**KHAI PHÁ VÀ TỔNG KẾT QUAN ĐIỂM KHÁCH HÀNG**

Minqing and Bing Liu

Department of Computer Science

University of Illinois at Chicago

815 South Morgan Street

Chicago, IL 60607-7053

{mhu1, [liub}@cs.uic.edu](mailto:liub%7d@cs.uic.edu)

**TÓM TẮT**

Các doanh nghiệp bán các sản phẩm trên Web thường hỏi các khách hàng của họ đánh giá về các sản phẩm mà họ đã mua và các dịch vụ lên quan. Với thương mại điện tử ngày càng phổ biến, số lượng đánh giá của khách hàng về sản phẩm tăng lên nhanh chóng. Với một sản phẩm phổ biến, số lượng đánh giá có thể lên đến hàng trăm, thậm chí hàng nghìn. Điều này gây khó khăn cho khách hàng khi muốn đọc các đánh giá này để quyết định xem có nên mua sản phẩm không. Mặt khác nó cũng gây khó khăn cho nhà sản xuất khi muốn theo dõi và quản lý các quan điểm của khách hàng. Với nhà sản xuất, lại càng thêm khó khăn khi họ sản xuất nhiều sản phẩm và sản phẩm của họ lại được bán trên nhiều trang web thương mại. Trong nghiên cứu này, mục đích của chúng tôi nhằm khai thác và tổng hợp tất cả đánh giá của khác hàng về 1 sản phẩm. Phần tổng kết ở đây sẽ khác với tổng kết văn bản truyền thống, ở đây chúng tôi chỉ khai thác đặc trưng sản phẩm mà khách hàng bày tỏ quan điểm và quan điểm đó tích cực hay tiêu cực. Chúng tôi không tổng kết đánh giá bằng cách chọn một tập con hay viết lại một vài câu gốc từ các đánh giá để nắm bắt được các điểm chính như trong tổng kết văn bản cổ điển. Nhiệm vụ của chúng tôi sẽ được thực hiện trong 3 bước: (1) khai thác đặc trưng sản phẩm được khác hàng nhận xét; (2) xác định câu mang quan điểm và xác định câu mang quan điểm tích cực hay tiêu cực; (3) tổng hợp kết quả. Trong bài báo này, chúng tôi để xuất một vài kỹ thuật mới để thực hiện các nhiệm vụ trên. Kế quả thực nghiệm của chúng tôi sử dụng các đánh giá của một số sản phẩm được bán online để chứng minh tính hiệu quả của các kỹ thuật.

**XẾP LOẠI VÀ MIỂU TẢ CHỦ ĐỀ**

H.2.8 [Database Management]: Database Applications – *data mining*.

I.2.7 [Artificial Intelligence]: Natural Language Processing – *text analysis.*

**THUẬT NGỮ CHUNG**

Algorithms (Thuật toán), Experimentation (Thực nghiệm), Human Factors (Các yếu tố con người).

**TỪ KHÓA**

Text mining, sentiment classification (phân loại cảm tính), summarization, reviews.

Cho phép tạo ra các bản sao số hoặc bản cứng của tất cả hay 1 phân của sản phẩm này cho cá nhân hoặc phòng đọc sử dụng để cung cấp miễn phí và các bản sao không được làm hay phân phối vì lợi nhuận hay lợi thế thương mại, các bản sao phải có thông báo này và có đầy đủ sự trích dẫn như ở trang đầu. Sao chép các khác, hay tái bản, đưa lên server hay tái phân phối vào danh sách đòi hỏi phải có sự cho phép từ trước và hoặc có lệ phí.

KDD’04, August 22–25, 2004, Seattle, Washington, USA.

Copyright 2004 ACM 1-58113-888-1/04/0008…$5.00.

1. **GIỚI THIỆU**

Với việc mở rông nhanh chóng của thương mại điện tử, ngày càng nhiểu sản phẩn được bạn trên Web, và ngày càng nhiều người cũng mua các sản phẩm online. Để nâng cao sự hài lòng của khách hàng và kinh nghiệm mua sắm, nó đã trở thành một thực tế phồ biến cho các nhà bán hàng online cho phép khách hàng của mình có thể đánh giá hay bày tỏ quan điểm về sản phẩm mà họ đã mua. Cùng với ngày càng nhiều người sử dụng thông thường cảm thấy thoải mái với Web, con số những người viết đánh giá ngày càng tăng. Như một hệ quả, số lượng đánh giá của một sản phẩm nhận được tăng lên nhanh chóng. Một vào sản phẩm phổ biến có thể lấy được hàng trăm đánh giá trên một vài website bán hàng lớn. Hơn nữa, nhiều đánh giá dài nhưng lại chỉ có một vài câu mang quan điểm về sản phẩm. Điều đó tạo ra khó khăn cho một khách hàng tiềm năng đọc chúng để có quyết định về việc có nên mua sản phẩm hay không. Nếu anh/chị ấy chỉ đọc một vài đánh giá thì anh/chị ấy có thể sẽ có cái nhìn không chính xác. Một số lượng lớn đánh giá cũng gây ra khó khăn cho nhà sản xuất sản phẩm theo dõi quan điểm của khách hàng về sản phẩm của mình. Đối với một nhà sản xuất, sẽ có thêm khó khăn khi nhiều trang web thương mại có thể bán sản phẩm của mình, và nhà sản xuất có thể cũng cấp nhiều loại sản phẩm.

Trong nghiên cứu này, chúng tôi nghiên cứu vấn đề tìm ra *đặc trưng dựa trên tổng hợp* các đánh giá của khách hàng về sản phẩm đã được bạn online. Ở đây, *đặc trưng* rộng hơn có nghĩa là tính năng của sản phẩm hay thuộc tính hay chức năng của nó. Với một tập hợp đánh giá của khách hàng về một sản phẩm cụ thể, công việc ở đây sẽ liên quan đến 3 việc nhỏ sau: (1) xác định các đặc trưng của sản phẩm mà khách hàng bày tỏ trên quan điểm của mình (gọi là product feature); (2) với mỗi đặc trưng, xác định câu mang quan điểm mà cho quan điểm tích cực hay tiêu cực; (3) tạo một bản tổng kết dựa trên thông tin đã tìm kiếm được.

Cùng chúng tôi sử dụng một ví dụ để minh họa bản tổng kết dựa trên đặc trưng. Giả sử rằng chúng tôi tổng hợp các đánh giá của một camera kỹ thuật số cụ thể, camera\_digital\_1. Bản tổng kết sẽ trông giống như sau:

Digital\_camera\_1:

Feature: picture quality

Positive: 253

< indidual review sentences>

Negative: 6

< indidual review sentences>

Feature: size

Positive: 134

< indidual review sentences>

Negative: 10

< indidual review sentences>

…

Hình 1: Một ví dụ bản tổng kết.

Trong hình 1, picture quality và (camera) size là các đặc trưng của sản phẩm. Có 253 đánh giá của người dùng bày tỏ quan điểm tích cực về chất lượng hình ảnh, và chỉ có 6 đánh giá bày tỏ quan điểm tiêu cực. Và < indidual review sentences> điểm liên kết đến các câu cụ thể và/hoặc toàn bộ đánh giá đưa ra quan điểm tích cực hay tiêu cực về đặc trưng.

Với một bản tổng kết dựa trên đặc trưng như vậy, một khách hàng tiềm năng có thể dễ dàng thấy khách hàng hiện tại cảm thấy như thế nào về camera kỹ thuật số. Nếu anh/chị ấy tích thú về một đặc trưng cụ thể, anh/chị ấy có thể xem trong < indidual review sentences> và nó cho thấy tại sao khách hàng hiện tại thích nó hay phàn nàn những gì về nó. Với một nhà sản xuất, nó cũng có thể kết hợp đánh giá của nhiều website thương mại tạo thành một bản tổng hợp duy nhất cho mỗi sản phẩm của họ.

Nhiệm vụ của chúng tôi khác với tổng hợp văn bản truyền thống [15, 39, 36] trong một số cách. Đầu tiên, một bản tổng kết của chúng tôi có cấu trúc và ngắn hơn bất kì tài liệu văn bản miễn phí từ các hệ thống tổng hợp văn bản. Thứ hai, chúng tôi chỉ quan tâm đến đặc trưng của sản phẩm mà khách hàng bày tỏ quan điểm, cũng như quan điểm đó là tích cực hay tiêu cực. Chúng tôi không tổng hợp các đánh giá bởi lựa chọn hay viết lại một tập con các câu gốc từ các đánh giá để nắm bắt các điểm chính như trong tổng hợp văn bản truyền thống.

Như đã nói ở trên, nhiệm vụ của chúng tôi được thực hiện trong 3 bước chính:

1. Khai phá đặc trưng sản phẩm được khách hàng đánh giá. Chúng tôi sử dụng kết hợp kỹ thuật data mining và natural language processing để thực hiện công việc này. Một phần của nghiên cứu này được báo cáo tại [19]. Tuy nhiên, để đầy đủ, chúng tôi sẽ tóm tắt các kỹ thuật của mình có trong tờ báo này và cũng đưa ra đánh giá so sánh.
2. Xác định câu mang quan điểm trong mỗi đánh giá và quyết định xem mỗi quan điểm đó là tích cực hay tiêu cực. Chú ý rằng mỗi câu quan điểm ở trên phải chứ một hoặc nhiều hơn đặc trưng sản phẩm được xác định ở trên. Để quyết định *hướng quan điểm* của mỗi câu (quan điểm được thể hiện trong câu là tích cực hay tiêu cực), chúng tôi thực hiện 3 việc nhỏ: Đầu tiên, một tập hợp tính từ (là những từ bình thường sử dụng để bày tỏ quan điểm) được xác định bằng phương pháp xử lý ngôn ngữ tự nhiên. Các từ này cũng được gọi là *opinion words*  trong tờ báo này. Thứ hai, với mỗi từ quan điểm, chúng tôi xác định hướng ngữ nghĩa của nó, v.v., tích cực hay tiêu cực. Một công nghệ bootrapping được đế suất để thực hiện công việc này là sử dụng WordNet [29, 12]. Cuối cùng, chúng tôi quyết định hướng quan điểm cho mỗi câu. Một thuật toán hiệu quả cũng được đưa ra cho mục đích này.
3. Tổng hợp kết quả. Bước này tập hợp kết quả của bước trước đó và trình bày chúng theo như định dàng ở Hình 1.

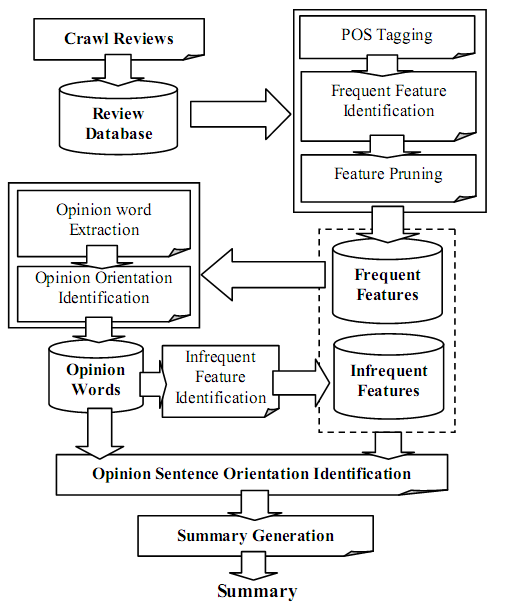
Phần 3 trình bày chi tiết công nghệ cho việc thực hiện các công việc. Một hệt thống, gọi là FBS (Feature-Based Summarization), cũng đã được thực hiện. Kết quả thực nghiệm của chúng tôi với một số lượng lớn đánh giá của khách hàng cho 5 sản phẩm đã bán online cho thấy FBS và kỹ thuật của nó có hiệu quả cao hơn.

1. **CÁC CÔNG VIỆC LIÊN QUAN.**

Công việc của chúng ra liên quan chặt chẽ đến công việc của Dave, Lawrence và Pennock [9] trong phân lớp ngữ nghĩa các đánh giá. Sử dụng các ngữ nghĩa được học sẵn từ một vài Web sites. ở đây mỗi đánh giá đều đã có một lớp (vd, like và dislike, hay một một vài xếp hạng khác hoặc xếp hạng nhị phân), chúng được thiết kế và thử nghiệm với một số phương thức cho xây dựng phân loại cảm tính. Chúng cho thấy rằng phân loại như vậy thực hiện rất tốt với các test gồm các câu đánh giá. Chúng cũng được sử dụng phân loại của chúng để phân lớp các câu thu được từ kết quả của Web tìm kiếm. với những gì thu được từ một Máy tìm kiếm sử dụng một tên sản phẩm như truy vấn tìm kiếm. Tuy nhiên, việc thực hiện bị giới hạn bởi một câu chứa ít thông tin hơn nhiều so với câu đánh giá…..

1. **CÁC KỸ THUẬT ĐỂ XUẤT.**

Hình 2 chỉ ra tổng quan kiến trúc của hệ thống tổng hợp quan điểm.



**Hình 2: Tổng kết quan điểm dựa trên đặc tính**

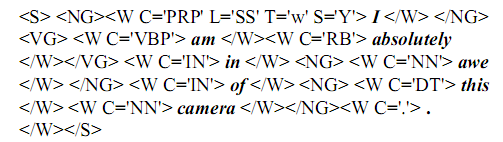
Đầu vào của hệ thống là tên một sản phẩm và một mục của tran web có chứa tất cả đánh giá về sản phẩm. Đẩu ra của hệ thống là bản tổng kết các quan điểm như đã thấy ở hình 1.

Hệ thống thực hiện tổng hợp trong 3 bước chính (như đã thảo luận ở trên): (1) khai phá đặc trưng sản phẩm được khách hàng đánh giá; (2) xác định câu quan điểm trong mỗi đánh giá và quyết định mỗi câu mang quan điểm tích cực hay tiêu cực; (3) tổng hợp kết quả. Các bước được thực hiện bởi nhiều bước con.

Với đầu vào, hệ thống download tất cả đánh giá, và đưa chúng vào database chứa câu đánh giá (review database). Nó sẽ tìm thấy những đặc trưng “hot” hay phổ biến mà được nhiều người để cập đến trong quan điểm của họ. Sau đó, các từ quan điểm được trích chọn dựa trên đặc trưng phổ biến được nhắc tới, và hướng ngữ nghĩa của từ quan điểm cũng được xác định với sự trợ giúp của WordNet. Bằng các sử dụng trích chọn các từ mang quan điểm, hệ thống sau đó sẽ tìm thấy những đặc trưng không phổ biến. Trong hai bước cuối cùng, hướng của mỗi câu quan điểm sẽ được xác định và một bản tổng kết cuối cùng được tạo ra. Ghi chú răng POS tagging là gán nhãn từ loại [28] từ xử lý ngôn ngữ tự nhiên, mà nó giúp chúng tôi tìm đặc trưng quan điểm. Phía dưới, chúng tôi sẽ thảo luận lần lượt các bước nhỏ.

**3.1 Part-of-Speech Tagging (POS).**

Đặc trưng sản phẩm thông thường là danh từ hoặc cụm danh từ trong câu quan điểm. Do đó gán nhãn từ loại là rất quan trọng. Chúng tôi đã sử dụng phân tích cú pháp ngôn ngữ NLProcessor [31] để phân tích mỗi đánh giá và phân tách văn bản bên trong các câu và gắn thẻ từ loại cho mỗi từ ( với từ đó là danh từ, động từ, tính từ, v.v). Quá trình xử lý cũng xác định danh tư đơn và nhóm danh từ. Dưới đây cho thấy một câu với POS tags.



NLProcessor tạo ra file XML. Ví dụ, <W C= ‘NN’> chỉ một danh từ và <NG> chỉ một cụm danh từ, hay một nhóm danh từ.

Mỗi câu được lưu lại trong review database cùng với thẻ POS cho mỗi từ trong câu. Một file giao dịch sẽ được tạo ra cho các đặc trưng phổ biến trong bước tiếp theo. Trong file này, mỗi dòng chứa các từ từ một câu, mà chỉ bao gồm danh từ và cụm danh từ của câu đã được xác định. Các thành phần khác của câu khó có thể là đặc trưng của sản phẩm. Một vài bước tiền xử lý các từ cũng được thực hiện, mà bao gồm loại bỏ các từ stopword, làm rõ hay làm mờ phù hợp. Làm mở phù hợp được sử dụng để đối phó với biến thể từ và lỗi chính tả [19]

**3.2 XÁC ĐỊNH ĐặC TRƯNG THƯỜNG XUYÊN.**

Đây là bước con xác định đặc trưng sản phẩm mà nhiều người đưa ra trong quan điểm của họ. Trước khi thảo luận về xác định đặc trưng phổ biến, đầu tiên chúng ra đưa ra một vài câu ví dụ từ một vài đánh giá để mô tả nhưng loại quan điểm mà chúng tôi sẽ xử lý. Kể từ khi hệ thống của chúng tôi nhằm mục đích tìm nhưng gì mọi người thích và không thích về một sản phẩm nhất định. Như thế nào để tìm đặc trưng sản phẩm mà mọi người nói về là một bước quan trọng. Tuy nhiên, do khó khăn trong việc hiểu ngôn ngữ tự nhiên, một vào kiểu câu rất khó đối phó. Chúng tôi hãy cùng xem một câu đơn giản và một câu phức tạp từ đánh giá của một camera kỹ thuật số:

“Các bức ảnh đều rất sáng.”

Trong câu này, người sử dụng hài lòng với chất lượng bức ảnh của camera, picture là đặc trưng mà người dùng nói tới. Trong khi đặc trưng của câu này được đề cập rõ ràng trong câu, thì một vào đặc trưng lại tiềm ẩn và khó tìm. Ví dụ,

“Thật không dễ dàng để cho nó vào túi áo.”

Khách hàng này nói về kích thước của camera, nhưng từ size không xuất hiện trong câu. Trong trường hợp này, chúng ra tập trung vào tìm kiếm đặc trưng xuất hiện rõ ràng như danh từ hay cụm danh từ trong các đánh giá. Chúng tôi tạm thời để việc tìm đặc trưng tiềm ẩn cho nghiên cứu trong tương lai.

Ở đây, chúng tôi tập trung và tìm dặc trưng phổ biến, v.v . những tính năng mà được nhiều khách hàng nói đến (tìm đặc trưng ít phổ biến sẽ được thảo luận sau). Với mục đích này, chúng tôi sử dụng association mining [1] để tìm tất cả các tập hợp phổ biến. Trong bối cảnh này, một tập phổ biến đơn giản chỉ là một tập các từ hay một cụm từ xuất hiện cùng nhau trong một vài câu.

Lý do chính của việc sử dụng khai phá kết hợp (association mining) là bởi vì dựa theo sự theo dõi. Nó là phổ biến khi một đánh giá của khách hàng có chứa nhiều ý mà nó không liên quan trực tiếp đến đặc tính của sản phẩm. Nhưng khách hàng khác nhau thường sẽ có các câu chuyện khác nhau. Tuy nhiên, khi họ đánh giá một đặc tính của sản phẩm, những từ mà họ dùng sẽ hội tụ. Do đó sử dụng khai phá kết hợp để tìm các tập phổ biến là thích hợp bởi vì các tập phổ biến phổ biến xuất hiện sẽ giống như là một đặc tính của sản phẩm. Nhưng danh từ, cụm danh từ không phổ biến xuất hiện coi như là không phải đặc tính của sản phẩm.

Chúng tôi chạy hệ thống khai phá liên kết CBA [26], nó dự trên thuật toán Apriori trong [1] trên tập danh từ/cụm danh từ thực hiện được tạo ra từ bước trước. Mỗi kết quả tập phổ biến phổ biến là một đặc tính khả thi. Trong công việc của chúng tôi, chúng tôi định nghĩa một tập phổ biến là thương xuyên nếu chúng xuất hiện nhiều hơn 1% trong tổng số các câu đánh giá. Các tập phổ biến phổ biến được tạo ra cũng được gọi là *đặc tính phổ biến* tạm thời trong bài báo này.

Tuy nhiên, không phải tất cả các đặc tính phổ biến tạm thời được tạo ra bởi khai phá kết hợp đều là đặc tính chuẩn. Có hai kiểu loại trừ được sử dụng để loại bỏ các đặc tính không thích hợp sau**:**

**Lược bỏ cơ bản:** Phương thức này kiểm tra đặc tính có chứa ít nhất 2 từ, mà chúng tôi gọi chúng là đặc tính kiểu cụm, và loại bỏ chúng nếu chúng thực sự vô nghĩa.

Thuật toán khai phá kết hợp không xem xét đến vị trí của từ trong một câu. Tuy nhiên, trong một câu, các từ xuất hiện cùng nhau theo một trật tự nhất định thì nhiều khả năng nó là một cụm có nghĩa. Vì thế, một vài cụm đặc tính phổ biến được tìm thấy bởi khai phá liên kết cũng có thể không phải đặc tính chuẩn. Loại bỏ cơ bản tập trung loại bỏ các đặc tính tạm thời có các từ không xuất hiện cùng nhau theo một trật tự nhất định. Xem [19] để có định nghĩa chi tiết về tính súc tích cũng như thủ tục lược bỏ.

**Lược bỏ dư thừa:** Trong bước này, chúng tôi tập trung loại bỏ tính năng rườm rà mà có chứa các từ đơn. Để mô tả ý nghĩa của tính năng rườm rà, chúng tôi sử dụng ý tưởng của *p-support* (*pure support*). *P-support* của tính năng *ftr* là số lượng câu mà *ftr* xuất hiện như là một danh từ hay cụm danh từ, và các câu này chứa không chứa đặc tính kiểu cụm nhưng là một tập hợp cha của *ftr*.

Chúng tôi sử dụng một giá trị p-support nhỏ nhất để lược bỏ các đặc tính dư thừa. Nếu một đặc tính có p-support thấp hơn p-support cực tiểu (trong hệ thống của chúng tôi, chúng tôi đặt nó là 3) và đặc tính là một tập con của đặc tính kiểu cụm khác. (điều này cho thấy tính năng đơn lẻ có thể không gây chú ý), nó đã được lược bỏ. Ví dụ: life bản thân nó không phải là đặc tính hữu dung trong khi battery life là một đặc tính kiểu cụm có nghĩa. Xem [19] cho nhiều ví dụ hơn.

**3.3 TRÍCH CHỌN TỪ MANG QUAN ĐIỂM.**

Bây giờ chúng tôi xác định các từ mang quan điểm. Các từ mà chủ yếu được sử dụng để bày tỏ quan điểm chủ quan. Rõ ràng hơn, nó liên quan đến công việc đang tồn tại trên các câu phân tách được sử dụng để bày tỏ quan điểm chủ quan từ từ các câu được sử dụng để mô tả khách quan một số thông tin thực tế [43]. Các nghiên cứu trước đây với chủ quan [44, 4] đã tạo lên một thống kê tích cực liên quan đáng kể đến sự xuất hiện của tính từ. Do đó sự xuất hiện của tính từ rất hữu dụng để dự đoán một câu có là chủ quan, v.v, bày tỏ một quan điểm. Bài báo này sử dụng tính từ như là từ mang quan điểm. Chúng tôi cũng giới hạn từ mang quan điểm trích chọn từ những câu mà có chứa nhiều hơn một đặc tính của sản phẩm, và cũng chỉ chú ý tới quan điểm của khách hàng trên các đặc tính của sản phẩm. Đầu tiên chúng tôi hãy cùng định nghĩa một câu mang quan điểm.

**Định nghĩa**: *Câu quan điểm*.

Nếu một câu chưa một hoặc nhiều hơn đặc tính của sản phẩm và một hoặc nhiều hơn từ mang quan điểm, thì câu đó được gọi là *câu quan điểm*.

Chúng tôi trích chọn các từ mang quan điểm theo cách sau đây (Hình 3):

for mỗi câu mang trong cơ sở dữ liệu chứa các đánh giá.

if (nó có chứa một đặc tính, trích chọn toàn bộ tính từ như là từ mang quan điểm)

for mỗi đặc tính có trong câu

Tính từ gần nhất được ghi nhận như là quan điểm hữu dụng của nó. /\* Một tính từ gần đó đề cập đến một tính từ liền kề mà làm thay đổi danh từ/cụm danh từ thì đó là một đặc tính phổ biến.\*/

**Hình 3: Trích chọn từ mang quan điểm.**

Ví dụ: horrible là quan điểm hữu dụng của strap trong “The strap is horrible and gets in the way of parts of the camera you need access to.” Quan điểm hữu dụng sẽ có ích khi chúng tôi tiên đoán hướng của câu quan điểm.

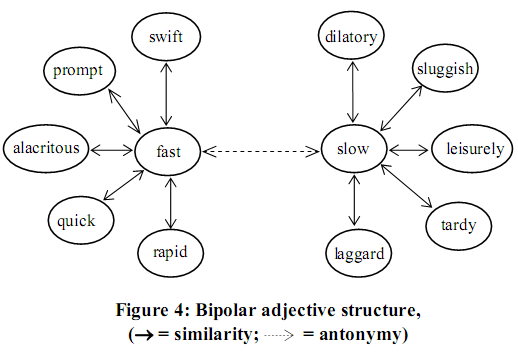
**3.4 XÁC ĐỊNH HƯỚNG CỦA TỪ MANG QUAN ĐIỂM.**

Với mỗi từ mang quan điểm, chúng tôi cần phải xác định hướng ngữ nghĩa của nó, nó sẽ được sử dụng để dự đoán hướng ngữ nghĩa của mỗi câu quan điểm. Hướng ngữ nghĩa cả một từ chỉ ra sự định hướng mà từ lệch khỏi tiêu chuẩn đối với nhóm ngữ nghĩa của nó. Các từ đó được mã hóa một trạng thái mong muốn (VD đẹp, kinh hoàng) mang hướng tích cực, trong khi các từ đó đại diện cho một trạng thái không mong muốn thì sẽ mang hướng tiêu cực (VD thất vọng). Trong khi các định hướng được áp dụng cho nhiều tính từ, cũng có nhiều tính từ mà không mang định hướng (VD bên ngoài, thuộc về số) [17]. Trong nghiên cứu này, chúng tôi chỉ chú trọng vào hướng tích cực và tiêu cực.

Thật đáng tiếc, các từ điển và các nguồn tương tự, vd WordNet [29] không có thông tin hướng ngữ nghĩa cho mỗi từ. Hatzivassiloglou và McKeown [16] dùng một thuật toán học có giám sát để suy luận hướng ngữ nghĩa của tính từ từ các rằng buộc trên các liên từ. Mặc dù phương pháp của họ đạt được sự chính xác cao, nó dựa trên tập văn bản lớn, và cần một số lượng lớn dữ liệu đào tạo đã được gán nhãn thủ công. Trong nghiên cứu của Turney [42], hướng ngữ nghĩa của một cụm từ được dùng như thông tin qua lại để suy đoán giữa cụm từ đã cho và từ “tốt” mang thông tin dấu “-” với những cụm từ đã cho và từ “xấu”. Thông tin qua lại được ước lượng bởi các dưa ra các truy vấn tới máy chủ tìm kiếm và chú trọng vào số lượng kết quả đúng. Bài báo [42], tuy không trình bày các kết quả hướng ngữ nghĩa từng từ/cụm từ riêng rẻ. Thay vào đó nó chỉ trình bày phân lớp các kết quả của các đánh giá. Chúng tôi không sử dụng các kỹ thuật trong bài biết này cũng như cả hai nghiên cứu dựa trên thông tin thống kê từ một tập văn bản lớn. Phương pháp của họ cũng không hiệu quả. Ví dụ, trong [42], với mỗi từ hay cụm từ, một Web tìm kiếm và một xử lý đáng kể các kết quả được trả về là cần thiết.

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp đơn giản nhưng hiệu quả sử dụng tập tính từ đồng nghĩa và trái nghĩa trong WordNet [29] để dự đoán hướng ngữ nghĩa của các tính từ.

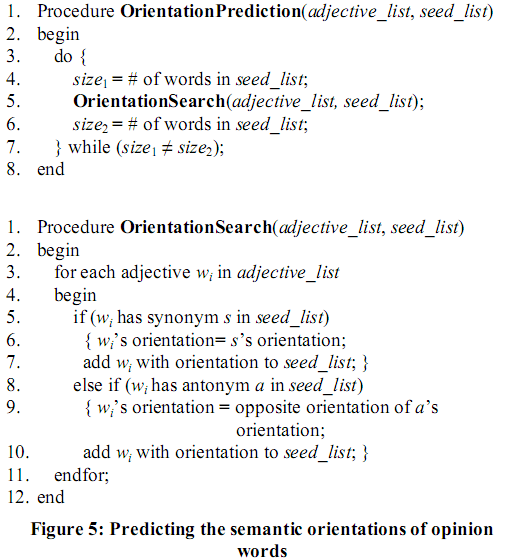
Trong WordNet, tính từ được tổ chức trong các cụm lưỡng cực, như minh họa trong Hình 4. Cụm fast/slow, bao gồm hai nữa cụm, một cho nghĩa fast và một cho nghĩa slow. Mỗi nửa cụm được đại diện bởi một *từ đại diện*, trong trường hợp này fast and trái nghĩa của nó là slow. Theo sau từ đại diện là các *từ vệ tinh* má ý nghĩa đại diện của chúng tương tự như ý nghĩa của từ đại diện. Nửa cụm còn lại được đại diện bởi từ còn lại của cập từ trái nghĩa slow/fast, các từ vệ tinh theo sau nó mang ý nghĩa của từ slow [12].



Nói chung, các tính từ mang hướng giống nhau như là đồng nghĩa của chúng và hướng đối diện như là trái nghĩa của chúng. Chúng tôi sử dụng ý tưởng này để dự đoán hướng ngữ nghĩa của một tính từ. Để làm điều đó, một synset của tính từ đã cho và tập trái nghĩa cần được tìm kiếm. Nếu đồng nghĩa/trái nghĩa đã biết hướng ngữ nghĩa, thì hướng của tính từ đã cho có thể suy luận tương ứng. Như một synset của một tính từ luôn luôn bao gồm một nghĩa mà liên kết đến một đại diện của synset, phạm vi tìm kiếm là khá lớn. Với đủ các tính từ gốc với hướng của nó, chúng ra có thể gần như dự đoán được hướng của tất cả các tính từ có trong tập đánh giá.

Vì thế, chiến lượng của chúng tôi là sử dụng một tập các tính từ gốc, mà chúng tôi đã biết hướng của nó và sau đó phát triển bộ này chẳng cách tìm kiếm trong WordNet. Để có một phạm vi rộng hợp lý các tính từ (trong thử nghiệm, chúng tôi sử dụng 30) như là danh sách từ gốc, vd tính từ tích cực: tuyệt vời, đẹp… và tính từ tiêu cực: tồi, đần độn. Sau đó chúng tôi nhờ đến WordNet để dự đoán hướng của tất cả các tính từ trong danh sách từ mang quan điểm. mỗi hướng của tính từ được dự đoán được, nó được thêm vào danh sách từ gốc. Do đó, dánh sách sẽ lớn dần theo quá trình.

Thuật toán hoàn chỉnh để dự đoán hướng ngữ nghĩa cho tất cả tính từ trong danh sách tính từ mang quan điểm được chỉ ra trong Hình 5.



Thủ tục *OrientationPrediction* lấy tính từ trong danh sách từ gốc và một danh sách các từ quan điểm cần được xác định hướng.

Thủ tục OrientationPrediction được gọi lặp đi lặp lại cho đến khi không có từ quan điểm mới được thêm vào danh sách từ gốc. Mỗi lần một tính từ với định hướng của nó được thêm vào dánh sách từ gốc, danh sách từ gốc được cập nhật sau đó gọi OrientationSearch nhiều lần để khai thác thông tin mới nhất được bổ sung.

Thủ tục OrientationSearch tìm kiếm WordNet và danh sách từ gốc cho mỗi tính từ đích để dự đoán hướng của nó (dòng 3 - 11). Trong dòng 5, nó tìm kiếm synset của tính từ đích trong WordNet và kiển tra nếu bất kỳ từ nào đồng nghĩa đã biết hướng ngữ nghĩa. Nếu có, hướng đích được xác định giống hướng của từ đồng nghĩa (dòng 6) và tính từ địch cùng với hướng của nó được thêm vào danh sách từ gốc (dòng 7). Ngoài ra, hàm tiếp tục tìm kiếm tập trái nghĩa với từ địch trong WordNet và kiểm tra nếu bất kỳ từ trái nghĩa đã biết hướng (dòng 8). Nếu có hướng địch sẽ được xác định đối lại của từ trái nghĩa (dòng 9) và tính từ đích với hướng của nó được thêm vào danh sách từ gốc (dòng 10). Nếu không có bất kỳ đồng nghĩa hay trái nghĩa nào của từ đã biết nghĩa, hàm vừa xong tiếp tục xử lý tương tự với tính từ tiếp theo cho đến khi hướng của từ này được tìm thấy trong những lần gọi sau của thủ tục với một danh sách từ gốc được update.

Với các tính từ mà WordNet không thể nhận ra, chúng bị loại bỏ như chúng có thể là những từ không hợp lệ. Với những từ mà không thể xác định hướng, chúng cũng sẽ bị loại bỏ khỏi danh sách từ quan điểm và người sử dụng sẽ được thông báo để chú ý. Nếu người sử dụng cảm thấy rằng từ đó là một từ quan điểm và biết cảm tính của nó, anh/chị ấy cso thể cập nhật nó vào danh sách từ gốc. Trong thử nghiệm của chúng tôi, không có sự tham gia của người sử dụng (những từ quan điểm bị loại bỏ sẽ bỏ hẳn). Đối với trường hợp mà đồng nghĩa/trái nghĩa của một tính từ có hướng ngữ nghĩa khác nhau, chúng tôi sử dụng hướng đầu tiên tìm thấy như là hướng cho tính từ đã cho.

**3.5 XÁC ĐỊNH TÍNH NĂNG KHÔNG PHỔ BIẾN**

Đặc tính không phổ biến là những đặc tính “hot” mà mọi người bình luận nhiều về sản phẩm đã cho. Tuy nhiên, có một vài tính năng chỉ có một số ít người nói đến. Có nhiều tính năng cũng có thể được chú ý tới một số khách hàng tiềm năng và nhà sản xuất ra sản phẩm. Câu hỏi là làm như thế nào để trích chọn được tính năng không phổ biến (association mining không thể xác định các tính năng này)? Cùng xem xét các câu sau đây:

“Các bức ảnh hoàn toàn tuyệt vời.”

“Phần mềm đi xem với nó thật tuyệt vời.”

Câu 1 và 2 cùng có từ mang quan điểm giống nhau “tuyệt vời” nhưng miêu tả hai tính năng khác nhau: câu 1 nói về bức ảnh, và câu 2 nói về phần mềm. Từ một số tính từ có thể được sử dụng để miêu tả các đối tượng khác nhau, chúng tôi cũng có thể sử dụng các từ mang quan điểm để tìm đặc tính mà không được tìm thấy bằng phương pháp tìm kiếm đặc tính phổ biến bằng association mining.

Chúng tôi trích chọn đặc tính không phổ biến sử dụng thuật toán như trong Hình 6:

for mỗi câu trong cơ sở dữ liệu đánh giá

if (nó không chứa tính năng phổ biến nhưng có một hay nhiều hơn các tính từ mang quan điểm)

{ tìm danh từ/cụm danh từ gần nhất xung quanh từ mang quan điểm. Danh từ/cụm danh từ đó được lưu lại là một đặc tính không phổ biến. }

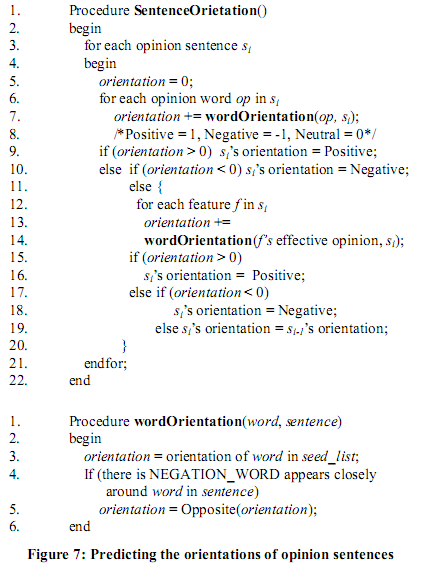
**Hình 6: Trích chọn tính năng không phổ biến.**

Chúng tôi sử dụng danh từ/cụm danh từ gần nhất như là danh từ/cụm danh từ mà từ mang quan điểm có ảnh hưởng bởi vì nó hay xảy ra nhiều nhất trong tất cả các trường hợp. Phỏng đoán đơn giản tỏ ra làm việc tốt trong thực tế.

Một vấn đề với việc xác định đặc trưng không phổ biến sử dụng từ mang quan điểm là chúng cũng tìm ra nhưng danh từ/cụm danh từ không liên quan đến sản phẩm đã cho. Lý do của vấn đề này là mọi người có thể sử dụng tính từ chung để miêu tả nhiều đối tượng, bao gồm cả những đặc tính mà chúng tôi muốn lẫn đặc tính không thích hợp. Tuy nhiên, vấn đề này không phải mà một vấn đề nghiêm trọng bởi vì số lượng tính năng không phổ biến, so với sô lượng tính năng phổ biến là nhỏ. Chúng chiếm khoảng 15-20% trên tổng số đặc tính của sản phẩm thu được trong thử nghiệm của chúng tôi. Tính năng không phổ biến được tạo ra cho đầy đủ. Hơn nữa, đặc tính phổ biến quan trọng hơn đặc tính không phổ biến. Từ việc chúng tôi xếp hạng đặc tính theo p-support, nhưng đặc tính không phổ biến lỗi không tốt sẽ được xếp hạng rất thấp và cũng sẽ không ảnh hưởng nhiều đến hầu hết người sử dụng.

**3.6 DỰ ĐOÁN HƯỚNG QUAN ĐIỂM CỦA CÂU.**

Bây giờ chúng tôi đến bước dự đoán hướng của câu quan điểm, vd tích cực, tiêu cực. Tổng quan, chúng tôi sử dụng hướng vượt trội của các từ quan điểm trong câu để xác định hướng quan điểm của câu. Đó là, nếu quan điểm tích cực/tiêu cực chiếm ưu thế thì quan điểm của câu cói như là tích cực/tiêu cực. Trong trường hợp số lượng từ mang quan điểm tích cựu và tiêu cực trong câu bằng nhau, chúng tôi dự đoán hướng quan điểm sử dụng hướng trung bình của *quan điểm hữu dụng* hoặc hướng của quan điểm của câu trước đó (nhắc lại *quan điểm hữu dụng* là từ mang quan điểm gần nhất của một tính năng trong mọt câu quan điểm). Đây là một phương pháp hiệu quả được chỉ ra trong kế quả thử nhiệm của chúng tôi. Thuật toán chi tiết được miêu tả trong Hình 7:



Thủ tục SentenceOrientation giải quyết 3 tình huống trong dự đoán hướng ngữ nghĩa của câu ngữ nghĩa:

1. Người sử dụng thích hoặc không thích hầu hết hay tất cả các tính năng trong một câu. Các từ mang quan điểm hầu hết tích cực hoặc tiêu cực, vd.., Có hai từ mang quan điểm tích cực, tốt và đặc biệt trong “hầu như nó là một máy ảnh tốt với thực tế bức ảnh rõ ràng và đặc biệt là khả năng chụp cận cảnh.”

2. Người sử dụng thích hoặc không thích hầu hết các đặc tính trong 1 câu nhưng có số lượng từ mang quan điểm tích cực và tiêu cực tương đương. Vd “Chế độ chế độ quay phim tự động hay bằng tay đều rất dễ sử dụng, nhưng phần mềm không trực quan.” Có một từ tích cực “dễ” và một từ tiêu cựu “không trực quan”, mặc dù người dùng thích hai đặc tính và không thích một đặc tính.

3. Các trường hợp còn lại.

Với trường hợp 1, hướng chiếm ưu thế có thể dễ dàng xác định (dòng 5-10 trong thủ tục SentenceOrientation). Đây là trường hợp phổ biến khi mọi người bày tỏ quan điểm của mình. Với trường hợp 2, chúng tôi sử dụng hướng trung bình của quan điểm hữu dụng với tính năng của nó. (dòng 12-18). Quan điểm hữu dụng được giả sử hầu hết có liên quan đến một tính năng. Với trường hợp 3, chúng tôi đưa ra hướng của câu quan điểm giống hướng của câu trước đó (dòng 19). Chúng tôi sử dụng thông tin ngữ cảnh để dự đoán hướng của câu bởi vì trong hầu hết các trường hợp mọi người bày tỏ quan điểm tích cực/tiêu cực của họ cùng với nhau trong một đoạn văn bản,…, một vài câu liên tiếp.

Đối với câu có chứa mệnh đề “nhưng” (một câu con bắt đầu với nhưng, tuy nhiên, v.v), chúng chỉ ra sự thay đổi cảm tính của đặc tính trong mệnh đề, đầu tiên chúng tôi sử dụng quan điểm hữu dụng trong mệnh đề và quyết định hướng của đặc tính. Nếu không có quan điểm xuất hiện trong mệnh đề, định hướng của câu sẽ bị đổi ngược lại.

Chú ý rằng trong thử tục wordOrientation, chúng tôi không đơn giản là lấy hướng ngữ nghĩa của từ quan điểm từ tập các từ quan điểm như hướng của nó trong câu cụ thể. Chúng tôi cũng xem xét đến các từ phủ định như “không”, “not”, “yet”, xuất hiện gần nhất từ quan điểm. Nếu vậy, hướng quan điểm của câu là ngược lại với hướng cơ bản của nó. (dong 4-5). Xuất hiện gần nhất, chúng tôi định nghĩa rằng khoảng cách các từ giữa từ phủ định và từ mang quan điểm không quá một ngưỡng (trong thử nghiệm, chúng tôi đặt chúng là 5). Phương thức đơn giản giải quyết với câu kiểu như “Máy ảnh không dễ sử dụng”, và “nó sẽ đẹp hơn nếu không thấy dấu hiệu của phóng to trên các cạnh”. Phương thức này thực sự có hiệu quả trong hầu hết các trường hợp.

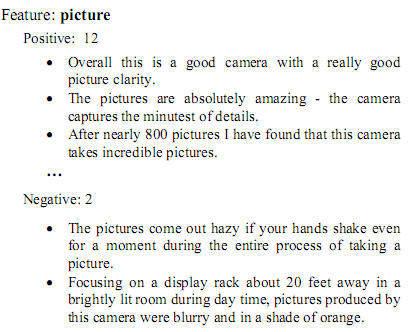
**3.7 ĐƯA RA TỔNG KẾT.**

Sau khi thực hiện tất cả các bước trên, chúng tôi sẵn sàng để tạo ta bản tổng hợp đánh giá dựa trên tính năng cuối cùng, nó không phức tạp lắm và được thực hiện theo các bước sau đây:

- Với mỗi tính năng được tìm thấy và các câu quan điểm liên quan được đặt vào nhóm tích cực và tiêu cực được phân loại theo hướng của câu quan điểm. Một số đếm được tính toán để cho thấy có bao nhiêu đánh giá cho quan điểm tích cực/tiêu cực của đặc trưng.

- Tất cả tính năng được xếp hạng theo tuần suất xuất hiện của chúng trong đánh giá. Các đặc tính là cụm danh từ đặt trước đặc tính là danh từ đơn như các cụm thông thường gây chú ý nhiều hơn đến người sử dụng. Có nhiều cách khác xếp hạng cũng khả thi. Ví dụ, chúng tôi cũng có thể xếp hạng tính năng theo số lượng đánh giá được bày tỏ quan điểm tích cực/tiêu cực.

Sau đây là bản tổng kết ví dụ cho tính năng “bức ảnh” của máy ảnh kỹ thuật số. Chú ý rằng các câu quan điểm riêng lẻ có thể bị ẩn dưới một siêu liên kết được đặt trước để người sử dụng có thể xem tổng quan bản đánh giá một cách dễ dàng.



**4. ĐÁNH GIÁ THỰC NHIỆM**

Một hệ thống, được gọi là FBS (Feature-Based Summarization), dựa trên kỹ thuật đã được đề xuất được cài đặt bằng C++. Bây giờ chúng tôi đánh giá FBS dựa trên 3 yếu tố:

1. Hiệu quả trích chọn các đặc tính.

2. Hiệu quả trích chọn câu mang quan điểm.

3. Độ chính xác của hướng ngữ nghĩa được dự đoán của câu mang quan điểm.

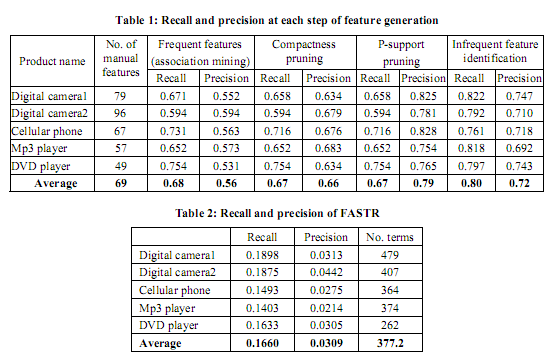
Chúng tôi tiến hành thí nghiệm của mình sử dụng các đánh giá của khác hàng trên 5 sản phẩm điên từ: 2 máy ảnh kỹ thuật số, 1 đầu DVD, 1 máy MP3, và một điện thoại di động. Các đánh giá được lựa chọn từ **Amazon.com** và **Cnet.com**. Các sản phẩm trong trang này có số lượng đánh giá lớn. Mỗi đánh giá bao gồm một đoạn văn bản và một tiêu đề. Thông tin bổ sung có sẵn nhưng không được sử dụng bao gồm ngày, giờ, tên người viết, vì trí (đối với Amazon), và xếp hạng.

Với mỗi sản phẩm, đầu tiên chúng tôi thu thập và tải về 100 đánh giá đầu tiên. Chúng được xem xét và loại bỏ các thẻ HTML. Sau đó, NLProcessor [31] được sử dụng để gán nhãn từ loại. Hệ thống cuat chúng tôi sau đó được áp dụng để thực hiện tổng kết.

Để đánh giá, chúng tôi đọc tất cả các đánh giá. Với mỗi câu trong một đánh giá, nếu nó có quan điểm của người sử dụng, tất cả tính năng có trong nó được người đánh giá bày tỏ quan điểm của họ sẽ được gán nhãn. Với quan điểm tích cực hay tiêu cực cũng đều được xác định. Nếu người sử dụng không có quan điểm trong câu, thì câu đó không được gán nhãn, vì chúng tôi chỉ quan tâm đến câu mang quan điểm trong nghiên cứu này. Với mỗi sản phẩm, chúng tôi tạo ra một danh sách đặc tính bằng tay. Cột “No. of manual features” trong bảng 1 chỉ ra số lượng đặc tính của mỗi sản phẩm. Tất cả kết quả được tạo ra bởi hệ thống của chúng tôi sẽ được so sanh với kết quả đã được gán nhãn bằng tay. Gán nhãn là khá đơn giản cho cả đặc tính sản phẩm và quan điểm. Một vấn đề nhỏ khi gán nhãn đặc tính là đặc tính đó có thể rõ ràng hay tiềm ẩn trong câu. Hầu hết các đặc tính xuất hiện rõ ràng trong câu quan điểm, vd *bức ảnh* trong *“Các bức ảnh hoàn toàn tuyệt vời”* Một vài đặc tính có thể không xuất hiện trong câu. Chúng tôi gọi cũng gán như thế với đặc tính tiềm ẩn, vd *kích thước* trong *“Nó vừa khít với cái túi áo khoác”.* Cả hai đặc tính rõ ràng và tiểm ẩn đểu dễ dàng xác định bởi người gán nhãn.

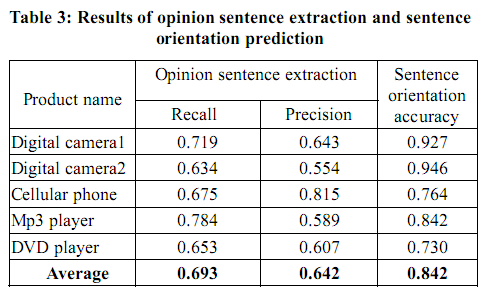
Một vấn đề khác là việc xác định quan điểm này trong các đánh giá có phần chủ quan. Nó thực sự dễ dàng để xác định quan điểm là tích cực hay tiêu cực nếu một câu bày tỏ quan điểm một cách rõ ràng. Tuy nhiên, quyết định một câu đưa ra một quan điểm hay không cũng gây tranh cãi. Đối với những trường hợp khó, một sự đồng thuận phải đạt được giữa người gán nhãn đầu tiên, và người gán nhãn thứ hai.

Bảng 1 cho thấy độ chính xác và kết quả đạt được của chức năng trích chọn đặc tính của FBS. Chúng tôi đã đánh giá kết quả qua tại mỗi bước của thuật toán. Trong bảng này, cột 1 liệt kê các sản phẩm, cột 3 và 4 cho thấy kết quả và độ chính xác của đặc tính phổ biến được tạo ra cho mỗi sản phẩm, chúng sử dụng association mining. Kết quả chỉ ra rằng đặc tính phổ biến bao gồm nhiều lỗi. Sử dụng bước này một mình cho thấy kết quả kém, độ chính xác thấp… Cột 5 và 6 chỉ ra kết quả tương ứng sau khi loại bỏ cơ bản được thực hiện. Chúng tôi có thể thấy rằng độ chính xác được cải thiện rõ rằng với loại bỏ này. Kế quả cho thấy vẫn ổn định. Cột 7 và 8 cho kết quả sau khi sử dụng p-support. Có nhưng cải thiện khác đáng kể trong độ chính xác. Kết quả ở các mức độ gần như không có sự thay đổi. Kết quả từ cột 4-8 cho thấy hiệu quả rõ ràng của hai kỹ thuật loại bỏ này. Cột 9 và 10 cho kết quả sau khi hoàn tất xác định đặc tính không phổ biến. Kết quả cải thiện đột ngột. Độ chính xác giảm vài phần trăm so với trung bình. Tuy nhiên, nó không phải là vấn đề lớn bời vì đặc tính không phổ biến được xếp loại thấp hơn, và cũng sẽ không ảnh hưởng đến hầu hết người sử dụng.



Để minh họa thêm hiệu quả của bước trích chọn đặc tính của chúng tôi, chúng tôi đã so sánh các đặc tính trích chọn được bởi phương pháp của chúng tôi với các từ được tìm thấy phổ biến và các từ trích chọn công bố công khai và bảng mục lục hệ thống, FASTR [11] của Christian Jacquemin. Bảng 2 cho thấy kế quả và độ chính xác của FASTR.

Chúng tôi nhận thấy rằng cả kết quả trung bình và độ chính xác của FASTR đều thấp hơn phương pháp của chúng tôi đáng kể. Sau khi kết thúc kiểu tra các từ tạo ra bởi FASTR, chúng tôi thấy rằng có hai lý do chính dẫn tới kết quả kém. Đầu tiên, FASTR tạo ra một số lượng lớn các từ, như đã thấy ở cột 4 “No. terms” trong bảng 2. Trung bình số từ tìm được bởi FASTR là 377. Hầu hết các từ này đều không phải đặc tính của sản phẩm (mặc dù nhiều từ là cụm danh từ). Thứ hai, FASTR không tìm các từ đơn, mà chỉ tìm các cụm từ mà có chứa 2 hoặc nhiều hơn 2 từ. Phương thức trích chọn đặc tính của chúng tôi tìm cả các từ đơn và cụm từ. So sánh kế quả trong bảng 1 và bảng 2, chúng tôi rõ ràng nhận thấy rằng phương pháp đã đề xuất có nhiều hiệu quả trong công việc này.



Bảng 3 cho thấy đánh giá kết quả của hai thủ tục: trích chọn câu mang quan điểm và dự đoán hướng ngữ nghĩa của câu. Kết quả trung bình của trích chọn câu mang quan điểm gần 70%. Độ chính xác trung bình của trích chọn câu mang quan điểm là 64%. Chú ý rằng việc xác định câu mang quan điểm chỉ ra trước đó mang tính chủ quan. Phân tích kết quả của chúng tôi chỉ ra rằng mọi người thích miêu tả “câu chuyện” của họ về sản phẩm giống như: họ thường để cập đến hoàn cảnh mà họ sử dụng sản phẩm, đặc tính chi tiết của sản phẩm được sử dụn, và cũng như kết quả họ đưa ra. Trong khi người gán nhãn không quan tâm đên những câu quan điểm mà không chỉ ra rằng họ thích sản phẩm hay không, hệ thống của chúng tôi gán nhãn các câu là câu quan điểm nếu nó có chứa đặc tính và một vài tính từ mang quan điểm. Điều này làm giảm độ chính xác. Mặc dù các câu không thể chỉ ra mạnh mẽ quan điểm của người dùng hướng tới đặc tính của sản phẩm, chúng vẫn sẽ có lợi và hữu ích.

Hệ thống của chúng tôi có độ chính xác tốt trong dự đoán hướng của câu: độ chính xác trung bình của 5 sản phẩm là 84%. Điều này cho thấy rằng phương pháp sử dụng WordNet của chúng tôi để dự đoán hướng ngữ nghĩa của tính từ và hướng ngữ nghĩa của là có hiệu qua cao.

Tổng kết, chúng tôi có thể thấy rằng kỹ thuật của chúng tôi rất triển vọng, đặc biệt cho xác định hướng ngữ nghĩa của câu. Chúng tôi tin rằng chúng sẽ được sử dụng trong thực tế. Chúng tôi cũng chú ý 3 giới hạn quan trọng của hệ thống của chúng tôi là: (1) Chúng tôi không xử lý với quan điểm của câu mà cần xác định đại danh từ [40]. Ví dụ, “Nó trầm lặng nhưng mạnh mẽ”. Để hiểu “Nó” đại diện cái gì, xác định đại danh từ cần được thực hiện. Xác định đại danh từ là một vấn đề xử lý tốn kém và phức tạp trong xử lý ngôn ngữ tự nhiên (NLP). Chúng tôi lên kế hoạch gắn kết một số kỹ thuật đã có từ NLP để phù hợp với nhu cầu. (2) Chúng tối chỉ sử dụng tính từ để chỉ thị hướng quan điểm của câu. Tuy nhiên, động từ và danh từ cũng có thể được sử dụng với mục đích này, vd “Tôi thích sự cảm nhận của máy ảnh”, “Tôi khuyến kích sử dụng máy ảnh”. Chúng tôi lên kế hoạch giải quyết vấn đề này trong tương lai. (3) Nó cũng rất quan trong cần được nghiên cứu là mức độ của quan điểm. Một số quan điểm rất mạnh mẽ và một số quan điểm lại mờ nhạt. Nêu bật mức độ quan điểm (mức độ thích hay không thích) có thể rất hữu dụng cho cá nhân người mua sắm và các nhà sản xuất.

**5. KẾT LUẬN**

Trong bài báo này, chúng tôi để cập đến một danh sách các kỹ thuật để khai phá và tổng kết đánh giá của khách hàng dựa trên khai phá dữ liệu và xử lý ngôn ngữ tự nhiên. Đối tượng. Mục tiêu để cung cấp một bản tổng hợp dựa trên đặc tính của một số lượng lớn các đánh giá của khách hàng cho sản phẩm được bán online. Kết quả thực nghiệm của chúng tôi chỉ ra rằng công nghệ được đề xuất rất triển vọng trong thực hiện công việc này. Chúng tôi tin rằng vấn đề này sẽ ngày càng trở nên quan trọng với nhiều hơn nữa những người đang mua hàng và bày tỏ quan điểm của họ trên Web. Bản tổng kết các đánh giá không chỉ hữu ích cho đa số nhưng người mua hàng mà còn quan trọng đối với nhà sản xuất.

Trong công việc tương lai của chúng tôi. Chúng tôi kế hoạch cải thiện hơn nữa và tinh chỉnh kỹ thuật, và giải quyết các vấn để nội bật đã được xác định như: xác định đại danh từ, xác định mức độ của quan điểm, và nghiên cứu quan điểm được bày tỏ với trạng từ, động từ, danh từ. Cuối cùng, chúng tôi cũng sẽ tiếp tục xem xét các đánh giá của khách hàng. Chúng tôi tin rằng việc theo dõi sẽ thật sự hữu ích cho nhà sản xuất bời vì họ muốn biết bất kì nhận xét tích cực hay tiêu cực mới của khách hàng trên sản phẩm của họ bất cứ khi nào có thể. Từ khóa ở đây là “Mới”. Mặc dù một đánh giá mới có thể được thêm, nó cũng có thể không chứa bất kỳ thông tin mới nào.