MỘT SỐ KHÁI NIỆM VÀ HƯỚNG TIẾP CẬN PHÂN TÍCH CẢM XÚC - ÁP DỤNG CHO TIẾNG VIỆT

Trần Khải Thiện, Tiếu Phùng Mai Sương
Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ-Tin học TP.HCM
thientk@huflit.edu.vn, suong.tpm@huflit.edu.vn

TÓM TẮT

Phân tích cảm xúc (hay còn gọi là phân tích ý kiến, khai phá quan điểm) hiện đang là bài toán nhận được rất nhiều sự quan tâm trong nghiên cứu và cả trong doanh nghiệp. Cho đến nay, các công trình về phân tích cảm xúc trong tiếng Việt đã có những đóng góp đáng kể cho cộng đồng khoa học và cho kinh tế. Bài báo này thực hiện việc khảo sát và đưa ra những phân tích về các vấn đề xoay quanh bài toán phân tích cảm xúc tiếng Việt, bao gồm: các công trình đáng chú ý, nguồn tài nguyên, và những ứng dụng điển hình.

Từ khóa: *phân tích ý kiến, phân tích cảm xúc, khai phá quan điểm, tiếng Việt.*

1. Giới thiệu

Theo B. Liu [1], một cảm xúc hay quan điểm được định nghĩa bằng một bộ gồm 5 thành phần:

$$(E_i, A_{ij}, S_{ijkl}, H_k, T_l) \quad (1)$$

với:

- E_i tên của thực thể,
- Aij khía cạnh Ei,
- Sijkl là ý kiến cảm xúc về khía cạnh Aij của thực thể Ei cho bởi chủ thể Hk tại thời điểm Tl,
- Hk là chủ thể thể hiện ý kiến,
- Tl thời gian thể hiện ý kiến của chủ thể H_k.

Trong định nghĩa trên, S_{ijkl} có thể là ý kiến tích cực, tiêu cực, trung lập, hoặc cũng có thể là một độ đo mô tả mức độ của tình cảm trong nhận xét như thang độ 1-5 sao của đánh giá Amazon. Thực thể E_i có thể là sản phẩm, dịch vụ, sự kiện hay các chủ đề.

Ví dụ: Một người dùng tên Nam tạo một nhận xét vào ngày 23/10/2017 như sau: "Tôi mua con Macbook vài ngày trước. Nó quả là cái máy vi tính tuyệt vời. Màn hình cực nét. Tuy nhiên, giá lại hơi cao." Có ba bộ cảm xúc trong đoạn nhận xét này, thể hiện bởi Bảng 1.

Bảng 1: *Ví dụ về định nghĩa quan điểm, cảm xúc*

Đối tượng/ khía cạnh (E/A)	**Tình cảm (S)**	**Chủ thể (H)**	**Thời gian (T)**
Macbook	Tích cực	Nam	23/10/2017
Màn hình	Tích cực	Nam	23/10/2017
Giá tiền	Tiêu cực	Nam	23/10/2017

Dựa trên định nghĩa về quan điểm/cảm xúc, phân tích cảm xúc nhắm tới việc phát hiện các bộ cảm xúc trong văn bản mà vì thế các bài toán phân tích cảm xúc được sinh ra xoay quanh việc phát hiện khai thác 5 thành phần của bộ cảm xúc. Ví dụ như phân tích cảm xúc mức câu, văn bản hướng đến thành phần thứ ba là ý kiến cảm xúc (tích cực, tiêu cực, trung lập) mà không quan tâm đến các thành phần khác. Trích xuất các cảm xúc thang độ mịn quan tâm đến 4 thành phần đầu tiên của bộ cảm xúc. Trong khi đó, phân tích cảm xúc mức khía cạnh lại chỉ quan tâm đến thành phần thứ 2 và thứ 3.

Trong bài báo này, chúng tôi giới thiệu về các nghiên cứu đáng chú ý, các nguồn tài nguyên, ứng dụng và đưa ra các đánh giá, phân tích cũng như trình bày các cơ hội và thách thức của phân tích cảm xúc tiếng Việt.

Phần còn lại của bài báo được trình bày như sau: Mục II sẽ giới thiệu về ba bài toán chính trong phân tích cảm xúc. Mục III đề cập đến các tài nguyên cho phân tích cảm xúc. Mục IV nhóm tác giả sẽ nói về các ứng dụng của phân tích cảm xúc và cuối cùng, các kết luận được đề cập tại Mục V.

2. Ba bài toán chính trong phân tích quan điểm

2.1. Phân tích cảm xúc mức từ, cụm từ, xây dựng từ điển

Phân tích cảm xúc mức từ, cụm từ liên quan đến việc xác định độ đo cảm xúc cho từ, cụm từ trong văn bản. Phân tích cảm xúc mức từ, cụm từ là nhiệm vụ then chốt, cung cấp ngữ liệu cho các bài toán phân tích cảm xúc các mức độ khác như mức khía cạnh hay mức văn bản.

Bảng 2: *So sánh các công trình phân tích quan điểm mức từ*

Công trình	Phương pháp	Hạn chế	Điểm mạnh
T. T. Vu & cộng sự [2]	Sử dụng từ điển xây dựng thủ công từ SentiWordnet tiếng Anh.	Từ điển xây dựng thủ công. Số lượng từ ít (1.179 từ). Điểm số cảm xúc của từ được gán từ điểm số của từ tiếng Anh tương ứng.	Cho kết quả cao trong miền các nhận xét về sản phẩm điện thoại di động.
S. Trinh & cộng sự [3]	Dựa vào phân tích cảm xúc tiếng Anh và điều chỉnh lại cho phù hợp với tiếng Việt.	Điểm số cảm xúc của từ và cụm từ được gán hoàn toàn từ điểm số của từ, cụm từ tiếng Anh tương ứng.	Quan tâm đến việc tính toán cảm xúc cho cả cụm từ. Quan tâm đến các từ loại (danh từ, động từ, tính từ, phó từ)
H. Nam & cộng sự [4]	Dịch SentiWordnet Anh-Việt. Sau đó sử dụng từ điển Việt-Việt để lọc bỏ từ gây nhiễu. Sử dụng WordNet score propagation algorithm để gán điểm số các term.	Phụ thuộc vào nhóm ngành cụ thể.	Có xử lý slang words và từ thuộc từng nhóm ngành. Số lượng từ lớn (hơn 26.000 từ)
H. Q. V. Vo & cộng sự [5]	Chuyển ngữ từ 3 từ điển Tiêng Anh sau đó sử dụng luật và kiểm tra thủ công.	Xử lý thủ công.	Quan tâm đến ngữ cảnh của từ trong văn bản.
T. K. Tran & cộng sự [6]	Sử dụng kết hợp các phương pháp thủ công, hồi quy logistic và tính toán logic mờ dựa trên đặc trưng ngôn ngữ tiếng Việt	Phụ thuộc vào các công cụ tiền xử lý. Chưa quan tâm đến danh từ, cụm danh từ mang cảm xúc.	Phân lớp mịn. Quan tâm đến các từ loại, từ lóng. Xử lý được các cụm từ. Số lượng từ lớn.

SentiWordNet [7] hiện là từ điển cảm xúc lớn nhất, giúp xác định giá trị cảm xúc của từ cho nhiều ngôn ngữ phổ biến như tiếng Anh, Pháp, Hoa. SentiWordNet miễn phí cho mục đích nghiên cứu, từ điển này được xây dựng dựa trên WordNet [8] bằng học máy bán giám sát. SentiWordNet đã từng được phát triển cho tiếng Việt qua công trình của T. T. Vu và cộng sự [2] với khoảng 1.000 từ mang cảm xúc. Công trình của H. Nam và cộng sự [4] xây dựng từ điển cảm xúc cho miền sản phẩm dựa vào sự kết hợp giữa các phương pháp thống kê, dịch máy và bản thể luận (ontology) WordNet. Tác giả S. Trinh và cộng sự [3] xây dựng từ điển cảm xúc tiếng Việt gồm năm từ điển nhánh cho danh từ, động từ, tính từ, trạng từ, và các đặc trưng khác, trên cơ sở từ các từ điển nhánh đã được nhóm tác giả nước ngoài phát triển cho tiếng Anh. Công trình [5] thực hiện việc chuyển ngữ từ ba từ điển tiếng Anh thành tiếng Việt kết hợp với sử dụng luật và kiểm tra thủ công để xây dựng từ điển cảm xúc tiếng Việt với khoảng 6.000 từ. T. K. Tran và cộng sự [6] đề xuất các luật mờ để tính toán giá trị cảm xúc cho cụm từ tính từ và động từ khi mà trước đó nhóm tác giả xây dựng từ điển lõi qua việc kết hợp chuyển ngữ từ SentiWordnet và hồi quy logistic.

Các công trình trên đã có đóng góp cho cộng đồng khoa học trong nước tuy nhiên còn tồn tại một số vấn đề như: lượng từ vựng không nhiều [2], [5]; phụ thuộc miền ứng dụng [4]; chưa quan tâm tính toán giá trị cảm xúc cho cụm từ [2], [4] và điều này được [3] xử lý nhưng cách mà các tác giả tính toán giá trị cảm xúc cho các hedges (rào cản ngôn ngữ) tiếng Việt lại dựa hoàn toàn vào cách tính của tiếng Anh. Công trình [6] chưa đưa ra giải pháp cho cụm danh từ khi mà các cụm từ này cũng đóng góp vị trí quan trọng trong phân tích cảm xúc tiếng Việt.

2.2. Phân tích cảm xúc mức văn bản

Trong hai thập kỷ qua, các phương pháp dựa vào học máy đã thống trị trong hầu hết các bài toán phân tích cảm xúc, đặc biệt là bài toán mức văn bản bởi việc biểu diễn các đặc trưng có tác động lớn đến hiệu năng của giải thuật học máy [9]. Các nghiên cứu đã tập trung vào việc tạo ra tập đặc trưng hiệu quả dựa vào hiểu biết về miền và các kỹ thuật chuyên biệt điển hình như các công trình [10], [11]. Trong đó các tác giả sử dụng ba phương pháp học máy truyền thống là Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME) học trên tập dữ liệu với đặc trưng n-gram cho kết quả thực nghiệm tương đối cao ngay cả với phân lớp nhị phân hay phân lớp nhiều cấp độ.

Tuy nhiên, công việc này hiện có thể được thực hiện tốt bởi các giải thuật học biểu diễn (representation learning) hay còn gọi là học đặc trưng (feature learning) như các hướng tiếp cận theo học sâu, kỹ thuật tự động phân biệt và giải thích các biểu diễn văn bản từ dữ liệu. Học sâu đã nổi lên do khả năng tạo ra các biểu diễn dữ liệu ở nhiều cấp độ. Trong đó phải kể đến công trình của D. Nguyen và cộng sự [12] khi tận dụng các văn bản tiếng Anh được dịch làm dữ liệu huấn luyện, các văn bản tiếng Việt được dịch bằng máy sang các văn bản tiếng Anh rồi được rút trích đặc trưng. Công trình sử dụng mạng CNN (Convolutional Neural Network) để phân lớp văn bản. Kết quả thực nghiệm đạt 84.40% trong tổng số 25.991 nhận xét sản phẩm tiếng Việt. Trong [13], tác giả Q. Vo và cộng sự sử dụng kết hợp hai mạng LSTM (Long Short-Term Memory) và CNN để phân lớp 17.500 nhận xét tiếng Việt theo ba mức khen, chê và trung tính khi nhận thấy rằng CNN hoạt động tốt trong việc bắt được các mối quan hệ lân cận nhau trong văn bản và LSTM với cơ chế nhớ-quên có thể xử lý được các mối phụ thuộc ở khoảng cách xa trong văn bản.

Bảng 3: *So sánh các công trình phân tích quan điểm mức văn bản*

Công trình	Phương pháp	Mô hình/ kỹ thuật	Bộ phân lớp
N. T. Duyen và cộng sự [10]	Học máy	Sử dụng 3 kỹ thuật SVM, Naïve Bayes, và Maximum Entropy. Đặc trưng n-gram.	2 lớp khen, chê.
T. K. Tran và cộng sự [11]	Học máy	Sử dụng 3 kỹ thuật SVM, Naïve Bayes, và Maximum Entropy. Đặc trưng n-gram.	3 lớp khen, chê, trung tính.
D. Nguyen và cộng sự [12]	Dịch máy kết hợp học sâu	Word embedding, CNN.	2 lớp khen, chê.
Q. Vo và cộng sự [13]	Học sâu	Word embedding, LSTM-CNN.	3 lớp khen, chê, và trung tính.

2.3. Phân tích cảm xúc mức khía cạnh

Phân tích cảm xúc mức khía cạnh là phân tích cảm xúc của người dùng nhắm vào các đối tượng là các khía cạnh, đặc trưng hay thuộc tính của một hay nhiều thực thể trong một văn bản cho trước. Phân tích quan điểm mức khía cạnh được nhiều sự quan tâm hơn cả với một số đông công bố, điển hình như các công trình của T. T. Nguyen và cộng sự [14] sử dụng phương pháp lai, T. K. Tran và cộng sự [15], [16] sử dụng cú pháp phụ thuộc, L. Mai và cộng sự [17] cũng như D. Nguyen và cộng sự [18] sử dụng học sâu. Việc xác định khía cạnh và từ mang cảm xúc là nhiệm vụ trọng tâm của bài toán phân tích cảm xúc mức khía cạnh. Vì vậy phân tích ở mức khía cạnh có hai tác vụ chính: 1) xác định

và rút trích các khía cạnh được nhận xét, và 2) xác định trị cảm xúc liên quan đến các khía cạnh tương ứng.

-Xác định và rút trích khía cạnh

Nhóm tác giả T. T. Nguyen và cộng sự [14] đã đề xuất phương pháp tiếp cận lai là xây dựng mô hình phân tích cảm xúc bao gồm đồ thị khái niệm (concept graph), ontology, 64 luật cảm xúc dựa trên biểu thức chính quy và phương pháp học máy để phân lớp khen hay chê. Đồ thị khái niệm và ontology phục vụ cho quá trình phân tích tự động các cấu trúc đơn giản của ngôn ngữ tự nhiên. Trong khi đó các luật cảm xúc giúp cho hệ thống hiểu được các thành phần của ngôn ngữ, giúp xử lý được các dạng câu so sánh, rút trích được một số các khía cạnh không tường minh. Trong quá trình rút trích, có thể xuất hiện nhiều từ khác nhau chỉ cùng một khía cạnh, ví dụ như 'phòng', 'phòng ốc', 'căn phòng' cùng đề cập về một khía cạnh là 'phòng', khi này hệ thống phải thực hiện việc 'gom nhóm' các khía cạnh này. Tác vụ gom nhóm khía cạnh thường được giải quyết nhờ vào xây dựng ontology cho miền chuyên biệt như công trình [14] hoặc sử dụng học máy bán giám sát để gom nhóm các khía cạnh như trong công trình [2].

-Xác định cảm xúc

Xác định cảm xúc cho từng khía cạnh là tác vụ thứ hai trong quá trình phân tích cảm xúc mức khía cạnh. Các tác giả thường sử dụng tập các từ cảm xúc, khía cạnh trong mỗi câu có thể được xác định bằng cách cộng dồn các điểm số của từ mang cảm xúc liên quan, nếu tổng điểm là lớn hơn 0 thì khía cạnh đó mang cảm xúc tích cực và ngược lại nhỏ hơn 0 là khía cạnh tiêu cực như đề xuất của Taboada cùng cộng sự với phương pháp SO-CAL [19] (The Semantic Orientation CALculator) sử dụng từ điển của các từ được gán nhãn cảm xúc cùng trọng số. SO-CAL tỏ ra hiệu quả trong nhiệm vụ phân lớp cảm xúc cho văn bản. Một số công trình tận dụng phương pháp này như [2], [16], [20].

-Khai thác mối quan hệ phụ thuộc của các từ để rút trích đồng thời khía cạnh và từ cảm xúc

Bên cạnh việc thực hiện độc lập hai tác vụ rút trích khía cạnh và xác định giá trị cảm xúc, nhiều công trình lựa chọn cách tiếp cận rút trích đồng thời khía cạnh và từ mang cảm xúc.

Với phương pháp dựa trên luật có thể kể đến công trình của T. K. Tran và cộng sự trong [15] lấy ý tưởng của công trình G. Qiu và cộng sự [21], [22] đề xuất giải thuật "truyền kép" (double propagation) để rút trích khía cạnh và từ cảm xúc từ việc quan sát mối quan hệ phụ thuộc giữa chúng. Các mối quan hệ này được xác định bởi bộ phân tích cú pháp của văn phạm phụ thuộc. Ban đầu các tác giả sử dụng từ mồi (seed word) để rút trích các từ mang cảm xúc cùng các khía cạnh. Các từ mang cảm xúc và các khía cạnh mới này lại được dùng để rút trích các từ mang cảm xúc và các khía cạnh tiếp theo. Quá trình được tiếp diễn cho đến khi không tìm được từ mang cảm xúc nào khác nữa. Trong [16], các tác giả đề xuất một mô hình phân tích cảm xúc mức khía cạnh cho các nhận xét tiếng Việt, kết hợp từ điển cảm xúc và các luật văn phạm phụ thuộc để rút trích các cặp từ, cụm từ mang mối quan hệ (cảm xúc - khía cạnh). T. T. Nguyen và cộng sự [14] đã xây dựng 64 luật rút trích khía cạnh và cảm xúc tương ứng dựa trên biểu thức chính quy. Hệ thống xử lý được nhiều dạng cấu trúc câu, phát hiện được nhiều khía cạnh (aspect) không tường minh, và các trường hợp có sự dịch chuyển giá trị cảm xúc trong câu có quan điểm. Các tác giả đã tiến hành các thử nghiệm và cho kết quả tốt hơn so với các kỹ thuật của khai phá dữ liệu (như vector máy học-SVM).

Gần đây, phương pháp học sâu (deep learning) cho phân tích cảm xúc mức khía cạnh đã nổi lên như một mô hình học máy mạnh và tạo được các kết quả rất thuyết phục. Với phương pháp này, có thể kể đến các công trình như L. Mai và cộng sự [17], Đính và cộng sự [18]. Nhóm tác giả trong [17] đề xuất mô hình gọi tên là BRNN-CRF gồm thành phần gán nhãn chuỗi kết hợp với mạng BRNN (Bidirectional Recurrent Neural Networks) và CRF (Conditional Random Fields) để rút trích các đối tượng mang cảm xúc cùng các yếu tố tình cảm tương ứng trong các nhận xét về sản phẩm điện thoại di động. Trước hết dữ liệu được biểu diễn dạng từ nhúng rồi làm đầu vào cho mạng BRNN với tầng truyền ngược (backward layer) để thu thập các thông tin từ quá khứ và tầng truyền thẳng (forward layer) để thu thập các thông tin tương lai. Tiếp theo, lớp CRF sẽ xử lý các thông tin trên như là các đặc trưng để đưa ra các dự đoán. Nhóm tác giả trong [18] sử dụng kết hợp giữa CNN và LSTM. CNN với 64 cửa sổ nhân (kernel windows) mang vai trò lọc ra 64 khía cạnh/ đối tượng được đề cập trong mẫu tin trong khi LSTM để xử lý mẫu tin dài dễ gây nhiễu.

Bảng 4: *So sánh các công trình phân tích quan điểm mức khía cạnh*

Công trình	Phương pháp	Đặc điểm	Điểm mạnh
T. T. Vu & công sự [2]	Sử dụng luật, học máy bán giám sát và từ điển xây dựng thủ công.	Sử dụng luật cú pháp để trích các đặc trưng và từ mang cảm xúc. Các đặc trưng sau đó được gom nhóm bởi học máy bán giám sát HAC kết hợp SVM-kNN. Sử dụng từ điển xây dựng thủ công.	Phát hiện được các đặc trưng không tường minh và đồng tham chiếu bởi tập luật.
T. T. Nguyen và cộng sự [14]	Sử dụng kết hợp nhiều phương pháp.	Luật biểu thức chính quy, đồ thị khái niệm, ontology, và học máy SVM.	Xử lý được nhiều dạng cấu trúc câu, phát hiện được được khía cạnh không tường minh, và các trường hợp có sự dịch chuyển giá trị cảm xúc trong câu có quan điểm.
T. K. Tran và cộng sự [15], [16]	Sử dụng luật	Luật cú pháp phụ thuộc, ontology, từ điển cảm xúc.	Phát hiện được mối liên hệ ngữ nghĩa giữa các từ trong câu văn bản tiếng Việt. Từ đó phát hiện được khía cạnh và từ mang cảm xúc thông qua mối quan hệ phụ thuộc.
L. Mai và cộng sự [17]	Học sâu, Sequence labelling	Gán nhãn chuỗi, Word embedding, BRNN-CRF.	Xử lý được các câu dài.
D. Nguyen và cộng sự [18]	Học sâu	Word embedding LSTM-CNN.	Phát hiện tốt 64 khía cạnh và cảm xúc tương ứng. Xử lý được các văn bản dài, lọc nhiễu tốt.

3. Tài nguyên

Tài nguyên cho phân tích cảm xúc tiếng Việt hiện là vấn đề thách thức của giới nghiên cứu trong lĩnh vực này do còn nhiều hạn chế và do nhiều nhóm nghiên cứu chưa công bố lên mạng để cho phép tải về. Các nguồn tài nguyên quan trọng cần cho bài toán phân tích cảm xúc bao gồm: dữ liệu nhận xét, mô hình từ nhúng (word embedding) được huấn luyện sẵn, và từ điển cảm xúc tiếng Việt.

-Về dữ liệu nhận xét đã gán nhãn: Năm 2015, T.N.Duy và cộng sự [23] đã giới thiệu bộ dữ liệu gồm 4.000 câu về lĩnh vực thiết bị điện tử và nghiên cứu thực nghiệm ý nghĩa của các câu so sánh bao gồm hai bài toán là xác định các câu so sánh và ghi nhận mối quan hệ giữa chúng. Năm 2016, cuộc thi về phân tích cảm xúc do VLSP-2016 tổ chức đã cung cấp 5.000 mẫu nhận xét cho việc huấn luyện và 1.000 nhận xét cho việc kiểm thử về lĩnh vực thiết bị điện tử. Các mẫu nhận xét này được gán nhãn theo ba lớp tích cực, tiêu cực và trung tính (vlsp.org.vn/vlsp2016/eval/sa). Đến năm 2018 có bộ dữ liệu VLSP 2018 datasets về lĩnh vực nhà hàng khách sạn của workshop Vietnamese Language and Speech Processing (vlsp.org.vn/vlsp2018/). Và mới đây, ngữ liệu về nhận xét của sinh viên được tác giả N.L.T. Ngan và cộng sự cung cấp có tên là UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis [24]. Về dữ liệu ẩm thực được nhóm www.streetcodevn.com thu thập từ Foody.vn và cung cấp cho cộng đồng bộ ngữ liệu gán nhãn gồm 50.000 mẫu bình luận.

-Về mô hình từ nhúng Word2Vec cho tiếng Việt đã được huấn luyện sẵn: có các thư viện Word2VECVN của tác giả Vũ Xuân Sơn (github.com/sonvx/word2vecVN), thư viện Word2Vector Vietnamese của nhóm Streetcodevn hay công trình [25] của nhóm P.T.Tuoi và cộng sự.

-Về từ điển cảm xúc tiếng Việt: có các nguồn đã được công bố cho phép tải về như VietSentiWordNet [26], VietSentiLex [5].

4. Ứng dụng

Phân tích cảm xúc giúp chủ thể có thể lắng nghe và hiểu được những gì đang được nói về mình. Các thông tin có thể là:

- Các xu hướng xã hội;

- Thông tin cảm nhận về thương hiệu;

- Thông tin về ngành hàng;

- Phản ứng về sản phẩm;

- Tâm lý hành vi người sử dụng, người mua.

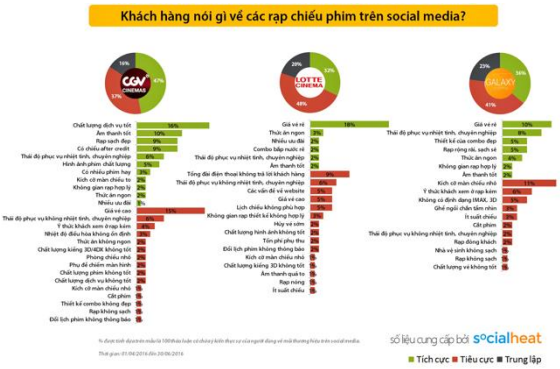
Nắm bắt được các thông tin này giúp chủ thể có thể đo lường các hoạt động tiếp thị, bán hàng cũng như đưa ra được các nghiên cứu thống kê, báo cáo các xu hướng. Một quy trình của một ứng dụng phân tích cảm xúc có thể như sau:

- Thu thập dữ liệu từ mạng xã hội, diễn đàn, và đo lường theo thời gian thực.

- Tự động phân tích và đánh giá các thông tin.

- Hiển thị tự động và đưa ra các phân tích báo cáo.

Các doanh nghiệp, thương hiệu đáng chú ý hoạt động trong lĩnh vực này có thể kể đến YouNet Media (www.younetmedia.com), DAZIKZAK (www.dazikzak.com), và SMCC (www.smcc.vn). Hình 1 mô tả phản hồi của người tiêu dùng về rạp chiếu phim được thống kê từ một công cụ Social Listening của YouNet Media. Giữa ba rạp CGV, Lotte Cinema và Galaxy thì CGV được đánh giá là có chất lượng dịch vụ tốt nhất, tuy nhiên giá vé lại khá cao. Lotte có nhiều chương trình ưu đãi với giá bắp nước rẻ, nhưng lại bị khách hàng phàn nàn về thái độ phục vụ của nhân viên.



Hình 1: *Ý kiến người dùng rạp chiếu phim*
(nguồn: younetmedia.com)

5. Kết luận

Trong bài báo này, chúng tôi đã đề cập đến tình hình nghiên cứu cũng như các nguồn tài nguyên, các ứng dụng về phân tích cảm xúc tiếng Việt. Có thể nói đây là một trong những bài toán nhận được quan tâm nhiều nhất trong cộng đồng nghiên cứu bởi tính ứng dụng và thực tiễn. Chúng tôi cho rằng các kết quả nghiên cứu sẽ tốt hơn nữa khi tài nguyên cho phân tích cảm xúc tiếng Việt phát triển, nhất là trong thời đại bùng nổ về dữ liệu như hiện nay.

TÀI LIỆU THAM KHẢO

[1] B. Liu. Sentiment Analysis and Opinion Mining, *Synth. Lect. Hum. Lang. Technol.*, **5**: 1:1–167, doi: 10.2200/S00416ED1V01Y201204HLT016. May 2012.

[2] T.T. Vu, H. T. Pham, C.T. Luu, Q.T. Ha. A Feature-Based Opinion Mining Model on Product Reviews in Vietnamese, *Springer*, Berlin, Heidelberg, 23–33, 2011.

[3] S. Trinh, L. Nguyen, M. Vo. Combining Lexicon-Based and Learning-Based Methods for

Sentiment Analysis for Product Reviews in Vietnamese Language, Springer, Cham, 57–75, 2018.
[4] H. Nam Nguyen, T. Van Le, H. Son Le, and T. Vu Pham. Domain Specific Sentiment Dictionary for Opinion Mining of Vietnamese Text, Springer, Cham, pp. 136–148, 2014.
[5] H. Q. V. Vo, Kazuhide Yamamoto. VietSentiLex: a sentiment dictionary by considering the polarity of ambiguous sentiment words - Google Search, *The 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, 2018.
[6] T. K. Tran and T. T. Phan. A hybrid approach for building a Vietnamese sentiment dictionary, *J. Intell. Fuzzy Syst.*, **35**:1,967–978, doi: 10.3233/JIFS-172053, Jul. 2018.
[7] F. S. Stefano Baccianella, Andrea Esuli. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," *Lrec*, vol. 10, pp. 2200–2204, 2010.
[8] C. Fellbaum. *WordNet : an electronic lexical database*. MIT Press, 1998.
[9] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning, *Nature*, **521**:436–444, doi: 10.1038/nature14539, May 2015.
[10] N. T. Duyen, N. X. Bach, and T. M. Phuong. An empirical study on sentiment analysis for Vietnamese, *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, pp. 309–314, doi: 10.1109/ATC.2014.7043403, 2014.
[11] T. K. Tran, T. T. Phan. Multi-Class Opinion Classification for Vietnamese Hotel Reviews, *Int. J. Intell. Technol. Appl. Stat.*, **9**:1:7–18, doi: 10.6148/IJITAS.2016.0901.02, Mar 2016,
[12] D. Nguyen, K. Vo, D. Pham, M. Nguyen, and T. Quan. A Deep Architecture for Sentiment Analysis of News Articles, Springer, Cham, 129–140, 2018.
[13] Q. H. Vo, H.T. Nguyen, B. Le, M.L. Nguyen. Multi-channel LSTM-CNN model for Vietnamese sentiment analysis, *9th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 24–29, doi: 10.1109/KSE.2017.8119429, 2017.
[14] T. T. Nguyen, T. Thanh Quan, and T. Thi Phan. Sentiment search: an emerging trend on social media monitoring systems, *Aslib J. Inf. Manag.*, **66**:5:553–580, doi: 10.1108/AJIM-12-2013-0141, Sep 2014.
[15] T. K. Tran and T. T. Phan. Mining opinion targets and opinion words from online reviews, *Int. J. Inf. Technol.*, **9**:3:239–249, doi: 10.1007/s41870-017-0032-9, Sep. 2017,
[16] T. P. TK Tran. Towards a sentiment analysis model based on semantic relation analysis, *Int. J. Synth. Emot.*, **9**:2:54–75.
[17] L. Mai, B. Le, Aspect-Based Sentiment Analysis of Vietnamese Texts with Deep Learning, Springer, Cham, pp. 149–158, 2018
[18] D. Nguyen, K. Vo, D. Pham, M. Nguyen, and T. Quan, A Deep Architecture for Sentiment Analysis of News Articles," Springer, Cham, pp. 129–140, 2018
[19] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-Based Methods for Sentiment Analysis, *Comput. Linguist.*, **37**:2:267–307, doi: 10.1162/COLI_a_00049, Jun. 2011,
[20] T. K. Tran and Tuoi Thi Phan, An upgrading SentiVoice - a system for querying hotel service reviews via phone, *2015 International Conference on Asian Language Processing (IALP)*, pp. 115–118, doi: 10.1109/IALP.2015.7451545, 2015.
[21] G. Qiu, B. Liu, J. Bu, and C. Chen, Expanding domain sentiment lexicon through double propagation, *Proceedings of the 21st international jont conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc., pp. 1199–1204, 2009.
[22] G. Qiu, B. Liu, J. Bu, and C. Chen, Opinion Word Expansion and Target Extraction through Double Propagation, *Comput. Linguist.*, **37**: 1:9–27, doi: 10.1162/coli_a_00034, Mar. 2011,
[23] T. N. Duy and P. T. Bach NX, Van PD. Mining Vietnamese comparative sentences for sentiment analysis, *Seventh International Conference on Knowledge and Systems Engineering - 2015 (KSE)*, pp. 162–167, 2015.
[24] N. L.-T. Ngan, Kiet V N, Vu Duc Nguyen, Phu Xuan-Vinh Nguyen, Tham Thi-Hong Truong. "UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis, *10th International Conference on Knowledge and Systems Engineering (KSE 2018)*, 2018.
[25] P. T. Tuoi. and P. Do Nguyen Ngoc Duy, A Data Preprocessing Method to Classify and Summarize Aspect-Based Opinions using Deep Learning, *11th Asian Conference on Intelligent Information and Database Systems*, 2019.
[26] X.-S. Vu and S. B. Park, Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary, Dec. 2014.

A SURVEY ON SENTIMENT ANALYSIS FOR VIETNAMESE

Tran Khai Thien, Tieu Phung Mai Suong
Department of Information Technology, HUFLIT
thientk@huflit.edu.vn, suong.tpm@huflit.edu.vn

Abstract: Sentiment analysis (or opinion mining) is an important new field of research that has attracted the attention not only of researchers, but also businesses and organizations. In this article, the authors conduct a survey for sentiment analysis for Vietnamese. First, the remarkable work is introduced. Then the resources and the notable applications are presented.

Keywords: *sentiment analysis, opinion mining, Vietnamese*.

ThS. Trần Khải Thiện tốt nghiệp thủ khoa Thạc sĩ tại trường ĐH Công nghệ thông tin, ĐHQG-HCM. Ông hiện đang làm nghiên cứu sinh ngành Khoa học máy tính tại trường ĐH Bách Khoa, ĐHQG-HCM và là giảng viên công tác tại khoa Công nghệ thông tin trường ĐH Ngoại ngữ - Tin học TPHCM. Hướng nghiên cứu chính của ông là Xử lý ngôn ngữ tự nhiên/ Trí tuệ nhân tạo. ThS. Thiện là bình duyệt viên và là tác giả của nhiều công bố trong các tạp chí SCIE uy tín như Journal of Intelligent & Fuzzy Systems, Applied Sciences, hay IEEE Access.

ThS. Tiếu Phùng Mai Sương nhận học vị Thạc sĩ chuyên ngành Khoa học máy tính vào năm 2017 tại trường Đại học Khoa học tự nhiên, ĐHQG - HCM. Hiện tại Thạc sĩ Sương đang là giảng viên tại Công nghệ thông tin tại trường Đại học Ngoại ngữ - Tin học TPHCM (Huflit). Lĩnh vực nghiên cứu Machine Learning, Data Mining.