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

TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2016

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.

Người tiêu dùng khi đánh giá về một sản phẩm dịch vụ nào đó, họ 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" 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 được tổ chức thành 5 chương 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

Chương 5: Kết luận

Chương 1. Tổng quan về khai phá quan điểm

1.1. Giới thiệu

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.

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

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].

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].

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

- Những người khác nhau có phong cách viết khác nhau
- Quan điểm thay đổi theo thời gian
- Độ mạnh của quan điểm
- Quan điểm theo ngữ cảnh
- Các câu đánh giá có sự pha trộn
- Quan điểm mang tính châm biếm mỉa mai
- Xử lý ngôn ngữ tự nhiên trong câu quan điểm

1.3. Các ứng dụng của khai phá quan điểm

- Nghiên cứu thị trường dành cho người mua và bán
- Cải thiện chất lượng của sản phẩm dịch vụ
- Hệ thống gợi ý
- Chính quyền thông minh
- Hỗ trợ đưa ra quyết định

1.4. Các bài toán trong khai phá quan điểm

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
- Tổng hợp quan điểm.

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

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

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.

Tính 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" hoặc "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]

2.2. Trích xuất khía cạnh

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

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.

Mô hình này cũng được nhóm tác giả Hà Quang Thụy nghiên cứu và thực hiện đối với các đánh giá của người dùng Việt đối với sản phẩm [27]

2.3. Nhóm các từ cùng chỉ về một tính năng

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 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 [27].

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].

2.5. Loại bỏ quan điểm Spam

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

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,

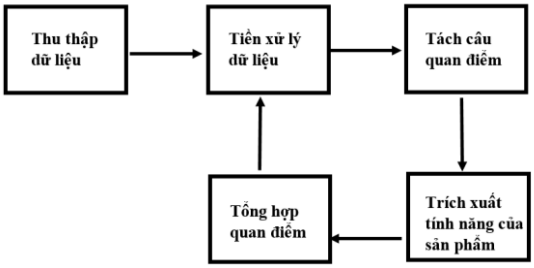
nhưng chỉ nhận xét về các nhãn hiệu hoặc nhà sản xuất của sản phẩm. Chúng được coi như là Spam,chúng không nhắm vào các sản phẩm cụ thể và thường sai lệch. Ví dụ, một đánh giá cho một máy in HP cụ thể nói: “Tôi ghét HP. Tôi không bao giờ mua bất kỳ sản phẩm của chúng”.

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, nhà phê bình giả, và nhóm phê bình giả.

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

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

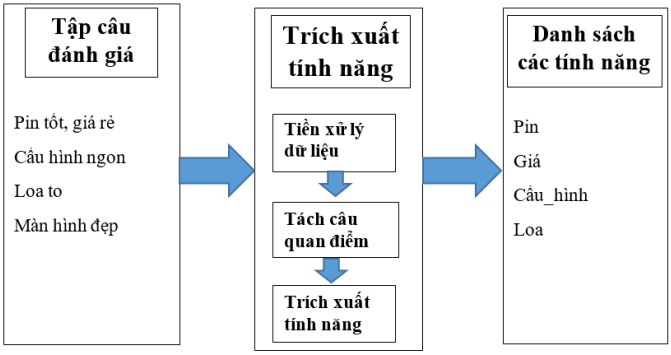


Hình 3.1 Mô hình hệ thống

Hình 3.1 mô tả khái quát các công việc chúng tôi thực hiện trong quá trình tổng hợp quan điểm trực tuyến của người tiêu dùng theo tính năng của sản phẩm. Hệ thống của chúng tôi gồm các công việc sau:

- Thu thập dữ liệu: Chúng 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: Chúng 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.

3.1. Trích xuất tính năng



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

Trong hình 3.2, chúng 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

Bước 1: Tiền xử lý dữ liệu

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

Chúng tôi sử dụng bộ công cụ JvnTextPro[1] dành cho xử lý các câu trong Tiếng Việt để phân đoạn câu, tách câu và gán nhãn cho các từ loại cho các từ, cụm từ trong câu

Tuy nhiên không phải tất cả các câu thu được đều là câu quan điểm. Câu quan điểm là câu có chứa *từ quan điểm*. Hệ thống thực hiện loại bỏ các câu không phải là câu quan điểm trong bộ dữ liệu thu thập được.

Bước 2: Tách câu quan điểm

Từ những câu quan điểm đã được gán nhãn từ loại trong bước trước, chúng 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.

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

Sau khi thu thập dữ liệu trên các diễn đàn chúng 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.

Chúng tôi thực hiện tách câu phức và câu ghép dựa trên luật trong câu dựa vào cấu trúc ngữ pháp của câu mà chúng tôi thu được khi phân tách câu tiếng Việt. Để đơn giản mà vẫn đảm bảo được tính chính xác, chúng 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.

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 3.1. 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, chúng tôi đưa ra một số luật áp dụng trong việc tách câu khi chúng 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

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

Bước 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].

Chúng 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. Chúng 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 chúng tôi không gán nhãn.

Dựa vào các bộ từ quan điểm đã xây dựng, chúng 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

3.2. Nhóm các từ cùng nói về một tính năng

Ngôn ngữ tiếng Việt vốn đa dạng và phong phú, cùng mô tả về một tính năng nhưng đối với những người đánh giá khác nhau thì họ dùng những từ khác nhau để nêu lên quan điểm của mình về chất lượng của sản phẩm.

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. Chúng 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 . 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 chúng tôi đưa các danh từ chỉ tính năng lại thành một nhóm.

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 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 chúng 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, chúng tôi thực hiện lược bỏ thủ công một số các danh từ mà được nhầm lẫn sang miêu 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, chúng 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.

Phân lớp câu quan điểm

Trong phần này, chúng 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, chúng 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.

Một số trường hợp riêng:

- Đối với các câu đánh giá có chứa từ phủ định như không, chẳng, chưa, chả thì chúng 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 chúng tôi gán cho câu quan điểm nhãn -.
- Đối với từ quan điểm nhãn - thì chúng tôi không gán nhãn cho câu quan điểm.
- Đối với từ quan điểm không có nhãn thì chúng tôi gán nhãn - cho câu quan điểm.

3.4. Độ đo tính chính xác của hệ thống

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. Dữ liệu thực nghiệm và cài đặt

Chúng 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.

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. Kết quả thực nghiệm và phân tích

Dữ liệu được đưa qua bộ tách câu quan điểm để tách các câu phức và câu ghép thành các câu đơn mà chúng tôi xây dựng dựa trên luật (đã trình bày ở chương 3). Chúng 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. Sau khi tách câu, chúng tôi thu được bộ dữ liệu với số câu. Kết quả trả về là các câu đơn chỉ phát biểu về một tính năng (gồm một danh từ và một tính từ).

Bảng 4.2. 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 %

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 (gồm khoảng 150 từ quan điểm). Chúng 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 4.1. 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%

Trong danh sách các tính năng chúng 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 chúng 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, chúng tôi chọn độ hỗ trợ tối thiểu (minsup = 4), các danh từ có tần số xuất hiện <4 được hệ thống loại bỏ đi

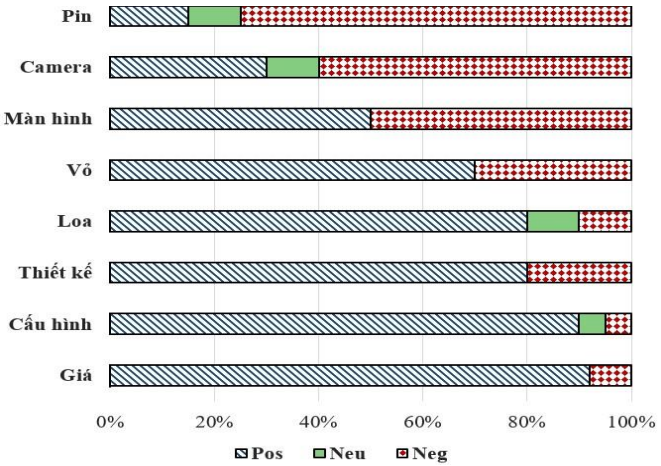
Cuối cùng, hệ thống thực hiện phân lớp các câu quan điểm theo từng tính năng (nhóm tính năng) mà đã được xử lý trong giai đoạn trước. Nhãn của từ quan điểm được lấy làm nhãn cho câu đánh giá. Trong phần này, tôi chỉ sử

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 4.4.

Bảng 4.4. 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%

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

Chương 5. Kết luận

5.1. Những vấn đề giải quyết được 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. Công việc nghiên cứu 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
- Trích xuất thực thể của các tính năng trong câu đánh giá

- 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
- Xử lý được các câu quan điểm so sánh khi người tiêu dùng so sánh các sản phẩm với nhau.