

TRÍCH XUẤT CÂU GỢI Ý TỪ DIỄN ĐÀN TRỰC TUYẾN

Ngô Thị Lan Phương*, Ngô Thị Lan, Phạm Thị Liên

Trường Đại học Công nghệ Thông Tin và Truyền Thông – ĐH Thái Nguyên

TÓM TẮT

Các phương tiện truyền thông xã hội phát triển đã tạo điều kiện thuận lợi cho người dùng chia sẻ quan điểm, kinh nghiệm và thảo luận các sự kiện trong thế giới thực. Các nội dung thảo luận, chia sẻ có nhiều thông tin hữu ích, đặc biệt chúng sẽ là các gợi ý cho người dùng khác. Trong bài báo này, chúng tôi nghiên cứu phương pháp trích xuất tự động các câu gợi ý trên diễn đàn trực tuyến. Kết quả thực nghiệm trên bộ dữ liệu tiếng Việt trong miền lĩnh vực du lịch, sử dụng mô hình máy học Maximum Entropy là 77,68%. Kết quả đủ để áp dụng cho các ứng dụng thực tế như hệ thống tư vấn, hệ thống hỏi đáp thông minh. Đây là nghiên cứu đầu tiên về phân tích gợi ý trong ngôn ngữ tiếng Việt.

Từ khóa: Khai phá gợi ý, trích xuất gợi ý, phân tích gợi ý, trích xuất lời khuyên, diễn đàn trực tuyến

GIỚI THIỆU

Trong thời đại internet ngày nay, người dùng có cùng sở thích có thể dễ dàng chia sẻ ý tưởng, thái độ và ý kiến của mình trên các cộng đồng trực tuyến (diễn đàn, nhóm trên mạng xã hội, website cho phép bình luận, v.v.). Hơn nữa, hầu hết mọi người thường tạo thành cộng đồng để giúp đỡ lẫn nhau trong việc tìm kiếm, quảng cáo và ra quyết định. Do đó, có rất nhiều thông tin hữu ích trên các cộng đồng trực tuyến cho các mục đích cụ thể như: thực phẩm, du lịch, sức khỏe, trò chơi, phim ảnh, sản phẩm, dịch vụ, v.v. Các thành viên của cộng đồng có thể trao đổi kinh nghiệm từ các trải nghiệm thực tế của họ về một chủ đề nào đó. Ngoài ra, các thành viên khác có thể bình luận, nhận xét về kinh nghiệm vừa được đưa ra. Cụ thể sẽ có một số ý kiến đồng ý hoặc có các thông điệp khuyến cáo và có một số thông tin phản hồi sau khi họ sử dụng kinh nghiệm đã đưa ra đó. Đây là các thông tin giá trị, là các gợi ý tốt cho những người dùng khác đang tìm kiếm thông tin hoặc giải pháp về vấn đề đang được thảo luận. Đồng thời, nó cũng là thông tin hữu dụng, những gợi ý tốt cho phép các chủ thương hiệu cải tiến sản phẩm/dịch vụ để nâng cao chất lượng sản phẩm/dịch vụ của mình. Mặc dù có những gợi ý, những kinh

nghiệm được chia sẻ trên internet nhưng thông tin thường tản mát, không dễ để người dùng có thể tìm kiếm được các thông tin gợi ý liên quan đến vấn đề của mình, thường họ sẽ mất rất nhiều thời gian lang thang trên các diễn đàn khác nhau và sẽ không tổng hợp hết được các gợi ý phù hợp. Do đó, hệ thống có thể tự động phân tích thông tin gợi ý (lời khuyên, kinh nghiệm, mẹo) từ các ý kiến của người dùng chia sẻ trên diễn đàn để cung cấp cho người dùng khác hoặc chủ thương hiệu có thể tìm kiếm dễ dàng là cần thiết.

Kết quả trích xuất tự động các câu gợi ý từ diễn đàn có thể được tích hợp trong ứng dụng tìm kiếm thông minh và trong các hệ thống tư vấn người dùng. Vấn đề phân tích gợi ý được xem như bài toán phân loại câu thành hai nhóm: gợi ý hoặc không phải gợi ý.

Bài báo này trình bày phương pháp để trích xuất các câu gợi ý trong các bài viết của diễn đàn. Về nguyên lý có thể áp dụng với các miền lĩnh vực khác nhau, nhưng trong phạm vi nghiên cứu của mình chúng tôi lựa chọn lĩnh vực du lịch. Trích xuất các câu gợi ý cho việc lên kế hoạch du lịch của một khách hàng hoặc các câu gợi ý giúp cho chủ thương hiệu cải tiến chất lượng sản phẩm/dịch vụ của họ. Chúng tôi khảo sát hiệu quả của hướng tiếp cận máy học (machine learning) sử dụng mô hình Maximum Entropy (MaxEnt) trên bộ ngữ liệu (corpus) tiếng Việt được xây dựng từ

* Tel: 0975 272359, Email: ntlphuong@ictu.edu.vn

các bài viết (post) trên diễn đàn xã hội trực tuyến (forum).

Đóng góp chính trong nghiên cứu của chúng tôi có thể tóm lược theo hai khía cạnh sau:

- Thứ nhất, về mặt dữ liệu, chúng tôi đã xây dựng một kho ngữ liệu được gán nhãn gợi ý hay không gợi ý cho các nội dung thảo luận, hoặc chia sẻ bằng ngôn ngữ tiếng Việt.
- Thứ hai, chúng tôi đã mô hình hoá nhiệm vụ trích xuất gợi ý và triển khai thực nghiệm việc xác định câu gợi ý hay không phải câu gợi ý bằng phương pháp máy học ME. Kết quả thực nghiệm của chúng tôi cho thấy, nhiệm vụ trích xuất các câu gợi ý từ diễn đàn trực tuyến là khả thi. Theo chúng tôi được biết, đây là nghiên cứu đầu tiên về khai phá gợi ý cho ngôn ngữ tiếng Việt.

Bài báo có cấu trúc như sau: Sau phần giới thiệu, chúng tôi trình bày sơ lược về các nghiên cứu liên quan. Tiếp theo, chúng tôi phát biểu bài toán xác định câu gợi ý. Phần tiếp theo là phương pháp học máy Maximum Entropy. Phần kế tiếp chúng tôi chỉ ra cách xây dựng dữ liệu, kết quả thực nghiệm và đánh giá kết quả đó. Cuối cùng là một số kết luận và hướng phát triển trong tương lai.

CÁC NGHIÊN CỨU LIÊN QUAN

Trên thế giới, trong những năm gần đây, đã có một số nghiên cứu về vấn đề phân tích gợi ý. Trong lĩnh vực hiểu văn bản nói, gợi ý được coi là một loại hành động ngôn từ (speech act) hay hành động hội thoại (dialog act). Tuy nhiên, không có nghiên cứu nào riêng về ý định gợi ý. Trong khai phá văn bản thể hiện quan điểm, nghiên cứu toàn diện về ý định gợi ý của người dùng đã được Sapna Negi thực hiện. Sapna Negi là người đầu tiên đưa vấn đề phân tích gợi ý thành một lĩnh vực nghiên cứu riêng và sử dụng thuật ngữ “khai phá gợi ý” (suggestion mining) để chỉ các nghiên cứu trong lĩnh vực này [1-4].

Trong vấn đề phát hiện đề xuất, Goldberg và cộng sự [5] lần đầu tiên đã đưa ra khái niệm về một gợi ý nhằm cải tiến sản phẩm/dịch vụ trong bối cảnh của lĩnh vực khai phá quan điểm. Các gợi ý thể hiện mong muốn cải tiến

được nhấn mạnh bởi Ramanand và cộng sự [6]. Sau đó, Brun và các cộng sự [7] xác định đề xuất cải tiến sản phẩm bằng cách sử dụng các quy tắc ngôn ngữ được làm thủ công. Dong và các cộng sự [8] đã thực hiện phân loại các bài viết trên mạng xã hội Twitter là gợi ý hay không. Wicaksono và cộng sự [9] đã thực hiện trích xuất lời khuyên trong các diễn đàn du lịch. Ynzhong Liu và cộng sự [10] đã nghiên cứu về các câu gợi ý cho bệnh nhân và thực hiện việc khai thác kinh nghiệm tự động có nhận biết được ngữ cảnh từ các diễn đàn y tế trực tuyến. Weber và cộng sự [11] và Guy cùng cộng sự [12] đề cập đến mẹo là “một phần ngắn gọn của lời khuyên thiết thực không rõ ràng, có thể thường dẫn đến một hành động”. Tuy nhiên trong ngôn ngữ Việt Nam chưa có nghiên cứu nào về vấn đề phân tích gợi ý từ văn bản cả.

BÀI TOÁN XÁC ĐỊNH CÂU GỢI Ý

Theo từ điển tiếng Việt trực tuyến [13], “**gợi ý**” là “*gợi ra để tự suy nghĩ, tự có ý kiến hoặc tự giải quyết một việc nào đó*”. Các từ đồng nghĩa với gợi ý là “lời khuyên”, “mẹo”. Trong bài báo này, chúng tôi xem các gợi ý trên diễn đàn gồm 2 loại sau:

1) Các gợi ý hướng tới người dùng là khách hàng, những người đang muốn tìm kiếm giải pháp hay lên kế hoạch cho một vấn đề tương tự đang được những khách hàng đi trước (những người đã trải qua, đã sử dụng sản phẩm/dịch vụ) post lên. Những câu mang ý định gợi ý thường là lời khuyên, mẹo, các trải nghiệm, kinh nghiệm của người dùng. Một ví dụ về câu gợi ý và không phải câu gợi ý được thể hiện trong bảng 1. Ví dụ (1), “*Đi Sầm Sơn không nên đi vào dịp 30/4 đông người lắm*” là lời khuyên gợi ý rằng người đọc trên diễn đàn không nên đi chơi Sầm Sơn vào dịp 30/4, thể hiện rõ ý định gợi ý. Ví dụ (2), “*Sầm Sơn 30/4 đông người lắm, họ phục vụ không chu đáo được*”, thể hiện gợi ý cùng với nội dung ở ví dụ (1) nhưng người đọc phải suy luận ra ý định gợi ý đó nên nó được gọi là gợi ý ẩn. Ví dụ (3), là kinh nghiệm của một người dùng trong việc chăm sóc con nhỏ và là gợi ý ẩn cho những người dùng có con nhỏ khác.

Ngoài ra các nhận xét tốt (tích cực) cũng gợi ý cho người dùng nên sử dụng sản phẩm/dịch vụ hoặc kinh nghiệm đó như ví dụ (5). Các nhận xét không tốt (tiêu cực) về sản phẩm/dịch vụ hoặc kinh nghiệm cũng là gợi ý cho người dùng khác không nên sử dụng (ví dụ (6)).

Bảng 1. Ví dụ về câu gợi ý và không gợi ý

STT	Câu	Đối tượng
1	Đi Sầm sơn bạn không nên đi vào 30/4 vì đông lắm	Khách hàng
2	Đi Sầm Sơn vào 30/4 đông người lắm, họ phục vụ sẽ không chu đáo	Khách hàng
3	Bé nhà tớ bị ho chỉ uống mỗi lá hẹ hấp mật ong thôi, uống thuốc kháng sinh sớm hại cho trẻ con lắm.	Khách hàng
4	Hãy chuẩn bị đầy đủ kem chống nắng khi đi du lịch biển nhé.	Khách hàng
5	Vinpearland gofl đi nhé phòng ốc sạch sẽ và đẹp hơn nhiều khu kia nhé.	Khách hàng
6	Phòng ốc của khách sạn này vệ sinh kém lắm	Khách hàng & chủ
7	Giả mà nhân viên tươi cười hơn thì quán của anh Trường Giang sẽ thu hút được nhiều khách hơn	Chủ
8	Đậu phụ là một trong những nguyên liệu quen thuộc trong các món chay, tuy nhiên nếu không chế biến đúng cách sẽ làm mất đi hương vị thơm ngon vốn có.	Không gợi ý
9	Bạn ơi cho mình hỏi chút mình đi Nha Trang 4 ngày 3 đêm thì không biết là buổi tối có những địa điểm nào để vui chơi thể bạn, và nên đi những điểm nào bạn cho mình xin ít kinh nghiệm với.	Không gợi ý

2) Các gợi ý hướng tới người chủ sản phẩm/dịch vụ, gợi ý cho người chủ cải thiện sản phẩm/dịch vụ của mình. Các bài viết loại gợi ý này, ngoài các lời khuyên, gợi ý rõ ràng thì nó thường ở dạng ẩn, là các nhận xét không tốt (tiêu cực) hoặc là những ước muốn về một tính năng còn thiếu hoặc chưa tốt của sản phẩm dịch vụ (ví dụ (6) và (7)).

Việc thực hiện trích xuất gợi ý tự động từ văn bản được mô hình hoá thành vấn đề xác định một câu đầu vào có phải là câu gợi ý hay không. Nhiệm vụ này được xem như bài toán phân lớp câu vào 2 lớp {gợi ý, không gợi ý}.

TRÍCH XUẤT CÂU GỢI Ý SỬ DỤNG PHƯƠNG PHÁP MAXIMUM ENTROPY

Bài toán phân lớp câu gợi ý hay không gợi ý được phát biểu cụ thể như sau:

Tập các câu $S = \{x_1, x_2, \dots, x_n\}$ và tập các nhãn $L = \{l_1, l_2\}$, xác định nhãn l_i tương ứng cho từng x_i . Trong đó $l_1 = \text{"gợi ý"}$, $l_2 = \text{"không gợi ý"}$.

Chúng tôi tiếp cận theo hướng máy học có giám sát sử dụng mô hình MaxEnt. Chi tiết về MaxEnt được trình bày trong [14]. Chúng tôi đã cài đặt mô hình phân lớp câu gợi ý sử dụng phương pháp máy học MaxEnt [14] và sử dụng giải thuật BSGF [15] để tìm trọng số của mô hình.

Đặc trưng được sử dụng trong mô hình học MaxEnt là n-gram (1, 2, 3- gram). Các đặc trưng trong mô hình được biểu diễn trong Bảng 2.

THỰC NGHIỆM

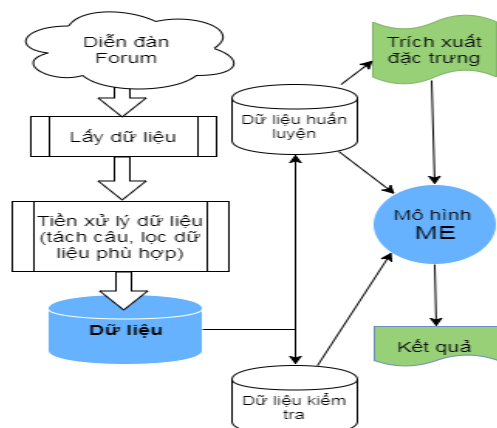
Quá trình thực hiện trích xuất câu gợi ý trong nghiên cứu của chúng tôi có thể tóm lược qua các bước trong Hình 1. Nói chung, nó gồm 2 quá trình chính là xây dựng dữ liệu và chạy mô hình MaxEnt với dữ liệu xây dựng được.

Xây dựng dữ liệu

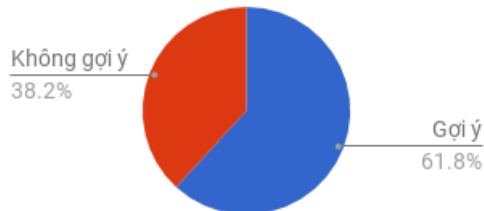
Để xây dựng dữ liệu thực nghiệm, chúng tôi lấy các bài viết (post) đăng trên diễn đàn lamchame.com. Về nguyên lý chung chúng ta có thể thực hiện trên các miền dữ liệu khác nhau. Để minh hoạ chúng tôi lựa chọn các bài viết trong chủ đề du lịch. Các bài viết từ diễn đàn được lấy về tự động, sau đó chúng tôi lọc các bài viết theo chủ đề quan tâm và thực hiện tách câu cho các post. Một nhóm sinh viên được yêu cầu để gán nhãn gợi ý hay không gợi ý bằng tay cho các câu đã được lựa chọn. Kết quả dữ liệu xây dựng được 4318 câu trong đó có 2670 câu gợi ý và 1648 câu không gợi ý. Tỷ lệ phân bố dữ liệu được thể hiện như Hình 2.

Bảng 2. Các đặc trưng của mô hình MaXent phân lớp câu gợi ý

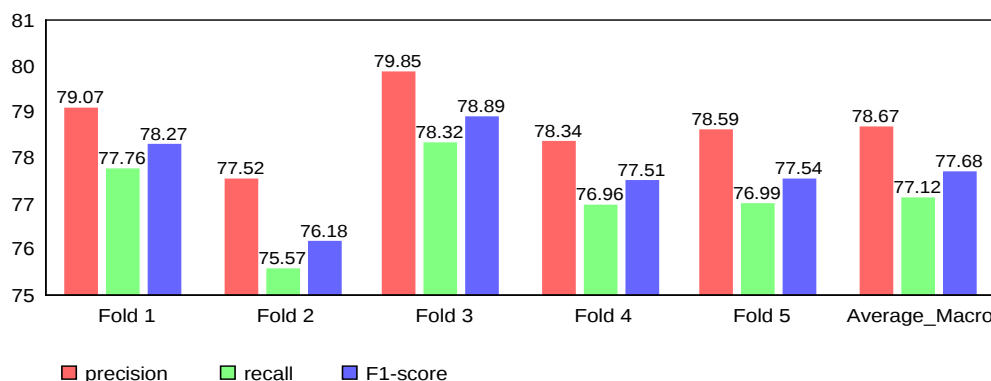
1-grams	$[w_{-2}], [w_{-1}], [w_0], [w_1], [w_2]$
2-grams	$[w_{-2}w_{-1}], [w_{-1}w_0], [w_0w_1], [w_1w_2]$
3-grams	$[w_{-2}w_{-1}w_0], [w_{-1}w_0w_1], [w_0w_1w_2]$
is-SI, is-none SI	1-tokens : $-2 \dots 2 \rightarrow [w_{-2}], [w_{-1}], [w_0], [w_1], [w_2]$ 2-tokens : $-2 \dots 2 \rightarrow [w_{-2}w_{-1}], [w_{-1}w_0], [w_0w_1], [w_1w_2]$ 3-tokens : $-2 \dots 2 \rightarrow [w_{-2} w_{-1} w_0], [w_{-1} w_0 w_1], [w_0 w_1 w_2]$... (n+1)-tokens: $-n \dots n \rightarrow [w_{-n} w_{-n+1} w_{-n+2} \dots w_{-1} w_0] \dots [w_0 w_1 \dots w_{n-1} w_n]$ (with (n + 1) is the length of the sentence)



Hình 1. Quá trình thực hiện thực nghiệm trích xuất gợi ý từ diễn đàn trực tuyến.



Hình 2. Phân bố dữ liệu trong kho ngữ liệu thực nghiệm



Hình 3. Kết quả độ chính xác trung bình trên 5 lần thực hiện của mô hình MaxEnt.

Kết quả

Để đánh giá mô hình phân loại, chúng tôi chia dữ liệu thành 5 phần, trong đó có 4 phần để huấn luyện mô hình và 1 phần còn lại để kiểm tra mô hình. Thực nghiệm được thực hiện kiểm tra chéo 5 lần. Bảng 3 trình bày kết quả thử nghiệm của lần tốt nhất. Cột “thực tế” là số lượng các câu gợi ý được gán nhãn thủ công. Cột mô hình là số lượng các câu mô hình dự đoán được. Cột tiếp theo là số lượng các câu mô hình đoán khớp với thực tế. Ba cột tiếp theo là độ chính xác (precision), độ hồi tưởng (recall) và độ đo F1 của mô hình. Các kết quả thực nghiệm của 5 lần được hiển thị trong Hình 3, độ chính xác F_1 trung bình trên 5 lần là 77,68% theo Macro (độ chính xác trung bình theo từng lớp). Kết quả này cho thấy mô hình có thể áp dụng sử dụng trong hệ thống thực tế.

Bảng 3. Kết quả trên một lần thực hiện tốt nhất

Nhãn	Thực tế	Mô hình	Mô hình khớp với thực tế	Độ chính xác	Độ hồi tưởng	F1
Gợi ý	528	567	463	81,66	87,69	84,57
Không gợi ý	335	296	231	78,04	68,96	73,22
Trung bình (Average – macro)				79,85	78,32	78,89
Trung bình (Average – micro)	863	863	694	80,42	80,42	80,42

KẾT LUẬN

Việc trích xuất thông tin gợi ý tự động từ các bài viết trên diễn đàn trực tuyến có nhiều ứng dụng trong các hệ thống gợi ý, hệ thống hỏi đáp thông minh để hỗ trợ việc ra quyết định của người dùng bằng cách cung cấp các gợi ý thích hợp, giảm quá tải thông tin cho người dùng. Bài báo này đã trình bày đề xuất của nhóm tác giả sử dụng mô hình MaxEnt để phân lớp câu thành câu gợi ý và câu không gợi ý. Kết quả thực nghiệm là rất đáng hứa hẹn, có thể áp dụng vào các hệ thống thực tế. Trong tương lai, chúng tôi sẽ nghiên cứu các đặc trưng về cú pháp, ngôn ngữ sâu hơn để cải thiện hiệu quả của mô hình.

LỜI CẢM ƠN

Bài báo là sản phẩm của đề tài NCKH có mã số T2017-07-01, được hỗ trợ bởi Trường Đại học Công nghệ thông tin & Truyền thông - Đại học Thái Nguyên. Nhóm tác giả xin chân thành cảm ơn sự hỗ trợ của quý Trường.

TÀI LIỆU THAM KHẢO

1. Negi S., Buitelaar P. (2015), "Towards the Extraction of Customer-to-Customer Suggestions from Reviews", *Conference on Empirical Methods in Natural Language Processing*, pp. 2159-2167.
2. Negi S., Asooja K., Mehrotra S., Buitelaar P. (2016), "A Study of Suggestions in Opinionated Texts and their Automatic Detection", *Joint Conference on Lexical and Computational Semantics*, pp. 170-178.
3. Pozzi, F. A., Fersini E., Messina E., Liu B. (2016) "Sentiment Analysis in Social Networks", *Morgan Kaufmann*.
4. Negi S. (2016), "Suggestion Mining from Opinionated Text", *In \textit{proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (ACL) - Student Research Workshop}, pp 119-125.

5. A.B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, X. Zhu(2009), "May all your wishes come true: a study of wishes and how to recognize them", *NAACL'09, ACL*, pp. 263-271.

6. J. Ramanand, Krishna Bhavsar, Niranjan P.edanekar(2010), "Wishful thinking - finding suggestions and 'buy' wishes from product reviews", *NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 54-61.

7. Brun, C., Hagege, C.(2013), "Suggestion Mining: Detecting Suggestions for Improvement in Users' Comments", *Research in Computing Science*, pp. 199-209.

8. Li Dong, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu.(2013), "The automated acquisition of suggestions from tweets", *AAAI Press*.

9. Alfian Farizki Wicaksono and Sung-Hyon (2015), "Myaeng Automatic extraction of advice-revealing sentences for advice mining from online forums", *K-CAP*, pp. 97-104. ACM.

10. Liu, Y., Chen, Y., Tang, J., Liu, H. "Context-aware experience extraction from online health forums", *Healthcare Informatics (ICHI)*, pp 42-47.

11. I. Weber, A. Ukkonen, and A. Gionis(2012), "Answers, not links: Extracting tips from yahoo! answers to address how-to web queries", *WSDM*, pp 613-622.

12. Guy I., Mejer A., Nus, A., Raiber, F.(2017), "Extracting and Ranking Travel Tips from User-Generated Reviews", *International Conference on World Wide Web*, pp 987-996.

13. <http://tratu.soha.vn/>

14. Berger A. Pietra, S. A. D. Pietra (1996), "V.J.D.: A maximum entropy approach to natural language processing", *Computational Linguistics*, 22(1), 39-71.

15. Liu D., Nocedal J. (1989), "On the limited memory BFGS method for large-scale optimization", *Mathematical Programming*, 45, pp. 503-528.

SUMMARY

SUGGESTION EXTRACTION FROM TEXTS ON ONLINE FORUM**Ngô Thị Lan Phương^{*}, Ngô Thị Lan, Phạm Thị Lien***University of Information and Communication Technology - TNU*

Social media development allows users to share their opinions, experiences and discuss real-world events. The contents provided have useful information as suggestion for other users. In this paper, we study the method of automatically extracting suggestion from posts in online forums. The experimental results on Vietnamese corpus in travel domain, employing the machine learning Maximum entropy, is 77.68%. The results are enough to apply to practical applications such as recommendation systems, smart question answering systems. This is the first work of suggestion mining in Vietnamese.

Key words: *Suggestion mining, Suggestion analysis, Vietnamese suggestion, maximum entropy, Vietnamese suggestion detection, advice extraction, online forum*

Ngày nhận bài: 08/9/2017; Ngày phản biện: 10/10/2017; Ngày duyệt đăng: 30/11/2017

^{*} Tel: 0975 272359, Email: ntlphuong@ictu.edu.vn