BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐỀ CƯƠNG LUẬN VĂN THẠC SỸ

Chuyên ngành: Công nghệ Thông tin

Mã ngành: 60480201

ỨNG DỤNG KHAI THÁC DỮ LIỆU VÀO LĨNH VỰC GIÁO DỤC

GVHD: TS. LÊ THỊ NGỌC THƠ

HVTH: VÕ MINH QUÂN

MSHV: 1741086036

Lớp: 17SCT21

TP. HCM, tháng 03/2019

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. HCM, ngày …… tháng …… năm 2019

Giáo viên hướng dẫn

NHẬN XÉT CỦA HỘI ĐỒNG XÉT DUYỆT

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. HCM, ngày …… tháng …… năm 2019

Hội đồng xét duyệt

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Viết tắt	Tiếng Anh	Tiếng Việt
1	CSDL	Database	Cơ sở dữ liệu
2	SVM	Support Vector Machines	
3	PU	Positive - Unlabeled	Tích cực - Chưa gán nhãn
4	DP	Double Propagation	

1. GIỚI THIỆU TỔNG QUAN

1.1. Đặt vấn đề

Trong nhiều lĩnh vực của xã hội hiện nay, việc thu thập ý kiến, cảm nhận, phản hồi đánh giá của con người là một việc rất phổ biến mà ta có thể nhờ vào đó để đưa ra những đánh giá, nhận xét liên quan. Một vài nguồn tài nguyên phổ biến cho việc thu thập và sử dụng ý kiến phản hồi được kể ra như dưới đây:

- Kinh nghiệm cá nhân và ý kiến về bất cứ điều gì trong đánh giá, diễn đàn, blog, v.v.
- Nhận xét về bài viết, vấn đề, chủ đề, bài đánh giá, v.v.
- Thông tin đăng tại các trang web mạng xã hội, ví dụ: Facebook.

Vậy tại sao những ý kiến này lại quan trọng đến như vậy? Những luận điểm dưới đây sẽ giải đáp những vấn đề này:

- "Ý kiến" là những yếu tố ảnh hưởng quan trọng đến hành vi của một người.
- Những ý kiến đánh giá là một phần quan trọng để đo lường và đánh giá chất lượng sản phẩm hay dịch vụ.
- Thực tế khi chúng ta cần đưa ra quyết định, chúng ta thường tìm kiếm ý kiến của người khác. Với cá nhân sẽ tìm kiếm ý kiến từ bạn bè và gia đình, còn với tổ chức sử dụng khảo sát ý kiến, tư vấn.

Những ứng dụng từ việc phân tích ý kiến cũng được áp dụng rộng rãi trong nhiều lĩnh vực. Đối với các doanh nghiệp và tổ chức, việc phân tích ý kiến hỗ trợ việc cung ứng ra thị trường các sản phẩm phù hợp nhu cầu và xu hướng. Đối với cá nhân, việc phân tích ý kiến có thể hỗ trợ người dùng trong quá trình ra quyết định sử dụng dịch vụ, thu thập các ý kiến liên quan đến môi trường xã hội xung quanh.

Tuy nhiên hiện nay việc phân tích xử lý các dữ liệu này phần lớn còn được thực hiện một cách thủ công dưới sự đánh giá trực tiếp từ con người. Vì vậy những hệ thống phân tích ý kiến tự động và đưa ra những tổng hợp đánh giá là cần thiết. Ứng

dụng của những hệ thống này hứa hẹn sẽ mang lại nhiều giá trị to lớn trong nhiều lĩnh vực. Trong lĩnh vực giáo dục việc áp dụng một hệ thống phân tích ý kiến tự động những ý kiến đánh giá của học sinh, sinh viên về chất lượng giảng viên trong các khóa học, chương trình đào tạo sẽ giúp tiết kiệm một lượng lớn nguồn nhân lực cũng như thời gian đánh giá.

1.2. Tính cấp thiết của đề tài

Sau một thời gian tìm hiểu và phân tích chúng tôi nhận thấy việc thu thập ý kiến đánh giá chất lượng giảng dạy của sinh viên trong mỗi học kì ở Trường Đại học Công nghệ TP.HCM hiện nay là một bài toán thực tế và có thể áp dụng được mô hình phân tích và đánh giá ý kiến. Với một lượng dữ liệu rất lớn về việc ý kiến đánh giá của sinh viên trong mỗi học kì thì việc tổng hợp và đánh giá thủ công thông qua con người sẽ tốn rất nhiều thời gian và chi phí.

Vì vậy luận văn này sẽ thực hiện nghiên cứu và áp dụng phân tích, tổng hợp các ý kiến đánh giá một cách tự động. Mục tiêu của nghiên cứu này là giúp rút ngắn thời gian thực hiện đánh giá, phân tích bên cạnh đó sẽ hỗ trợ đánh giá chất lượng được khách quan hơn.

2. MỤC TIÊU, NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mục tiêu:

Mục tiêu nghiên cứu chính của luận văn là tìm hiểu về các phương pháp phân tích ý kiến và cách tổng hợp những dữ liệu thực tế. Bên cạnh đó, nghiên cứu cũng sẽ so sánh độ hiệu quả giữa các phương pháp phân tích ý kiến thông qua bài toán phân tích ý kiến đánh giá giảng viên.

Đối với bài toán phân tích ý kiến đánh giá của sinh viên về chất lượng giảng dạy tại Trường Đại học Công nghệ TP.HCM, chúng tôi dự kiến tạo được một hệ thống phân tích các ý kiến thu thập được một cách tự động, xác định được cụ thể ý kiến là đánh giá tích cực hay tiêu cực hoặc là một ý kiến trung tính.

Ngoài ra luận văn này có thể phát triển thêm ở việc xác định khía cạnh đánh giá của ý kiến, hướng phát triển này phụ thuộc vào độ hiệu quả của việc đánh giá ý trước trước đó.

2.2. Nội dung nghiên cứu:

Đề tài cần nghiên cứu các nội dung chính sau:

- Nghiên cứu về phân lớp chủ quan về phân lớp cảm nghĩ.
- Nghiên cứu về tóm tắt ý kiến.
- Nghiên cứu về phương pháp phân lớp dữ liệu.
- Nghiên cứu về phân loại ý kiến dựa trên học không giám sát.
- Thực nghiệm và đánh giá trên CSDL khảo sát sinh viên năm học 2016-2017.

2.3. Phương pháp nghiên cứu

Tìm hiểu các tài liệu về phân tích ý kiến, cảm xúc thông qua các từ khóa phổ biến như: opinion mining, data mining opinion, data mining and education, v.v.

Tìm hiểu các phương pháp liên quan đến khai thác văn bản, ý kiến, phân lớp dữ liệu, học giám sát, học không giám sát so sánh độ hiệu quả giữa các phương pháp thông qua các ứng dụng thực tế đã có và chọn lưa phương pháp hướng đi phù hợp trong luận văn nhằm mang lại độ hiệu quả cao nhất.

Cài đặt đặt các thuật toán trong nội dung nghiên cứu.

Thực nghiệm các dữ liệu đánh giá giảng viên trên các thuật toán và kiểm chứng các kết quả thu được

3. TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU

3.1. Phân tích ý kiến (opinion mining)

Vào năm 2006 Jindal và Liu đã đưa ra một nhận xét thì ý kiến thường xuyên có 2 loại là: **cảm xúc** và **ý kiến** [1].

Trong khi đó Hu và Liu thì lại cho rằng một ý kiến có cấu trúc bao gồm **thực thể** và **khía cạnh** [2].

Sau đó vào năm 2010 để phân tích ý kiến Liu đã đưa các ý kiến về theo một cấu trúc gồm năm thành phần [3]:

$$e_{j,}a_{jk}, so_{ijkl}, h_i, t_l$$

Trong đó:

- e_j là một thực thể đích.
- a_{jk} là một khía cạnh/tính năng của thực thể e_j.
- so_{ijkl} là giá trị cảm xúc của ý kiến từ người giữ ý kiến h_i về tính năng a_{jk} của thực thể e_j tại thời gian t_l.
- h_i là người đưa ra ý kiến.
- t_l là thời điểm đưa ý kiến.

Ngoài ra còn có một số ý kiến mang tính chất riêng biệt như ý kiến so sánh.

3.2. Phân loại cảm xúc

Cảm xúc là suy nghĩ chủ quan của một con người về một khía cạnh nào đó. Theo nghiên cứu của Parrott [4], con người có sáu cảm xúc chính: tình yêu, niềm vui, bất ngờ, giận dữ, buồn bã và sợ hãi.

Phân loại cảm xúc có thể dựa trên những phương pháp như:

- Theo phương pháp phân lớp không giám sát [5].
- Theo phương pháp phân lớp có giám sát [6]. Kỹ thuật chủ yếu dùng là Naïve Bayes hoặc SVM (support vector machines).
- Phân tích cảm xúc dựa trên khía cạnh. Một số kỹ thuật tiêu biểu của phương pháp này là dựa trên từ vựng [7].
- Phân loại cảm xác dựa trên chủ đề [8].

3.3. Phân lớp câu chủ quan

Câu chủ quan là một câu thể hiện về cảm xúc hoặc ý kiến rõ ràng của một cá nhân. Phải phân biệt rõ 1 câu là thể hiện ý kiến khách quan hay chủ quan.

Ở kỹ thuật này ta dùng đa số bằng phương pháp Navie Bayes [9]. Với kỷ thuật trên ý kiến không được đưa về bộ 5 thành phần như ở phân tích ý kiến. Ngoài ra còn một số phương pháp khác như:

- Phân loại câu chủ quan sử dụng mẫu [10].
- Câu chủ quan và phân lớp cảm xúc [11].

3.4. Tóm tắt ý kiến

Tóm tắt ý kiến là một phần không thể thiếu của việc phân tích ý kiến, phương pháp được sử dụng phổ biến là tóm tắt ý kiến dựa trên khía cạnh. Một số phương pháp đã được nghiên cứu như:

- Chọn và sắp xếp các câu [12].
- Tính thông minh và độ dễ đọc [13].

3.5. Từ vựng ý kiến

Từ vựng ý kiến là một danh sách các từ và biểu thức được sử dụng để thể hiện cảm xúc chủ quan của con người. Có 3 các chính để biên soạn ra bộ từ vựng này là: thủ công, dựa trên corpus hay dựa trên từ điển.

Một số nghiên cứu liên quan như:

- Phương pháp Double Propagation [7].
- Sử dụng thông tin từ WordNet [15].

4. TIẾN ĐỘ THỰC HIỆN ĐỀ TÀI

Tháng (năm 2019) / **Dự kiến nội dung thực hiện**	**2**	**3**	**4**	**5**	**6**	**7**
Thực hiện đề cương luận văn						
Nghiên cứu cơ sở lý thuyết và viết báo cáo sơ lược						

- Xử lí dữ liệu - Chạy thực nghiệm dữ liệu - Phân tích số liệu thu được, so sánh, đánh giá kết quả						
Kiểm tra và hoàn chỉnh báo cáo theo các yêu cầu đề ra						
Hoàn thiện luận văn						

5. BỐ CỤC DỰ KIẾN CỦA LUẬN VĂN

Luận văn sẽ dự kiến thực hiện 5 chương:

- Chương 1: Giới thiệu tổng quan:
  - 1.1 Giới thiệu về phân tích ý kiến.
  - 1.2 Giới thiệu về phân loại cảm xúc.
  - 1.3 Đặt vấn đề bài toán phân tích ý kiến đánh giá giảng viên .
- Chương 2: Cơ sở lý thuyết:
  - Trình bày các thuật toán sử dụng để phân tích ý kiến, phân loại cảm xúc, phân lớp câu chủ quan, tóm tắt ý kiến.
- Chương 3: Phương pháp thực hiện:
  - 3.1: Lựa chọn các phương pháp phân tích ý kiến, phân loại cảm xúc, phân lớp câu chủ quan để triển khai.
  - 3.2: So sánh các phương pháp với nhau với nhau.
- Chương 4: Thực nghiệm và đánh giá:
  - 4.1: Trình bày về các công cụ dùng trong thực nghiệm bài toán phân tích ý kiến đánh giá giảng viên.
  - 4.2: Trình bày kết quả thực nghiệm trên tập dữ liệu đánh giá sinh viên. Phân tích so sánh kết quả thu được trên nhiều khía cạnh và phương pháp tiếp cận khác nhau.
- Chương 5: Kết luận và hướng phát triển.

TÀI LIỆU THAM KHẢO

[1] B. Jindal & B. Liu, Mining Comparative Sentences and Relations, American Association for Artificial Intelligence, Pages 1331-1336, 2006.

[2] M. Hu & B. Liu, Mining and summarizing customer reviews, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 168-177, 2004.

[3] B. Liu , Sentiment analysis and subjectivity, Handbook of Natural Language Processing, 2010.

[4] J. Parrott & A. Bourne & R. Akien & J. Irvine, Self-Optimizing Continuous Reactions in Supercritical Carbon Dioxide, Angewandte Chemie International Edition, Pages 3788-3792, 2010.

[5] B. Pang & L. Lee & S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP, Page 79-86, 2002.

[6] D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Page 417-424, 2002.

[7] G. Qiu & B. Liu & J. Bu & C. Chen, Opinion word expansion and target extraction through double Propagation, Journal Computational Linguistics, Page 9-27, 2011.

[8] X. Ding & B. Liu & S. Yu, A holistic lexicon approach to opinion mining, Proceedings of the 2008 International Conference on Web Search and Data Mining, Page 231-240, 2008.

[9] J. Wiebe & F. Bruce & P. O'Hara, Development and use of a gold standard data set for subjectivity classifications, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Page 246-253, 1999.

[10] E. Riloff & J. Wiebe, Learning Extraction Patterns for Subjective Expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Page 105-112, 2003.

[11] H. Yu & V. Hazivassiloglou, Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Page 129-136, 2003

[12] S. Tata & B. Di Eugenio, Generating Fine-Grained Reviews of Songs From Album Reviews, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Page 1376-1385, 2003.

[13] H. Nishikawa & T. Hasegawa & Y. Matsuo & G. Kikui, Optimizing Informativeness and Readability for Sentiment Summarization, Proceedings of the ACL 2010 Conference Short Papers, Page 325-330, 2010.

[14] A. Andreevskaia & S. Bergler, Sentiment Tagging of Adjectives at the Meaning Level, Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence, Page 336-236, 2006.