Tóm tắt văn bản trên cơ sở phân loại ý kiến độc giả của báo mạng tiếng Việt

- **Nguyễn Ngọc Duy** [(1)]
- **Phan Thị Tươi** [(2)]

[(1)] Học viện Công nghệ Bưu chính Viễn thông Cơ sở Tp.HCM

[(2)] Trường Đại học Bách khoa, ĐHQG-HCM

(Bản nhận ngày 01 tháng 03 năm 2016, hoàn chỉnh sửa chữa ngày 06 tháng 09 năm 2016)

ABSTRACT

Muốn biết ý kiến của người dùng về một mặt hàng, hoặc của cộng đồng về một vấn đề nóng trong xã hội,..., cách tốt nhất trong thời đại bùng nổ thông tin trên internet và mạng xã hội, là khai thác thông tin một cách hiệu quả từ những nguồn này. Mỗi ý kiến không chỉ là một thông tin đơn thuần, mà còn chứa cả cảm xúc của người viết. Do đó, chúng có thể tạo nên một luồng dư luận tác động đến cộng đồng mạng. Đây thật sự là nguồn tài nguyên khổng lồ, có ý nghĩa to lớn đối với nhiều lĩnh vực – từ kinh tế, chính trị đến văn hóa xã hội – nếu có phương pháp khai thác thông tin hiệu quả. Một hệ thống tự động để phân loại ý kiến dựa trên cảm xúc là rất cần thiết để khai thác hiệu quả nguồn tài nguyên này. Để hỗ trợ người sử dụng khai thác thông tin hiệu quả hơn, vấn đề tóm tắt thông tin cần được nghiên cứu giải quyết, nhất là ở khía cạnh quan điểm và cảm xúc trong mỗi ý kiến.

Hướng đến mục tiêu khai thác hiệu quả nguồn tài nguyên, bài báo này sẽ giới thiệu mô hình tóm tắt văn bản, không chỉ dựa vào ngữ nghĩa mà còn dựa trên yếu tố cảm xúc. Chúng tôi đã xây dựng một mô hình tổng quát để giải quyết bài toán này. Từ các phương pháp phân tích ý kiến và tóm tắt văn bản mà nhiều công trình nghiên cứu đã sử dụng, bài báo đã kết hợp và phát triển các phương pháp để tóm tắt văn bản tiếng Việt trên cơ sở phân loại cảm xúc. Các văn bản được tóm tắt là các trang báo mạng tiếng Việt.

Từ khóa: *phân loại ý kiến, phân tích ý kiến, tóm tắt ý kiến, tóm tắt văn bản.*

1. GIỚI THIỆU

Internet và mạng xã hội phát triển mang lại nhiều lợi ích cho người dùng. Trên mạng xã hội mọi người thể hiện ý kiến, cảm xúc, ... của mình mà ít bị ràng buộc bởi yêu cầu chuẩn mực về từ và văn phạm. Đó là nguồn tài nguyên lớn để khai thác cho những cá nhân, tổ chức hoạt động trong các lĩnh vực liên quan đến cộng đồng trước khi ra quyết định. Nếu nhận diện được ý kiến của cộng đồng đối với một đối tượng, một vấn đề thì chúng ta có thể có những quyết định hiệu quả trong các hoạt động kinh tế, chính trị, xã hội, …

Khai thác nguồn tài nguyên này cần có sự hỗ trợ của khoa học, công nghệ với những công cụ tự động trong thống kê và tóm tắt thông tin, hỗ trợ hiệu quả quá trình ra quyết định. Từ nhu cầu đó, chúng tôi đã nghiên cứu xây dựng hệ thống "Tóm tắt ý kiến trên cơ sở phân loại cảm xúc". Đối tượng chúng tôi xử lý là ý kiến độc giả các trang báo mạng.

Phần tiếp theo của bài báo được tổ chức như sau: Phần 2 chúng tôi sẽ thảo luận về các công trình liên quan; phần 3 nói về phương pháp tiếp cận của chúng tôi để tóm tắt ý kiến dựa trên phân tích cảm xúc; phần 4 là kết quả thử nghiệm; và phần 5, chúng tôi sẽ có kết luận của mình và đưa ra hướng phát triển tiếp theo.

2. CÔNG TRÌNH LIÊN QUAN

Nội dung bài báo liên quan đến nhiều vấn đề đã và đang được nghiên cứu trên thế giới. Phân loại văn bản, tóm tắt văn bản đã được nghiên cứu nhiều, trong đó có tiếng Việt. Tương tự là bài toán gán nhãn, xác định các đối tượng trong văn bản, … Hướng phân tích ý kiến (cảm xúc) hiện đang được quan tâm. Hướng nghiên cứu này với tiếng Anh bắt đầu từ đầu những năm 2000, có nhiều kết quả rất tốt [1]. Các lĩnh vực được nghiên cứu theo hướng này như giải trí (bình phim), thương mại (bình sản phẩm), xã hội (việc làm), …

Các tác giả [2] thực nghiệm phân thành ba mức (cao, trung bình và thấp) cho các cảm xúc: tích cực, tiêu cực (positive, negative). Kho ngữ liệu của [2] là 51 bài blog tiếng Anh. Kết quả có độ chính xác khá cao, trên 90%. Hệ phân tích cảm xúc và hệ tóm tắt văn bản tách biệt, xử lý phân tích cảm xúc trước khi tóm tắt.

Công trình [4] và [5] thực hiện phân loại cảm xúc các văn bản tiếng Việt. [5] là bản cải tiến của [4], phân loại những đánh giá cho từng đặc tính kỹ thuật (mạng, màn hình, giá, …) của một số điện thoại thông minh (smartphone). Các đánh giá được phân vào các lớp positive, neutral (trung hòa) và negative, dựa vào các tính từ gắn với mỗi đặc tính kỹ thuật. Sau đó, thống kê các ý kiến ở mỗi lớp cho từng đặc tính kỹ thuật của mỗi điện thoại. [5] có kết quả khá tốt: độ chính xác đạt 56.60% - 77.12%, độ truy hồi (recall) đạt 48% - 78.23% và độ F đạt 52.30% - 77.45% tùy mỗi loại điện thoại. Tuy nhiên, [4] và [5] mới chỉ dừng lại ở việc thống kê theo kết quả phân cực cảm xúc, chưa hỗ trợ khai thác thông tin hiệu quả bằng việc tóm tắt các đánh giá dựa trên cảm xúc đó.

3. TÓM TẮT Ý KIẾN TRÊN CƠ SỞ PHÂN LOẠI Ý KIẾN

3.1 Phân loại ý kiến

Ý kiến (cảm xúc) thường được thể hiện một cách tinh tế, nên xác định cảm xúc sẽ khó hơn xác định chủ đề của văn bản. Không đơn giản xác định được cảm xúc mà chỉ dựa vào một câu, hay một thành phần của văn bản. Việc xác định cảm xúc thường không dựa vào tần suất xuất hiện của các thực thể cảm xúc, mà dựa trên nhiều yếu tố như từ loại, ngữ cảnh xuất hiện chúng, …

Bài báo đề xuất mô hình tóm tắt ý kiến trên cơ sở phân loại cảm xúc, như ở hình 1. Các ý kiến là những văn bản thể hiện suy nghĩ chủ quan của độc giả đối với vấn đề hoặc đối tượng trong *Bài báo*. Các văn bản này ít sự chuẩn mực về từ và văn phạm. Vì vậy, mô hình phải tiến hành chuẩn hóa văn bản cho *Tập văn bản thông tin* và tách câu bằng mô đun *Tiền xử lý*. Sau đó, chúng được rút trích các đặc trưng cảm xúc. Mô đun *Phân loại cảm xúc* sẽ phân chúng vào các lớp cảm xúc positive, negative và neutral. Giá trị cảm xúc cần khai thác là positive và negative. Mô đun *Tóm tắt* sẽ tóm tắt ý kiến thuộc hai lớp này dựa trên đặc

trưng cảm xúc của chúng, kết hợp đặc trưng ngữ nghĩa được rút trích dựa theo *Bài báo*. Bài báo kết hợp đặc trưng cảm xúc và đặc trưng ngữ nghĩa để bản tóm tắt có sự súc tích bên cạnh chiều cảm xúc, giúp khai thác hiệu quả thông tin.

Do mỗi ý kiến là của từng độc giả khác nhau, chúng tôi sẽ tóm tắt độc lập từng ý kiến, không liên hệ chúng với nhau về ngữ nghĩa, hoặc cảm xúc. Hệ thống đề xuất là một quá trình khép kín, nên mô đun *Tóm tắt* sẽ xử lý tóm tắt văn bản (ý kiến) đã được phân cực cảm xúc, mà không xét tính đúng đắn của quá trình phân cực này.

3.1.1 Từ điển cảm xúc

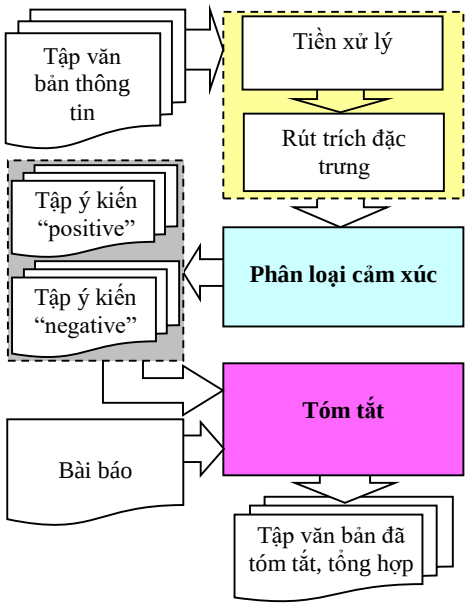
Đây là thành phần rất quan trọng trong hệ thống phân tích cảm xúc. Tuy nhiên, việc xây dựng từ điển cảm xúc là một công việc cần nhiều công sức và thời gian. Hiện chưa có bộ từ điển cảm xúc tiếng Việt chuẩn dùng cho nghiên cứu.

Để có bộ từ điển cảm xúc, chúng tôi đã chuyển ngữ sang tiếng Việt bộ từ điển cảm xúc tiếng Anh của công trình [3], có mở rộng từ điển trong quá trình thực nghiệm. Từ điển này có hơn 21.000 mục từ được gán trọng số cảm xúc.

3.1.2 Rút trích đặc trưng cảm xúc

Để rút trích đặc trưng cảm xúc, bài báo hiện thực phương pháp *Đối sánh thực thể dài nhất* (Maximum Matching) [6]. Đây là phương pháp tương đối dễ cài đặt, tốc độ cao, độ chính xác chấp nhận được, nhất là với đối tượng văn bản không chuẩn như những ý kiến trên mạng xã hội.

Phương pháp này dựa trên một từ điển tiếng Việt, gồm những từ và cụm từ sau đây gọi chung là thực thể. Có hai phương pháp *Đối sánh thực thể dài nhất* là đối sánh từ trái qua phải và đối sánh từ phải qua trái. Bài báo này sử dụng phương pháp *Đối sánh thực thể dài nhất* từ phải qua trái, dựa vào từ điển mô tả ở 3.1.1. Qua thực



Hình 1: Mô hình hệ thống Tóm tắt ý kiến trên cơ sở phân loại cảm xúc.

nghiệm, phương pháp này cho thấy khá hiệu quả với tiếng Việt.

3.1.3 Phân loại cảm xúc

Đầu tiên, mỗi câu trong ý kiến sẽ được phân loại cảm xúc bằng phương pháp Naïve Bayes. Sau đó, mô hình hóa tập đặc trưng cảm xúc của mỗi câu thành các vector. Tiếp theo, chuẩn hóa các vector về chiều, và tổng hợp thành vector đặc trưng cho mỗi lớp cảm xúc bằng cách tính tổng các vector trong đó. Cuối cùng là xây dựng vector đặc trưng cảm xúc cho cả văn bản.

Quá trình sẽ chuẩn hóa ba vector:

– Vector tổng (G): là vector chứa tất cả các đặc trưng cảm xúc của ý kiến. Các phần tử cảm xúc của G có thứ tự như trong văn bản gốc.

– Vector lớp tích cực P (positive): là vector tập hợp tất cả các đặc trưng cảm xúc có thứ tự như trong văn bản gốc, trong đó các phần tử của các vector lớp negative suy biến bằng 0.

– Vector lớp tiêu cực N (negative): là vector

tập hợp tất cả các đặc trưng cảm xúc có thứ tự như trong văn bản gốc, trong đó các phần tử của các vector lớp positive suy biến bằng 0.

Để phân cực cảm xúc cho văn bản, chúng tôi tính độ tương đồng của G, P và N theo từng cặp: Sim(G, P) và Sim(G, N) theo công thức (3.1):

$$Sim(X,Y)=Cos\sin(X,Y)=\frac{X.Y}{|X|.|Y|}=\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (3.1)$$

Trong đó X, Y là 2 vector với:

$X = \{x1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\}$.

So sánh hai giá trị độ tương đồng của các cặp vector trên để xác định G tương đồng với P hoặc N. Vector G sẽ tương đồng với vector thành phần có giá trị độ tương đồng lớn hơn. Nếu hai giá trị là bằng nhau chúng tôi xếp ý kiến vào lớp trung hòa (neutral).

3.2 Tóm tắt văn bản

Mỗi ý kiến là của mỗi độc giả khác nhau và thường ngắn gọn. Nhưng số lượng ý kiến rất lớn. Do đó, việc tóm tắt các ý kiến là rất cần thiết cho khai thác thông tin. Chúng tôi sẽ dùng phương pháp tóm tắt rút trích (Extract Summarization) để tóm tắt các ý kiến. Việc lượng giá độ quan trọng sẽ dựa theo trọng số của các thực thể đặc trưng cảm xúc và đặc trưng ngữ nghĩa. Bản tóm tắt là những câu có độ quan trọng cao, số lượng câu tùy thuộc tỷ lệ rút gọn thiết lập cho hệ thống.

3.3. Tóm tắt ý kiến dựa trên phân loại cảm xúc

Theo mô hình ở hình 1, mô đun *Tóm tắt* làm việc sau quá trình phân cực cảm xúc. Mô đun này sẽ rút gọn những ý kiến có nội dung dài, để có thể khai thác chúng hiệu quả hơn, mà không làm thay đổi kết quả phân lớp của quá trình trước đó.

Khi tóm tắt văn bản cảm xúc, chúng tôi đánh giá độ quan trọng của câu không chỉ bằng trọng số cảm xúc của mục từ trong từ điển cảm xúc, mà các thành phần ngữ nghĩa cũng được xem xét. Mô hình đề xuất sẽ thử nghiệm phương pháp kết hợp cả yếu tố cảm xúc và yếu tố ngữ nghĩa (nội dung) của ý kiến để đánh giá độ quan trọng của câu trong ý kiến. Từ đó, mô hình chọn được những câu quan trọng nhất cho bản tóm tắt.

Để xác định yếu tố ngữ nghĩa, các thực thể quan trọng của *Bài báo* được rút trích bằng phương pháp mô tả phần 3.1.2. Các thực thể được rút trích là những đối tượng có tên và các thực thể được chúng tôi xem là quan trọng, xuất hiện từ 2 lần trở lên, làm thành tập thực thể có yếu tố ngữ nghĩa, đặc trưng cho đối tượng chủ đề dùng cho tóm tắt ý kiến ở khía cạnh nội dung.

Rút trích các đối tượng có tên là một bài toán khá phức tạp. Chúng tôi xem các bài báo là những văn bản chuẩn mực. Tức là, xác suất rất cao các đối tượng có tên sẽ được viết hoa. Do đó, chúng tôi sẽ rút trích các đối tượng được đặt tên theo nguyên tắc là các từ viết hoa. Do đặc trưng đặt và gọi tên trong tiếng Việt, một đối tượng có tên là một cụm từ thì có thể được gọi bằng một hoặc hai từ sau cùng, tính từ phải qua trái.

Ví dụ: Đối tượng "Vũ Lê Ngô" có thể được gọi là "Ngô", "Lê Ngô", hay đầy đủ là "Vũ Lê Ngô". Do đó, với mỗi đối tượng có tên, chúng tôi sẽ tạo một tập con gồm các từ và cụm từ kết hợp từ phải qua trái. Cụ thể, với "Vũ Lê Ngô" thì tập con sẽ là {"Ngô", "Lê Ngô", "Vũ Lê Ngô"}.

Để rút trích các câu có nội dung quan trọng cho bản tóm tắt, chúng tôi dựa vào hai tiêu chí:

Tiêu chí về ngữ nghĩa của thực thể

– Các đại từ như: anh, chị, ông, bà, anh ấy, ông ấy, chúng nó, họ, …được bổ sung vào tập

thực thể đặc trưng cho các đối tượng có tên của Bài báo. Cùng với các đối tượng có tên, các đại từ này nếu xuất hiện trong câu sẽ làm nội dung ý kiến hướng đến đối tượng chủ đề rõ ràng hơn.

– Chúng tôi không quan tâm đến tần suất xuất hiện của thực thể ngữ nghĩa trong ý kiến. Mỗi thực thể xuất hiện được gán giá trị một (1) vào tập thực thể đặc trưng ngữ nghĩa của câu.

– Các câu có nhiều yếu tố nội dung (liên kết đến bài báo chủ đề) cũng cần được đánh giá cao trong chọn lựa để rút trích.

Tiêu chí về trọng số cảm xúc

Do trọng số cảm xúc của một thực thể có thể có giá trị âm hoặc dương, nên khi tóm tắt, các câu có nhiều thực thể cảm xúc (dương hoặc âm), sẽ được ưu tiên chọn. Tiêu chí này đạt được khi hệ thống chỉ lấy độ lớn của trọng số cảm xúc.

Cụm từ có mức cảm xúc cao là rất quan trọng. Chủ đề có thể được nhấn mạnh bởi sự xuất hiện thường xuyên của từ khóa nhất định, còn cảm xúc tổng thể có thể không tăng lên nếu lặp lại sự xuất hiện của một số thực thể. Do đó, câu có số lượng ít các thực thể cảm xúc, nhưng chúng lại có vai trò lớn (trọng số cảm xúc cao) cần được chọn cho bản tóm tắt để cung cấp thêm thông tin về cảm xúc. Để hệ thống ghi nhận yếu tố này, cần khuếch đại các trọng số cảm xúc bằng phép bình phương mỗi trọng số cảm xúc trước khi tính tổng.

Từ các tiêu chí phân tích ở trên, bài báo đề xuất công thức (3.2) tính độ quan trọng của câu:

Gọi: $x_1, x_2, \ldots, x_n$ là các trọng số ngữ nghĩa của các thực thể trong câu,

$y_1, y_2, \ldots, y_n$ là các trọng số cảm xúc của các thực thể trong câu.

Độ quan trọng của câu xác định theo công thức:

$$W = \left(\sum_{i=1}^{n} x_i\right)^2 + \sum_{i=1}^{n} y_i^2 \qquad (3.2)$$

Xét một ý kiến ví dụ về chiếc điện thoại Passport của hãng BlackBerry:

"Chiếc Passport cực kỳ ấn tượng ngay từ cái nhìn đầu tiên. Phong cách thiết kế lịch lãm, cuốn hút và cá tính làm cho chiếc BlackBerry này không lẫn với ai. "

Các đặc trưng ngữ nghĩa, cảm xúc và trọng số của chúng trong mỗi câu của ý kiến như sau:

– Câu 1: x1 = 1 ("Passport"), y1 = 3 ("cực kỳ ấn tượng").

– Câu 2: x1 = 1 ("BlackBerry"); y1 = 1 ("lịch lãm"), y2 = 1 ("cuốn hút"), y3 = 1 ("cá tính").

Tổng trọng số đặc trưng cả hai câu đều là 4; tổng trọng số đặc trưng cảm xúc cả hai câu đều là 3. Độ quan trọng tính bằng công thức (3.2) cho mỗi câu lần lượt là: $W_1 = 10$, $W_2 = 4$. Độ quan trọng của câu 1 cao hơn do thực thể cảm xúc "cực kỳ ấn tượng" có trọng số bằng 3 thể hiện vai trò khi được khuếch đại.

Sau khi tính độ quan trọng cho tất cả các câu của ý kiến, chúng sẽ được xếp theo thứ tự giảm dần của trọng số *W*. Hệ thống sẽ chọn từ trên xuống số câu theo tỷ lệ tóm tắt người dùng mong muốn. Với các ý kiến chỉ có một câu thì sẽ mặc nhiên được chọn, không cần qua mô đun *Tóm tắt*.

4. KẾT QUẢ THỰC NGHIỆM

Với mô hình trình bày ở hình 1. Chúng tôi tiến hành thử nghiệm trên tập dữ liệu gồm 220 ý kiến đối với 7 bài báo thuộc chủ đề Kinh doanh và chủ đề Xã hội, như phân loại của trang VNExpress, địa chỉ http://www.vnexpress.net.

Đây là trang báo mạng có lượng người đọc rất lớn. Với những vấn đề được quan tâm, có bài báo được hàng nghìn độc giả đưa ý kiến tranh luận.

4.1 Nguồn ngữ liệu thực nghiệm

Số liệu dữ liệu thử nghiệm như trong bảng 1.

Bảng 1: Số liệu nguồn ngữ liệu thực nghiệm

Bài báo	**Số lượng**	**Số ý kiến**
Chủ đề xã hội	3	79
Chủ đề kinh doanh	4	141
Tổng	**7**	**220**

Bài báo có nhiều ý kiến nhất là 59, và ít nhất là 14 ý kiến. Trung bình mỗi bài báo có khoảng 30 ý kiến. Lượng dữ liệu thử nghiệm này không lớn, nhưng phù hợp để có thể kiểm nghiệm kỹ sự vận hành của hệ thống trong giai đoạn đầu.

4.2 Phương pháp đánh giá thực nghiệm

Để đánh giá hiệu quả của mô hình đề xuất, chúng tôi sử dụng độ chính xác và độ truy hồi.

* *Độ chính xác (Precision).*

Được tính bởi công thức: $precision = \frac{c}{b}$ (4.1)

* *Độ truy hồi (Recall)*

Được tính bởi công thức: $recall = \frac{c}{a}$ (4.2)

Với *a* là số câu đúng của bản tóm tắt (theo tập tóm tắt mẫu), *b* là số câu của bản tóm tắt do máy tính thực hiện và *c* là số câu giao giữa *a* và *b*.

4.3 Phân loại cảm xúc

Bảng 2 trình bày kết quả thực nghiệm:

Bảng 2: Kết quả đánh giá ý kiến.

Lớp	Độ đúng đắn (%)	Độ chính xác (%)	Độ truy hồi (%)
Positive	74,57	80,41	75,73
Neutral		68,18	50,00
Negative		65,63	38,89

Từ bảng 2, chúng tôi có một số nhận xét sau:

– Độ đúng đắn (Accuracy) đạt 74,57% cho thấy mô hình đề xuất là hiệu quả. Kết hợp Naïve Bayes và Vector Space Model là mô hình khá triển vọng cho phân tích cảm xúc.

– Độ chính xác (Precision) trong cả 3 lớp có kết quả khá tốt, trên 65%. Độ chính xác của lớp negative thấp hơn nhiều so với lớp positive.

– Độ truy hồi (Recall) lớp position có kết quả khá tốt. Lớp neutral và negative có kết quả khá thấp, nhất là negative.

– Nguyên nhân Precision và Recall thấp là do từ điển cảm xúc còn hạn chế, chưa phủ đầy đủ các cách diễn đạt cảm xúc, nhất là dạng phủ định.

– Khi xét riêng từng chủ đề, kết quả thực nghiệm được thể hiện ở bảng 3 và bảng 4.

Bảng 3: Kết quả đánh giá ý kiến chủ đề kinh doanh

Lớp	Độ đúng đắn (%)	Độ chính xác (%)	Độ truy hồi (%)
Positive	72,28	82,50	61,11
Neutral		64,71	57,89
Negative		66,67	48,65

Bảng 4: Kết quả đánh giá ý kiến chủ đề xã hội

Lớp	Độ đúng đắn (%)	Độ chính xác (%)	Độ truy hồi (%)
Positive	77,78	78,95	91,84
Neutral		80,00	36,36
Negative		60,00	17,65

Các kết quả này thể hiện:

– Độ đúng đắn của phân cực cảm xúc các bài ý kiến chủ đề xã hội tốt hơn chủ đề kinh doanh khá nhiều, đến trên 5.5%. Có thể nguyên nhân chính dẫn đến điều này là sự phù hợp của từ điển cảm xúc đối với chủ đề. Bên cạnh đó, với chủ đề xã hội, chúng tôi nhận thấy độc giả thể hiện cảm xúc nhiều hơn so với chủ đề kinh doanh. Bài báo chủ đề kinh doanh nói về điện thoại thông minh (smartphone), máy tính mới, nên thu hút nhiều độc giả trẻ. Còn các vấn đề xã hội có nhiều thành phần và độ tuổi khác nhau quan tâm.

– Độ truy hồi của cả hai chủ đề là khá thấp. Trong đó, độ truy hồi của lớp negative là thấp hơn nhiều so với lớp positive và neutral. Nguyên nhân như chúng tôi nói ở trên (mục 4.3) có thể cũng là yếu tố chính trong trường hợp này.

4.4 Tóm tắt trên cơ sở phân loại cảm xúc

Tóm tắt văn bản thường dùng hai phép đo phổ biến để đánh giá hiệu năng là độ chính xác (P) và độ truy hồi (R) (công thức 4.1 và 4.2). Chúng tôi sẽ dùng hai độ đo này để đánh giá hiệu năng của mô đun *Tóm tắt.* Kết quả trình bày ở bảng 5.

Để đánh giá sự hiệu quả của phương pháp đề xuất với công thức 3.2, ngoài thử nghiệm với phương pháp trên, bài báo còn thử nghiệm tóm tắt chỉ dựa trên yếu tố cảm xúc. Độ quan trọng của câu được lượng giá bằng trọng số của các thực thể cảm xúc. Các trọng số cảm xúc được bình phương trước khi tính tổng. Nguyên tắc này tương đương công thức 3.2, nhưng triệt tiêu yếu tố ngữ nghĩa ($\left(\sum_{i=1}^{n} x_i\right)^2$). Kết quả thể hiện ở bảng 6.

Bảng 5: Kết quả đánh giá quá trình tóm tắt ý kiến kết hợp ngữ nghĩa và cảm xúc.

Bài báo	*Số câu*	*a*	*b*	*c*	*P (%)*	*R (%)*
Chủ đề xã hội	128	82	84	76	90,48	92,68
Chủ đề kinh doanh	247	165	167	150	88,76	90,91
Tổng hợp	375	247	251	226	90,04	91,50

Bảng 6: Kết quả đánh giá quá trình tóm tắt ý kiến chỉ dựa vào cảm xúc.

Bài báo	*Số câu*	*a*	*b*	*c*	*P (%)*	*R (%)*
Chủ đề xã hội	128	82	84	73	86,90	89,02
Chủ đề kinh doanh	247	165	167	146	87,43	88,48
Tổng hợp	375	247	251	219	87,25	88,66

Thực nghiệm tóm tắt ý kiến theo tỷ lệ rút gọn 50%. Trong đó, ý kiến chỉ có một câu sẽ được giữ nguyên, không qua mô đun *Tóm tắt* để xử lý.

Một số nhận xét từ kết quả ở hai bảng 5 và 6:

– Số câu đúng trong bản tóm tắt a không đổi do tập dữ liệu không đổi. Số câu của bản tóm tắt do máy tính thực hiện b là như nhau do chúng tôi không thay đổi thiết lập tỷ lệ rút gọn ý kiến.

– Thành phần c, trong mọi trường hợp đánh

giá, đều có kết quả cao hơn nếu kết hợp cả yếu tố cảm xúc và yếu tố ngữ nghĩa (bảng 5) khi so với trường hợp chỉ sử dụng yếu tố cảm xúc (bảng 6).

– Mô hình đề xuất đánh giá độ quan trọng của câu ở phần 3.3, với công thức 3.2 (bảng 5) cho kết quả tốt hơn trong mọi trường hợp. Bằng kết quả thực nghiệm có thể kết luận phương pháp đề xuất bài báo đã trình bày có tính hiệu quả hơn.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đưa ra cách tiếp cận tóm tắt ý kiến dựa trên phân tích cảm xúc bằng cách kết hợp hai mô đun độc lập là *Phân loại cảm xúc* và *Tóm tắt* cho văn bản tiếng Việt. Với mô đun *Tóm tắt*, bài báo đã đề xuất tiêu chí đánh giá độ quan trọng của câu dựa trên ngữ nghĩa của thực thể và trọng số cảm xúc của câu (mục 3.3) để rút trích cho bản tóm tắt. Kết quả thu được từ thực nghiệm cho thấy đây là cách tiếp cận khá triển vọng cho nhu cầu khai thác ý kiến một cách hiệu quả với một đối tượng, hoặc vấn đề trên mạng.

Trong tương lai, chúng tôi sẽ nâng cao khả năng phân loại cảm xúc bằng cách phân chia nhiều mức trong mỗi lớp cảm xúc. Khả năng tóm tắt ý kiến cũng được cải thiện hơn bằng việc khai thác ngữ nghĩa của thực thể kết hợp với xác định cảm xúc một cách hiệu quả hơn. Vấn đề rút trích sẽ được quan tâm nghiên cứu hướng đến đặc trưng của ngôn ngữ tiếng Việt nhằm nâng cao hiệu quả của mô hình đã đề xuất.

Text summarization based on sentiment classification of comments from online Vietnamese newspaper

- **Nguyen Ngoc Duy** [(1)]
- **Phan Thi Tuoi** [(2)]

[(1)] Posts and Telecommunications Institute of Technology
[(2)] Ho Chi Minh city University of Technology,VNU-HCM

ABSTRACT

To know opinions of consumers regarding products or public about important problems in society, then the best and most effective way is to exploit information of community from Internet and social network. Today is an era of information explosion through Internet and

social networking, so we are able to exploit effectively information from the huge sources. The opinion of individuals is not only objective information but also contains emotions of the author. It through Internet has big power to make a stream of public opinion that will impact on network community. This is really an enormous subjective information resource, then it will have great meaning for many areas, such as economics, politics, society and culture if we have methods and techniques to exploit it effectively. An automatic system classifying comments based on sentiment is really necessary to exploit efficiently this resource. In order to support users have more concise and appropriate information, then question of summary information should be studied and solved, especially on side of the views and sentiments of each opinion.

To exploit the resource effectively to summary information, the paper will propose a text Vietnamese summary model, not only based on semantics but also based on sentiment features. We have built a base model to solve this problem. We have exploited and developted methods summarizing and sentiment analysing for our proposed model. Our system can draw Vietnamese comments from online Vietnamese newspaper, analyze the sentiments of comments, classify them and make a summary of opinions effectively.

Keywords: Sentiment *Analysis, Opinion Mining, Text Summarization, Sentiment Classification.*

TÀI LIỆU THAM KHẢO

[1]. Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, 2, 1-2, 1–135 (2008).

[2]. Balahur, A.; Kabadjov, M.;Steinberger, J.; Steinberger, R.; Montoyo, A., "Summarizing Opinions in Blog Threads", *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 606-613 (2009).

[3]. Vo Ngoc Phu, Phan Thi Tuoi, "Sentiment Classification using enhanced Contextual Valence Shifters", *Proceedings of International Conference on Asian Language Processing, Malaysia* (2014).

[4]. Tien-Thanh Vu, Huyen-Trang Pham, Cong-To Luu, Quang-Thuy Ha, "A Feature-based Opinion Mining Model on Product Reviews in Vietnamese", *Workshop on Semantic Methods for Knowledge Discovery and Communication*, 23-33 (2011).

[5]. Quang-Thuy Ha, Tien-Thanh Vu, Huyen-Trang Pham, Cong-To Luu, "An Upgrading Feature-based Opinion Mining Model on Vietnamese Product Reviews", *Proceedings of the 7th International Conference on Active Media Technology*, 173-185 (2011).

[6]. Tung-Hui Chiang, Jing-Shin Chang, Ming-Yu Lin, Keh-Yih Su, "Statistical Models for Word Segmentation and Unknown Word Resolution", *Proceedings of 1992 R.O.C. Computational Linguistics Conference (ROCLING V),* 121-146 (1992).