XÂY DỰNG TẬP NHÃN TỪ SO SÁNH ĐỂ PHÂN TÍCH CẢM XÚC NGƯỜI DÙNG TỪ NHỮNG BÌNH LUẬN TIẾNG VIỆT

CONSTRUCTION SET LABELS OF COMPARISON SENTENCE TO SENTIMENT ANALYSIS OF THE USER FROM VIETNAMESE COMMENTS

***LÝ THỊ HUYỀN CHÂU*(*)**

TÓM TẮT: *Câu so sánh đóng vai trò quan trọng trong việc thể hiện cảm xúc của người viết về vấn đề họ đang quan tâm bằng cách so sánh với các đối tượng khác nhằm đưa ra quan điểm đánh giá đối tượng là tốt hoặc không tốt. Bài viết xây dựng tập nhãn để xác định câu so sánh trong những bình luận tiếng Việt thuộc một miền cụ thể (trang web bán điện thoại di động) và tập nhãn từ so sánh được ứng dụng để đưa ra kết quả phân tích cảm xúc của người dùng dựa trên các bình luận của họ. Việc xây dựng này được thực hiện từng bước bằng cách phân tích trên một miền dữ liệu cụ thể, đồng thời ứng dụng các chương trình xử lý ngôn ngữ và kho từ vựng phong phú của Từ điển cảm xúc tiếng Việt để đưa ra kết quả phân tích với độ chính xác cao. Hiệu quả của phương pháp này được thể hiện thông qua một chương trình ứng dụng được xây dựng để đánh giá độ chính xác của tập nhãn xác định câu so sánh trong bình luận tiếng Việt.*

Từ khóa: *nhãn, so sánh, cảm xúc, điểm tích cực, điểm tiêu cực.*

ABSTRACT*: Comparison sentences have important role in presenting the writer's emotions about the issues they are concerned by comparison with other objects in order to evaluate whether the object is good or bad. This paper builds set labels to identify the comparison sentences in the Vietnamese comments in a specific domain (website selling mobile phones) and the collective label for comparison used to analyze the emotions of users based on their comments. The construction is carried out gradually by analyzing data of a specific domain, and applying special programs to processing language and by referring to the rich vocabulary of the Vietnamese emotional dictionary in order to arrive at highly accurate results of analysis. The effectiveness of this method is manifestedthrough an application program which is built to evaluate the accuracy of the collective label in determining comparison sentences of Vietnamese comments.*

Key words: *label, comparative/comparison, emotions, positive points, negative points.stu*

1. ĐẶT VẤN ĐỀ

Sự gia tăng của các thiết bị sử dụng web cho phép con người có thể giao tiếp với nhau trong cộng đồng web bằng nhiều hình thức khác nhau như diễn đàn, mạng xã hội, blog. Do đó một số lượng lớn các dữ

(*) ThS. Trường Đại học Văn Lang, *Email: lythihuyenchau@vanlanguni.edu.vn*

liệu không đồng nhất được tạo ra bởi những người sử dụng trong các cộng đồng, trong đó những câu bình luận của người dùng là một nguồn tài nguyên vô cùng lớn và có ý nghĩa thực tiễn. Hiện nay, các doanh nghiệp luôn sử dụng các mạng xã hội trực tuyến để quảng bá kinh doanh của công ty, cũng như sử dụng các dịch vụ vốn có của một trang mạng truyền thông xã hội đang có để phục vụ cho hoạt động kinh doanh của họ.

Trong thời đại phát triển của mạng xã hội, thông qua những câu bình luận dạng so sánh, người dùng mạng xã hội muốn trình bày thái độ của mình về sản phẩm mình quan tâm, hoặc muốn tìm hiểu về sản phẩm (điện tử công nghệ như máy tính, điện thoại) thông qua những bình luận trước đó của người sử dụng đã từng tìm hiểu. Về phía doanh nghiệp, họ muốn biết được đánh giá của người dùng về sản phẩm của công ty từ các bình luận có tính chất so sánh đó, chúng thể hiện sự đánh giá của người bình luận về sản phẩm của công ty dựa vào một sản phẩm khác được so sánh, có thể tốt hơn hoặc tệ hơn và gây ảnh hưởng đến tâm lý, cảm xúc của nhiều người đọc khác.

2. THỰC TRẠNG NGHIÊN CỨU PHÂN TÍCH CẢM XÚC TỪ CÁC BÌNH LUẬN SO SÁNH

Nhận thấy được tầm quan trọng của việc rút trích quan điểm từ những bình luận có tính chất so sánh, nghiên cứu [10] đưa ra phương pháp xác định cảm xúc của người dùng bằng cách đưa ra người nắm giữ quan điểm, đồng thời xác định các từ cảm xúc đã tạo nên nhiều cảm xúc trong một câu. Tuy nhiên, việc xác định người nắm giữ quan điểm không đạt được kết quả tốt khi trong câu có nhiều hơn một người nắm giữ chủ đề.

Một nghiên cứu khác của Jindal và Liu [7, tr.244-251] cho thấy việc xác định câu so sánh hữu ích cho việc phân tích câu trong tài liệu. Nhận thấy được tầm quan trọng của câu so sánh, bài báo đưa ra những vấn đề của việc xác định câu so sánh, phân loại các câu so sánh, đưa ra các nhãn và sau đó tiếp cận phương pháp học giám sát để xác định câu so sánh từ tài liệu bằng việc kết hợp phương pháp CSR (Class Sequential Rules) và học máy (Machine Learning).

Ngoài ra, bài viết [4, tr.417-422] nghiên cứu xây dựng từ điển cảm xúc dựa trên bộ từ vựng tiếng Anh với các trọng số điểm tích cực và tiêu cực. Nghiên cứu xây dựng tập từ điển từ vựng SentiWordNet làm nguồn tài nguyên công khai cho các nghiên cứu khai thác quan điểm khác.

Một nghiên cứu khác tương tự như Jindal và Liu [8, tr.1331-1336] phân loại các loại câu so sánh, xác định các đặc điểm riêng của chúng, cách xác định vị trí của các thực thể để đưa ra kết quả khai thác quan điểm chính xác. Tuy nhiên, chưa xác định được các đối tượng khác trong câu so sánh và chỉ thực hiện trên ngôn ngữ tiếng Anh.

Khai phá quan điểm trên mức độ câu và cụm câu được thực hiện trong [5, tr.201-248]. Nghiên cứu đề xuất được các giải pháp để giải quyết các vấn đề tồn đọng ở nghiên cứu trước của tác giả. Với những kết quả đạt được là nguồn tham khảo tốt liên quan đến khai phá quan điểm.

Trong nghiên cứu [14, tr.230-235], nhóm tác giả phân tích để thực hiện những công việc chính của việc khai phá quan điểm từ những bình luận trên web của khách hàng về sản phẩm và dịch vụ mà họ quan tâm sử dụng. Kết quả là nghiên cứu cung cấp một cái nhìn tổng quan khi đưa ra nhiều công việc và kỹ thuật đáp ứng việc khai phá quan điểm.

Một nghiên cứu khác, [6, tr.211-217] thực hiện việc khai thác quan điểm từ những tiểu blog trên internet bằng cách rút trích các tính từ thuộc một lĩnh vực cụ thể, đồng thời đưa ra cách tiếp cận mới bằng phương pháp tự động trích xuất tính từ để đưa ra quan điểm người dùng từ những tài liệu thu thập được trên internet.

Nhận thấy khai thác quan điểm là nhiệm vụ của việc trích xuất từ một tập hợp các tài liệu, nghiên cứu [2, tr.523-526] đánh giá cách tiếp cận việc sử dụng dấu ngoặc chú thích trích từ tin tức được cung cấp bởi công cụ thu thập tin tức Europe Media Monitor (EMM). Nghiên cứu này thực hiện trên dữ liệu đặc biệt (bảng báo giá), sẽ làm đa dạng việc khám phá quan điểm người tiêu dùng.

Việc phân tích cảm xúc trên mức độ câu được thực hiện trong nghiên cứu [9, tr.153.153] bằng cách xây dựng hệ thống phân tích cảm xúc dựa trên quy tắc bằng cách sử dụng Framework Gate. Nghiên cứu này cho thấy kết quả phân tích cảm xúc cho một vài sản phẩm trên dữ liệu training và dữ liệu test đạt kết quả chính xác cao, đồng thời tạo tiền đề để khai phá những vấn đề liên quan đến phân tích cảm xúc tiếng Việt.

Ngoài ra, trong [1, tr.17-23] trình bày việc xây dựng từ điển từ vựng SentiWordNet giúp người dùng phân loại cảm xúc và trích xuất quan điểm. Tuy nhiên, các từ vựng trong từ điển chưa đầy đủ và chỉ đáp ứng trong một miền cụ thể.

Dùng dữ liệu thu thập được từ Twitter, [11, tr.538-541] nghiên cứu các tiện ích của tính năng ngôn ngữ để phát hiện cảm xúc của các thông điệp Twitter. Đây là đánh giá về nguồn tài nguyên sử dụng, thực sự hữu ích cho nhiều nghiên cứu sử dụng để khai phá quan điểm.

Nhận thấy tầm quan trọng của từ khóa trong việc rút trích quan điểm, nghiên cứu [3, tr56-59] tập trung xác định tập từ khóa để phân loại và rút trích quan điểm. Nghiên cứu đưa ra tập từ khóa phân loại cảm xúc và đánh giá tính hiệu quả của tập từ khóa đó góp phần cho các nghiên cứu khai phá quan điểm sau này.

Việc rút trích chính kiến của người dùng trong các văn bản trên mạng xã hội nên được thực hiện trong [12, tr.538-547] cung cấp một phương pháp phát hiện chính kiến của người dùng dựa trên những ý kiến cá nhân họ trình bày trên mạng xã hội Twitter. Đây là nghiên cứu cung cấp một thuật toán mới cho việc phát hiện chính kiến của chủ thể trong văn bản.

Phân tích cảm xúc dựa vào từ điển cảm xúc tiếng Việt được thực hiện trong [15, tr.136-148]. Từ điển khá chính xác khi được xây dựng dựa trên từ điển SentiWordNet và từ cảm xúc được rút trích từ các trang mạng xã hội trong một miền cụ thể. Đây là nghiên cứu cung cấp một từ điển cảm xúc tiếng Việt với số từ vựng khá lớn giúp ích cho việc khai phá quan điểm.

Trong việc xử lý ngôn ngữ tự nhiên, nghiên cứu [16] cho rằng bản chất của quá

trình rút trích cảm xúc người dùng trên mạng xã hội là một quá trình máy học. Nghiên cứu thông qua những bình luận, những tiểu blog trên mạng xã hội, nghiên cứu đánh giá được hành vi của con người thể hiện rất nhiều qua ngôn ngữ, và cần phải được ghi nhớ.

Qua nhiều nghiên cứu về phân tích cảm xúc có thể thấy đa số quan điểm được rút trích từ các bình luận tiếng Anh và chưa tập trung trên các câu so sánh nên việc xây dựng tập nhãn để xác định câu so sánh từ những bình luận so sánh tiếng Việt trong một miền cụ thể để đưa ra kết quả phân tích cảm xúc đang là một vấn đề đang rất được người dùng quan tâm.

3. TÌM HIỂU PHẦN MỀM GÁN NHÃN TỪ LOẠI VÀ TỪ ĐIỂN CẢM XÚC TIẾNG VIỆT

3.1. Phần mềm gán nhãn từ loại tiếng Việt

vnTagger là phần mềm mã nguồn mở của Lê Hồng Phương dùng để tách từ và gán nhãn từ loại cho văn bản tiếng Việt. Nghiên cứu [13, tr.12] đã mô tả tập nhãn được dùng trong chương trình vnTagger bao gồm 18 nhãn từ loại. Phiên bản chúng tôi sử dụng là phiên bản 4.2.0 được công bố vào tháng 4/2010.

3.2. Từ điển cảm xúc tiếng Việt

Sử dụng từ điển để trích xuất cảm xúc là một trong những cách tiếp cận chính để khai thác quan điểm. Trong [15], nhóm nghiên cứu đã dựa trên nguồn từ vựng tiếng Anh của SentiWordNet để xây dựng một Từ điển tiếng Việt với 26,186 từ cảm xúc thuộc loại tính từ, trạng từ, danh từ và động từ, trong đó mỗi từ cảm xúc sẽ có một trọng số điểm tích cực và tiêu cực. Ngoài ra, từ điển này được xây dựng dựa trên một miền cụ thể là các bình luận được thu thập từ các trang web thương mại đặc biệt là điện thoại di động và máy tính nên rất phù hợp với mục đích của nghiên cứu. Đồng thời, vì từ điển này đã được xây dựng dựa trên SentiWordNet và WordNet nên nghiên cứu này chỉ dùng ngữ liệu SentiWordNet như là cơ sở dữ liệu để kiểm tra tính chính xác của từ điển. Trong [1] mô tả các thành phần của SentiWordNet như sau:

Synset: là một bản ghi, cấu tạo bởi 6 cột, các cột phân cách bởi dấu <tab>:

- POS: từ loại của từ
- ID: mã đại diện cho synset
- PosScore: trọng số tích cực của từ
- NegScore: trọng số tiêu cực của từ
- SynsetTerms: là những từ nhận định trong synset.

SynsetTerms: là những từ nhận định trong synset. Một synset có thể chứa nhiều từ, và các từ này là từ đồng nghĩa với nhau. Một từ có thể có nhiều ngữ cảnh khác nhau và trọng số Pos(s)/Neg(s) sẽ khác, do đó các từ này sẽ được gán kèm theo số hiệu để phân biệt các từ.

POS	ID	PosScore	NegScore	SynsetTerms
v	1984570	0.125	0	ngay#3
v	1988080	0.125	0	không_gian#1
v	1988330	0.5	0	nguyên_soái#1
v	1993670	0.125	0.125	mứt#2
v	2002720	0.125	0	đuổi#1đuổi#1
v	2006710	0.25	0	đến#16
v	2007680	0.125	0	lũ#1
v	2020410	0.125	0	tàn_phá#1

Hình 1. Một vài dòng dữ liệu trong Từ điển cảm xúc tiếng Việt

4. ĐỀ XUẤT PHƯƠNG PHÁP PHÂN TÍCH CẢM XÚC DỰA TRÊN TỪ ĐIỂN CẢM XÚC TIẾNG VIỆT

4.1. Xác định các loại so sánh tiếng Việt

Tiếng Việt giống tiếng Anh về các loại so sánh được mô tả chi tiết trong [5]. Các câu bình luận tiếng Việt thường thuộc một trong ba loại câu so sánh sau, các câu bình luận còn lại thuộc dạng câu thông thường hoặc câu bất thường:

Câu so sánh nhất: là những câu so sánh lớn hơn hoặc nhỏ hơn tất cả các đối tượng còn lại. Trong câu thường có các từ như: nhất, số 1,…

Ví dụ: "*iPhone là dòng điện thoại đẹp nhất"*

Câu so sánh bằng: là những câu so sánh sự tương đương về một số đặc điểm giữa các đối tượng. Trong câu thường có các từ như: như nhau, giống,…

Ví dụ: "*iPhone và Android là hai dòng điện thoại cảm ứng tốt như nhau"*.

Câu so sánh hơn: là những câu so sánh sự lớn hơn hoặc nhỏ hơn, sự sắp xếp có thứ tự giữa các đối tượng. Trong câu thường có các từ như: hơn, thua,….

Ví dụ: "*iPhone chụp hình đẹp hơn Nokia"*.

Câu thông thường: là câu bình luận thông thường không chỉ ra sự so sánh, cũng như không đưa ra thứ tự giữa các đối tượng.

Ví dụ*: "Điện thoại iPhone cảm ứng rất tốt"*.

Câu bất thường: là bao gồm những câu tiếng lóng, không dấu, hoặc viết theo thuật ngữ thanh thiếu niên, theo thuật ngữ mạng xã hội,…

Ví dụ: *"Điện thoại iPhone thì chuẩn cơm mẹ nấu"*.

Nghiên cứu này tập trung phân tích các bình luận tiếng Việt dạng so sánh nên trong nghiên cứu này có thể bỏ qua các câu thông thường và câu bất thường, tuy nhiên chúng vẫn được thu thập để đánh giá mức độ chênh lệch giữa câu so sánh và câu thông thường của các bình luận được thu thập từ các trang web thương mại. Bảng 1 sau đây cho biết danh sách các loại câu so sánh mà chúng tôi tập trung nghiên cứu.

Bảng 1. Danh sách loại câu so sánh

TT	Loại câu so sánh	Nhãn
1	So sánh nhất	N
2	So sánh hơn	H
3	So sánh bằng	B

4.2. Xác định bộ tập từ theo loại câu so sánh

Dựa trên các bình luận được thu thập từ các trang web thương mại, người nghiên cứu tự xác định các câu bình luận so sánh và xây dựng bộ tập từ theo từng loại so sánh. Kết quả khởi tạo có 16 từ loại được xác định (trong đó các nhãn: N: so sánh nhất, H: so sánh hơn, B: so sánh bằng).

Bảng 2. Danh sách khởi tạo từ theo loại so sánh

TT	Nhãn	Từ thể hiện
1	N	nhất
2	N	no 1
3	N	number 1
4	N	số 1
5	N	số một
6	N	number one
7	H	hơn
8	H	thua
9	H	kém

10	B	giống
11	B	same
12	B	cỡ
13	B	y xì
14	B	như
15	B	bằng
16	B	đều

Đánh giá độ chính xác của Thuật toán với 16 từ khởi tạo này được thống kê cụ thể trong Bảng 3. Thống kê này được thực hiện trên 705 câu bình luận, được lấy từ 5 chủ đề ngẫu nhiên.

Bảng 3. Kết quả thống kê độ chính xác của thuật toán xác định câu so sánh và gán nhãn so sánh

TT	**Chủ đề**	**Câu bình luận**	**Đúng**	**Độ chính xác**
1	Điện thoại nào có camera chụp hình đẹp hơn iPhone 6?	98	88	89%
2	Dùng iPhone 6 Plus rồi thì chuyển sang Note 4 hay HTC One M9?	246	231	94%
3	Galaxy Note 4 hay iPhone 6 Plus phù hợp hơn với việc thư ký?	105	100	96%
4	Pin Galaxy S6 tốt hơn iPhone 6	67	63	94%
5	Galaxy S6 Edge và iPhone 6 Plus đọ khả năng chống rung	189	172	91%

Quan sát Bảng 3, có thể thấy với bộ tập từ khởi tạo gồm 16 từ ở Bảng 2, độ chính xác trung bình của thuật toán xác định câu so sánh và gán nhãn so sánh là 92.8%.

Độ sai số của thuật toán chủ yếu tập trung trên các cụm từ có gắn liền với từ "như" trong Bảng 3, mặc dù có từ "như" nhưng câu lại không mang ý nghĩa so sánh bằng, ví dụ: hầu như, như vậy thôi, mong như thế, giá như, như kiểu của em, đơn cử như, như thế là, như sau, như cách nhìn,… Mặt khác với từ "hơn" có thể dẫn đến một vài trường hợp sai, như: hơn 1 năm,…

Sau quá trình tính độ chính xác và quan sát trên tập từ dẫn đến kết quả sai, người nghiên cứu nhận thấy cần bổ sung một số từ vào bộ từ khởi tạo, với lý do, tần suất xuất hiện thường xuyên của các từ này và các từ đúng chuẩn "tiếng Việt".

Hiện tại, bộ tập từ loại so sánh bao gồm 26 từ, sau khi thực thi thuật toán mới để xác định câu so sánh và gán nhãn so sánh trên bộ tập từ mới này, kết quả với 1720 câu bình luận thì có 457 câu thuộc dạng so sánh. Danh sách đầy đủ của bộ từ khởi tạo và từ bổ sung sau quá trình phân tích được thể hiện trong Bảng 4.

4.3. Các bước thực hiện chính

Bước 1: Thu thập và tiền xử lý dữ liệu bình luận: là bước thu thập dữ liệu bình luận tự động từ các trang web thương mại (sử dụng công cụ Craw Tool của Website Internet Marketing Ninjas), sau đó dữ liệu

sẽ được chuẩn hóa và tách câu để phù hợp với mục đích phân tích.

Bước 2: Xác định câu bình luận tiếng Việt dạng so sánh: là bước dựa vào tập danh sách các từ xác định câu so sánh để xác định và gán nhãn câu so sánh. Tiếp theo, sử dụng chương trình vnTagger để gán nhãn từ loại tiếng Việt, sau đó rút trích danh sách và vị trí của các từ được gán nhãn theo yêu cầu phân tích.

Bảng 4. Danh sách từ theo từng loại so sánh sau quá trình phân tích

TT	Nhãn	Từ thể hiện
1	N	nhất
2	N	no 1
3	N	number 1
4	N	số 1
5	N	số một
6	N	number one
7	N	vô đối
8	N	trên cả tuyệt vời
9	N	khó ai vượt qua
10	N	xuất sắc
11	N	hoàn hảo
12	N	làm gì có đối thủ
13	N	chưa có đối thủ
14	N	đỉnh của đỉnh
15	N	ăn đứt hết
16	H	hơn
17	H	thua
18	H	kém
19	B	giống
20	B	same
21	B	cỡ
22	B	y xì
23	B	như
24	B	bằng
25	B	đều
26	B	ngang

Bước 3: Sử dụng từ điển cảm xúc tiếng Việt để tính điểm trọng số tích cực, tiêu cực: bước này sẽ kiểm tra câu bình luận có thuộc dạng câu phủ định, sau đó dựa vào Từ điển cảm xúc tiếng Việt và danh sách các từ gán nhãn để tính điểm tích cực và tiêu cực.

Điểm tích cực của tính từ và động từ được tính theo công thức:

$$pos = \sum P_i \quad (1)$$

Trong đó:

pos: Điểm tích cực

P_i*:* Điểm tích cực của tính từ/động từ thứ i

Điểm tiêu cực của tính từ và động từ được tính theo công thức:

$$neg = \sum N_i \quad (2)$$

Trong đó:

neg: Điểm tiêu cực

N_i*:* Điểm tiêu cực của tính từ/động từ thứ i

Ví dụ: "Note/N 4/M chụp/V đẹp/A hơn/R ip/N 6/M".

Kết quả: Với câu trên, tính từ trong câu là "*đẹp*", với tính từ này khi tìm trong Từ điển cảm xúc tiếng Việt theo công thức (1), (2), kết quả điểm tích cực của tính từ "*đẹp*": *pos* = 6.75, điểm tiêu cực của tính từ "*đẹp*" *neg* = 0.5.

Nếu trong câu có xuất hiện từ phủ định và vị trí xuất hiện của từ phủ định trước ngay vị trí của của tính từ/động từ thì điểm số tích cực và tiêu cực của tính từ/động từ đó được tính theo công thức sau:

$$\begin{aligned} fpos &= neg \\ fneg &= pos \end{aligned} \quad (3)$$

Trong đó:

fpos: Điểm tích cực của tính từ/động từ có phủ định

fneg: Điểm tiêu cực của tính từ/động từ có phủ định.

Ví dụ: "Note/N 4/M chụp/V không/R đẹp/A hơn/R ip/N 6/M".

Kết quả: Điểm tích cực và tiêu cực của tính từ "*đẹp*" đã được tính ở Bước (3). Vậy điểm tích cực và tiêu cực của tính từ này sau khi có từ phủ định "*không*" kèm phía trước như sau:

- Điểm tích cực: *pos* = 0.5
- Điểm tiêu cực: *neg* = 6.75

Điểm tích cực và tiêu cực cho từng đối tượng chủ đề tùy thuộc vào vị trí xuất hiện của đối tượng đó so với vị trí của từ loại so sánh và câu so sánh.

Bảng 5. *Bảng ưu tiên đối tượng chủ đề trong câu so sánh*

TT	Câu so sánh	Cú pháp	Đối tượng ưu tiên
1	Nhất	$O_{1..n}$ + Adj/V + char Ví dụ: *iPhone 6 là đẹp nhất.*	$O_{1..n}$
2	Nhất	Adj/V + char + $O_{1..n}$ Ví dụ: *đẹp nhất là iPhone 6.*	$O_{1..n}$
3	Bằng	O_1 + O_2 + Adj/V + char Ví dụ: *iPhone 6 và Z3 đẹp như nhau.*	O_1, O_2
4	Bằng	O_1 + Adj/V + char + O_2 Ví dụ: *iPhone 6 đẹp như Z3.*	O_1, O_2
5	Hơn	O_1 + Adj/V + char + O_2 Ví dụ: *iPhone 6 đẹp hơn Z3.*	O_1
6	Hơn	O_1 + Adj/V + char Ví dụ: *iPhone 6 đẹp hơn.*	O_1
7	Hơn	O_1 + char + O_2 + Adj/V Ví dụ: *iPhone 6 hơn Z3 về chụp đẹp.*	O_1

Điểm tích cực cho toàn bộ chủ đề được tính theo công thức:

$$spos = \sum pos_j \quad (4)$$

Trong đó:

spos: Tổng điểm tích cực của chủ đề.

pos_j: Điểm tích cực của đối tượng thứ j.

Điểm tiêu cực cho toàn bộ chủ đề được tính theo công thức:

$$sneg = \sum neg_j \quad (5)$$

Trong đó:

sneg: Tổng điểm tiêu cực của chủ đề.

neg_j: Điểm tiêu cực của đối tượng thứ j.

Sau khi tính được tổng điểm tích cực và tiêu cực của các đối tượng, tiến hành so sánh kết quả và phân tích:

Nếu *spos* > *sneg*: Đối tượng trong chủ đề được đánh giá tốt

Nếu *spos* < *sneg*: Đối tượng trong chủ đề được đánh giá không tốt.

Nếu *spos* = *sneg*: Đối tượng trong chủ đề được đánh giá bình thường.

Bước 4: Phân tích cảm xúc người dùng dựa trên bình luận tiếng Việt dạng so sánh: bước này sẽ xác định vị trí của các đối

tượng chủ đề trong câu so sánh để tính tổng điểm tích cực, tiêu cực cho mỗi đối tượng, sau đó tổng hợp so sánh và đưa ra nhận xét.

5. KẾT QUẢ VÀ ĐÁNH GIÁ

Các bình luận: được thu thập từ các trang Web thương mại, đặc biệt là trong lĩnh vực điện thoại, với số lượng chủ đề bài báo: 25 (nguồn: sohoa.vnexpress.net), số lượng bình luận: 2,185 (bao gồm câu thông thường và câu so sánh).

Số lượng 2,185 bình luận trên được chuẩn hóa để các bình luận đáp ứng được yêu cầu nghiên cứu. Sau khi chuẩn hóa, số bình luận còn lại đáp ứng được yêu cầu là 17 chủ đề và 795 bình luận.

Từ 17 chủ đề trên, có 25 đối tượng chủ đề cùng với 427 từ các từ viết tắt/các dạng khác có thể có của đối tượng chủ đề.

Kết quả khi ngắt câu bình luận sau khi áp dụng tập ký tự là 1,720 câu được tách từ 795 bình luận đã được chuẩn hóa. Trong danh sách 1,720 câu có 457 câu thuộc dạng câu bình luận so sánh và được gán nhãn.

Sau khi áp dụng tập nhãn xác định câu so sánh tiếng Việt, chúng tôi đo lường độ chính xác của chúng bằng cách tiến hành kiểm tra kết quả phân tích cảm xúc của người dùng trên từng câu bình luận (1,720 câu), những vấn đề kiểm tra gồm có:

Gán nhãn từ loại (chỉ xét những từ nhãn quan tâm: danh từ, động từ, tính từ,…).

Xác định loại câu so sánh

Tính điểm tích cực

Tính điểm tiêu cực

Độ chính xác trung bình của việc áp dụng tập nhãn xác định câu so sánh để phân tích cảm xúc tổng hợp là 74.7%. Qua quá trình đánh giá, chúng tôi nhận thấy rằng tập nhãn này có độ chính xác có thể ứng dụng được, tập nhãn này hoạt động tốt khi những bình luận thu được trong một miền cụ thể. Việc thực thi sẽ bị giảm đi nếu những đánh giá đến từ các lĩnh vực khác nhau. Trong ngôn ngữ Việt, có một số trường hợp mà một từ quan điểm có thể được giải thích bởi nhiều ý nghĩa khác nhau. Một chữ "dài" là một cảm xúc tích cực nếu nó đề cập đến pin nhưng sẽ trở thành cảm xúc tiêu cực khi chúng ta nói điều gì đó về thời gian chờ đợi trong một nhà hàng. Đây là lý do tại sao độ chính xác của tập nhãn này không cao như mong đợi, nhưng độ chính xác của phương pháp này sẽ góp phần tạo tiền đề cho các nghiên cứu khác về việc phân tích cảm xúc của người dùng dựa trên các bình luận tiếng Việt.

6. KẾT LUẬN VÀ KIẾN NGHỊ

Bài viết đã tìm hiểu và xây dựng được các loại so sánh và tập nhãn xác định câu so sánh, sau đó đưa ra cách xác định và gán nhãn các câu so sánh. Đồng thời, nghiên cứu này đã đưa ra được thuật toán ngắt câu và áp dụng chương trình vnTagger để gán nhãn từ loại tiếng Việt. Từ đó, dựa trên từ điển cảm xúc tiếng Việt, nghiên cứu này chỉ ra được kết quả phân tích cảm xúc dựa trên tập danh từ, danh từ riêng, danh từ đơn vị, số từ và dựa trên điểm tích cực, tiêu cực của tính từ, động từ.

Để thể hiện độ chính xác của tập nhãn xác định câu so sánh, người nghiên cứu đã xây dựng chương trình thử nghiệm. Chương trình được thực hiện trên 17 chủ đề, 795 bình luận, 1,720 câu và dựa trên Từ điển cảm xúc tiếng Việt gồm 26,186 từ. Kết quả của phương pháp ứng dụng tập nhãn

xác định câu so sánh tiếng Việt là khả quan với độ chính xác trung bình đạt 74.7%.

Trong thời gian tới, ngoài việc tiếp tục giải quyết các vấn đề còn tồn tại, một số nghiên cứu tiếp theo dự kiến sẽ thực hiện: Nghiên cứu thêm về quy luật gán nhãn của vnTagger để bao phủ hết tất cả các trường hợp gán nhãn; Bổ sung danh sách cách viết khác của các đối tượng có thể có từ các trang mạng xã hội; Cải tiến phương pháp xác định câu so sánh bằng cách hoàn thiện tập từ thể hiện các loại so sánh.

TÀI LIỆU THAM KHẢO

1. Baccianella, S., A. Esuli, and F. Sebastiani (2010), *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, Proceedings of the International Conference on Language Resources and Evaluation.
2. Balahur, A. et al. (2009), *Opinion Mining on Newspaper Quotations*. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.
3. Baumgarten, M. et al. (2013), *Keyword-Based Sentiment Mining using Twitter*, International Journal of Ambient Computing and Intelligence.
4. Esuli, A. and F. Sebastiani (2006), *Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining*, Proceedings of the 5th Conference on Language Resources and Evaluation.
5. Ganapathibhotla, M. and B. Liu (2008), *Mining Opinions in Comparative Sentences*. Proceedings of the 22nd International Conference on Computational Linguistics.
6. Harb, A., et al. (2008). *Web Opinion Mining: How to extract opinions from blogs?* Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology.
7. Jindal, N. and B. Liu (2006), *Identifying Comparative Sentences in Text Documents*, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.
8. Jindal, N. and B. Liu (2006), *Mining Comparative Sentences and Relations*, Proceedings of the 21st national conference on Artificial intelligence.
9. Kieu, B.T. and S.B. Pham (2010), *Sentiment Analysis for Vietnamese*, Proceedings of the 2010 Second International Conference on Knowledge and Systems Engineering.
10. Kim, S.-M. and E. Hovy (2004), *Determining the Sentiment of Opinions*, Proceedings of the 20th international conference on Computational Linguistics, no. 1367.
11. Kouloumpis, E., T. Wilson, and J.D. Moore (2011), *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, Proceedings of the Fifth International Conference on Weblogs and Social Media.

12. Kwon, A. and K.-S. Lee (2013), *Opinion Bias Detection Based on Social Opinions for Twitter*, Journal of Information Processing Systems.
13. Le-Hong, P., et al. (2010), *An Empirical Study of Maximum Entropy Approach for part-of-Speech Tagging of Vietnamese Texts*, Traitement Automatique des Langues Naturelles.
14. Lee, D., O.-R. Jeong, and S.-g. Lee (2008), *Opinion Mining of Customer Feedback Data on the Web*, Proceedings of the 2nd international conference on Ubiquitous information management and communication.
15. Nguyen, H.N., et al. (2014), *Domain Specific Sentiment Dictionary for Opinion Mining of Vietnamese Text*, Proceedings of the 8th International Workshop on Multi-disciplinary Trends in Artificial Intelligence.
16. Thakkar, H. and D. Patel (2015), *Approaches for Sentiment Analysis on Twitter: A State-of-Art study*, arXiv preprint arXiv:1512.01043.

Ngày nhận bài: 08/11/2016. Ngày biên tập xong: 17/02/2017. Duyệt đăng: 21/3/2017