Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00070

MỘT PHƯƠNG PHÁP PHÂN TÍCH QUAN ĐIỂM ĐÁNH GIÁ CỦA NGƯỜI DÙNG ĐỐI VỚI CHẤT LƯỢNG SẢN PHẨM DỰA TRÊN CÁC NHẬN XÉT CÁ NHÂN

Nguyễn Thị Ngọc Tú[1], Nguyễn Đức Long[1], Nguyễn Khắc Giáo[2], Nguyễn Thị Thu Hà [1], Nguyễn Việt Anh[2]

1 Khoa CNTT, Đại học Điện lực, 235 Hoàng Quốc Việt, Từ Liêm, Hà Nội,

[2]Viện CNTT, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

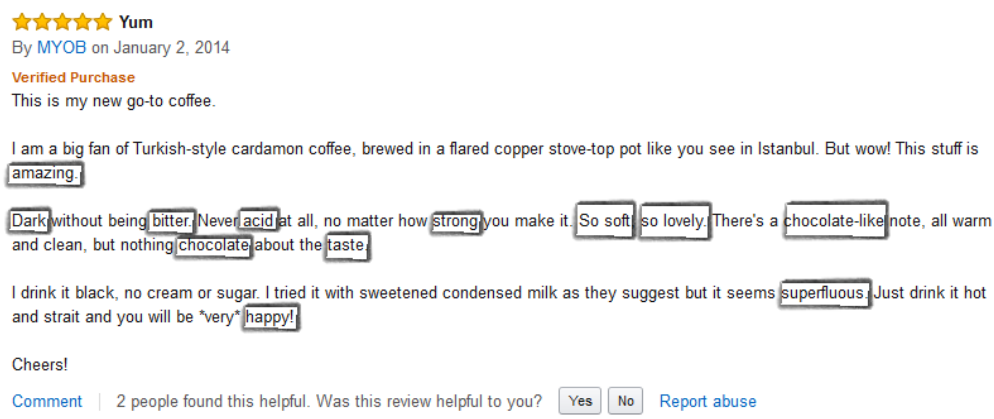
{ hantt, tuntn, }@epu.edu.vn, {giaonk@ioit.ac.vn, nva.nguyen@gmail.com}

TÓM TẮT: *Trong bài báo này chúng tôi trình bày một phương pháp phân tích quan điểm người dùng dựa trên các nhận xét cá nhân. Chúng tôi tập trung vào giải quyết ba nhiệm vụ của bài toán phân tích quan điểm: Nhận dạng và trích rút nội dung theo từng khía cạnh; Khám phá việc người dùng xếp hạng trên từng khía cạnh đối với sản phẩm; Dự đoán trọng số xếp hạng của các khía cạnh trong mỗi nhận xét. Đối với nhiệm vụ đầu tiên, chúng tôi sử dụng từ chủ đề bao gồm danh từ và cụm danh từ để trích rút các khía cạnh được người dùng đề cập đến trong bài viết của họ. Phương pháp của chúng tôi được thực hiện dựa trên thuật toán bootstrap kết hợp mô hình từ chủ đề dựa trên xác suất có điều kiện. Nhiệm vụ thứ hai và thứ ba được chúng tôi giải quyết dựa trên học có giám sát theo Naïve Bayes. Kết quả thực nghiệm trên ba bộ dữ liệu cà phê, bia, khách sạn cho thấy độ chính xác của phương pháp đề xuất là khá tốt cho cả bài toán trích rút khía cạnh cũng như cho bài toán dự đoán xếp hạng khía cạnh.*

Từ khóa: *Khai phá quan điểm, phân tích ngữ nghĩa, khía cạnh, nhận xét của người dùng, xếp hạng khía cạnh, tập từ chủ đề, học giám sát,…*

I. GIỚI THIỆU

Internet đã mang tới sự bùng nổ thông tin trong xã hội hiện đại, các hoạt động kinh doanh, thương mại dựa trên nền tảng Internet ngày càng phổ biến. Khách hàng có thể tự do bày tỏ quan điểm, đánh giá hay nhận xét về một mặt hàng, sản phẩm, dịch vụ trên các trang thương mại điện tử, các trang mạng xã hội... Việc khai phá những dòng nhận xét cá nhân này sẽ hữu ích cho những khách hàng khác khi tìm mua một loại sản phẩm hoặc có tính tích cực trong việc giúp cho nhà sản xuất nhận biết được nhu cầu của khách hàng để cải tiến sản phẩm dịch vụ tốt hơn. Ngày nay, đối với mỗi sản phẩm, nhiều trang web đã tổng hợp các đánh giá tổng thể của người dùng và hiển thị nó trên trang. Tuy nhiên, điều này là chưa đầy đủ vì nó không thể cung cấp một cách chi tiết những đánh giá của người dùng trên từng khía cạnh của sản phẩm. Ví dụ, ta hãy xem xét một bình luận về sản phẩm cà phê Trung Nguyên trên trang Amazon (Hình 1).


★★★★★ Yum
By MYOB on January 2, 2014
Verified Purchase
This is my new go-to coffee.

I am a big fan of Turkish-style cardamon coffee, brewed in a flared copper stove-top pot like you see in Istanbul. But wow! This stuff is amazing.

Dark without being bitter. Never acid at all, no matter how strong you make it. So soft, so lovely. There's a chocolate-like note, all warm and clean, but nothing chocolate about the taste.

I drink it black, no cream or sugar. I tried it with sweetened condensed milk as they suggest but it seems superfluous. Just drink it hot and strait and you will be *very* happy!

Cheers!

Comment | 2 people found this helpful. Was this review helpful to you? Yes No Report abuse

Hình 1. Nhận xét sản phẩm cà phê trên trang Amazon.com

Bình luận này đề cập đến nhiều khía cạnh của chất lượng cà phê như độ đậm (body), hương vị (taste), mùi thơm (aroma) và tính axít/vị chua (acidity). Nhưng xem xét kỹ, bài nhận xét này chỉ đưa ra một đánh giá tổng thể là cấp độ 5/5 mà không chỉ ra cụ thể đã đề cập những khía cạnh nào của sản phẩm và chúng được đánh giá như thế nào, cũng như không chỉ ra mức độ coi trọng của người dùng đối với từng khía cạnh đó. Bài toán của chúng tôi hướng tới việc phát hiện, trích rút ra các câu thảo luận cho từng khía cạnh được đánh giá trong bài viết, đồng thời dự đoán xếp hạng và mức độ coi trọng của người dùng đối với từng khía cạnh đó. Đây là một dạng của bài toán phân tích quan điểm, chúng sẽ giúp người dùng lĩnh hội các đánh giá cá nhân một cách hiệu quả và có được cái nhìn sâu sắc hơn về chất lượng sản phẩm. Trong bài báo này, chúng tôi đề xuất ra 02 giai đoạn của quá trình phân tích như sau:

- Giai đoạn thứ nhất, trích rút tự động các khía cạnh trong một bài nhận xét cá nhân. Mỗi khía cạnh có thể được xem là một chủ đề, số các chủ đề là xác định từ trước kèm theo các từ lõi cho mỗi chủ đề.

- Giai đoạn thứ hai, dự đoán xếp hạng và trọng số trên từng khía cạnh của sản phẩm.

Nhiệm vụ trích rút các khía cạnh của sản phẩm trong bài nhận xét được chúng tôi thực hiện thông qua tập các từ chủ đề và để dự đoán xếp hạng khía cạnh chúng tôi thực hiện trích rút từ ngữ nghĩa kết hợp sử dụng Naïve Bayes.

Phần còn lại của bài báo này được cấu trúc như sau. Phần II giới thiệu các nghiên cứu liên quan. Phần III trình bày phương pháp đề xuất. Phần IV đưa ra các đánh giá thực nghiệm của phương pháp dựa trên phần mềm đã được cài đặt. Cuối cùng là kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Các bài nhận xét trên các trang online cùng với xếp hạng đã được chứng minh là nguồn dữ liệu có giá trị trong nhiều hệ thống ứng dụng như khuyến nghị sản phẩm [1], khai thác đặc trưng [13], phân tích ngữ nghĩa [18]. Do tính đa dạng của các ứng dụng khai thác dữ liệu này đã dẫn đến sự cần thiết của các phương pháp có hiệu quả tốt hơn [5], [7], [12]. Một trong những nhánh quan trọng của khai phá quan điểm là bài toán trích rút khía cạnh và dự đoán xếp hạng khía cạnh [14].

Một số cách tiếp cận phổ biến trong bài toán trích rút khía cạnh tiềm ẩn và phân lớp quan điểm bao gồm tiếp cận theo mô hình chủ đề (topic model), tiếp cận dựa trên thuật ngữ và tần suất (terms and their frequency), tiếp cận dựa trên từ và cụm từ ngữ nghĩa (sentiment words and phrases).

Phương pháp trong [2], đầu tiên các tác giả sử dụng mô hình chủ đề để khám phá ra các khía cạnh. Sau đó với mỗi khía cạnh được trích rút ra tất cả các tính từ có liên quan và xây dựng một đồ thị kết nối. Thuật toán lan truyền nhãn [21] được sử dụng trên đồ thị để học điểm phân cực ngữ nghĩa của các tính từ. Sauper and Barzi-lay [16] cũng tiếp cận dựa trên mô hình chủ đề nhưng sử dụng từ lõi (seed word) để xác định các trật tự bất đối xứng. Phương pháp này có hiệu quả cao trong phân loại chủ đề ngữ nghĩa tích cực/tiêu cực, nhưng với phân lớp ngữ nghĩa nhiều mức thì nó chưa thực sự hiệu quả.

Phương pháp tiếp cận dựa trên tần suất từ cũng là một trong những phương pháp được dùng phổ biến và hiệu quả cao [9]. Theo phương pháp này các khía cạnh được diễn tả thông qua tần suất xuất hiện của các danh từ và cụm danh từ. Điển hình là [7], [13], [11], [21], [10]. Hu and Liu [7] đã sử dụng một thuật toán khai phá dữ liệu. Danh từ và cụm danh từ được xác định bằng gán nhãn từ loại. Tần suất xuất hiện của chúng được đếm và chỉ những từ có tần suất cao được giữ lại. Việc xác định ngưỡng tần suất được thực hiện nhờ thực nghiệm. Mặc dù, phương pháp này rất đơn giản nhưng nó thực sự khá hiệu quả [9]. Một số các công ty thương mại hiện nay đang sử dụng phương pháp này với một vài cải tiến. Một phương pháp khác theo hướng này là nghiên cứu của Moghaddam và Ester [11], trong đó nhóm tác giả đã sử dụng tần số nghịch đảo từ (TF-IDF) kết hợp với việc thêm vào một bộ lọc dựa trên các mẫu (pattern-based filter) để loại bỏ một vài thuật ngữ phi khía cạnh (non-aspect terms). Long, Zhang and Zhu [10] trích rút các khía cạnh dựa trên tần suất và khoảng cách thông tin. Đầu tiên, các từ lõi của khía cạnh được phát hiện sử dụng phương pháp dựa trên tần suất. Sau đó, các từ liên quan khác đối với khía cạnh được tìm ra dựa trên khoảng cách thông tin [3]. Trong bài báo này chúng tôi cũng đi theo hướng tiếp cận này, tuy nhiên chúng tôi không sử dụng độ đo khoảng cách thông tin mà sử dụng xác suất có điều kiện kết hợp với kỹ thuật bootstrap để tìm ra các thuật ngữ có liên quan đối với khía cạnh.

Trong phân lớp ngữ nghĩa, những từ ngữ nghĩa thường là yếu tố quyết định. Tuy nhiên rất khó để chúng ta hình dung ra các từ và cụm từ ngữ nghĩa nào có thể sử dụng trong phương pháp học bán giám sát. Phương pháp học trong [18] là một trong các phương pháp để giải quyết vấn đề này. Turney và các cộng sự đưa ra cách giải quyết vấn đề phân lớp ngữ nghĩa dựa trên việc trích rút từ và cụm từ theo một số mẫu cú pháp nhất định. Ngoài ra, để thực hiện học phân lớp nhóm tác giả còn đưa ra độ đo xác định khuynh hướng ngữ nghĩa (SO) của cụm từ dựa trên độ đo hỗ trợ thông tin liên quan (PMI). Một cách tiếp cận không giám sát khác là phương pháp dựa trên từ vựng (lexicon-based method) [24]. Phương pháp này sử dụng một từ điển của các từ và cụm từ ngữ nghĩa có kèm theo xu hướng và mức độ ngữ nghĩa liên quan. Đồng thời phương pháp cũng kết hợp các từ chỉ mức độ tăng cường hay giảm nhẹ và các từ phủ định để tính toán điểm ngữ nghĩa cho các văn bản [17]. Phương pháp này ban đầu đã được sử dụng trong phân lớp ngữ nghĩa mức câu và khía cạnh trong các công bố [4] [7] [8]. Trong phương pháp của chúng tôi để trích chọn đặc trưng chúng tôi áp dụng các mẫu cú pháp của Turney và sử dụng phương pháp Naïve Bayes để thực hiện học đa phân lớp.

III. PHƯƠNG PHÁP

A. Xác định vấn đề

Phần này cung cấp một số các khái niệm cơ bản được sử dụng trong bài báo.

Định nghĩa 1. Tập các nhận xét (comments)**:**

Tập các nhận xét được ký hiệu là $D=\{d_1, d_2, \ldots, d_N\}$ là một tập các bài viết nhận xét về một loại sản phẩm.

Định nghĩa 2. Khía cạnh (aspect)**:**

Khía cạnh là một tập các giá trị đặc tả cho một loại sản phẩm, ký hiệu là $A=\{a_1, a_2, \ldots, a_K\}$

Một sản phẩm có thể có nhiều thuộc tính khác nhau, ví dụ sản phẩm cà phê có thể là hương thơm (aroma), mùi vị (taste), hay độ đậm (body). Mỗi thuộc tính này khi người dùng nhận xét về chúng có thể coi là một khía cạnh được đề cập trong bài viết. Chúng tôi biểu diễn a_k ={t|t Є V, a(t) = k} trong đó a(.) là một hàm ánh xạ từ một từ tới một nhãn khía cạnh.

Định nghĩa 3. Xếp hạng khía cạnh (aspect ranking)**:**

Xếp hạng khía cạnh là sự sắp xếp trật tự theo một tiêu chí nào đó tương ứng với mỗi nhận xét và được mô tả bởi R_i=[r_{i1} r_{i2} … r_{iK}]. Một xếp hạng khía cạnh R_i là một vector có K chiều. Trong đó, chiều thứ *k* là một giá trị biểu thị mức độ đánh giá (sự hài lòng) của khách hàng thể hiện trong nhận xét d_i đối với khía cạnh a_k, r_{ik} ∈[r_{min}, r_{max}].

Định nghĩa 4. Trọng số khía cạnh (aspect weighting)**:**

Trọng số khía cạnh biểu thị mức độ quan tâm của người dùng đối với các khía cạnh sản phẩm trong nhận xét d_i biểu diễn bởi Wt_i=[wt_{i1} wt_{i2} … wt_{iK}]. Trọng số khía cạnh Wt_i là một vector K chiều, trong đó chiều thứ *k* biểu diễn mức độ quan trọng mà khách hàng đưa ra cho khía cạnh a_k trong đánh giá của mình. Để dễ dàng tính toán chúng tôi xác định wt_{ik} ∈[0,1] và một điều kiện ràng buộc cho mỗi nhận xét d_i là $\sum_{k=1}^{K} wt_{ik} = 1$.

Phân tích đánh giá quan điểm người dùng

Đối với mỗi sản phẩm chúng ta thường có một tập các nhận xét D, mà trong đó mỗi một nhận xét d_i lại đi kèm với một xếp hạng R_i . Bài toán phân tích đánh giá quan điểm người dùng chính là tìm ra trong các nhận xét đó người dùng đã đề cập tới khía cạnh A_k nào của sản phẩm, sau mỗi khía cạnh đó thì họ đánh giá R_i như thế nào và họ thường chú trọng tới những thuộc tính/khía cạnh nào của sản phẩm (xác định Wt_i). Xem xét lại ví dụ 1, trong bài viết đề cập đến các khía cạnh như mùi vị, và tính axít. Chúng ta cần trích rút các câu "Dark without being bitter" và "There's a chocolate-like note, all warm and clean, but nothing chocolate about taste" thuộc về khía cạnh mùi vị, hoặc câu "Never acid at all, no matter how strong you make it" thuộc về khía cạnh tính axít. Mặt khác, chúng ta chỉ biết người dùng xếp hạng một thang điểm tổng cho tất cả những khía cạnh này là 5, vì vậy chúng ta cần ước lượng mức độ xếp hạng của người dùng cho từng khía cạnh là như thế nào? Và mức độ coi trọng của họ đối với hai khía cạnh này ra sao?

B. Trích rút khía cạnh

Mục đích của bước đầu tiên này là tìm ra một tập con các câu tương ứng với từng khía cạnh trong tập các nhận xét. Chúng tôi xác định tập các khía cạnh cho trước tương ứng với mỗi loại sản phẩm. Giả sử rằng, trong sản phẩm cà phê gồm 4 khía cạnh là hương thơm (aroma), mùi vị (taste), độ đậm (body), tính axit (acidity). Việc xác định khía cạnh dựa trên các thuộc tính nổi bật của sản phẩm và thông thường chúng đã được thống kê từ người dùng hay các chuyên gia cùng lĩnh vực. Ban đầu chúng tôi giả định rằng chỉ có một vài từ đặc biệt được dùng để mô tả cho mỗi khía cạnh, các từ này chúng tôi gọi là từ lõi (core term) khía cạnh. Sau đó, chúng tôi sử dụng xác suất có điều kiện trong mô hình từ chủ đề kết hợp thuật toán bootstrap để tìm ra tập các từ chủ đề liên quan đến mỗi khía cạnh. Các từ này chúng tôi gọi là từ chủ đề (topic word). Các bước thực hiện trích rút khía cạnh như sau:

Bước tiền xử lý: Tất cả các văn bản nhận xét được coi như đầu vào bài toán. Các văn bản này được tách thành các câu.

Bước 1: Gán nhãn khía cạnh a_k cho câu nếu trong câu đó có chứa từ lõi thuộc khía cạnh a_k. Nếu có nhiều mối ràng buộc thì gán nhãn câu với nhiều khía cạnh.

Bước 2: Cập nhật lại các từ lõi khía cạnh và xác định tập các từ chủ đề dựa trên tập các câu đã được gán nhãn.

Bước 3: Gán lại nhãn khía cạnh cho câu, nếu trong câu có chứa từ lõi thuộc khía cạnh a_k thì gán nhãn khía cạnh a_k cho câu. Ngược lại (trong câu không chứa bất cứ một từ lõi thuộc bất cứ khía cạnh nào), đếm số từ chủ đề tương ứng với mỗi khía cạnh a_k có trong câu và ghi vào Count(k). Gán nhãn cho câu một khía cạnh nào đó bởi a_k=argmax Count(k). Nếu có nhiều mối ràng buộc thì gán nhãn câu với nhiều khía cạnh.

Bước 4: Nếu danh sách các từ lõi khía cạnh là không thay đổi hoặc số lần lặp vượt L thì chuyển sang bước 5, ngược lại chuyển qua bước 2.

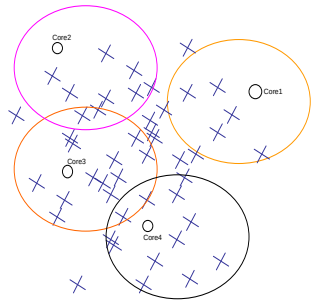
Bước 5: Đầu ra là tập các câu chú thích với việc gán nhãn khía cạnh.

Để mở rộng tập từ lõi, chúng tôi sử dụng mô hình chủ đề dựa trên xác suất có điều kiện.

1. Tập từ chủ đề

Mỗi khía cạnh của sản phẩm mà người dùng đề cập đến, chúng ta có thể coi như một chủ đề. Các chủ đề này được biểu diễn bởi một tập các từ, khi các từ này xuất hiện sẽ gợi ý cho người đọc liên tưởng đến chủ đề đó. Trong các phương pháp truyền thống, tập các từ này thường được xây dựng dựa trên Bayes hay mô hình Markov ẩn. Trong

phương pháp của chúng tôi, tập các từ chủ đề này được xây dựng dựa trên mô hình xác suất có điều kiện [6] từ tập dữ liệu huấn luyện được gán nhãn tự động theo thuật toán bootstrap. Hình 2 dưới đây mô tả một số khía cạnh trong tập không gian K khía cạnh. Trong đó các kí hiệu o biểu diễn cho từ lõi (core term), còn các kí hiệu x biểu diễn cho các từ khóa xuất hiện trong không gian K khía cạnh.



Hình 2. Mô hình tập từ chủ đề dựa trên xác suất có điều kiện

Giả sử A={a_1, a_2,..., a_K} là một không gian K chiều các khía cạnh. Mỗi không gian a_k bao gồm tập các từ thuộc a_k nếu như khả năng xuất hiện của nó trong a_k là khác 0 (hoặc một ngưỡng B). Các không gian a_k và $a_{k'}$ có thể giao nhau, do vậy, các từ thuộc a_k có thể cũng thuộc một không gian $a_{k'}$ khác.

Giả sử ta lấy một từ gọi là từ lõi của không gian a_k (từ này được coi là từ có trọng số rất cao), khoảng cách của các từ còn lại trong không gian a_k chỉ cần so với từ lõi. Để tính được khoảng cách của các từ đó so với từ lõi, chúng tôi sử dụng cách tính xác suất có điều kiện.

Update(C_k, TW_k) - THUẬT TOÁN TÌM TẬP TỪ CHỦ ĐỀ KHÍA CẠNH

Đầu vào:
- S_k: Tập các câu đã được gán nhãn tương ứng với các khía cạnh a_k
- V: Tập từ điển
- C_k : Tập từ lõi thuộc khía cạnh a_k

Đầu ra:
- TW_k : Tập các từ chủ đề được gán nhãn tương ứng với khía cạnh a_k.
- C_k: Tập từ lõi được cập nhật thuộc khía cạnh a_k.

Khởi tạo:
$TW_k = \emptyset$; // tập từ chủ đề tương ứng với chủ đề thứ k
n=0; N=0;
1. For each $w_i \in S_k$ do
1.1 if $w_i \in V$ then n(i)⟵ n(i) +1; // Nếu từ w_i thuộc tập từ V đếm số lần xuất hiện w_i trong khía cạnh a_k
2. N=argmax(n(i)); // Lấy tần suất lớn nhất của từ w_i trong khía cạnh a_k
3. For each $w_i \in S_k$ do
3.1 if $w_i \in V$ & n(i)==N & $w_i \notin C_k$ then $C_k \longleftarrow w_i$ // Cập nhật từ lõi với từ có trọng số cao nhất
3.2 if $w_i \in V$ & $w_i \notin C_k$ & Pr(w_i |C_k) >=B then $TW_k \longleftarrow w_i$; // cho các từ w_i vào tập TW_k của a_k

2. Thuật toán trích rút khía cạnh

THUẬT TOÁN PHÂN ĐOẠN KHÍA CẠNH

Đầu vào:
- D: D={d_1, d_2, ..., d_N} Tập văn bản nhận xét về một sản phẩm;
- A: A={a_1, a_2, ..., a_K} Tập các khía cạnh;
- C: C ={C_1, C_2, ..., C_K} Tập các từ lõi của các khía cạnh tương ứng
- V: tập từ điển
- chọn số bước lặp L
- chọn ngưỡng B

Đầu ra:
- S_k Tập các câu được gán nhãn tương ứng với mỗi a_k (k=1, K)

Khởi tạo:
S=∅; S_k=∅;// k=1-K
loop=1; // Số lần lặp nếu vượt quá L thì dừng
1. For each $d_i \in D$
1.1 S ⟵ segment(d_i); // Tách các văn bản thành các câu
2. For each $s_i \in S$
2.1. For k=1 to K do
2.1.1. For j=1 to legth(s_i) do

2.1.1.1 if $w_{ij} \in C_k$ then { $s_i \leftarrow$label(a_k); $S_k \leftarrow s_i$;}
3. For k=1 to K do
3.1 Update(C_k, TW_k);
4.1. loop++;
4.2. For each $s_i \in S$
4.2.1 For k=1 to K do
4.2.1.1. { m(k)=0;
4.2.1.2. For j=1 to legth(s_i) do
4.2.1.2.1. {if $w_{ij} \in C_k$ then { $s_i \leftarrow$label(a_k); $S_k \leftarrow s_i$;}
else {if $w_{ij} \in TW_k$ then m(k) $\leftarrow$m(k)+1; }}
4.2.1.3. if m(k)=argmax(m(k)) then {$s_i \leftarrow$ label(a_k); $S_k \leftarrow s_i$; }
5. If not Update(C_k, TW_k) || loop>=L then Break;
Else go to 3.

Sau khi phân đoạn khía cạnh, chúng tôi sẽ có được K phần của tập nhận xét D đã được tách thành các câu. Bước tiếp theo, chúng tôi thực hiện trích rút đặc trưng từ tập các câu đã được gán nhãn khía cạnh này. Thông thường, trong các bài nhận xét về một sản phẩm, người dùng đánh giá chúng thông qua các từ biểu thị cảm xúc như "good", "very good", "terrible", "clean", "great", "nice", "nothing special".

C. Mô hình phân tích quan điểm dựa trên xác suất Naïve Bayes

Trước tiên chúng ta hãy xem xét hành vi đánh giá của một khách hàng đối với một sản phẩm P. Để có được một đánh giá chung cho sản phẩm P, thông thường trước khi viết nhận xét khách hàng này sẽ nghĩ đến những thuộc tính mà anh/cô ấy cho là nổi trội của sản phẩm. Sau đó họ sẽ xác định mức độ quan trọng (thực chất là mức độ quan tâm của cá nhân khách hàng) của từng thuộc tính sản phẩm và sắp xếp chúng theo một thứ tự ưu tiên nào đó trong bài viết của chính họ. Bằng việc xác định như vậy người bình luận (khách hàng) bắt đầu lựa chọn các từ khóa có thể biểu thị ý kiến cá nhân của anh/cô ấy trong từng nhận định trên từng thuộc tính của sản phẩm. Đến cuối cùng người bình luận sẽ đưa ra một mức độ đánh giá tổng thể Y dành cho sản phẩm. Để đưa ra được chỉ số này người bình luận sẽ dựa vào từng đánh giá riêng lẻ của từng thuộc tính với trọng số tương ứng được xác định theo công thức [19]:

$$Y = \sum_{k=1}^{K} r_{ik} wt_{ik} \quad (1)$$

Đến đây để làm rõ vấn đề xác định các tham số r_{ik} và w_{ik} của bài toán, chúng tôi thực hiện trích rút đặc trưng là các từ và cụm từ ngữ nghĩa dựa trên các mẫu cú pháp được đưa ra trong [18]. Các mẫu cú pháp trích rút dựa trên gán nhãn POS được chỉ ra trong bảng 1.

Bảng 1. Các mẫu của gán nhãn POS cho trích rút cụm hai từ [18]

Từ đầu tiên	Từ thứ hai	Từ thứ ba (không trích rút)
JJ	NN or NNS	Bất cứ từ nào
RB, RBR, or RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN nor NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, or VBG	Bất cứ từ nào

Sau khi phân đoạn các khía cạnh A_k đối với mỗi sản phẩm P (mục 3.2) thu được một tập các câu đã được gán nhãn. Các câu này lại được kết nối với các bài nhận xét cụ thể. Mỗi bài nhận xét trong kho dữ liệu đã được xác định các điểm xếp hạng của từng khía cạnh. Sau các bước trên, xác định được xác suất có điều kiện của các đặc trưng (cụm từ ngữ nghĩa được trích rút) trong tập dữ liệu huấn luyện đã được gán nhãn. Chúng tôi dự đoán điểm xếp hạng cho từng khía cạnh theo lý thuyết Naïve Bayes. Xem xét một nhận xét d trong khía cạnh A_k với q đặc trưng (q cụm từ ngữ nghĩa thuộc khía cạnh A_k) thì xác suất để r_{ik} thuộc tập C_R (R=1-5) hay không theo công thức sau:

$$P(r_{ik} \in C_R | F_1, F_2, \dots, F_q) = \frac{P(F_1, F_2, \dots, F_q | r_{ik} \in C_R) x P(r_{ik} \in C_R)}{P(F_1, F_2, \dots, F_q)} \quad (2)$$

Giả thuyết rằng các đặc trưng là độc lập công thức 2 chuyển thành:

$$P(r_{ik} \in C_R | F_1, F_2, \dots, F_q) = \frac{\prod_{j=1}^{q} P(F_j | r_{ik} \in C_R) x P(r_{ik} \in C_R)}{\sum_{j=1}^{q} P(F_j)} \quad (3)$$

Trong đó:

$P(r_{ik} \in C_R) = n_{Ak}(R)/n_{Ak}$ là số các câu có gán nhãn khía cạnh A_k được tính điểm R và tổng số các câu có gán nhãn khía cạnh A_k trong tập huấn luyện.

$P(F_j|r_{ik} \in C_R) = n_{Ak}(F_j, R)/n_{Ak}(R)$ là số lần xuất hiện của F_j trong các câu được tính điểm R có gán nhãn A_k và số các câu có gán nhãn khía cạnh A_k được tính điểm R.

Để làm trơn công thức (3) bài báo sử dụng phương pháp Laplace, (3) chuyển thành:

$$P(F_j|r_{ik} \in C_R) = \frac{n_{Ak}(F_j,R)+1}{n_{Ak}(R)+|V|+1} \quad (4)$$

Với: |V| là tập tất cả các từ và cụm từ đặc trưng có trong khía cạnh A_k.

Sau khi xác định được điểm xếp hạng của mỗi khía cạnh trong sản phẩm, tiếp theo xác định trọng số của khía cạnh dựa trên tần suất xuất hiện của các từ đặc trưng thuộc các khía cạnh trong toàn bộ văn bản nhận xét. Trọng số của mỗi khía cạnh được tính theo công thức:

$$wt_{ik} = \frac{\sum_{j=1}^{q} w_{j|w_j \in A_k}}{\sum w} \quad (5)$$

Trong đó:

$\sum_{j=1}^{q} w_{j|w_j \in A_k}$ là tổng số lần xuất hiện của các từ đặc trưng có trong văn bản thuộc khía cạnh A_k.

$\sum w$ là tổng số lần xuất hiện của tất cả các từ đặc trưng (đối với tất cả các khía cạnh) có trong d.

IV. KẾT QUẢ THỰC NGHIỆM

A. Tập dữ liệu thử nghiệm

Trong thực nghiệm, chúng tôi sử dụng hai bộ dữ liệu đã được công bố là các bài viết trên trang TripAdvisor về khách sạn và trang BeerAdvocate về bia. Đối với hai bộ dữ liệu này đã có đánh giá xếp hạng của người dùng về các khía cạnh cụ thể của sản phẩm. Ngoài hai bộ dữ liệu trên, chúng tôi còn thực hiện thu thập một tập các bài nhận xét về cà phê (một lĩnh vực mà hiện nay chưa có dữ liệu) trên trang amazon.com. Sau khi thu thập dữ liệu, chúng tôi thực hiện bước làm sạch dữ liệu (loại bỏ một số lỗi chính tả, loại bỏ các ký hiệu dưa thừa không mang ý nghĩa trong phân tích dữ liệu) và thực hiện xếp hạng bằng tay cho các bài nhận xét trên các khía cạnh: hương thơm (aroma), tính axít (acidity), độ đậm (body), mùi vị (taste). Dữ liệu về cà phê của chúng tôi bao gồm 17 loại sản phẩm với hơn 1.200 bài nhận xét.

Trong dữ liệu về khách sạn là 7 khía cạnh (rooms, location, cleanliness, check in/front desk, service, business service), và trong sản phẩm bia là 5 khía cạnh (aroma, palate, taste, appearance, overall). Các so sánh dựa trên phương pháp đánh giá tại [22] và [23].

B. Đánh giá kết quả

Chúng tôi đánh giá phương pháp đề xuất dựa trên hai nhiệm vụ là phân đoạn khía cạnh và dự đoán xếp hạng khía cạnh. Yêu cầu của phân đoạn khía cạnh là chúng tôi phải dự đoán nhãn khía cạnh cho mỗi câu trong mỗi bài nhận xét của kho dữ liệu. Trong khi đó, xếp hạng khía cạnh lại xét trên từng bài nhận xét (không phải cho từng câu). Do vậy đối với xếp hạng khía cạnh chúng tôi thực hiện dự đoán chung trên tất cả các câu được gán cùng nhãn khía cạnh có trong bài nhận xét.

Chúng tôi chia bộ dữ liệu ngẫu nhiên thành tập dữ liệu huấn luyện và tập test.

Phân đoạn khía cạnh:

Trong bảng 2, 3, 4 chỉ ra các từ lõi khía cạnh và từ chủ đề khía cạnh được trích rút theo từng khía cạnh.

Bảng 2. Tập các từ lõi và từ chủ đề học từ tập dữ liệu về khách sạn

Khía cạnh	core term	Topic word
Value	value, price, worth	Hotel, charge, cost, discount, dollars
Room	room, rooms	Bathroom, bathrooms, bed, beds, bath, floor, floors, chair, chairs, balcony, shower, lobby, noise, pool, queen, couple, sheraton, coffee, desk, hotel, suite, tv, view, water, window, carpet, closet, doors, furniture, king, pillows, sink, toilet, tub, toiletries,
Location	location	Airport, area, center, downtown, hotel, market, place, places, restaurant, shop, shops, shopping, showreview, street, view, views , neighborhood, square, waterfront,
Cleanliness	dirty, smelled, clean	Hotel, floor, shelf, desk, chair, bag, door, lobby, stairs,
Check in/ front desk	Staff	Desk, clerk, lounge, luggage, reception, checkout
Service	service, breakfast, food	Bar, bars, coffee, concierge, food, park, parking, restaurant, wine, buffet
Business service	internet, wifi	Tv, television,

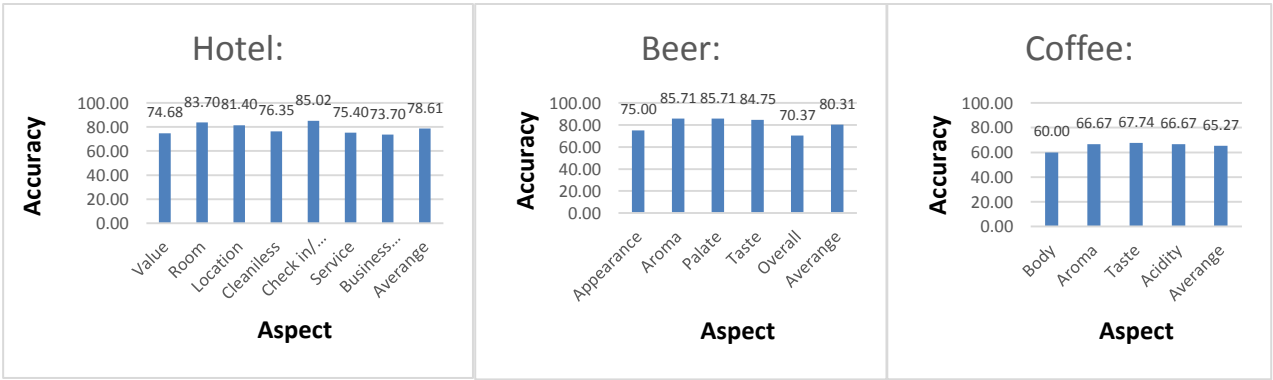
Bảng 3. Tập các từ lõi và từ chủ đề học từ tập dữ liệu về bia

Khía cạnh	Core term	Topic word
Appearance	Appearance, color, colors, coloring, head, foam	Black, Body, brown, bubble, bud, copper, lace, lacing, dots, drip, dust, back, finger, fizzy, fluff, golden, half, layer, orange, straw, surface, top, white, yellow…
Aroma	Aroma, aromas, smell, smelling	Bacon, banana, basil, caramel, cheese, cream, dry citrus, fruitiness, honey, light, malt, malts, meat, mint, nose, pear, perfume, pill, pine, roast, sandlewood, smoke, smoky, spice, sweet, sweetness, yeast…
Palate	Palate, mouth, feel, mouthfeel	Alcohol, body, carbonation, cream, Drinkability, dry, hoppy, light, round, spring, summer, …
Taste	Taste, tastes, aftertasste, in the end, finish, finishing	Alcohol, avalanche, balance, bitterness, body, bread, burn, caramel, carbonation, cheese, chocolate, clear, cocoa, coffee, complexity, flavors, flavors, fruit, fruitiness, ginger, grains, malt, maltiness, meat, Medium-dry, oats, pear, roast, smoke, smoothness, spring, subtleties, summer, sweet, throat, toffee, tongue, vanilla, wood,...
Overall	Overall	Beer, beers, bottle, drink, beverage, style,…

Bảng 4. Tập các từ lõi và từ chủ đề học từ tập dữ liệu về cà phê

Khía cạnh	Core term	Topic word
Aroma	Aroma, aromas, smell, smelling, flavor, flavors,	Bran, brew, butter, charr, chocolate, citrus, fruit, honey, love, lover, organic, press, quality, smooth, lemon, smoke, stuff,
Taste	Taste, tastes, aftertasste, finish, finishing, mouthfeel	Bitter, bitterness, chocolate, honey, salt, freshness, brew, love, lover, mild, organic, press, quality, roaster, smooth, soft, sour, stuff, sweet, sweetness, syrup,
Acidity	Acid, acidity	
Body	Body, aged, vintage	Love, press, smooth, richness, thick, thin, soft,

Hình 3 thể hiện độ chính xác của nhiệm vụ trích rút khía cạnh trong phương pháp của chúng tôi trên ba tập dữ liệu.

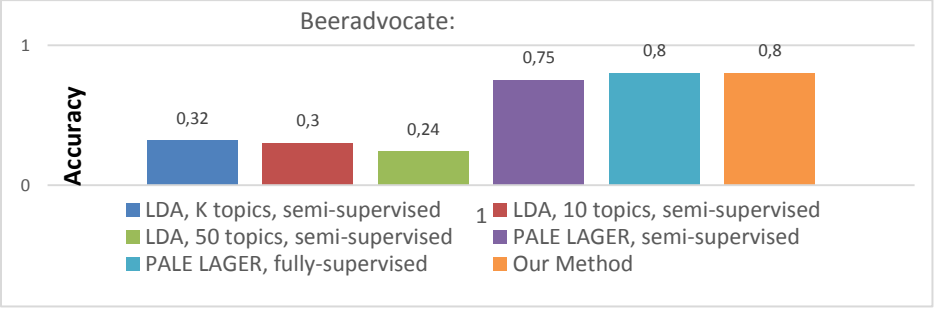


Hình 3. Độ chính xác trích rút khía cạnh

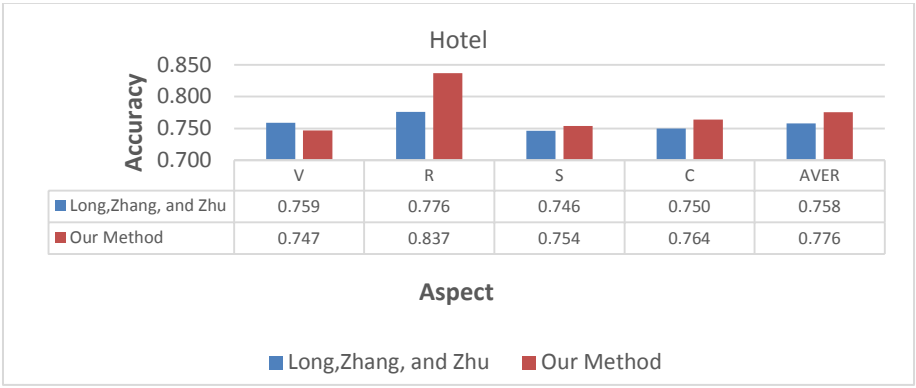
Phương pháp học trong nhiệm vụ trích rút khía cạnh sử dụng học bán giám sát. Kết quả thực nghiệm trong hình 3 cho thấy phương pháp của chúng tôi đề xuất là hiệu quả. Tuy nhiên đối với bộ dữ liệu cà phê kết quả trung bình thấp hơn so với kết quả thực hiện trên hai bộ khách sạn và bia bởi một số nguyên nhân sau:

- Dữ liệu cà phê mới chỉ được tiền xử lý đơn giản (raw preprocessing) như loại bỏ các kí hiệu dư thừa và một số lỗi chấm câu. Những phần lỗi chính tả trong văn bản chưa được xử lý.
- Các bài nhận xét của người dùng đa phần là các nhận xét rất ngắn, thường chỉ gồm hai tới ba câu, thậm chí chỉ có một câu. Với dữ liệu như vậy rất khó có được độ chính xác cao trong nhiệm vụ phân đoạn khía cạnh.
- Vấn đề thứ ba trong tập dữ liệu mà chúng tôi thu thập là có nhiều bài viết có xu hướng chú trọng đánh giá tổng thể mà không thảo luận khía cạnh (một phần lý do vì bài viết của họ quá ngắn).

Để đánh giá đầy đủ hơn hiệu xuất của phương pháp mà chúng tôi đề xuất, chúng tôi tiến hành thử nghiệm và so sánh kết quả với hai phương pháp cơ sở là LDA [22] và PALE LAGER [23] trên bộ dữ liệu bia (beeradvocate), và so sánh với kết quả trong [10] trên bộ dữ liệu khách sạn (TripAdvisor). Kết quả cho thấy hiệu suất phương pháp của chúng tôi cao hơn của LDA và PALE LAGER trong học bán giám sát và tương đương với PALE LAGER trong học giám sát. So với cách tiếp cận của Long, Zhang and Zhu phương pháp đề xuất cũng cho thấy một kết quả khả quan hơn.



Hình 4. Độ chính xác trích rút khía cạnh so sánh với LDA, PALE LAGER



Hình 5. Độ chính xác trích rút khía cạnh so với phương pháp của Long, Zhang and Zhu

Dự đoán xếp hạng khía cạnh:

Trong các tập dữ liệu mà chúng tôi xem xét, chỉ có xếp hạng tổng thể là bắt buộc. Trong khi đó các xếp hạng khía cạnh là tùy chọn. Chính vì vậy để tiến hành quá trình học phân lớp chúng tôi bắt buộc phải khôi phục các xếp hạng khía cạnh thiếu sót. Để đo lường được nhiệm vụ này chúng tôi chia dữ liệu làm hai phần. Một phần chúng tôi sử dụng trong quá trình huấn luyện và phần còn lại chúng tôi tiến hành dự đoán các xếp hạng khía cạnh. Đương nhiên chúng tôi đảm bảo rằng không có bất cứ một văn bản dự đoán xếp hạng khía cạnh nào thuộc trong lớp huấn luyện kể cả trong quá trình học trích rút khía cạnh.

Độ đo MSE (Mean squared erro) được sử dụng trong đánh giá kết quả dự đoán xếp hạng khía cạnh và xếp hạng tổng thể. Kết quả dự đoán được thể hiện trong bảng 5 và 6:

Bảng 5. Sai số dự đoán xếp hạng khía cạnh

Hotel		**Beer**		**Coffee**	
Khía cạnh	**MSE**	**Khía cạnh**	**MSE**	**Khía cạnh**	**MSE**
Value	0,15	Appearance	0,11	Body	0,15
Room	0,08	Aroma	0,07	Aroma	0,11
Location	0,09	Palate	0,9	Taste	0,12
Cleaniless	0,14	Taste	0,6	Acidity	0,15
Check in/ front desk	0,07				
Service	0,10				
Bussiness service	0,09				
Average	**0,101**	**Average**	**0,083**	**Average**	**0,133**

Bảng 6. Sai số dự đoán xếp hạng tổng thể

Xếp hạng tổng thể	**MSE**
Hotel	0,1456
Beer	0,1423
Coffee	0,1904

V. KẾT LUẬN

Trong bài báo này chúng tôi đã nghiên cứu các hệ thống đánh giá của người dùng, trong đó người dùng cung cấp các xếp hạng cho nhiều khía cạnh của sản phẩm. Bằng cách học những từ mô tả (từ chủ đề, từ ngữ nghĩa) khía cạnh, phương pháp của chúng tôi có thể xác định được phần nào của một bài nhận xét tương ứng với mỗi xếp hạng khía cạnh. Bên cạnh đó, chúng tôi cũng xác định được mức độ quan tâm của người dùng trên các khía cạnh đó thể hiện

qua việc học trọng số của khía cạnh. Trọng số này được học dựa trên chính các từ mà người dùng sử dụng trong bài viết của họ. Phương pháp của chúng tôi cũng có thể tạo ra bộ từ điển ngữ nghĩa và có thể dễ dàng mở rộng với kho dữ liệu thực.

LỜI CẢM ƠN

Chúng tôi trân trọng gửi lời cảm ơn tới đề tài CS 17.12 "**Nghiên cứu một số phương pháp theo dõi và giám sát thương hiệu trên mạng xã hội tại Việt Nam**", Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã hỗ trợ một phần kinh phí cho nghiên cứu này. Chúng tôi cũng cảm ơn các chuyên gia phòng Khoa học Dữ liệu và Ứng dụng - Viện CNTT - Viện Hàm lâm Khoa học và Công nghệ Việt Nam đã đóng góp ý kiến giúp nghiên cứu đạt hiệu quả nhất có thể.

TÀI LIỆU THAM KHẢO

[1]. J. Bennett and S. Lanning. The Netflix prize. In KDD Cup and Workshop, 2007.

[2]. Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10, Stroudsburg, PA, USA. 2010 pages 804–812.

[3]. Cilibrasi, Rudi L. and Paul M. B. Vitanyi. The google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007. 19(3): p. 370-383.

[4]. Ding, Xiaowen, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008.

[5]. G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. Twelfth International Workshop on the Web and Databases (WebDB 2009). June 28, 2009.

[6]. Ha Nguyen Thi Thu, Tinh Dao Thanh, Thanh Nguyen Hai, Vinh Ho Ngoc. Building Vietnamese Topic Modeling Based on Core Terms and Applying in Text Classification. Proc. Of The Fifth IEEE International Conference on Communication Systems and Network Technologies, 2015, pp: 1284 -1288.

[7]. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, New York, NY, USA. ACM, 2004, pages 168–177.

[8]. Kim, Soo-Min and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2004). 2004.

[9]. Bing Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. May 2012.

[10. Long, Chong, Jie Zhang, and Xiaoyan Zhu. A review selection approach for accurate feature rating estimation. In Proceedings of Coling 2010: Poster Volume. 2010.

[11]. Moghaddam, Samaneh and Martin Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In Proceeding of the ACM conference on Information and knowledge management (CIKM-2010). 2010.

[12]. R. Ng and A. Pauls. Multi-document summarization of evaluative text. In ACL. 2006.

[13]. Popescu and O. Etzioni. Extracting product features and opinions from reviews. HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing Pages 339-346.

[14]. Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge - Based Systems, 89.2015, pp.14-46.

[15]. Santorini, Beatrice. Part-of-speech tagging guidelines for the Penn Treebank Project, University of Pennsylvania, School of Engineering and Applied Science, Dept. of Computer and Information Science. 1990.

[16]. Christina Sauper and Regina Barzilay. Auto-matic aggregation by joint modeling of aspects and values. J. Artif. Int. Res. January 2013, 46(1):89-127.

[17]. Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Computational Linguistics, 2011. 37(2): p. 267-307.

[18]. P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Pages 417-424

[19]. Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, New York, NY, USA. ACM. 2010, pages 783-792.

[20]. Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propa-gation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University. 2002.

[21]. Zhu, Jingbo, Huizhen Wang, Benjamin K. Tsou, and Muhua Zhu. Multi-aspect opinion polling from textual reviews. in Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009). 2009.

[22]. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. JMLR. 2003.

[23]. Julian McAuley, Jure Leskovec, Dan Jurafsky. Learning Attitudes and Attributes from Multi-Aspect Review. International Conference on Data Mining (ICDM). 2012.

[24]. F. Wogenstein, J. Drescher, D. Reinel, S. Rill, J. Scheidt. Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM). 2013.

A METHOD OF OPINION MINING BY CONSUMERS FOR PRODUCT QUALITY BASED ON REVIEWS

Nguyen Thi Ngoc Tu, Nguyen Duc Long, Nguyen Khac Giao, Nguyen Thi Thu Ha, Nguyen Viet Anh

ABSTRACT: *In this paper we present a method of sentiment analysis based on customer's review. We focus on solving the three tasks of sentiment analysis problem: identifying and extracting aspect; Discover user ratings on each aspect that is referred to; Predict the rating weight of the aspects in each review. For the first task, we use topic words to include nouns and noun phrases to extract aspects that are mentioned by the user in their review. Our method is based on a combination of conditional probability and bootstrap algorithm. The second and third tasks we solve based on supervised learning with Naïve Bayes. The experimental results on three data sets (coffee, beer and hotel) show that the accuracy of the proposed method is good for both the aspect extraction and the aspect ranking prediction task.*