tổng hợp quan điểm không dựa trên khía cạnh

Bài toán này đi vào tổng hợp quan điểm của đối tượng, đầu ra là một bản tổng hợp tóm tắt chung chung về đối tượng mà không có các đánh giá cụ thể về từng khí cạnh của đối tượng. Ví dụ như "*Chiếc điện thoại Sony này là chưa tốt, khách hàng chưa hài lòng về nó, chúng ta cần phải cải tiến thêm".*

Chương 2. Các phương pháp tiếp cận bài toán tổng hợp quan điểm theo tính năng của sản phẩm

Hầu hết các ứng dụng khai thác quan điểm cần phải nghiên cứu ý kiến của một số lượng lớn quan điểm từ nhiều người khác nhau. Một ý kiến từ một người duy nhất thường là không đủ để đưa ra quyết định. Chính vì vậy, cần thiết phải có một công cụ tự động thực hiện tổng hợp quan điểm từ nhiều người, nhiều đối tượng khác nhau . Như đã nói ở trên, khi đưa ra quan điểm về một đối tượng hay một sự kiện nào đó, ngoài những ý kiến đưa ra đánh giá về đối tượng, mọi người thường hay đưa ra ý kiến đánh giá về một khía cạnh nào đó của đối tượng. Vì vậy, tổng hợp quan điểm theo khía cạnh được áp dụng rãi trong các ngành công nghiệp (Hình 2.1).Thực tế, các ý kiến phát hiện có thể được lưu trữ trong các bảng cơ sở dữ liệu. Sau đó, dữ liệu có thể được hiển thị để người sử dụng dễ hình dung kết quả theo các cách khác nhau như dạng biểu đồ dạng cột hoặc biểu đồ hình tròn để có thể biết được tổng quan về các ý kiến được người tiêu dùng đưa ra đánh giá như thế nào.

Các nhà nghiên cứu cũng đã nghiên cứu tổng hợp ý kiến có thể được thực hiện theo cách truyền thống tức là đưa ra một bản tóm tắt văn bản ngắn [3] hay còn gọi là phương pháp tổng hợp ý kiến không theo khía cạnh. Một bản tóm tắt này cung cấp cho người đọc một cái nhìn tổng quát về những gì mọi người nghĩ về một sản phẩm hoặc dịch vụ. Một điểm yếu của một bản tóm tắt dựa trên văn bản như vậy là nó không phải là định lượng mà chỉ có chất lượng, và chúng thường không thích hợp cho mục đích phân tích. Ví dụ, một bản tóm tắt văn bản thông thường có thể đưa ra kết quả "*Hầu hết mọi người không thích sản phẩm này*". Tuy nhiên, một bản tóm tắt định lượng có thể nói rằng 60% số người không thích sản phẩm này và 40% trong số họ thích nó. Trong hầu hết các ứng dụng, việc định lượng là rất quan trọng. Thay vì tạo ra một bản tóm tắt văn bản trực tiếp từ đánh giá đầu vào, chúng ta cũng có thể tạo ra một bản tóm tắt văn bản dựa trên các kết quả khai thác từ các biểu đồ hình cột hoặc biểu đồ hình tròn [14].



Hình 2.1. Một ví dụ về tổng hợp quan điểm dựa trên tính năng của sản phẩm iPad [22]

Thông thường, tổng hợp quan điểm qua tính năng của sản phẩm gồm các bước sau [26]:

- Xác định đối tượng
- Trích xuất tính năng
- Nhóm các tính năng
- Phân lớp quan điểm
- Lọc quan điểm Spam

2.1. Xác định đối tượng

Trong khai phá quan điểm, việc đầu tiên là phải định nghĩa được các đối tượng (thực thể) trong các câu đánh giá. Vấn đề này cực kỳ quan trọng, vì nếu không xác định được đối tượng trong câu thì câu quan điểm đó dường như không có ý nghĩa. Hơn nữa, đối với các trang mạng xã hội, các ý kiến spam không phải là không có khi mọi người đang trao đổi về đối tượng này thì vẫn có những ý kiến trao đổi về một vài đối tượng khác xen vào. Hoặc có thể họ so sánh tính năng của đối tượng này với tính năng của đối tượng khác (quan điểm so sánh). Nếu như hệ thống không xác định được đối tượng của câu đánh giá thì kết quả thu được sẽ không chính xác.

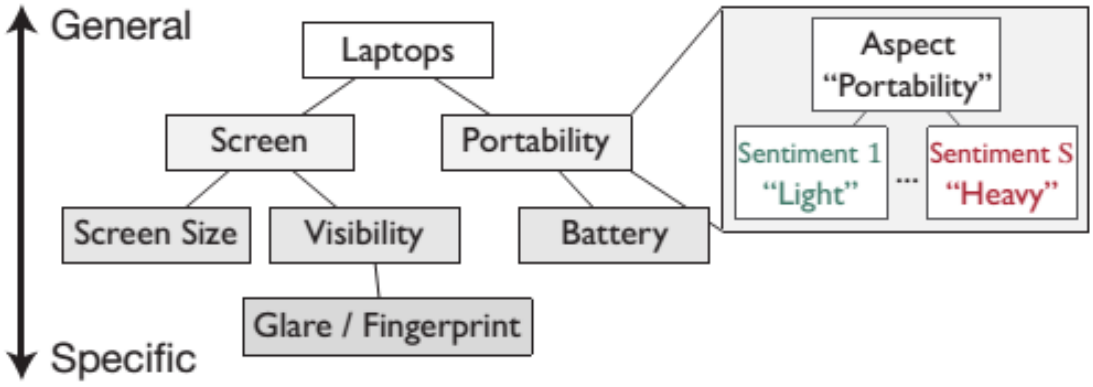
Đầu tiên, chúng ta cùng trao đổi về một vấn đề cụ thể của trích xuất tên thực thể (đối tượng) trong lĩnh vực khai phá quan điểm. Tên của một đối tượng, một tổ chức có thể được

người dùng gọi theo nhiều cách khác nhau. Ví dụ, "*Motorola*" có thể được viết là "*Moto*" hoặc "*Mot*". Việc dùng từ điển sẵn có để xác định đối tượng không phải là tối ưu vì đó là cách gọi người sử dụng, chúng ta không thể đưa ra được hết các trường hợp theo phương pháp thủ công. Vì thế, cần cho một hệ thống tự động phát hiện ra chúng từ trong cơ sở dữ liệu (trang web đánh giá, blog và các diễn đàn thảo luận).

Ding và Liu [12] đề xuất các vấn đề về giải pháp **coreference** (sự đồng nghĩa) đối với thực thể và khía cạnh. Nhiệm vụ nhằm mục đích xác định đề cập đến các thực thể hoặc khía cạnh. Bài báo sử dụng phương pháp học có giám sát. Những điểm chính là việc thiết kế và thử nghiệm hai tính năng quan điểm liên quan, nó chỉ ra rằng phân tích quan điểm đã được sử dụng cho mục đích giải quyết vấn đề coreference[13]. Chức năng đầu tiên được dựa trên phân tích tình cảm của câu thông thường và câu so sánh, và ý tưởng về sự thống nhất trong tâm lý. Ví dụ như "*Chiếc điện thoại Nokia là tốt hơn so với điện thoại Motorola. Nó rẻ quá*". Ở đây, "*nó*" có nghĩa là "*điện thoại Nokia*" vì trong câu đầu tiên, quan điểm về "*điện thoại Nokia*" theo chiều hướng dương (quan điểm tích cực), nhưng nó là chiều hướng âm (quan điểm tiêu cực) cho "*điện thoại Motorola*", và câu thứ hai là tích cực. Do đó, kết luận rằng "*Nó*" là "*điện thoại Nokia*" bởi vì người ta thường bày tỏ quan điểm một cách nhất quán. Ở đây, không chắc rằng "*Nó*" là "*điện thoại Motorola*". Tuy nhiên, nếu chúng ta thay đổi "*Nó rẻ quá*" đến "*Nó cũng đắt*". Trong trường hợp này, "*Nó*" có thể thay thế cho "*điện thoại Motorola*". Để có được tính năng này, hệ thống cần phải có khả năng xác định ý kiến tích cực và tiêu cực thể hiện ở cả câu thông thường và câu so sánh.

Chức năng thứ hai xem xét những gì các thực thể và các khía cạnh được sửa đổi theo những gì quan điểm bày tỏ. Ví dụ: "*Tôi đã mua một chiếc điện thoại Nokia ngày hôm qua. Chất lượng âm thanh tốt. Giá của nó rẻ quá."* Câu hỏi đặt ra là "*nó*" là "*chất lượng âm thanh*" hay "*điện thoại Nokia".* Rõ ràng, chúng ta biết rằng "*nó"* là "*điện thoại Nokia*" vì "*chất lượng âm thanh*" không có "*giá rẻ"*. Để có được chức năng này, hệ thống cần phải xác định những gì từ quan điểm thường được kết hợp với các thực thể hoặc các khía cạnh nào. Các mối quan hệ như vậy phải được khai thác từ các ngữ liệu. Tuy nhiên, hai chức năng này là đặc trưng ngữ nghĩa mà các phương pháp giải quyết coreference chung hiện nay chưa giải quyết được [14]

Kim & Zhang cũng đã sử dụng mô hình phân cụm phân cấp (*Hierarchical Aspect Sentiment Model* - **HASM**) [9]. Trong HASM, toàn bộ cấu trúc là một cây. Mỗi nút chính nó là một cây hai cấp, có nút gốc đại diện cho một khía cạnh và các nút lá đại diện cho chiều hướng tình cảm của nó.



Hình 2.2. Một phần cây phân cấp được khai thác từ mô hình HASM, ứng dụng cho việc khai phá laptop [9]

Theo như hình vẽ ta thấy, đối tượng "*laptop*" gồm có 2 tính năng là "*Screen*" và "*Portability*". "*Portability*" có thuộc tính con là "*Battery*" và có các quan điểm là "*Light*" và "*Heavy*". Như vậy, "*Portability*" vừa có thể là đối tượng cũng có thể là khía cạnh cho đối tượng.

Ngoài ra, còn có rất nhiều các kỹ thuật học máy khác dùng để nhận dạng đối tượng như việc sử dụng mô hình HMM [8,18] và CRF [10] để nhận dạng.

2.2. Trích xuất khía cạnh

Trên thế giới hiện nay, có một số phương pháp dùng để trích xuất khía cạnh cho đối tượng như sử dụng danh từ và cụm danh từ thường xuyên, luật lan truyền kép, mô hình chủ đề,... Chúng ta cùng tìm hiểu một số phương pháp sau:

2.2.1. Sử dụng danh từ và cụm danh từ thường xuyên

Hu và Liu [7] đã đề xuất một phương pháp trích xuất tính năng của sản phẩm dựa theo luật kết hợp. Ý tưởng của phương pháp này có thể được tóm tắt qua hai bước chính. Đầu tiên là tìm các danh từ và cụm danh từ và coi chúng như là các tính năng của sản phẩm, sau đó là sử dụng mối quan hệ của tính năng và từ quan điểm để định nghĩa lại các tính năng.

Bước 1: Tìm các danh từ và cụm danh từ. Danh từ và cụm danh từ được xác định bởi việc gán nhãn từ loại (POS tagger). Xác định tần số xuất hiện của các danh từ và cụm danh từ. Tần suất xuất hiện được xác định theo kinh nghiệm qua tập dữ liệu. Vì thông thường, các danh từ được người dùng đánh giá đến nhiều thường là các tính năng quan trọng. Nội dung trong các câu đánh giá rất đa dạng. Do đó, các danh từ ít xuất hiện thường là các tính năng không quan trọng. Trong trường hợp dữ liệu lớn thì có thể loại bỏ chúng đi.

Bước 2: Tìm các tính năng ít xuất hiện bằng cách khai thác mối quan hệ giữa các tính năng và các từ quan điểm. Trong bước 1 có thể hệ thống sẽ bỏ qua một số các tính năng mà thực tế chúng là các khía cạnh quan trọng. Trong bước này, hệ thống sẽ thực hiện tìm kiếm các khía cạnh đó. Ý tưởng của hệ thống được thể hiện như sau: Các từ quan điểm thể hiện quan điểm cho các khía cạnh thường xuyên cũng có thể sử dụng để thể hiện quan điểm cho các khía cạnh không thường xuyên. Ý tưởng sử dụng các danh từ và cụm danh từ để trích xuất khía cạnh là đơn giản nhưng hiệu quả.

Ví dụ:

"*Hình ảnh này trông rất đẹp*"

"*Màn hình đẹp*"

Giả sử chúng ta đã tìm thấy từ "*Hình ảnh*" là một tính năng cho sản phẩm ở bước 1 và "*đẹp*" là một từ quan điểm. Xét ví dụ thứ 2, dựa vào cấu trúc ngữ pháp của câu thì màn hình là một danh từ. Hơn nữa, "*bức ảnh*" và "*màn hình*" đều có sự kết hợp với từ quan điểm "đẹp" để tạo thành câu. Nên trong bước này, tìm được "*màn hình*" cũng là một khía cạnh của đối tượng.

2.2.2. Sử dụng mối quan hệ của từ quan điểm và khía cạnh

Năm 2011, Qiu [17] đã phát triển ý tưởng trên theo luật lan truyền kép. Phương pháp cần một bộ từ quan điểm làm điều kiện đầu vào. Từ quan điểm có thể được nhận ra bởi các khía cạnh và các khía cạnh có thể được định nghĩa bởi từ quan điểm đã biết. Những từ quan điểm và các khía cạnh đã được trích xuất được sử dụng để tìm từ quan điểm mới và khía cạnh mới. Quá trình lan truyền này kết thúc khi không thể tìm ra được thêm từ quan điểm và khía cạnh mới. Và quá trình này được gọi là lan truyền kép. Các quy luật trích xuất được phát hiện dựa trên mối quan hệ khác nhau giữa các khía cạnh và từ quan điểm. Mối quan hệ này thường được thể hiện bằng cấu trúc ngữ pháp trong câu. Phương pháp này chỉ sử dụng một quan hệ phụ thuộc gọi là phụ thuộc trực tiếp vào mối quan hệ hữu ích. Một phụ thuộc trực tiếp chỉ ra rằng một từ phụ thuộc vào một từ khác mà không có bất kỳ một từ khác xen vào trong mô hình phụ thuộc đó. Phương pháp này coi các từ quan điểm là các tính từ và các khía cạnh là danh từ và cụm danh từ.

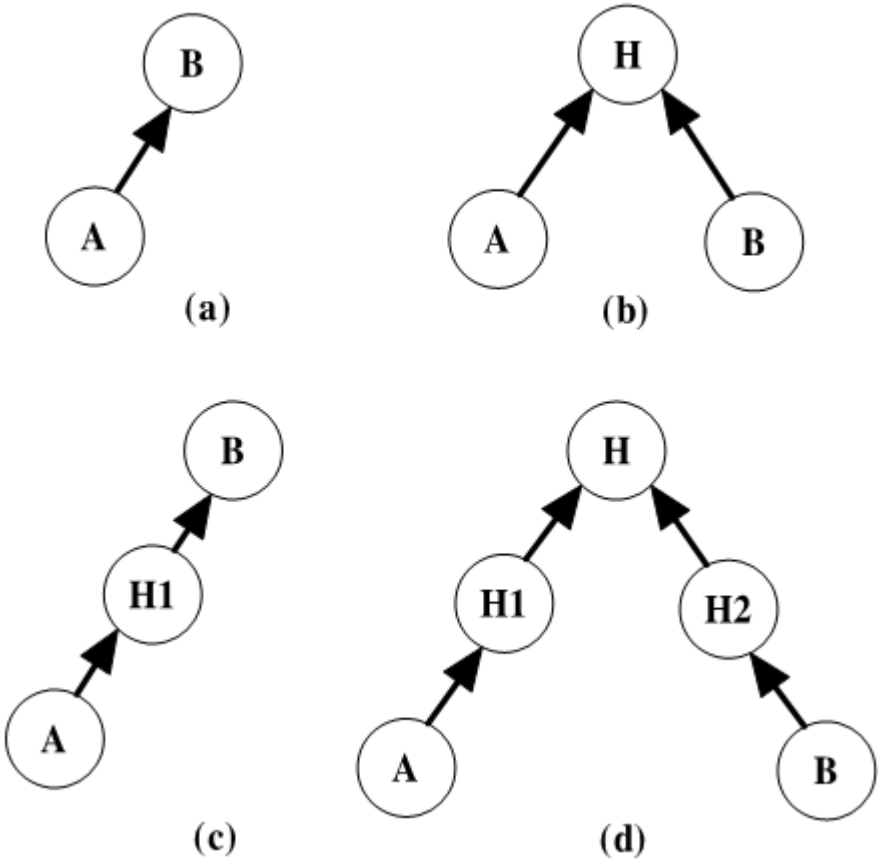
Tuy nhiên, luật lan truyền kép làm việc tốt trong tập dữ liệu trung bình nhưng đối với tập dữ liệu lớn và nhỏ thì phương pháp này có độ tin cậy và độ hồi tưởng thấp. Lý do là mô hình này dựa trên mối quan hệ trực tiếp, đối với tập dữ liệu lớn sẽ có nhiều dữ liệu nhiễu. [22]

Luật lan truyền kép phải thực hiện được 4 nhiệm vụ sau:

- Trích xuất khía cạnh từ các từ quan điểm
- Trích xuất khía cạnh từ các khía cạnh đã được trích xuất
- Trích xuất từ quan điểm từ các khía cạnh đã được trích xuất
- Trích xuất từ quan điểm từ các từ quan điểm đã được trích xuất

Như vậy, điểm mấu chốt là phải xác định được mối quan hệ phụ thuộc giữa các từ trong câu. Thông thường, trong câu có hai loại hình quan hệ cho các từ [17]. Đó là quan hệ trực tiếp và quan hệ gián tiếp

- Quan hệ phụ thuộc trực tiếp là quan hệ mà một từ có quan hệ trực tiếp với một từ mà không có từ thứ ba xen vào hoặc cả hai từ cùng có quan hệ với một từ thứ ba.
- Quan hệ phụ thuộc gián tiếp là quan hệ mà một từ quan hệ với một từ khác thông qua một từ thứ ba hoặc cả hai có cùng quan hệ với một từ thứ ba thông qua những từ khác.



Hình 2.3. Một ví dụ về quan hệ giữa từ A và từ B

Một vài trường hợp quan hệ từ trực tiếp và gián tiếp được thể hiện trong hình 4. Trường hợp (a) A và B có quan hệ trực tiếp với nhau; trong trường hợp (b) A và B đều có quan hệ trực tiếp với H nên (a) và (b) là quan hệ từ phụ thuộc trực tiếp.

Trường hợp (c) A quan hệ với B thông qua H1 còn (d) A và B có quan hệ với H thông qua H1 và H2 nên (c) và (d) là ví dụ minh họa cho quan hệ từ phụ thuộc gián tiếp.

	Observations	**Output**	**Examples**
$R1_1$ (OA-Rel)	$O \rightarrow O\text{-}Dep \rightarrow A$ **s.t.** $O \in \{O\}$, $O\text{-}Dep \in \{MR\}$, $POS(A) \in \{NN\}$	$a = A$	*The phone has a good "screen".* *good* **→mod→** *screen*
$R1_2$ (OA-Rel)	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow A\text{-}Dep \leftarrow A$ **s.t.** $O \in \{O\}$, $O/A\text{-}Dep \in \{MR\}$, $POS(A) \in \{NN\}$	$a = A$	*"iPod" is the best mp3 player.* *best* **→mod→** *player* **←subj←** *iPod*
$R2_1$ (OA-Rel)	$O \rightarrow O\text{-}Dep \rightarrow A$ **s.t.** $A \in \{A\}$, $O\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as $R1_1$ with *screen* as the known word and *good* as the extracted word
$R2_2$ (OA-Rel)	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow A\text{-}Dep \leftarrow A$ **s.t.** $A \in \{A\}$, $O/A\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as $R1_2$ with *iPod* is the known word and *best* as the extract word.
$R3_1$ (AA-Rel)	$A_{i(j)} \rightarrow A_{i(j)}\text{-}Dep \rightarrow A_{j(i)}$ **s.t.** $A_{j(i)} \in \{A\}$, $A_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(A_{i(j)}) \in \{NN\}$	$a = A_{i(j)}$	*Does the player play dvd with audio and "video"?* *video* **→conj→** *audio*
$R3_2$ (AA-Rel)	$A_i \rightarrow A_i\text{-}Dep \rightarrow H \leftarrow A_j\text{-}Dep \leftarrow A_j$ **s.t.** $A_i \in \{A\}$, $A_i\text{-}Dep = A_j\text{-}Dep$ OR ($A_i\text{-}Dep = subj$ AND $A_j\text{-}Dep = obj$), $POS(A_j) \in \{NN\}$	$a = A_j$	*Canon "G3" has a great len.* *len* **→obj→** *has* **←subj←** *G3*
$R4_1$ (OO-Rel)	$O_{i(j)} \rightarrow O_{i(j)}\text{-}Dep \rightarrow O_{j(i)}$ **s.t.** $O_{j(i)} \in \{O\}$, $O_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	*The camera is amazing and "easy" to use.* *easy* **→conj→** *amazing*
$R4_2$ (OO-Rel)	$O_i \rightarrow O_i\text{-}Dep \rightarrow H \leftarrow O_j\text{-}Dep \leftarrow O_j$ **s.t.** $O_i \in \{O\}$, $O_i\text{-}Dep = O_j\text{-}Dep$ OR ($O_i / O_j\text{-}Dep \in \{pnmod, mod\}$), $POS(O_j) \in \{JJ\}$	$o = O_j$	*If you want to buy a sexy, "cool", accessory-available mp3 player, you can choose iPod.* *sexy* **→mod→** *player* **←mod←** *cool*

Hình 2.4.Một ví dụ về trích xuất khía cạnh của đối tượng của Qiu

Luật lan truyền kép được thực hiện qua việc khai thác mối quan hệ giữa từ quan điểm và khía cạnh trong câu. Đầu tiên là hệ thống phải phân tích cú pháp của các từ trong câu để xác định từ loại của các từ (từ quan điểm là các tính từ còn tính năng là các danh từ và cụm danh từ trong câu). Sau đó hệ thống dựa vào mối quan hệ của tính từ và danh từ trong câu để trích xuất ra các từ quan điểm và các khía cạnh

Hình 2.4 là một ví dụ về trích xuất khía cạnh và từ quan điểm theo luật lan truyền kép. Cột 1 là mã luật. Cột 2 là quan hệ phụ thuộc của các từ trong câu. Cột 3 là đầu ra và cột 4 là ví dụ minh họa.

Trong đó:

- OA-Rel: Quan hệ từ giữa từ quan điểm và khía cạnh
- AA-Rel: Quan hệ từ giữa 2 khía cạnh
- OO-Rel: Quan hệ từ giữa 2 từ quan điểm
- Mỗi một quan hệ giữa OA-Rel, OO-Rel hoặc AA-Rel có thể được xây dựng như một bộ ba <POS (w_i), R, POS (w_j)>, trong đó POS (w_i) Là từ loại của từ w_i, và R là quan hệ.
- o (hoặc a) là viết tắt của các đầu ra của từ quan điểm (hay một khía cạnh).
- {O} (hoặc {A}) là tập hợp các từ ý kiến (hoặc tập hợp các khía cạnh) đã được trích xuất
- H có nghĩa là bất kỳ từ nào.
- POS (O (hoặc A)) và O (hoặc A)-Dep tiêu chuẩn cho thẻ POS và sự phụ thuộc liên quan của từ O (hoặc A) tương ứng.
- {JJ} và {NN} là tập hợp các thẻ POS của từ quan điểm và các khía cạnh tương ứng. {JJ}bao gồm JJ, JJR và JJS; {NN} bao gồm NN và NNS.
- {MR} bao gồm các mối quan hệ phụ thuộc mô tả mối quan hệ giữa từ quan điểm và các khía cạnh (mod, pnmod, subj, s, obj, obj2 và desc).
- {CONJ} chỉ chứa *conj*. Các mũi tên có nghĩa là phụ thuộc. Ví dụ, O → O-Dep → A có nghĩa là O phụ thuộc vào A đến một mối quan hệ cú pháp O-Dep. Cụ thể, nó sử dụng luật $R1_i$ để trích xuất các khía cạnh (a) sử dụng những từ quan điểm (O), $R2_i$ để trích xuất từ quan điểm (o) sử dụng các khía cạnh (A), $R3_i$ để trích xuất các khía cạnh (a) sử dụng các khía cạnh đã được trích xuất (A_i) và $R4_i$ để trích xuất ý kiến từ (o) sử dụng những từ quan điểm được biết đến (O_i).

Ví dụ trong quan hệ từ $R1_1$ của bảng trên. Đây là quan hệ từ sử dụng từ quan điểm để trích xuất ra các tính năng đối với các câu có cấu trúc O → O-Dep → A. Theo như ví dụ "*The phone has a good screen"*. Thì "*good"* ở đây là một từ quan điểm; "*good"* và "*screen"* có quan hệ trực tiếp với nhau; hơn nữa "*screen"* là một danh từ (NN) nên tìm ra được "*screen"* là một tính năng.

Giải thuật của luật lan truyền kép được Qiu đề xuất như trong hình 7. Để đảm bảo nội dung được khách quan, tránh sai sót và mất nghĩa tôi xin được trích dẫn nguyên văn bằng tiếng anh. Để dễ hiểu hơn, tôi xin trình bày lại giải thuật bằng ngôn ngữ tiếng Việt như trong hình 8

Input: Opinion Word Dictionary {O}, Review Data R
Output: All Possible Features {F}, The Expanded Opinion Lexicon {O-Expanded}
Function:
1. {O-Expanded} = {O}
2. $\{F_i\} = \emptyset, \{O_i\} = \emptyset$
3. for each parsed sentence in R
4. if(Extracted features not in {F})
5. Extract features $\{F_i\}$ using $R1_1$ and $R1_2$ based on opinion words in {O-Expanded}
6. endif
7. if(Extracted opinion words not in {O-Expanded})
8 Extract new opinion words $\{O_i\}$ using $R4_1$ and $R4_2$ based on opinion words in {O-Expanded}
9. endif
10. endfor
11. Set $\{F\} = \{F\} + \{F_i\}$, $\{O\text{-Expanded}\} = \{O\text{-Expanded}\} + \{O_i\}$
12. for each parsed sentence in R
13. if(Extracted features not in {F})
14. Extract features {F'} using $R3_1$ and $R3_2$ based on features in $\{F_i\}$
15. endif
16. if(Extracted opinion words not in {O-Expanded})
17. Extract opinion words {O'} using $R2_1$ and $R2_2$ based on features in $\{F_i\}$
18. endif
19. end for
20. Set $\{F_i\} = \{F_i\} + \{F'\}$, $\{O_i\} = \{O_i\} + \{O'\}$
21. Set $\{F\} = \{F\} + \{F'\}$, $\{O\text{-Expanded}\} = \{O\text{-Expanded}\} + \{O'\}$
22. Repeat 2 till size($\{F_i\}$) = 0, size($\{O_i\}$) = 0

Hình 2.5. Giải thuật lan truyền kép

Đầu vào: Bộ từ quan điểm {O}; dữ liệu {R}

Đầu ra: Tất cả các tính năng {F}, và bộ từ quan điểm mở rộng {O-E}

Giải thuật:

1. {O-E}={O}
2. $\{F_i\}=\varnothing$
3. Với mỗi câu được phân tách cú pháp trong R
   - 3.1. Nếu các tính năng đã được trích xuất không có trong F thì

   Trích xuất tính năng F_i sử dụng $R1_1$ và $R1_2$ dựa vào bộ từ quan điểm trong {O-E}

   - 3.2. Nếu các từ quan điểm đã được trích xuất không có trong {O-E}

   Trích xuất từ quan điểm mới O_i sử dụng $R4_1$ và $R4_2$ dựa vào bộ từ quan điểm {O-E}

4. Đặt $\{F\}= \{F\}+\{F_i\}$; $\{O\text{-}E\}=\{O\text{-}E\} +\{O_i\}$
5. Với mỗi câu được phân tách cú pháp trong R
   - 5.1. Nếu các tính năng đã được trích xuất không có trong F thì trích xuất các tính năng F’ sử dụng $R3_1$ và $R3_2$ dựa vào các tính năng trong F_i
   - 5.2. Nếu các từ quan điểm đã được trích xuất không có trong {O-E} thì trích xuất các từ quan điểm O’ sử dụng $R2_1$ và $R2_2$ dựa vào các tính năng trong F_i
6. $\{F_i\}=\{F_i\}+\{F'\}$; $\{O_i\}= \{O_i\}+ \{O'\}$
7. $\{F\} = \{F\} +\{F'\}$; $\{O\text{-}E\} =\{O\text{-}E\} + \{O'\}$
8. Lặp cho đến khi $\{F_i\}=0$, $\{O_i\}=0$

Hình 2.6. Giải thuật luật lan truyền kép (tiếng Việt)

Năm 2010, Zhang đã đề xuất một phương pháp được phát triển dựa theo luật lan truyền kép [23]. Phương pháp này gồm 2 bước là trích xuất ra các khía cạnh và xếp hạng các khía cạnh. Về việc trích xuất ra các khía cạnh, vẫn sử dụng luật lan truyền kép. Tuy nhiên, sau khi trích xuất, hệ thống sẽ xếp hạng các khía cạnh đã trích xuất, các khía cạnh quan trọng

sẽ được xếp hạng cao, còn các khía cạnh không quan trọng hoặc là nhiễu sẽ có thứ hạng thấp.

Luật lan truyền kép cũng được nhóm nghiên cứu của Hà Quang Thụy [27] nghiên cứu và thử nghiệm để trích xuất ra các tính năng của sản phẩm đối với các đánh giá của người dùng đối với sản phẩm điện thoại di động dành cho tiếng Việt với độ chính xác khoảng 87%.

2.2.3. Mô hình chủ đề.

Trong những năm gần đây, các mô hình thống kê theo chủ đề đã nổi lên như là một phương pháp khám phá các chủ đề từ một bộ ngữ liệu các tài liệu văn bản. Mô hình chủ đề là một phương pháp học không giám sát, nó cho rằng mỗi tài liệu bao gồm một hỗn hợp các chủ đề và mỗi chủ đề là một phân bố xác suất của các từ. Một mô hình chủ đề cơ bản là một mô hình sinh sản tài liệu trong đó quy định một xác suất mà các tài liệu có thể được tạo ra. Các đầu ra của mô hình chủ đề là một tập hợp các cụm từ. Mỗi cụm tạo thành một chủ đề và là một phân bố xác suất của các từ trong bộ ngữ liệu.

Có hai mô hình cơ bản chính, pLSA (ProbabilisticLatent Semantic Analysis) [6] và LDA (Latent Dirichlet Allocation) [2]. Mô hình chủ đề được áp dụng để trích xuất ra các khía cạnh và đã được phát triển trong các nghiên cứu của Gou, Moghadam and Ester, Titov and McDonald [4, 16, 19].

Titov and McDonald [19] đã đề xuất phương pháp MG-LDA (Multi Grain - Latent Drichlet Analysis) để trích xuất ra các khía cạnh bằng cách phát hiện ra các chủ đề chung và riêng. Ví dụ như trong câu "*Giao thông ở Luân Đôn khá thuận tiện, trạm dừng đi bộ khoảng 8 phút và mất khoảng 1,5$ để đi xe bus".* Mô hình đã phân biệt được *Luân Đôn* là chủ đề chung và *trạm dừng, xe bus* là các chủ đề con. Sau đó, nhóm tác giả đã cải tiến phương pháp này và đề xuất ra một phương pháp mới là Multi-Aspect Sentiment model (MAS) [20]. Nó bao gồm hai phần, phần đầu tiên là dựa trên MG-LDA để xác định chủ đề mà nó đại diện của các khía cạnh có thể đánh giá. Phần thứ hai là một bộ phân loại cho từng khía cạnh, nó cố gắng suy ra các ánh xạ giữa các chủ đề chung và các khía cạnh với sự trợ giúp của các khía cạnh cụ thể đã xếp hạng được cung cấp cùng với các văn bản đánh giá.

2.3. Nhóm các từ chỉ cùng một khía cạnh

Phân nhóm khía cạnh cho thấy các khía cạnh có sự tương đồng về ngữ nghĩa là rất cần thiết cho các ứng dụng quan điểm. Mặc dù từ điển WordNet và một số từ điển khác có thể hỗ trợ, nhưng chúng vẫn chưa đầy đủ do thực tế, nhiều từ đồng nghĩa là miền phụ thuộc

trong một lĩnh vực cụ thể nào đó. Ví dụ, *hình ảnh* và *phim* là từ đồng nghĩa trong đánh giá bộ phim, nhưng chúng không phải là từ đồng nghĩa trong đánh giá máy ảnh kỹ thuật số. *Hình ảnh* là có liên quan tới *ảnh*, trong khi *phim* đề cập đến *video*. Cũng cần lưu ý rằng mặc dù hầu hết các cách thể hiện khía cạnh khác nhau của một khía cạnh là từ đồng nghĩa trong một miền nào đó, nhưng chúng không phải là luôn luôn đồng nghĩa. Ví dụ, "*đắt*" và "*giá rẻ*" có thể đều nói đến khía cạnh *giá* nhưng chúng không phải là từ đồng nghĩa của giá cả.

Năm 2010, Zhai đã đề xuất một phương pháp học bán giám sát để nhóm các khía cạnh vào nhóm khía cạnh do người dùng định nghĩa [31]. Mỗi nhóm đại diện cho một khía cạnh cụ thể. Để phản ánh các nhu cầu của người sử dụng, đầu tiên họ gán nhãn bằng tay một số lượng nhỏ các khía cạnh cho mỗi nhóm. Hệ thống sau đó phân phần còn lại của các khía cạnh cho mỗi nhóm bằng cách tự động phát hiện ra các nhóm thích hợp sử dụng học bán giám sát dựa trên các mẫu có nhãn và các mẫu không có nhãn. Phương pháp này sử dụng thuật toán Expectation–Maximization (EM). Hai mảng kiến thức trước khi được sử dụng để cung cấp một khởi tạo tốt hơn cho EM, một là khía cạnh chia sẻ một số từ thông dụng có thể sẽ thuộc về cùng một nhóm, và hai là biểu hiện khía cạnh đó là những từ đồng nghĩa trong một từ điển có khả năng thuộc cùng một nhóm.

Năm 2012, Mauge đã sử dụng một dữ liệu ngẫu nhiên dựa trên thuật toán phân nhóm tối đa cho các khía cạnh nhóm trong một sản phẩm [30]. Đầu tiên, nó huấn luyện một bộ phân loại Maximum Entropy để xác định p xác suất mà hai khía cạnh là từ đồng nghĩa. Sau đó, một đồ thị vô hướng có trọng số được xây dựng. Mỗi đỉnh đại diện cho một khía cạnh. Mỗi trọng số cạnh tỉ lệ với *p* xác suất giữa hai đỉnh. Cuối cùng, phương pháp phân vùng đồ thị gần đúng được sử dụng cho việc nhóm các khía cạnh của sản phẩm

Năm 2011, nhóm nghiên cứu của Hà Quang Thụy cũng sử dụng phương pháp học bán giám sát sử dụng kết hợp mô hình phân cụm HAC (Hierarchical Agglomerative Clustering) và phân lớp SVM-kNN (Support Vecto Machine – k Nearest Neighbor) để nhóm các từ chỉ cùng một tính năng vào một nhóm [27]. Giải thuật được trình bày cụ thể trong hình 2.7.

Ban đầu, nhóm tác giả đã sử dụng giải thuật phân cụm HAC để tạo ra các mẫu huấn luyện. Thứ nhất, các dữ liệu thiết lập để nhóm là tất cả các câu đánh giá, trong đó một từ tính năng ẩn (hiện) xảy ra. Mỗi từ tính năng trích xuất được tương ứng với một túi của từ từ câu đánh giá, trong đó bao gồm các từ tính năng. Sau đó, các bộ dữ liệu đã được bổ sung bằng cách thêm tất cả các câu đánh giá cho tất cả các tính năng từ từ tiếng Việt trực tuyến

- Trang web từ điển tiếng Việt (http://www.tratu.vn). Mỗi câu đánh giá trong từ điển cũng được tương ứng với một túi của từ

Giải thuật phân cụm HAC được thực hiện với ngưỡng 0,5. Các cụm gồm ít nhất hai từ tính năng sẽ được xem xét tiếp. Độ đo tương tự trong thuật toán HAC là độ đo cosin cho

Đầu vào

L: bộ dữ liệu huấn luyện

C: Số nhóm

U: bộ từ tính năng chưa có nhãn

SVM: giải thuật SVM

kNN giải thuật kNN

s: số lượng vector hỗ trợ có trong 1 lớp

t: kích thước mong muốn của bộ huấn luyện

Giải thuật

1. Huấn luyện bộ dữ liệu L với SVM1
2. Lặp lại cho đến khi $\| L \| \geq t * \| L \cup U \|$
   - 2.1. Sử dụng SVM1 gán nhãn của tất cả các từ quan điểm trong U
   - 2.2. Chọn s vector hỗ trợ từ U để làm dữ liệu kiểm chứng giải thuật kNN
   - 2.3. Sử dụng giải thuật kNN với bộ huấn luyện được gán nhãn lại từ s dữ liệu test. Đặt bộ s mẫu được gán nhãn là New
   - 2.4. L<- L ∪ New; U<-U/New (Cập nhật lại L và U)
   - 2.5. Huấn luyện bộ dữ liệu L đã update với SVM2
   - 2.6. SVM1<-SVM2

Kết thúc

3. Sử dụng nhãn học được từ SVM2 cho tất cả các tính năng còn lại trong U.

Hình 2.7. Giải thuật bán giám sát SVM-kNN để nhóm các từ chỉ tính năng

các từ trong bộ túi của từ. Trong mỗi nhóm, chỉ có từ tính năng có số lần xuất hiện cao nhất được chọn là nhãn của nhóm. Bộ dữ liệu huấn luyện được tạo ra.

Độ đo Cosin được xác định như sau:

$$Sim\ (u, v) = \frac{u.v}{\| u \|.\| v \|}$$

trong đó:

$Sim(u, v)$: Khoảng cách (độ tương tự) giữa 2 từ u và v

2.4. Phân lớp chiều hướng quan điểm

Nhiệm vụ này xác định xem quan điểm về các tính năng là tiêu cực, tích cực hay trung lập. Cách thông thường là dựa vào từ quan điểm trong câu [26]. Đối với tiếng Anh, mọi người dựa vào từ điển SentimentWordNet để xác định được chiều hướng của từ quan điểm trong các câu đánh giá. Việc sử dụng SentimentWordnet có những hạn chế nhất định. Có những từ là từ quan điểm trong lĩnh vực này nhưng lại không phải là từ quan điểm trong một lĩnh vực khác. Ví dụ như "*Vỏ màu trắng này đẹp nhỉ*" thì từ *trắng* là một tính từ trong lĩnh vực màu sắc nhưng đối với câu đánh giá trên thì nó lại không phải là từ quan điểm mà nó chỉ là trợ động từ cho từ *vỏ*. Hơn nữa có những từ quan điểm mang tính tích cực trong một văn cảnh này nhưng lại mang tính tiêu cực trong một văn cảnh khác. Ta xét một ví dụ về từ quan điểm "*nhỏ*", nếu như ta nói "*chiếc máy ảnh này có kích thước nhỏ*" thì từ "*nhỏ*" ở đây có thể là có chiều hướng tích cực nhưng khi nói "*máy ảnh này có độ phân giải nhỏ*" thì từ "*nhỏ*" ở đây lại có nghĩa là tiêu cực. Vì vậy, vấn đề xác định chính xác chiều hướng quan điểm trong câu là một vấn đề hết sức khó khăn.

Hơn nữa, trong câu còn có thể chứa những từ quan điểm mang tính tích cực nhưng khi có từ phủ định đằng trước thì nó lại mang tính tiêu cực. Chẳng hạn như "*tốt*" là một từ quan điểm tích cực nhưng khi nói "*không tốt*" thì nó lại trở thành tiêu cực.

2.5. Loại bỏ quan điểm Spam

Ý kiến của phương tiện truyền thông xã hội đang ngày càng được dùng cho các cá nhân và tổ chức trong việc hỗ trợ ra quyết định mua hàng, tiếp thị và thiết kế sản phẩm. Các ý kiến tích cực thường sẽ mang lại lợi nhuận cho doanh nghiệp và các cá nhân. Vì vậy, mọi người có thể tạo ra các ý kiến giả để gia tăng uy tín cho doanh nghiệp mình và hạ uy tín của các đối thủ cạnh tranh. Và các quan điểm đó là không chính xác. Những người này đưa ra các đánh giá như vậy được gọi là Spammer (người đánh giá giả mạo) và các đánh giá của họ được coi là các quan điểm Spam [28, 29].

Thách thức chính của phát hiện quan điểm spam không giống như các hình thức khác của Spam, nó là rất khó, nếu không phải không thể để nhận ra ý kiến giả bằng cách thủ đọc chúng. Đây là một việc khó khăn để tìm thấy các dữ liệu quan điểm spam hỗ trợ cho việc thiết kế và đánh giá thuật toán phát hiện. Đối với các hình thức khác của spam, người ta có thể nhận ra chúng khá dễ dàng

Theo Jindal và Liu, có 3 loại quan điểm Spam [29]:

- Loại 1 (đánh giá giả mạo): Đây là những nhận xét sai sự thật được viết không dựa trên kinh nghiệm chính hãng của các nhà phê bình của việc sử dụng các sản phẩm hay dịch vụ, nhưng được viết dưới dạng ẩn. Họ thường có ý kiến tích cực không chính xác về một số đối tượng (các sản phẩm hoặc dịch vụ) nhằm quảng cáo cho các đối tượng ấy hoặc ý kiến tiêu cực sai lệch về một số đối tượng khác để làm tổn hại danh tiếng của họ.
- Loại 2 (đánh giá chỉ về thương hiệu): Những nhận xét không bình luận về các sản phẩm hoặc dịch vụ cụ thể mà chúng lại được cho là các nhận xét, nhưng chỉ nhận xét về các nhãn hiệu hoặc nhà sản xuất của sản phẩm. Chúng được coi như là Spam,chúng không nhắm vào các sản phẩm cụ thể và thường sai lệch. Ví dụ, một đánh giá cho một máy in HP cụ thể nói: "*Tôi ghét dòng sản phẩm của HP. Tôi không bao giờ mua bất kỳ sản phẩm của chúng*".
- Loại 3 (không đánh giá): Đây không phải là đánh giá. Có hai phân nhóm chính: (1) quảng cáo và (2) các văn bản liên quan khác có chứa không có ý kiến (ví dụ, các câu hỏi, câu trả lời, và các văn bản ngẫu nhiên).

Mục đích chính của việc loại bỏ các quan điểm Spam là xác định mọi *đánh giá giả mạo, nhà phê bình giả mạo, và nhóm phê bình giả mạo*. Ba khái niệm có liên quan rõ ràng là đánh giá giả mạo được viết bởi các nhà phê bình giả và phê bình giả có thể hình thành các nhóm phê bình giả. Việc phát hiện một loại có thể giúp phát hiện của người khác. Tuy nhiên, mỗi người lại có những đặc điểm riêng biệt của nó, có thể được khai thác để phát hiện.

Có 2 phương pháp chính để xác định quan điểm Spam đó là phương pháp học có giám sát và học bán giám sát.

Phương pháp học có giám sát: phương pháp phân lớp được sử dụng trong trường hợp này. Dữ liệu huấn luyện được phân chia thành hai lớp, một lớp chứa các quan điểm có nhãn Spam và một lớp chứa các quan điểm có nhãn là không Spam. Quá trình huấn luyện sẽ phát hiện ra được quan điểm là Spam hay không phải là Spam.

Phương pháp học bán giám sát: phương pháp này sử dụng bộ dữ liệu có nhãn và không có nhãn làm dữ liệu huấn luyện trong quá trình phát hiện ra quan điểm nào là quan điểm Spam.

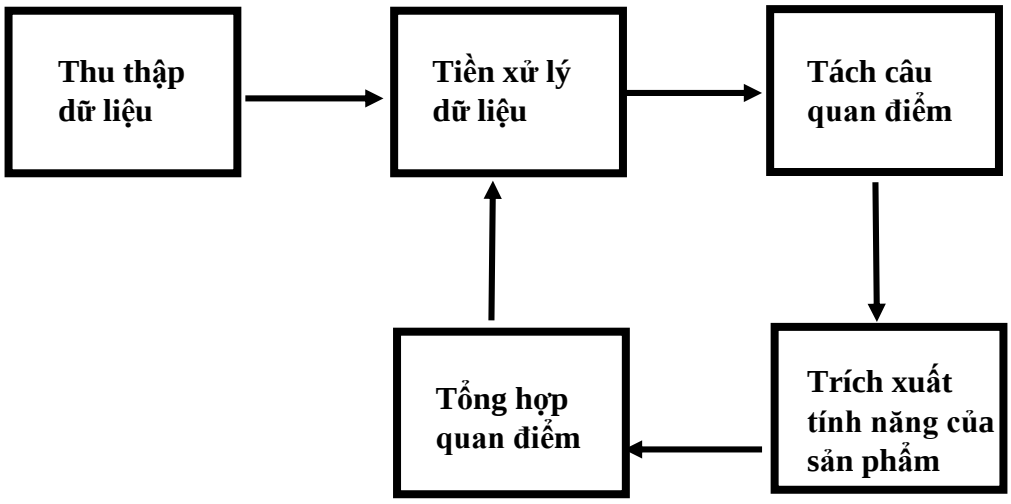
Chương 3. Tổng hợp quan điểm trực tuyến của người tiêu dùng Việt Nam theo tính năng của sản phẩm

Trong luận văn này, tôi hướng đến xây dựng hệ thống tổng hợp quan điểm của người tiêu dùng Việt Nam theo tính năng của sản phẩm. Đầu vào của hệ thống là tập dữ liệu đã được thu thập trên diễn đàn, qua quá trình tiền xử lý dữ liệu, hệ thống trích xuất được các tính năng cho sản phẩm. Kết quả trả về của hệ thống là bản tổng hợp các ý kiến đánh giá của người tiêu dùng theo từng tính năng của sản phẩm. Trong quá trình xây dựng hệ thống, ngoài các khó khăn về vấn đề khai phá dữ liệu như đã trình bày ở trên thì tôi còn gặp một khó khăn lớn về vấn đề xử lý ngôn ngữ tiếng Việt. Dữ liệu khi thu thập từ trên diễn đàn, có những trường hợp người dùng không chỉ đánh giá về một tính năng trên một câu mà họ thường đánh giá về 2, 3 tính năng hoặc nhiều hơn nữa. Tôi xây dựng bộ công cụ tách câu ghép và câu phức thành các câu đơn để thu được kết quả chính xác hơn. Khi đưa ra các ý kiến nhận xét của mình, mọi người thường dùng ngôn ngữ nói để đăng lên diễn đàn. Thông thường, ngôn ngữ nói không tuân theo chuẩn cấu trúc câu chung mà mỗi người sẽ có một cách viết khác nhau. Vì vậy, việc xử lý ngôn ngữ trong tiếng Việt là vô cùng khó khăn. Khi thực hiện tách câu, tôi chỉ quan tâm đến từ loại danh từ và tính từ và các từ nối trong câu, bỏ qua các từ loại khác.

Khi xây dựng hệ thống, tôi bỏ qua bước trích xuất thực thể cho đối tượng, và loại bỏ các quan điểm spam, coi tất cả các ý kiến đánh giá đều là các đánh giá về một đối tượng. Việc thực hiện loại bỏ các quan điểm Spam và các quan điểm đánh giá về các đối tượng khác tôi thực hiện một cách thủ công. Tôi thực hiện trích xuất khía cạnh theo luật lan truyền kép, phân cụm các tính năng và phân lớp quan điểm. Hệ thống của tôi có thể được khái quát hóa qua hình 3.1. Hình 3.1 mô tả khái quát mô hình hệ thống tổng hợp quan điểm trực tuyến của người tiêu dùng theo tính năng của sản phẩm. Hệ thống của tôi gồm các công việc sau:

- Thu thập dữ liệu: Tôi thực hiện thu thập tất cả các ý kiến đánh giá về dòng sản phẩm điện thoại trên nguồn dữ liệu tinhte.vn.
- Tiền xử lý dữ liệu: Tôi thực hiện gán nhãn từ loại cho các từ trong câu và loại bỏ đi các câu không phải là các câu quan điểm
- Tách câu quan điểm: Đầu vào là các câu đánh giá đã được gán nhãn từ loại và đầu ra là các câu chỉ chứa có một tính năng và một từ quan điểm
- Trích xuất tính năng của sản phẩm: Hệ thống thực hiện trích xuất các tính năng của sản phẩm từ các câu quan điểm

– Tổng hợp quan điểm theo tính năng của sản phẩm: Hệ thống dựa vào các tính năng đã được trích xuất làm căn cứ để tiến hành tổng hợp quan điểm theo tính năng của sản phẩm.

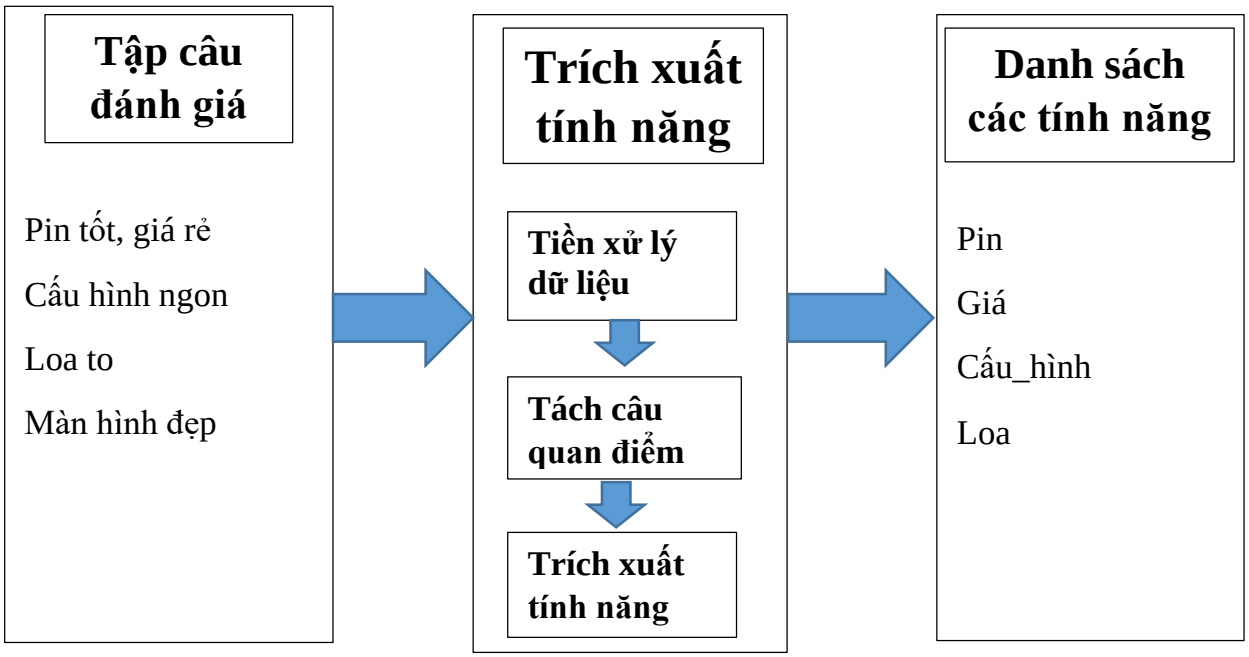


Hình 3.1. Mô hình hệ thống

3.1. Trích xuất tính năng của sản phẩm

Một sản phẩm bao gồm nhiều các tính năng khác nhau. Tính năng của sản phẩm là tất cả những bộ phận, chức năng, những thành phần cấu thành nên một sản phẩm [22]. Người tiêu dùng khi nhận xét về một sản phẩm, họ không chỉ nhận xét tổng quan về sản phẩm ấy mà thường họ nhận xét về từng tính năng của sản phẩm. Chính vì vậy, trong luận văn, tôi hướng đến việc phân tích và khai thác các tính năng của sản phẩm. Đối với các dữ liệu trên diễn đàn đa dạng và phong phú thì người tiêu dùng không chỉ đánh giá về tính năng mà nhà sản xuất đưa ra, đôi khi, nó là các tính năng mà được người tiêu dùng tự định nghĩa và cùng trao đổi. Ví dụ "*Cam hơi tệ*". "*Cam*" là một từ mà người tiêu dùng dành để đánh giá về tính năng *camera* của điện thoại chứ nó không phải là một tính năng có sẵn của sản phẩm được nhà sản xuất cung cấp. Nên trích chọn thông tin các tính năng của sản phẩm là một việc cần thiết. Trong phần này, tôi trình bày một phương pháp trích chọn tính năng của sản phẩm dựa vào từ quan điểm.

Thông thường trong các câu đánh giá, các tính năng cho sản phẩm thường là danh từ, cụm danh từ [11]. Ví dụ như câu "*Màn hình đẹp*" thì *màn hình* là một danh từ và cũng là một tính năng của sản phẩm. Để trích xuất ra các tính năng cho sản phẩm tôi thực hiện tách từ và gán nhãn từ loại sau đó dựa vào từ quan điểm để xác định các tính năng cho sản phẩm. Tuy nhiên, đối với dữ liệu trên các diễn đàn, blog, thường trong một ý kiến đánh giá, người dùng đánh giá về 2, 3 tính năng của sản phẩm trong cùng một câu. Tôi thực hiện tách các câu đánh giá đó về các câu có dạng chỉ gồm có một tính năng và một từ quan điểm nhằm tăng độ chính xác của bài toán.



Hình 3.2. Mô hình trích xuất tính năng của sản phẩm

Trong hình 3.2, tôi đưa ra mô hình trích xuất tính năng cho sản phẩm. Đầu vào là các câu đánh giá. Kết quả trả về của hệ thống là các tính năng được trích xuất trong câu. Hệ thống thực hiện trích xuất các tính năng qua 3 bước sau:

- Tiền xử lý dữ liệu
- Tách câu quan điểm
- Trích xuất các tính năng theo luật lan truyền kép

3.1.1. Tiền xử lý dữ liệu

Trong bước này, tôi thực hiện gán nhãn cho các từ loại và loại bỏ đi các câu không phải là câu quan điểm.

Gán nhãn cho các từ loại

Tôi sử dụng bộ công cụ JvnTextPro[3] dành cho xử lý các câu trong Tiếng Việt để tách câu, tách từ và gán nhãn cho các từ loại cho các từ, cụm từ trong câu bằng phương pháp sử dụng CRFs (Conditionnal Random Fields).

Bảng 3.1. Bảng từ viết tắt của các từ loại trong câu

1. N: Noun (danh từ) 2. Np: Personal Noun (danh từ riêng) 3. Nc: Classification Noun (danh từ chỉ loại) 4. Nu: Unit Noun (danh từ đơn vị) 5. V: verb (động từ) 6. A: Adjective (tính từ) 7. P: Pronoun (đại từ) 8: L: attribute (định từ) 9. M: Numeral (số từ) 10. R: Adjunct (phụ từ) 11. E: Preposition (giới từ)	12. C: conjunction (liên từ) 13. I: Interjection (thán từ) 14. T: Particle, modal particle (trợ từ, tiểu từ) 15. B: Words from foreign countries (Từ mượn tiếng nước ngoài ví dụ Internet, ...) 16. Y: abbreviation (từ viết tắt) 17. X: un-known (các từ không phân loại được) 18. Mrk: punctuations (các dấu câu)

Một số từ viết tắt của từ loại được mô tả trong bảng 3.1.

Ví dụ: *Màn hình đẹp. Giá quá ngon rồi. Mấy em zen xách chân mình nghĩ cũng ko lại được với nó*. Sau khi thực hiện tách câu, chúng ta thu được kết quả với hai câu được phân tách:

(a) Màn hình đẹp .

(b) Giá quá ngon rồi .

(c) *Mấy em zen xách chân mình nghĩ cũng ko lại được với nó .*

Dữ liệu được đưa qua bộ tách từ để tách các từ, cụm từ (các từ nào được ghép với nhau tạo thành một cụm từ được các định trong bước này)

(a) Màn_hình đẹp .

(b) Giá quá ngon rồi .

(c) *Mấy em zen xách chân mình nghĩ cũng ko lại được với nó .*

[3] http://jvntextpro.sourceforge.net/

Cuối cùng, dữ liệu được đưa qua bộ gán nhãn với phương pháp CRFs, chúng ta thu được nhãn của các từ, cụm từ như sau:

(a) Màn_hình/N đẹp/A.

(b) Giá/N quá/T ngon/A rồi/R

(c) *Mấy/L em/N zen/V xách/V chân/N mình/R nghĩ/V cũng/R ko/V lại/R được/V với/E nó/N*

Loại bỏ các câu không phải câu quan điểm

Khi thu thập các ý kiến đánh giá trên diễn đàn thì không phải tất cả các câu thu được đều là câu quan điểm. Câu quan điểm là câu có chứa *từ quan điểm*. Như trong ví dụ (b) phần trên, *Mấy em zen xách chân mình nghĩ cũng ko lại được với nó*, sau khi gán nhãn từ loại, ta thu được kết quả *Mấy/L em/N zen/V xách/V chân/N mình/R nghĩ/V cũng/R ko/V lại/R được/V với/E nó/N*. Theo nghiên cứu của Liu [7] thì các từ quan điểm trong câu thường là các tính từ (A). Trong câu trên không chứa tính từ nào nên có thể coi đó không phải là một câu quan điểm. Hệ thống thực hiện loại bỏ các câu không phải là câu quan điểm trong bộ dữ liệu thu thập được.

3.1.2. Tách câu quan điểm

Từ những câu quan điểm đã được gán nhãn từ loại trong bước trước, tôi tiến hành tách các câu chứa nhiều hơn một tính năng về các câu chỉ chứa có một tính năng và một từ quan điểm.

Sau khi thu thập dữ liệu trên các diễn đàn tôi nhận thấy, một người dùng khi đánh giá về một sản phẩm, trong một lần đánh giá, họ thường có ý kiến về hai, ba hay nhiều tính năng cùng một lúc.

Ví dụ: *Pin khỏe, nhạc hay*. Hoặc

Cấu hình cao nhưng loa không tốt

Như vậy, đối với các tính năng khác nhau, sẽ có các ý kiến đánh giá khác nhau kể cả trong trường hợp cùng một câu đánh giá của một người nêu quan điểm. Trong ví dụ thứ nhất, cả hai từ quan điểm là "*khỏe*" và "*hay*" đều mang chiều hướng tích cực. Nhưng khi xác định tính năng của từ quan điểm, thì có cả hai tính năng "*pin*" và "*nhạc*". Trong trường hợp này có thể gán cả 2 tính năng đều được người dùng đánh giá tốt. Tuy nhiên trong ví dụ thứ hai, thì một từ quan điểm mang chiều hướng tích cực và một từ mang nghĩa tiêu cực. Rõ ràng là câu đánh giá khen tính năng *cấu hình* và chê tín năng *pin*. Nếu như không phân

biệt được từ quan điểm nào dành cho tính năng nào thì sẽ dẫn đến giảm độ chính xác trong quá trình xử lý dữ liệu. Vì vậy, tôi xây dựng bộ công cụ tách các câu ghép và câu phức trong các câu đánh giá để đưa chúng về dạng câu đơn. Trong mỗi câu đơn chỉ chứa một tính năng và một từ quan điểm.

Tôi thực hiện tách câu phức và câu ghép dựa trên luật trong câu dựa vào cấu trúc ngữ pháp của câu mà tôi thu được khi phân tách câu tiếng Việt. Để đơn giản mà vẫn đảm bảo được tính chính xác, tôi bỏ qua các từ loại khác (động từ, trợ từ,…) trong câu mà chỉ quan tâm vào các danh từ (N) (từ chỉ tính năng) và các tính từ (A) (từ chỉ quan điểm) từ nối và các từ phủ định trong câu.

Khi sử dụng bộ công cụ JvnTextPro để gán nhãn dữ liệu thì các từ nối được định nghĩa như là một liên từ và các từ phủ định được định nghĩa như là một phụ từ trong câu.

Xét ví dụ sau:

(1) *Cấu hình cao nhưng loa không tốt* sau khi được gãn nhãn trở thành: *Cấu_hình/N cao/A nhưng/C loa/N không/R tốt/A*

(2) *Pin dùng thì cũng ngon* sau khi gán nhãn trở thành: *Pin/N dùng/V thì/C cũng/R ngon/A*

Xét ví dụ (1), như ta nhận thấy, câu đánh giá bao gồm 2 tính năng được đánh giá là tính năng *cấu hình* và tính năng *loa*. Sau khi gán nhãn, câu đánh giá có cấu trúc N – A – C – N - R – A. Hai đánh giá về hai tính năng *cấu_hình* và *loa* được nối bởi liên từ (C) *nhưng* và từ quan điểm *tốt* được phủ định bởi phụ từ (R) *không*.

Xét ví dụ (2), câu đánh giá chỉ đánh giá về một tính năng pin của sản phẩm. Sau khi gán nhãn, câu có cấu trúc N-V-C-R-A. Như vậy, trong câu này cùng có liên từ thì (C) và phụ từ cũng (R) nhưng thì không phải là một từ nối trong câu và cũng không phải là một từ phủ định.

Ở đây, nếu coi liên từ (C) là các từ nối và phụ từ (R) là các từ phủ định trong câu thì sẽ làm cho kết quả tách câu không chính xác. Vì vậy, tôi thực hiện xây dựng hai bộ từ điển bằng tay gồm các từ nối và các từ phủ định:

- Bộ từ nối (TN): và, nhưng, không những, mà còn, chỉ có “+”, “,”,…
- Bộ từ phủ định (PD): không, ko, chưa, chẳng, đâu có,…

Bảng 3.2. Một số luật trong câu

STT	Đầu vào	Đầu ra
1	N/Np1-TN-N/Np2- A	N/Np1 -A N/Np2 -A
2	N/Np1 – A1 - TN- N/Np2 –A2	N/Np1 – A1 N/Np2 –A2
3	N/Np1 – A1 - TN - N/Np2	N/Np1 – A1 N/Np2
4	N/Np1 – PD - A1 - TN- N/Np2 –A2	N/Np1 – PD - A1 N/Np2 –A2
5	N/Np1 – A1 - TN- N/Np2 - PD–A2	N/Np1 – A1 N/Np2 –PD - A2

Trong bảng 3.2, tôi đưa ra một số luật áp dụng trong việc tách câu khi tôi xử lý dữ liệu. Cột 1 là số thứ tự của các luật. Cột 2 là cấu trúc câu phức và câu ghép cần phân tích. Cột 3 là cấu trúc câu đơn nhằm thu được sau khi được phân tích.

Xét một số ví dụ sau:

Pin khỏe, nhạc hay sau khi phân tích được cấu trúc câu *Pin/N khỏe/A , nhạc/N hay/A.* Cấu trúc câu tương ứng với luật số 2 và được tách thành Pin/N khỏe/A và câu nhạc/N hay/A

Pin dùng thì cũng ngon sau khi gán nhãn từ loại ta thu được *Pin/N dùng/V thì/C cũng/R ngon/A.* Như đã nêu ở trên, khi xử lý, hệ thống chỉ quan tâm đến từ loại danh từ - N, tính từ -A, từ nối – TN, từ phủ định PD. Trong ví dụ này, hệ thống bỏ qua các từ loại khác và chỉ trả lại từ chỉ tính năng và từ quan điểm *Pin/N ngon/A.* Trong câu chỉ tồn tại một danh từ và một tính từ nên câu không được phân tách tiếp.

Giá thì ngon rồi chỉ có điều là chất lượng không ngon thôi. Sau khi phân tích cú pháp câu, ta được câu như sau: *Giá/N thì/C ngon/A rồi/T chỉ_có/TN điều/Np là/C Chất_lượng/N không/T ngon/A thôi/R.* Sau khi loại bỏ các từ loại không cần thiết, câu trở thành *Giá/N*

ngon/A chỉ_có/TN Chất_lượng/N không/PD ngon/A. Lúc này, câu có cấu trúc ***N-A –TN-N-PD- A***. Áp dụng luật số 5, hệ thống thu được 2 câu quan điểm là *Giá/N ngon/A* và *Chất_lượng/N không/PD ngon/A*

Phương pháp này tuy đơn giản và chưa tối ưu nhưng nó đã giải quyết được một vấn đề quan trọng trong bài toán tổng hợp quan điểm theo tính năng của sản phẩm với dữ liệu thực tế được lấy từ các diễn đàn đó là tách biệt các ý kiến đánh giá về các tính năng sản phẩm khác nhau. Kết quả thu được là mỗi câu đánh giá chỉ chứa một tính năng của sản phẩm

3.1.3. Trích xuất tính năng của sản phẩm

Phần tiếp theo, tôi trình bày phương pháp trích xuất tính năng của sản phẩm theo luật lan truyền kép, sử dụng từ quan điểm mà Qiu đã xây dựng năng 2011 [17].

Từ quan điểm là những từ ngữ mà người nêu quan điểm nêu lên ý kiến của mình về sản phẩm đó. Theo một nghiên cứu của Hu & Liu thì từ quan điểm thường là tính từ trong câu [7]. Các từ quan điểm tích cực như *tốt, bền, lâu, đẹp, ngon*; các từ quan điểm tiêu cực như *kém, thấp, tồi, dở*; các từ quan điểm mang nghĩa trung lập như *bình thường, cũng được*.

Tại Việt Nam, hiện đã có bộ từ điển VietSentiment WordNet. Tôi không sử dụng bộ từ điển VietSentiment Wordnet vào trong quá trình xử lý vì một số lý do sau:

- Bộ dữ liệu VietSentiment Wordnet là bộ từ quan điểm chung. Mà lĩnh vực tôi đang nghiên cứu là lĩnh vực dành cho điện thoại di động. Như đã trình bày trong phần trước, có những từ là từ quan điểm trong lĩnh vực này nhưng lại không phải là từ quan điểm trong lĩnh vực khác.
- Giá trị PosScore và NegScore chỉ mang tính tương đối, việc xác định từ quan điểm đó là tích cực hay tiêu cực dựa vào điểm số của PosScore và NegScore là rất khó khăn.

Ví dụ: *Vỏ màu trắng này đẹp nhỉ.* Trong ví dụ này tồn tại hai từ quan điểm nếu dựa vào VietSentiment là từ trắng và từ đẹp. Tuy nhiên, trong câu chỉ tồn tại một từ quan điểm là từ đẹp, còn trắng chỉ để bổ nghĩa cho từ vỏ. Nếu sử dụng cả hai từ quan điểm là trắng và đẹp thì sẽ làm mất đi tính đúng đắn của chiều hướng quan điểm

Để khắc phục được nhược điểm này, tôi thực hiện xây dựng bộ từ điển về từ quan điểm bao gồm các tính từ mà người tiêu dùng Việt Nam sử dụng khi đánh giá về chất lượng của một sản phẩm, kết hợp với việc gán nhãn từ loại. Tôi thực hiện gán nhãn thủ công trên các từ quan điểm. Các từ quan điểm mang tính tích cực được gán nhãn dương (+); các từ quan

điểm mang tính tiêu cực được gán nhãn âm (-); các từ quan điểm mang tính trung lập tôi không gán nhãn.

Việc gán nhãn từ loại cũng sẽ hỗ trợ việc xác định được từ quan điểm một cách chính xác hơn. Theo như ví dụ trên, sau khi gán nhãn ta thu được kết quả *vỏ/N màu_trắng/N này/P đẹp/A nhỉ/T*. Từ *trắng* kết hợp với từ *màu* để tạo thành một danh từ trong câu. Theo như Qiu [17] thì các từ quan điểm thường là các tính từ trong câu, trường hợp này có thể bỏ qua được *trắng* trong danh sách các từ quan điểm

Dựa vào các bộ từ quan điểm đã xây dựng, tôi thực hiện trích xuất ra các tính năng cho sản phẩm trong các câu đánh giá của người tiêu dùng theo luật lan truyền kép với một số quy tắc trong các câu đánh giá thường gặp đối với các diễn đàn Việt Nam.

Một số cấu trúc câu đánh giá:

N-A : Pin tốt

N-V-A: Pin dùng bình thường

N-R-A: Loa hơi bé

N-C-A: Giá thì ngon

Trong câu nhận xét *Cấu hình tốt* tuân theo quy tắc N-A. Dựa vào từ quan điểm *tốt* ta có thể tìm được tính năng *cấu hình* cho sản phẩm.

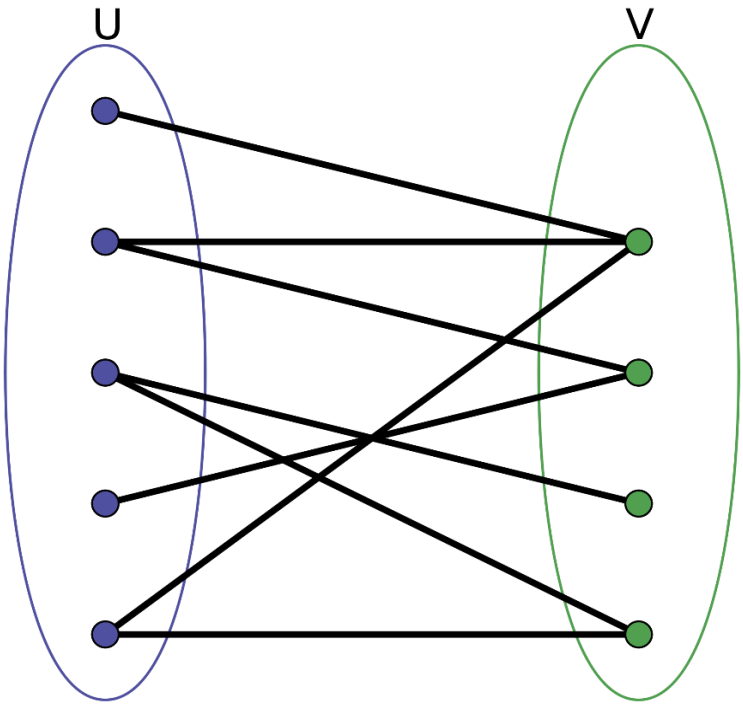
3.2. Nhóm các từ nói về cùng một tính năng

Như chúng ta đã biết, ngôn ngữ tiếng Việt vốn đa dạng và phong phú, cùng mô tả về một tính năng nhưng sẽ có nhiều cách điễn đạt. Đối với những người đánh giá khác nhau thì họ dùng những từ khác nhau để cùng nói về một đối tượng. Họ có thể dùng các dạng từ loại khác nhau như từ đồng nghĩa khác âm, từ viết tắt, từ đi mượn.

Ví dụ: *Máy ảnh tốt hoặc Camera nét*

Chúng ta có thể thấy, hai câu cùng phát biểu quan điểm về một tính năng là chất lượng của *camera* cho sản phẩm nhưng hai người dùng hai từ ngữ khác nhau là *máy ảnh* và *camera* để mô tả về tính năng của sản phẩm. Để giải quyết vấn đề này, tôi dựa vào tần số kết hợp giữa các tính từ và danh từ trong câu. Phần lớn, khi mô tả về một tính năng của sản phẩm thì người tiêu dùng thường dùng một số từ quan điểm nhất định. Tôi dựa trên kiến thức về đồ thị Bipartite Graph để thực hiện nhóm các từ quan điểm. Đồ thị Bipartite Graph là đồ thị mà trong đó tập các đỉnh có thể được chia thành hai tập không giao nhau thỏa mãn điều

kiện không có cạnh nối hai đỉnh bất kỳ thuộc cùng một tập[4]. Ví dụ khi mô tả về tính năng *pin*, người tiêu dùng thường dùng các từ quan điểm như *bền, tốt, lâu*. Khi hai hoặc nhiều danh từ đều được nhận xét bằng các từ quan điểm giống nhau trên 80% thì tôi đưa các danh từ chỉ tính năng lại thành một nhóm.



Hình 3.3. Mô hình đồ thị Bipartite Graph

Hình 3.3 mô tả mô hình đồ thị Bipartite Graph. Trong trường hợp này, tôi coi tập đỉnh U là tập các tính năng của sản phẩm. Còn tập đỉnh V là tập các từ quan điểm. Các liên kết giữa đỉnh trong tập U và đỉnh trong tập V thể hiện sự kết hợp của tính năng và từ quan điểm trong câu (người dùng sử dụng các từ quan điểm để đánh giá về tính năng của sản phẩm)

Mỗi người tiêu dùng khác nhau sẽ có các đánh giá khác nhau về các tính năng khác nhau của sản phẩm. Thông thường, các tính năng quan trọng sẽ thường xuyên được người tiêu dùng đánh giá về chất lượng. Ví dụ như tính năng về *giá, pin,tốc độ xử lý, chất lượng hình ảnh, tốc độ lướt Web* là các tính năng mà được nhiều người tiêu dùng nhận xét nhất khi đánh giá về sản phẩm điện thoại di động. Tôi căn cứ vào tần suất xuất hiện của các danh từ chỉ tính năng trong tập dữ liệu để tìm các tính năng thường xuyên được người tiêu dùng đánh giá và loại bỏ các tính năng mà ít được người tiêu dùng quan tâm. Trên thực tế, các

[4] https://en.wikipedia.org/wiki/Bipartite_graph

tính năng ít được người tiêu dùng đề cập đến thì thường chúng không quan trọng và không mang nhiều giá trị trong việc xử lý các bài toán có số lượng dữ liệu lớn.

Sau khi loại bỏ các danh từ chỉ tính năng ít được người tiêu dùng đưa ra quan điểm tôi thu được một bộ các tính năng của sản phẩm. Tuy nhiên, vẫn còn một số ít trường hợp mà danh từ mô tả tính năng mang nghĩa chung chung, không rõ ràng.

Ví dụ trong câu: *Em này quá ngon*. Sau khi phân tích ta được *Em_này*/NP *quá*/P *ngon*/A

Theo như luật tôi xây dựng thì dựa vào tính từ *ngon* có trong từ điển, tôi tìm ra *Em_này* là một tính năng của sản phẩm. Nhưng thực tế, *Em_này* không phải là một tính năng cho một sản phẩm. Để khắc phục vấn đề này, sau khi đã thu thập được các danh từ chỉ tính năng cho sản phẩm, tôi thực hiện lược bỏ thủ công một số các danh từ mà được nhầm lẫn sang các từ mô tả tính năng của sản phẩm.

3.3. Tổng hợp quan điểm

Phân cụm các câu đánh giá về cùng một tính năng

Các câu đánh giá cùng đưa ra ý kiến về một nhóm tính năng, tôi thực hiện nhóm các câu đánh giá lại với nhau để thực hiện tổng hợp ý kiến theo từng tính năng cho sản phẩm.

Ví dụ: Các đánh giá về pin của sản phẩm HTC One E8 như *Pin tốt, Pin kém, Pin khá, Pin trâu, Pin bình_thường, Pin đuối.*

Phân lớp câu quan điểm

Trong phần này, tôi thực hiện phân lớp các câu quan điểm trong nhóm đã phân loại từ bước trước theo ba chiều hướng tích cực, tiêu cực và trung lập. Để thực hiện nhiệm vụ này, tôi thực hiện giải thuật phân lớp dựa vào nhãn của từ quan điểm trong câu. Nhãn của câu sẽ tương ứng với nhãn của từ quan điểm trong câu.

Ví dụ:

Lướt Web nhanh. Trong câu trên, *nhanh* là từ quan điểm được gán nhãn + nên câu được gán nhãn +.

Pin kém, kém là từ quan điểm được gán nhãn - câu được gán nhãn -.

Một số trường hợp riêng:

- Đối với các câu đánh giá có chứa từ phủ định như *không, chẳng, chưa, chả* thì tôi thực hiện gán nhãn cho câu ngược lại với nhãn của từ quan điểm.

- Đối với từ quan điểm có nhãn +, nếu có từ phủ định đứng trước thì tôi gán cho câu quan điểm nhãn -. Ví dụ: *Màn hình cảm ứng không mượt*. Từ quan điểm ở đây là từ *mượt* có nhãn +. Tuy nhiên, từ không là từ mang nghĩa phủ định đứng trước nên câu này không phải là câu khen mà lại là câu chê, chúng ta phải gán cho câu vào lớp -.
- Đối với từ quan điểm nhãn - thì tôi không gán nhãn cho câu quan điểm. Ví dụ: *hình ảnh không xấu*; *không xấu* không mang nghĩa khen cũng không mang nghĩa chê nên không gán nhãn cho câu quan điểm.
- Đối với từ quan điểm không có nhãn thì tôi gán nhãn - cho câu quan điểm. Ví dụ: *bình thường ->không bình thường*
- Một trường hợp khác trong câu có từ không nhưng nó nằm trong cụm *không những...mà còn* thì nó lại mang hàm nghĩa ngược lại. Ví dụ: *Hình ảnh không những nét mà còn đẹp.* Trong câu nhận xét này cũng có xuất hiện từ không nhưng ý kiến đánh giá là cùng chiều với từ quan điểm nằm trong bộ từ điển mà tôi đã xây dựng. Chính vì vậy, ngoài việc dựa vào bộ từ điển đã được xây dựng để phân lớp, tôi cũng bổ sung thêm một số luật trong một số trường hợp đặc biệt câu có các liên từ như *không, không những ... mà còn…*

3.4. Độ đo tính chính xác của hệ thống

Để tính độ chính xác, độ hồi tưởng dựa trên số lượng các phần tử được dự đoán đúng ở lớp dương (true positive), số lượng các phần tử bị đoán nhầm từ lớp dương sang âm (false positive) và số lượng các phần tử dự đoán nhầm từ lớp âm sang lớp dương (false negative). Đối với từng lớp cần đánh giá ta có công thức như sau:

Độ chính xác P (*Percision*):

$$P = \frac{true\ positive}{true\ positive + false\ positive} \times 100\%$$

Độ hồi tưởng R (*Recall*):

$$R = \frac{true\ positive}{true\ positive + false\ negative} \times 100\%$$

Độ đo F (*F-measure*):

$$F = \frac{2 \times P \times R}{P + R}$$

Chương 4. Thực nghiệm và đánh giá

4.1. Chuẩn bị dữ liệu và cài đặt

Trong phần này, tôi tiến hành thực nghiệm và đánh giá kết quả thu được qua mô hình mà tôi đã xây dựng trong chương 3. Tôi thực hiện trên dữ liệu được thu thập từ trang *tinhte.vn* với các ý kiến trao đổi về dòng điện thoại HTC One E8, Sony Z3 và Sony Aqua M4. Các ý kiến sau khi đã thu thập được, tôi dựa vào cấu trúc thẻ của HTML để trích xuất ra các ý kiến đánh giá của người tiêu dùng, bỏ qua các thông tin không cần thiết khác như thông tin về ngày tháng, người nêu quan điểm. Tôi thực hiện sàng lọc thủ công, bỏ qua các ý kiến Spam và các ý kiến không phải là đánh giá về đối tượng mà tôi đang xử lý.

Bảng 4.1. Số ý kiến đánh giá chuẩn bị làm thực nghiệm

Sản phẩm	Số Review	Số câu
HTC One E8	300	389
Sony Z3	216	265
Sony Aqua M4	96	112

4.2. Tiến hành thực nghiệm và đánh giá

Dữ liệu được đưa qua bộ công cụ JnvTextPro để phân đoạn câu, tách câu, tách từ và gán nhãn từ loại. Hệ thống thực hiện loại bỏ đi các câu không phải là câu quan điểm. Sau khi loại bỏ, hệ thống thu được dữ liệu như bảng 4.2

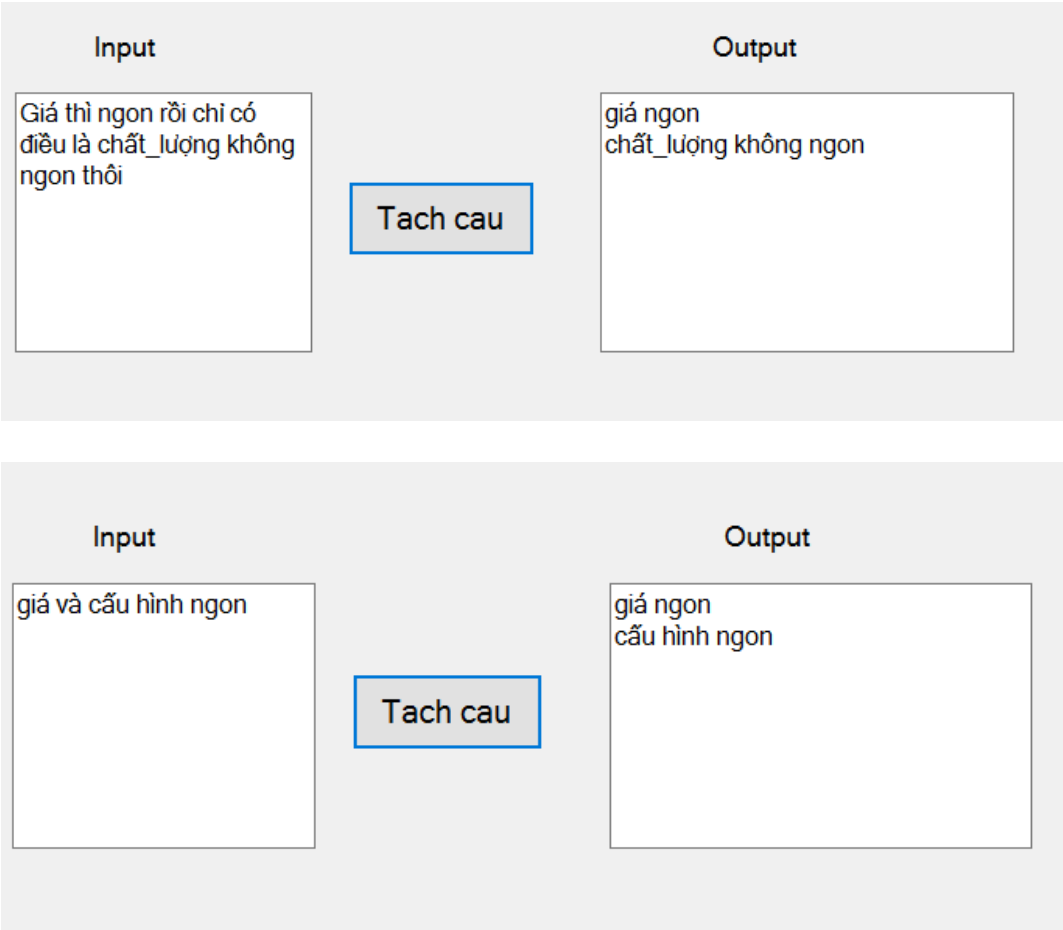
Bảng 4.2. Dữ liệu thu được sau tiền xử lý dữ liệu

Sản phẩm	Số câu	Số câu quan điểm
HTC One E8	389	354
Sony Z3	265	232
Sony Aqua M4	112	90

Dữ liệu được đưa qua bộ tách câu quan điểm để tách các câu phức và câu ghép thành các câu đơn mà tôi xây dựng dựa trên luật (đã trình bày ở chương 3). Tôi bỏ qua các từ loại

khác mà chỉ quan tâm đến tính từ và danh từ, các từ phủ định và các từ nối. Kết quả trả về là các câu đơn chỉ phát biểu về một tính năng (gồm một danh từ và một tính từ).

Trong hình 4.1 tôi trình bày một số ví dụ tách câu quan điểm từ câu phức, câu ghép thành các câu đơn.



Hình 4.1. Một số kết quả ví dụ tách câu quan điểm

Bảng 4.3. Kết quả dữ liệu thu được sau khi tách câu

Sản phẩm	Số câu tách qua hệ thống	Số câu được tách thực tế	P	R	F1
HTC One E8	525	562	93,3%	87,18%	90,15%
Sony Z3	332	316	96.02%	100%	97,9%

Sony Aqua M4	159	163	87,42%	85,27%	86,33%

Tôi thực hiện xây dựng bộ từ điển theo phương pháp thủ công gồm khoảng 150 từ quan điểm dùng cho đánh giá sản phẩm căn cứ vào các ý kiến đánh giá của người tiêu dùng Việt Nam trên các trang web đánh giá sử dụng kỹ thuật lan truyền kép tôi đã trình bày trong chương 3.

Hệ thống thực hiện trích xuất ra các tính năng của sản phẩm qua các luật trong câu được đưa vào hệ thống và dựa vào bộ từ điển đã xây dựng. Tôi thu được một danh sách gồm các tính năng của sản phẩm như *giá, pin, cấu hình, màn hình, loa, vỏ, camera, sóng, âm,,...* Kết quả đánh giá được thể hiện trong bảng 4.3.

Bảng 4.3. Kết quả thu được sau khi hệ thống trích chọn tính năng cho sản phẩm

Tên sản phẩm	Số lượng tính năng được trích xuất qua hệ thống	Số lượng tính năng thu được thực tế	P	R	F1
HTC One E8	45	36	77,78%	97,22%	86,40%
Sony Z3	21	16	80,9%	94,44%	87,18%
Sony Aqua M4	19	16	73,68%	87,5%	80%
Trung bình			77,45%	93,05%	84,53%

Năm 2011, nhóm tác giả Hà Quang Thụy đã có công trình nghiên cứu về việc trích xuất các tính năng cho sản phẩm của người tiêu dùng Việt Nam [27]. Trong bảng 6, tôi đưa ra kết quả trong phương pháp mà tôi đã thực hiện (PP1) với kết quả nghiên cứu của Hà Quang Thụy (PP2) đối với dữ liệu tiếng Việt

Bảng 4.4. Kết quả của PP1 và PP2 khi trích xuất tính năng cho sản phẩm

Phương pháp	P	R	F1
PP1	77,45%	93,05%	84,53%

PP2	87,56%	93,58%	90,32%

Kết quả mà tôi thu được có độ chính xác thấp hơn so với kết quả của nhóm tác giả Hà Quang Thụy đã nghiên cứu trước đó. Ở đây, tôi không đưa ra đánh giá phương pháp của tác giả Hà Quang Thụy có độ chính xác cao hơn vì nguồn dữ liệu đầu vào là khác nhau, và mỗi người đều có một cách đánh giá về sản phẩm là khác nhau.

Trong danh sách các tính năng tôi thu được có một số tính năng được người tiêu dùng mô tả bằng một số các danh từ khác nhau như *Camera* được mô tả bằng *Camera, máy ảnh.* Hệ thống thực hiện phân nhóm các danh từ chỉ tính năng. Áp dụng phương pháp GFN tôi thu được kết quả với độ chính xác là 76,6%. Phương pháp GFN có độ chính xác chưa cao vì số lượng dữ liệu chưa nhiều.

Tiếp theo, hệ thống dựa vào tần suất xuất hiện của các danh từ chỉ tính năng, tôi chọn độ hỗ trợ tối thiểu (*minsup* = 4), sau khi loại bỏ các danh từ mô tả tính năng ít xuất hiện hệ thống trả lại kết quả với danh sách gồm 38 tính năng thường xuyên xuất hiện trong các ý kiến đánh giá đối với sản phẩm HTC One E8, và thu được kết quả đạt 83% số danh từ còn lại là chỉ tính năng cho sản phẩm.

Bảng 4.5. Tần suất xuất hiện của một số tính năng của sản phẩm HTC One 8

Tính năng	**Số lần**	**Tính năng**	**Số lần**
Giá	49	Cấu hình	24
Pin	29	Màn hình	12
Loa	10	Camera	14
Vỏ	10	Thiết kế	12
Htc	11	Lướt Web	7
Sóng	8	Âm thanh	9

Tuy nhiên, bộ dữ liệu sau khi xử lý vẫn còn một số danh từ mà ko phải để miêu tả tính năng mang hàm ý chung chung khác như *em này, con này, máy này*. Để tăng tính chính xác của hệ thống tôi tiến hành lược bỏ thủ công các danh từ và các câu chứa danh từ đó.

Bảng 4.6. Kết quả sau khi loại bỏ còn số tính năng và số câu

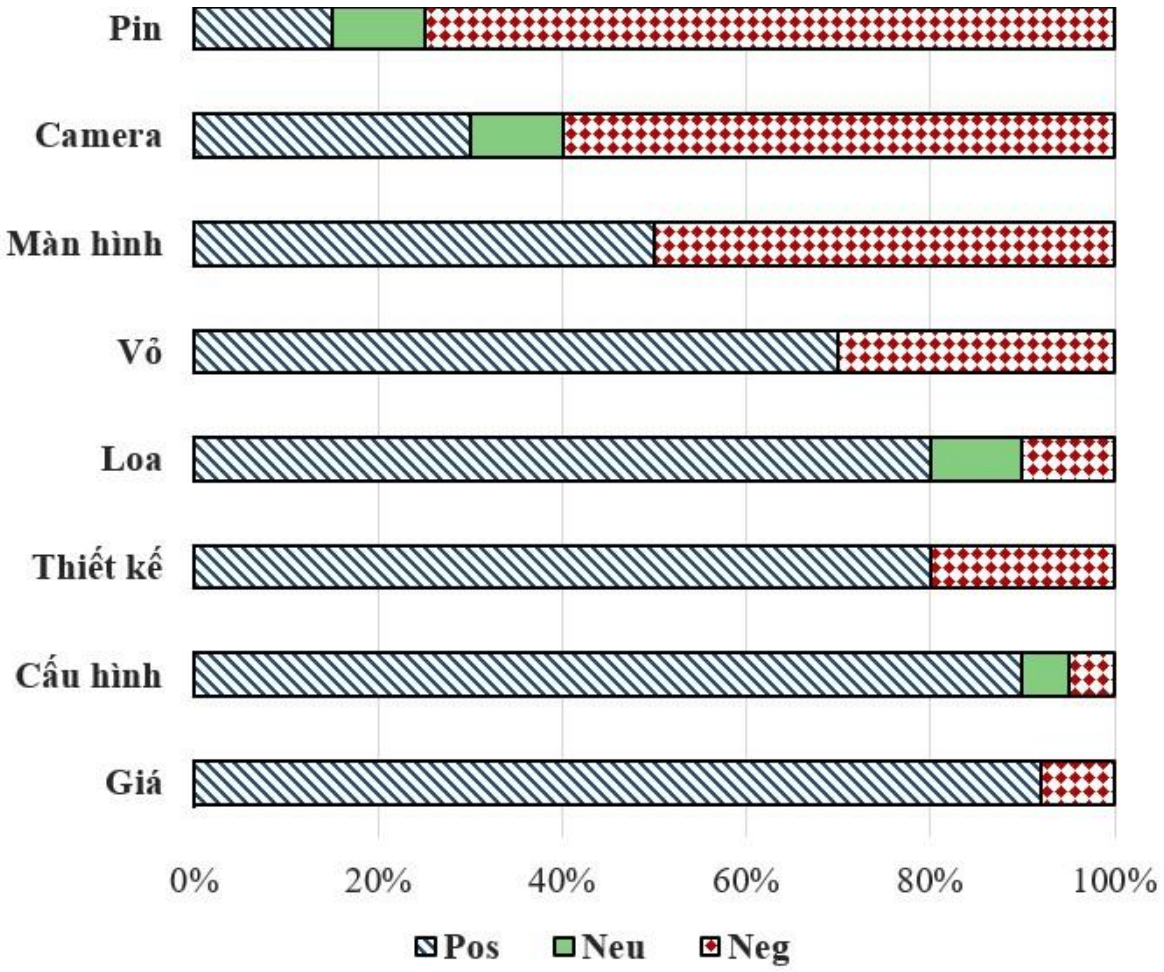
Tên sản phẩm	Số tính năng sau khi xử lý	Số câu sau khi xử lý
HTC One E8	26	497
Sony Z3	17	268
Sony Aqua M4	15	135

Bảng 4.7. Đánh giá kết quả tổng hợp ý kiến theo tính năng của sản phẩm

Tên sản phẩm	P	R	F1
HTC One E8	97,58%	100%	98,78%
Sony Z3	96,85%	100%	98,40%
Sony Aqua M4	97,03%	99,24%	98,12%

Cuối cùng, hệ thống thực hiện phân lớp các câu quan điểm theo từng tính năng (nhóm tính năng) mà đã được xử lý trong giai đoạn trước. Nhãn của từ quan điểm được lấy làm nhãn cho câu đánh giá. Trong phần này, tôi chỉ sử dụng phương pháp thống kê để đưa ra được bản tổng hợp quan điểm theo các tính năng của sản phẩm. Kết quả hệ thống phân lớp được mô tả qua bảng 4.7.

Bảng tổng hợp các ý kiến đánh giá của người tiêu dùng theo tính năng của sản phẩm HTC One E8 có thể được biểu diễn như hình 4.2.



Hình 4.2.Tổng hợp ý kiến theo tính năng của sản phẩm HTC One E8

Qua biểu đồ trên, chúng ta có thể thấy được *pin, camera* là 2 tính năng được người tiêu dùng đánh giá thấp nhất, còn *giá* và *cấu hình* được người tiêu dùng ủng hộ cao. Người mua hàng có thể căn cứ vào kết quả đánh giá sản phẩm của những người dùng trước và nhu cầu sử dụng của mình để lựa chọn sản phẩm phù hợp.

Chương 5. Kết luận

5.1. Những vấn đề đã giải quyết trong luận văn này

Luận văn đã tiến hành nghiên cứu bài toán khai phá quan điểm mà cụ thể là tổng hợp quan điểm theo tính năng của sản phẩm. Luận văn đã trình bày một số các phương pháp liên quan đến tổng hợp quan điểm theo tính năng của sản phẩm trên thế giới cũng như ở Việt Nam

Trong luận văn này, tôi đã trình bày một phương pháp tổng hợp ý kiến đánh giá trực tuyến của người tiêu dùng Việt Nam đối với các tính năng của sản phẩm. Hệ thống đã thực hiện trích xuất tính năng của sản phẩm dựa vào từ quan điểm. Đặc biệt, luận văn đã thực hiện tách các câu phức và câu ghép thành các câu đơn. Theo đó, mỗi câu đơn chỉ chứa một tính năng của sản phẩm và một từ quan điểm. Luận văn cũng thực hiện phân nhóm các câu quan điểm phát biểu về cùng một tính năng và tổng hợp quan điểm theo các từ quan điểm trong câu dựa vào nhãn của từ quan điểm theo chiều hướng tích cực, tiêu cực và trung lập.

Bên cạnh đó, trong phạm vi của luận văn, luận văn chưa thực hiện được việc trích xuất sản phẩm mà người tiêu dùng đánh giá trong mỗi câu quan điểm và lọc các quan điểm spam.

Trong quá trình thực hiện luận văn, tôi đã cố gắng tiếp cận phương pháp tổng hợp ý kiến theo tính năng của sản phẩm của người tiêu dùng Việt Nam và tham khảo các tài liệu liên quan cả về xử lý ngôn ngữ tự nhiên và học máy trên thế giới cũng như ở Việt Nam. Tuy nhiên do thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót nhất định. Do vậy tôi thật sự mong muốn nhận được những góp ý cả về kiến thức chuyên môn lẫn cách trình bày.

5.2. Hướng nghiên cứu tiếp theo trong tương lai

Khai phá quan điểm được khá nhiều nhà nghiên cứu trên thế giới quan tâm bởi nó được ứng dụng rộng rãi trong các lĩnh vực. Trong luận văn của tôi, tôi cũng chỉ chọn một hướng nhỏ để nghiên cứu.

Trong tương lai, tôi muốn mở rộng nghiên cứu của mình và cải thiện một số vấn đề còn tồn tại để cải thiện kết quả cho mô hình tổng hợp ý kiến theo tính năng của sản phẩm:

- Nghiên cứu phương pháp trích xuất thực thể (sản phẩm) trong các câu đánh giá để có hệ thống có kết quả tối ưu hơn

- Cải tiến mô hình trích xuất tính năng cho sản phẩm
- Cải tiến phương pháp tách câu ghép và câu phức thành các câu đơn
- Xử lý tốt hơn việc nhóm các từ chỉ về cùng một tính năng
- Thực hiện xử lý quan điểm Spam, loại bỏ các câu đánh giá không phải là các đánh giá dành cho sản phẩm mà hệ thống đang xử lý

Các công trình đã công bố

Vũ Thị Nhạn, Nguyễn Việt Anh, Nguyễn Khắc Giáo (2015) *Một phương pháp tổng hợp ý kiến đánh giá trên tính năng của sản phẩm của người tiêu dùng Việt Nam*, Kỷ yếu Hội thảo quốc gia lần thứ XVIII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, tr.185-190

TÀI LIỆU THAM KHẢO

1. Blair-Goldensohn, S.,Hannan, K., McDonald, R., Neylon, T., Reis,G.A., and Reyna,J. (2008), Building a sentiment summarizer for local service reviews. *In Proceedings of International Conference on World Wide Web Workshop of NLPIX*.
2. Blei, D.M., Ng, A.Y., Jordan,M.I.(2003), Latent dirichlet allocation.The Journal of Machine Learning Research. 3: p. 993-1022
3. Carenini, G., Ng, R., Pauls, A.. (2006), Multi-Document summarization of evaluative text. In Proceeding of Conference of the European Chapter of the ACL(EACL-2006).
4. Guo, H., Zhu, H., Guo, H., Zhang, X., Su, Z. (2009), Product feature categorization with multilevel latent semantic association. *In Proceedings of ACM International Conference on Information and Knowledge Management*.
5. H Lee, A Chang, Y Peirsman, N Chambers, M Surdeanu, D Jurafsky Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Journal Computational Linguistics (4),* December 2013 Pages 885-916) ()
6. Hofmann, Thomas. (1999), Probabilistic latent semantic indexing. *In Proceedings of Conference on Uncertainty in Artificial Intelligence* (UAI-1999).
7. Hu, M., Liu, B. (2004), Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.
8. Jin, Wei, Ho,H.H., (2009), A novel lexicalized HMM-based learning framework for web opinion mining. *In Proceedings of International Conference on Machine Learning* (ICML-2009).
9. Kim, S., Zhang, J., Chen, Z., Oh, A.H., Liu, S. (2013), "A hierarchical aspect – sentiment model for online reviews", AAAI
10. Lafferty, John, Andrew McCallum, and Fernando Pereira (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of International Conference on Machine Learning* (ICML-2001).
11. Liu, B. (2009), Handbook Chapter: "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing. *Marcel Dekker, Inc. New York, NY, USA.*
12. Liu, B. (2010), "Sentiment analysis and subjectivity", In *Handbook of Natural Language Processing, Second Edition*.
13. Liu, B. (2012), "Sentiment analysis and Opinion mining", University Of Illinois at Chicago.

14. Liu, B. (2012), Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

15. Moghaddam, S.,Ester, M. (2010), Opinion digger: an unsupervised opinion miner from unstructured product reviews. In *Proceedings of ACM International conference on Information and Knowledge Management*, 2010.

16. Moghaddam, S.,Ester, M. (2011), ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *46 Proceedings of ACM SIGIR International Conference on Information Retrieval.*

17. Qiu, G., Liu, B., Bu, J., Chen, C. (2011), Opinion word expansion and target extraction through double propagation. *Computational Linguistics*.

18. Rabiner, Lawrence R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): pp. 257-286

19. Titov, I., and McDonald, R.(2008a), Modeling online reviews with multi-grain topic models. In *Proceedings of International Conference on World Wide Web*.

20. Titov, I., and McDonald, R.(2008b), A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*.

21. Yu, J., Zha, Z., Wang, M., Wang, K.,Chua, T (2011b). Domain-Assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

22. Zhang, L., Liu, B.(2014), "Aspect and Entity Extraction for Opinion Mining", book chapter in Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenges, and Opportunities.

23. Zhang, L., Liu, B., Lim, S., O'Brien-Strain, E., (2010), Extracting and ranking product features in opinion documents. In *Proceedings of International Conference on Computational Linguistics* (COLING-2010).

24. Pang, B., Lee, B. (2008), Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2, 1-2, 1–135

25. Haseena,R.P. (2014) "Opinion Mining and Sentiment Analysis -Challenges and Applications", *International Journal of Application or Innovation in Engineering & Management* (IJAIEM)

26. Seerat, B., Azam, F. (2012), "Opinion Mining: Issues and Challenges", *International Journal of Computer Applications*.

27. Thuy, H.Q. , Thanh, V.T., Trang, P.H., To,. L.C. (2011) An upgrading feature-based opinion mining model on Vietnamese product reviews. In*: Active Media Technology, Lecture Notes in Computer Science, Springer Berlin Heidelberg*, pp. 173–185.

28. Jindal, Nitin, Liu, B.(2007) Review spam detection. *In Proceedings of WWW* (Poster paper).

29. Jindal, Nitin, Liu, B. (2008) Opinion spam and analysis. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008).*

30. Mauge, K., Rohanimanesh, K., Ruvini, J.D., (2012) Structuring e-commerce inventory. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2012).*

31. Zhai, Z., Liu, B., Xu, H., Jia, P. (2010) Grouping product features using semisupervised learning with soft-constraints. In *Proceedings of International Conference on Computational Linguistics (COLING-2010).*