ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM ĐỨC HỒNG

KHAI PHÁ VÀ PHÂN TÍCH QUAN ĐIỂM NGƯỜI DÙNG TRÊN MẠNG INTERNET

Chuyên ngành: Khoa học máy tính
Mã số: 62 48 01 01

TÓM TẮT LUẬN ÁN

Hà Nội - 2018

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội.

Người hướng dẫn khoa học:

PGS.TS. Lê Anh Cường

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án đã được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội vào hồi ... giờ ngày ... tháng ... năm 2018.

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

Chương 1

Tổng quan

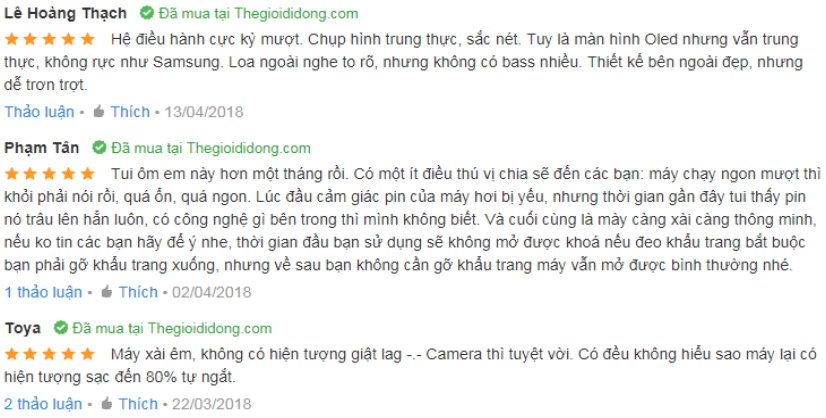
1.1 Giới thiệu

Trong khoảng 15 năm trở lại đây, cùng với sự phát triển của công nghệ Web 2.0, các hệ thống thương mại trực tuyến phát triển rất nhanh, tiêu biểu như hệ thống Amazon[1], Yelp[2], Tripadvisor[3] và Thegioididong[4]. Đặc điểm chung của các hệ thống thương mại là cho phép các khách hàng có thể đặt/mua hàng trực tuyến những sản phẩm mà họ yêu thích. Ngoài ra, các hệ thống cũng cho phép họ thể hiện ý kiến đánh giá về những sản phẩm mà họ quan tâm thông qua hệ thống. Những ý kiến đánh giá này là phần quan trọng của mỗi hệ thống, bởi nó cung cấp thông tin tới các nhà quản lý hệ thống thương mại cũng như với các khách hàng khác, giúp họ có sự hiểu biết nhất định về sản phẩm hay dịch vụ của hệ thống. Hình 1.1 là ví dụ ý kiến đánh giá của sản phẩm iPhone X 64GB trên hệ thống www.thegioididong.com. Nhằm hỗ trợ các hệ thống thương mại cung cấp thông tin hiệu quả tới người quản lý và khách hàng, một lĩnh vực mới của chuyên ngành xử lý ngôn ngữ tự nhiên đã ra đời trong giai đoạn này là "*Khai phá và phân tích quan điểm*".

Khai phá và phân tích quan điểm người dùng là nghiên cứu tính toán các quan điểm, đánh giá, thái độ và cảm xúc của con người đối với các thực thể và các khía cạnh của thực thể. Thực thể thông thường đề cập tới các sản phẩm, dịch vụ và các tổ chức riêng biệt, v.v. Các khía cạnh là các thuộc tính hoặc các thành phần của các thực thể. Ví dụ trong Hình 1 là các ý kiến khách hàng thảo luận về thực thể "iPhone X 64GB" với các khía cạnh là "Hệ điều hành", " Loa nghe", và "Pin".

Các bài toán phân tích quan điểm được thực hiện ở ba mức độ là mức văn bản,

[1]www.amazon.com
[2]www.yelp.com
[3]www.tripadvisor.com
[4]www.thegioididong.com



Hình 1.1: Ví dụ các ý kiến đánh giá sản phẩm iPhone X 64GB

mức câu, và mức khía cạnh. Trong đó, phân tích quan điểm mức văn bản là bài toán được nghiên cứu sớm và rộng rãi nhất (Pang và các cộng sự, 2002). Bài toán phân loại một văn bản đánh giá sản phẩm/dịch vụ bằng cách đưa ra quan điểm tổng thể là tích cực hay tiêu cực. Xem xét toàn bộ tài liệu như là một đơn vị thông tin cơ bản và nó giả thiết rằng tài liệu được biết là có quan điểm. Ở mức câu, việc phân loại quan điểm được áp dụng cho từng câu trong một tài liệu (Ellen và các cộng sự, 2005). Tuy nhiên, không phải bất kỳ câu nào trong văn bản đánh giá cũng có quan điểm. Do đó, nhiều nghiên cứu thực hiện bài toán xác định câu có thể hiện quan điểm của người dùng trước (Mihalcea và các cộng sự, 2007). Các câu có quan điểm xác định được sau đó được phân loại là câu thể hiện ý kiến quan điểm tích cực hoặc tiêu cực.

Mặc dù việc khai thác ý kiến ở mức văn bản và mức câu rất hữu ích trong nhiều trường hợp, nhưng chúng ta vẫn còn nhiều điều mong muốn hơn. Một văn bản đánh giá tích cực về một thực thể cụ thể không có nghĩa là người dùng có ý kiến tích cực về mọi khía cạnh của thực thể. Tương tự, một văn bản đánh giá tiêu cực cho một thực thể không có nghĩa là người dùng không thích tất cả mọi khía cạnh của thực thể đó. Ví dụ, trong một bài đánh giá sản phẩm, người đánh giá thường ghi cả khía cạnh tích cực và tiêu cực của sản phẩm, mặc dù quan điểm chung về sản phẩm có thể là tích cực hoặc tiêu cực. Để có được nhiều phân tích ý kiến tốt hơn, chúng ta cần phải nghiên cứu sâu về khía cạnh. Ý tưởng này dẫn đến việc khai thác ý kiến dựa trên khía cạnh, nó lần đầu tiên được gọi là khai phá và phân tích quan điểm theo khía cạnh trong công trình nghiên cứu của Hu và các cộng sự (2004).

1.2 Một số khái niệm và bài toán cơ bản trong phân tích quan điểm theo khía cạnh

1.2.1 Một số khái niệm

1.2.2 Một số bài toán

1.3 Các nghiên cứu liên quan

1.4 Tình hình nghiên cứu hiện nay

Trong những năm gần đây một số mô hình học biểu diễn đã đạt được nhiều kết quả xuất sắc trong lĩnh vực xử lý ngôn ngữ tự nhiên. Các mô hình học biểu diễn đã được đề xuất với các mức, như mức từ, mức câu, mức đoạn văn và mức cả văn bản.

Học biểu diễn (representation learning) hay còn gọi là học đặc trưng (feature learning)(Bengio và các cộng sự, 2014) là một lĩnh vực của học máy. Hầu hết các kỹ thuật học biểu diễn được xây dựng dựa trên mô hình mạng nơ-ron với nhiều tầng ẩn và làm việc thực hiện thông qua các hàm chuyển phi tuyến như hàm *tanh, sigmod*. Lĩnh vực xử lý tín hiệu và nhận dạng tiếng nói là lĩnh vực áp dụng kỹ thuật học biểu diễn sớm nhất (Bengio và các cộng sự, 1993), tiếp đến là lĩnh vực phân loại ảnh (Hinton và các cộng sự, 2006). Trong lĩnh vực xử lý ngôn ngữ tự nhiên, học biểu diễn được giới thiệu lần đầu vào năm 1986 bởi Hinton và các cộng sự và được phát triển vào năm 2003 với mô hình mạng nơ-ron ngôn ngữ của Bengio và các cộng sự. Tuy nhiên sự bùng nổ các kỹ thuật học biểu diễn cho lĩnh vực xử lý ngôn ngữ tự nhiên được bắt đầu từ năm 2013 đến nay. Một số mô hình tiêu biểu, học biểu diễn mức từ như Word2Vec (Mikolov và các cộng sự, 2013) và Glove (Pennington và các cộng sự, 2013). Học biểu diễn mức câu hay mức đoạn văn hoặc cả văn bản, có mô hình học không giám sát Paragraph (Quoc và các cộng sự, 2014), mô hình học biểu diễn câu giám sát thông qua một công việc cụ thể như mô hình mạng nơ-ron tích chập (Kim và các cộng sự, 2014).

Một số nghiên cứu khai phá và phân tích quan điểm dựa trện khía cạnh đã áp dụng các kỹ thuật biểu diễn để khắc phục điểm yếu về ngữ nghĩa của từ. Và đạt được mức độ ngữ nghĩa của câu, qua đó kết quả của các bài toán cũng đã được cải thiện như: (Pavlopoulos và các cộng sự, 2014) đã mở rộng phương pháp trích xuất khía cạnh của (Zhuang và các cộng sự, 2006) bằng cách dùng các véc-tơ từ. Poria và các cộng sự (2016) đề xuất mô hình mạng nơ-rơn tích chập nhiều tầng cho công việc trích xuất từ thể hiện khía cạnh. (Wang và các cộng sự, 2016) đề xuất

mô hình long short term memory và Tang và các cộng sự (2016) đề xuất mô hình mạng nơ-ron nhớ sâu cho bài toán phân loại quan điểm khía cạnh. Tuy nhiên, hầu hết các nghiên cứu giải quyết các bài toán ở mức câu và sử dụng các véc-tơ biểu diễn từ được học từ các mô hình học không giám sát Word2Vec hoặc GloVe. Cả hai mô hình Word2Vec và GloVe chỉ bắt được mối quan hệ ngữ nghĩa của các từ dựa trên các ngữ cảnh và bị trượt thông tin về khía cạnh và quan điểm khía cạnh. Hai thông tin khía cạnh và quan điểm khía cạnh là hai thông tin quan trọng, nó được thể hiện rõ thông qua các từ và câu trong các ý kiến đánh sản phẩm/dịch vụ. Theo hiểu biết của chúng tôi, chưa có nghiên cứu nào học biểu diễn mức khía cạnh cho bài toán xác định hạng và trọng số khía cạnh ẩn của sản phẩm/dịch vụ. Cũng chưa có nghiên cứu nào khai thác đa phiên bản véc-tơ biểu diễn từ cho các công việc của phân tích quan điểm theo khía cạnh, và hầu hết các mô hình học biểu diễn mức từ không bắt được ba loại thông tin: ngữ nghĩa, khía cạnh và quan điểm khía cạnh.

1.5 Các đóng góp của luận án

Luận án trình bày 05 kết quả đề xuất chính, góp phần giải quyết các vấn đề nêu trên.

- Thứ nhất, luận án coi từng khía cạnh được đề cập trong mỗi ý kiến đánh giá sản phẩm/dịch vụ gồm nhiều câu văn bản và coi nội dung văn bản của mỗi khía cạnh như là một đoạn văn. Sau đó luận án đề xuất mô hình học biểu diễn khía cạnh và thực hiện dự đoán hạng khía cạnh, trọng số khía cạnh ẩn dựa trên mô hình đã đề xuất. Đề xuất này đã được công bố trong kỷ yếu hội nghị quốc tế *Computational Social Network (CSoNet)* năm 2016.

- Thứ hai, để khai thác thông tin chung về mức độ quan trọng của các khía cạnh sản phẩm/dịch vụ cho các nhà quản lý sản phẩm/dịch vụ, luận án đề xuất mô hình xác định trọng số khía cạnh chung. Mô hình đề xuất giả thiết từng khía cạnh của các sản phẩm/dịch vụ chịu ảnh hưởng một trọng số, mức độ quan trọng chung. Mô hình đề xuất đã được công bố trong tạp chí quốc tế *Indian Journal of Science and Technology* năm 2016.

- Thứ ba, để tăng khả năng biểu diễn mức câu và mức khía cạnh phù hợp với tự nhiên nhất. Đồng thời hạn chế nhược điểm của các mô hình học biểu diễn không giám sát. Luận án đề xuất mô hình học biểu diễn đa tầng cho bài toán xác định hạng khía cạnh và trọng số khía cạnh ẩn. Mô hình đề xuất dựa trên giả thiết giữa các khía cạnh có mối quan hệ với nhau và trong mỗi biểu diễn

khía cạnh tồn tại mối quan hệ đó. Đề xuất đã được công bố trong tạp chí ISI-SCIE: *Data and Knowledge Engineering (DKE)* năm 2018.

- Thứ tư, luận án khắc phục những điểm yếu của các mô hình học biểu diễn từ không giám sát bằng đề xuất mô hình học giám sát. Đề xuất này giúp cho các véc-tơ biểu diễn từ bắt được ba loại thông tin: ngữ nghĩa, khía cạnh và quan điểm. Đề xuất này đã được công bố trong kỷ yếu hội nghị quốc tế *Text, Speech, and Dialogue (TSD)* năm 2017, và trong kỷ yếu hội nghị quốc tế *the Pacific Association for Computational Linguistics (PACLING)* năm 2017.

- Thứ năm, để khai thác hiệu quả sự kết hợp nhiều thông tin khác nhau, cụ thể là thông tin ở mức từ và mức ký tự cho các công việc của phân tích quan điểm theo khía cạnh. Luận án đề xuất mô hình khai thác đa phiên bản các véc-tơ biểu diễn từ và các véc-tơ biểu diễn ký tự. Mô hình đề xuất giả thiết rằng các véc-tơ biểu diễn được học từ các mô hình với các tập dữ liệu khác nhau có khả năng bắt được các khía cạnh khác nhau của ngôn ngữ. Đề xuất này đã công bố trong tạp chí ISI-SCI: *International Journal of Approximate Reasoning* năm 2018.

1.6 Bố cục của luận án

Ngoài phần mở đầu và kết luận, luận án được tổ chức thành 05 chương phù hợp với các công bố liên quan của luận án, với bố cục như sau:

- **Chương 1.** Giới thiệu tổng quan về các vấn đề nghiên cứu trong luận án. Luận án phân tích, đánh giá chung các công trình nghiên cứu liên quan; nêu ra một số vấn đề còn tồn tại mà luận án sẽ tập trung giải quyết; xác định các vấn đề và đóng góp mới của luận án.

- **Chương 2.** Trình bày các tiếp cận cho phân tích quan điểm theo khía cạnh và học biểu diễn.

- **Chương 3.** Trình bày nội dung, kết quả nghiên cứu mô hình đề xuất xếp hạng và xác định trọng số ẩn khía cạnh sản phẩm/dịch vụ. Bên cạnh đó, mô hình xác định trọng số khía cạnh chung cũng sẽ được trình bày.

- **Chương 4.** Trình bày nội dung, kết quả nghiên cứu hai mô hình học véc-tơ từ cho phân tích quan điểm theo khía cạnh.

- **Chương 5.** Trình bày nội dung, kết quả nghiên cứu mô hình khai thác đa véc-tơ biểu diễn từ và véc-tơ biểu diễn ký tự cho phân tích quan điểm theo khía cạnh.

Chương 2

Các tiếp cận cho phân tích quan điểm theo khía cạnh và học biểu diễn

2.1 Các tiếp cận cho phân tích quan điểm theo khía cạnh

2.1.1 Trích xuất khía cạnh

Sử dụng danh từ và cụm danh từ thường xuyên

Sử dụng mối quan hệ của từ thể hiện quan điểm và khía cạnh

Thuật toán phân đoạn khía cạnh (Aspect Segmentation)

2.1.2 Xếp hạng khía cạnh

Xếp hạng dựa trên thuật toán PRank

Thuật toán xếp hạng khía cạnh Good Grief

2.1.3 Thuật toán xác suất xếp hạng khía cạnh

2.2 Các mô hình học biểu diễn mức từ, câu, đoạn hoặc cả văn bản

2.2.1 Mô hình học biểu diễn véc-tơ từ Word2Vec

Mô hình Word2Vec là một mô hình học biểu diễn mỗi từ thành một véc-tơ có các phần tử mang giá trị diễn tả mối quan hệ giữa từ này với từ khác do Mikolov và các cộng sự (2013) đề xuất. Mô hình Word2Vec có khả năng làm việc với những tập dữ liệu và có hai kiến trúc mạng nơ-ron đơn giản: Mô hình túi từ liên tục (Continuous Bag-of-Words (CBOW)) và mô hình Skip-gram.

2.2.2 Mô hình véc-tơ Paragraph

Mô hình Word2Vec học được véc-tơ biểu diễn của một từ mà có thể bắt được ngữ nghĩa của từ đó. Trong mô hình véc-tơ paragraph, Lê Việt Quốc và các cộng sự (2014) mở rộng mô hình học biểu từ Word2Vec để có thể học biểu diễn mức cao hơn cho mức câu, mức đoạn văn, hoặc cả một văn bản. Thông qua kết quả thực nghiệm, các tác giả đã chỉ ra rằng mô hình véc-tơ Paragraph đạt được kết quả thực hiện tốt hơn các mô hình trước đó trong bài toán phân loại văn bản và phân tích ngữ nghĩa.

2.2.3 Mô hình mạng nơ-ron tích chập CNN

Mô hình mạng CNN lần đầu được giới thiệu vào năm 1988 bởi Lecun và các cộng sự. CNN là một mô hình học sâu gồm một số tầng tích chập kết hợp với các hàm kích hoạt phi tuyến (nonlinear activation function) như *ReLU* hay *Tanh* để tạo ra thông tin trừu tượng hơn (abstract/higher-level) cho các tầng tiếp theo, từng tầng tích chập tương ứng với một vài bộ lọc (filter) tích chập được áp dụng học đặc trưng (feature) cho đối tượng đầu vào được tốt hơn.

Tích chập (Convolution) trong ma trận câu

Trong lĩnh vực xử lý ngôn ngữ tự nhiên mức câu, khi áp dụng mô hình CNN thì công việc đầu tiên mô hình mạng CNN phải thực hiện là thực hiện phép toán tích chập trong ma trận câu. Giả sử chúng ta có một ma trận A_{mxn} biểu diễn cho một câu gồm có m từ, từng hàng biểu diễn cho một véc-tơ từ - n chiều thuộc câu đó. Khi đó, về hình thức, chúng ta có thể xem tích chập như một cửa sổ trượt (sliding window) w_{hxk} áp dụng lên ma trận A_{mxn}, điều kiện $h < m$.

Mô hình phân lớp câu quan điểm sử dụng mạng tích chập CNN

2.2.4 Mô hình véc-tơ kết hợp

Mitchell và các cộng sự (2008) đã sử dụng các luật kết hợp với các phép toán cộng và nhân véc-tơ biểu diễn từ để sinh ra mức biểu diễn tốt hơn, cho các mức cao hơn như mức câu, đoạn hoặc cả văn bản. Dựa trên các luật véc-tơ kết hợp, Hermann và các cộng sự (2014) đã giới thiệu hai hàm kết hợp các véc-tơ biểu diễn từ, tên là ADD và BI cho học biểu diễn câu và văn bản. Hàm ADD thực hiện biểu diễn câu bằng cách cộng tất cả các véc-tơ biểu diễn từ lại với nhau. Hàm BI được thiết kế để bắt lấy thông giữa các cặp từ kết hợp với nhau, họ sử dụng một hàm không tuyến tính (hàm *tanh*) thực hiện thông qua các cặp từ (bi-gram pairs).

Chương 3

Xác định hạng và trọng số khía cạnh của sản phẩm/dịch vụ sử dụng mô hình mạng nơ-ron

Trong chương này, luận án trình bày ba đề xuất liên quan đến bài toán Xác định hạng và trọng số khía cạnh của sản phẩm/dịch vụ. Ba đề xuất gồm: (1) mô hình mạng nơ-ron một lớp ẩn xác định hạng và trọng số ẩn của sản phẩm/dịch vụ sử dụng biểu diễn đặc trưng khía cạnh được học mô hình Paragraph; (2) mô hình mạng nơ-ron đa lớp ẩn xác định hạng và trọng số ẩn của sản phẩm/dịch vụ; (3) mô hình học trọng số khía cạnh chung được giám sát bởi hạng khía cạnh và hạng chung của sản phẩm/dịch vụ.

3.1 Xác định hạng và trọng số ẩn của sản phẩm/dịch vụ sử dụng mô hình mạng nơ-ron một lớp ẩn

Các công việc của bài toán xác định hạng và trọng số ẩn của riêng từng sản phẩm/dịch vụ gồm các công việc: (1) Tiền xử lý dữ liệu; (2) Phân đoạn khía cạnh; (3) Học biểu diễn khía cạnh; (4) Xác định hạng và trọng số khía cạnh.

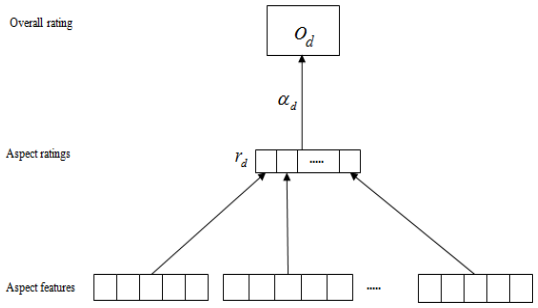
3.1.1 Phân đoạn khía cạnh (Aspect Segmentation)

3.1.2 Học biểu diễn khía cạnh bằng mô hình véc-tơ Paragraph

3.1.3 Xác định hạng và trọng số khía cạnh ẩn sử dụng mô hình mạng nơ-ron một lớp ẩn

Chúng tôi giả thiết rằng cả trọng số khía cạnh và hạng khía cạnh ẩn trong mô hình mạng nơ-ron và chúng tôi gọi mô hình này là mô hình mạng nơ-ron xếp hạng

ẩn (Latent Rating Neural Network Model (LRNN)). Hình 3.1. là một minh họa kiến trúc mô hình LRNN.



Hình 3.1: Minh họa mô hình mạng nơ-ron LRNN xếp hạng ẩn

Ký hiệu $\mathrm{w}_i = (\mathrm{w}_{i1},\mathrm{w}_{i2},...,\mathrm{w}_{in})$ là véc-tơ trọng số của khía cạnh A_i. Đối với văn bản đánh giá $d \in D$, hạng khía cạnh r_{di} của khía cạnh A_i được sinh tại tầng ẩn của mô hình mạng và được tính bởi công thức:

$$r_{di} = \mathrm{sigm}(\sum_{l=1}^{n} \mathrm{x}_{dil}\mathrm{w}_{il} + \mathrm{w}_{i0}) \tag{3.1.1}$$

với $\mathrm{sigm}(y) = 1/(1+e^{-y})$, w_{i0} là độ lệch, véc-tơ đặc trưng x_{di} được học từ mô hình véc-tơ Paragraph.

Hạng chung được sinh tại đầu ra của mô hình và nó được tính dựa trên tổ hợp tuyến tính của a_d và r_d như sau:

$$\hat{O_d} = \sum_{i=1}^{k} r_{di}\alpha_{di} \tag{3.1.2}$$

với điều kiện $\sum_{i=1}^{k} \alpha_{di} = 1$, $0 \le \alpha_{di} \le 1$, $i = 1, 2, \ldots, k$

Hàm giá cross entropy cho tập dữ liệu $\mathrm{D}= \{(\mathrm{X}_d, O_d)\}_{d=1}^{|D|}$ được trình bày như sau:

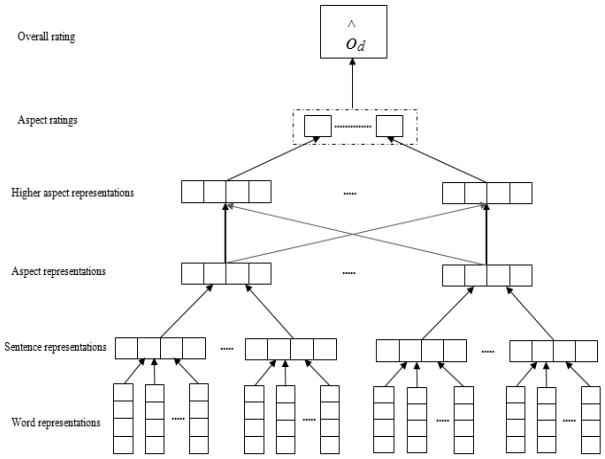
$$E(\mathrm{w}, \hat{\alpha}) = -\sum_{d\in \mathrm{D}} (O_d \log \hat{O_d} + (1-O_d)\log(1-\hat{O_d})) \tag{3.1.3}$$

Để học được mô hình LRNN, chúng tôi xây dựng 1 thuật toán lặp dựa trên thuật toán lan truyền ngược (backpropagation algorithm) để cực tiểu hàm mục tiêu 3.1.3.

3.2 Xác định hạng và trọng số ẩn của sản phẩm/dịch vụ sử dụng mô hình mạng nơ-ron đa lớp ẩn

Hình 3.2 minh họa kiến trúc mô hình mạng nơ-ron học đa tầng đề xuất của chúng tôi. Trong mô hình này, từng từ từ văn bản đánh giá sản phẩm/dịch vụ đầu vào được chuyển vào trong véc-tơ biểu diễn từ tương ứng bằng mô hình Word2Vec (Mikolov và các cộng sự, 2013). Sau đó chúng tôi kết hợp tất cả các từ trong một câu để sinh ra biểu diễn của câu bằng mô hình véc-tơ kết hợp *compositional vector model* (Hermann và các cộng sự, 2014).

Mô hình của chúng tôi là một mô hình mạng nơ-ron gồm sáu tầng: (1) biểu diễn từ; (2) biểu diễn câu; (3) biểu diễn khía cạnh; (4) biểu diễn khía cạnh mức cao; (5) hạng khía cạnh; (6) hạng chung. Chúng tôi đặt tên mô hình này là FULL-LRNN-ASR, với LRNN là mô hình mạng nơ-ron đánh giá khía cạnh ẩn chuẩn "Latent Rating Neural Network" và ASR ("Aspect Semantic Representation") có nghĩa là biểu diễn ngữ nghĩa khía cạnh. Mô hình FULL-LRNN-ASR có một phiên bản khác, tên là LRNN-ASR. Mô hình LRNN-ASR không có Tầng biểu diễn khía cạnh mức cao như mô hình FULL-LRNN-ASR.



Hình 3.2: Minh họa mô hình học biểu diễn đa tầng cho phân tích quan điểm theo khía cạnh

Đối với tập dữ liệu $D = \{d_1, d_2, ..., d_{|D|}\}$ chúng ta có hàm mục tiêu của mô hình như sau:

$$E(U, V, W, \hat{\alpha}) = -\sum_{d \in D} (O_d \log \hat{O_d} + (1 - O_d) \log(1 - \hat{O_d})) \quad (3.2.4)$$

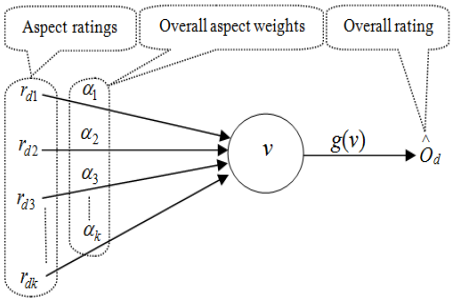
Trong đó, ký hiệu $U = [U_1^*, U_2^*, ..., U_k^*]$ là tập các tham số cho việc học các biểu mức câu tại tầng biểu diễn câu, tương tương ứng với k khía cạnh. $U_i^* = \{U_i, u_{i0}\}$

bao gồm ma trận trọng số và véc-tơ độ lệch tương ứng với khía cạnh A_i; ký hiệu $V = [V_1^*, V_2^*, ..., V_k^*]$ là tập các tham số cho việc học biểu diễn khía cạnh tại tầng biểu diễn khía cạnh, tương ứng với k khía cạnh. Trong đó, $V_i^* = \{V_i, v_{i0}\}$ gồm ma trận trọng số và véc tơ độ lệch tương ứng với khía cạnh A_i; ký hiêu $W = [w_1^*, w_2^*, ..., w_k^*]$ là tập tham số cho vệc xác định hạng khía cạnh, với $w_i^* = \{w_i, w_{i0}\}$ gồm véc-tơ trọng số w_i và độ lệch w_{i0} tương ứng với khía cạnh A_i, $i = 1, \ldots, k$.

Để học mô hình LRNN-ASR, chúng tôi xây dựng 1 thuật toán lặp dựa trên thuật toán lan truyền ngược (backpropagation) xác định cực tiểu hàm mục tiêu 3.2.4 .

3.3 Xác định trọng số khía cạnh chung của sản phẩm/dịch vụ

Hình 3.3 minh họa mô hình (được đặt tên là NNAWs *(Neural Network Aspect Weights)*) xác định trọng số khía cạnh chung của sản phẩm/dịch vụ mà luận án đề xuất.



Hình 3.3: Minh họa mô hình xác định hạng khía cạnh chung

Đầu vào là các véc-tơ hạng khía cạnh của từng sản phẩm/dịch vụ, đầu ra là các hạng chung tương ứng. Trọng số khía cạnh chung (overall aspect weights) được giả thiết là các trọng số của mô hình. Quá trình xác định trọng số khía cạnh chung là quá trình học mô hình dự đoán hạng chung (overall rating). Để học được mô hình NNAWs, chúng tôi xây dựng 1 thuật toán lặp dựa trên thuật toán lan truyền ngược.

3.4 Thực nghiệm

Tập dữ liệu được sử dụng trong thực nghiệm của chúng tôi được cung cấp bởi (Wang và các cộng sự. 2013)*http://times.cs.uiuc.edu/wang296/Data*. Gồm 174,615 ý kiến đánh giá của 1,768 khách sạn. Tập các ý kiến của dịch vụ khách sạn gồm

năm khía cạnh: *Value*, *Room*, *Location*, *Cleanliness*, và *Service*. Từng ý kiến đánh giá được gán với một hạng chung cho khách sạn và từng khía cạnh cũng được gán với một hạng khía cạnh. Các hạng được gán từ 1 sao đến 5 sao.

Đối với từng sản phẩm/dịch vụ khách sạn, chúng tôi xây dựng văn bản đánh giá của nó bằng cách hợp nhất tất các ý kiến đánh giá vào một văn bản chung. Hạng chung của văn bản được tính bằng trung cộng các hạng chung của các ý kiến đánh giá. Ngoài ra, để giá trị hạng chung và hạng khía cạnh theo giả thiết phù hợp với giá trị các hàm dự đoán theo mô hình đề xuất, chúng tôi chuẩn hóa hạng chung và hạng khía cạnh thành các số thực nằm trong đoạn $[0, 1]$ bằng cách lấy giá trị của hạng chia cho 5. Áp dụng thuật toán phân đoạn của (Wang và các cộng sự, 2013) để xác định khía cạnh và phân đoạn các văn bản đánh giá.

3.4.1 Đánh giá

Để đánh giá phương pháp đề xuất, luận án sử dụng biểu diễn đặc trưng của các khía cạnh theo các trường hợp sau: **Túi từ (Bag of words)**: gồm 3987 từ để biểu diễn các khía cạnh; **Trung bình véc-tơ từ (Word vector averaging)**: Sử dụng các véc-tơ biểu diễn từ được học từ mô hình Word2Vec[1] với cỡ ngữ cảnh là 7, tần số xuất hiện tối thiểu của các từ là 7. Từng khía cạnh của một văn bản đánh giá được biểu diễn bằng cách lấy trung bình cộng của các véc-tơ từ; **Trung bình véc tơ câu (Sentence vector averaging)**: áp dụng mô hình Sentence2Vec[2] với kích cỡ cửa sổ ngữ cảnh là 7, với từng khía cạnh trên văn bản đánh giá được biểu biểu diễn bằng cách lấy trung bình các véc-tơ câu; **Véc-tơ paragraph (Paragraph vector)**: Ap dụng mô hình Doc2Vec[3] với kích cỡ cửa sổ ngữ cảnh là 7, ngưỡng tần suất xuất hiện từ là 7, số chiều véc-tơ paragraph là 200 để học biểu diễn các khía cạnh.

Mô hình cơ sở là hồi quy đánh giá ẩn *Latent Rating Regression model (LRR)* (Wang và các cộng sự, 2013). Ba độ đo được sử dụng cho đánh giá dự đoán hạng khía cạnh, bao gồm: (1) Độ lệch trung bình bình phương của hạng khía cạnh, ký hiệu là Δ_{aspect} (Δ_{aspect} mà nhỏ hơn thì có nghĩa là tốt hơn), (2) Độ đo tương quan giữa các hạng khía cạnh (ρ_{aspect}, cao hơn có nghĩa là tốt hơn), (3) Độ đo tương quan của mỗi loại hạng khía cạnh thông qua toàn bộ tập dữ liệu đánh giá (ρ_{review} cao hơn có nghĩa là tốt hơn). Trong bảng 3.1 chúng tôi thể hiện ba độ đo đạt được của từng phương pháp trong từng trường hợp biểu diễn đặc trưng khía cạnh. Trong tất cả các trường hợp chúng ta thấy rằng cả mô hình LRR và LRNN thực hiện tốt

[1]https://github.com/piskvorky/gensim/
[2]https://github.com/klb3713/sentence2vec
[3]https://github.com/piskvorky/gensim/

Bảng 3.1: So sánh các mô hình xác định hạng khía cạnh ẩn trong bốn trường hợp biểu diễn khía cạnh

Feature kind	Method	Δ_{aspect}	P_{aspect}	P_{review}
Bag of words	LRR	0.752	0.341	0.621
	LRNN	0.817	0.445	0.587
Word vector averaging	LRR	0.756	0.398	0.644
	LRNN	0.753	0.459	0.641
Sentence vector averaging	LRR	0.781	0.432	0.646
	LRNN	0.770	0.465	0.645
Paragraph vector	LRR	0.747	0.424	0.658
	LRNN	0.742	0.432	0.667
	LRNN-ASR	**0.703**	**0.497**	**0.675**
	FULL-LRNN-ASR	**0.596**	**0.512**	**0.741**

nhất khia sử dụng mô hình véc-tơ paragraph học biểu diễn trực tiếp cho các khía cạnh.

Để đánh giá chất lượng mô hình NNAWs, chúng tôi thực hiện ba phương pháp liên quan. Thứ nhất là mô hình hồi quy xác suất LRR (Wang và các cộng sự, 2013), để xác định trọng số khía cạnh chung, chúng tôi lấy trung bình cộng trọng số khía cạnh của riêng từng khách sạn. Ký hiệu véc-tơ trọng số cho phương pháp này là $\overline{\alpha}_{PRR}$. Thứ hai là phương pháp xác định trọng số khía cạnh chung (Zha và các cộng sự, 2014) (ký hiệu là véc-tơ α_F) bằng cách thống kê tần xuất xuất hiện từ thể hiện quan điểm. Thứ ba là phương pháp LRNN, giống như với phương pháp *LRR*, chúng tôi tính trọng số khía cạnh chung trong phương pháp này bằng cách lấy trung bình cộng các trọng số khía cạnh riêng trên từng sản phẩm, ký hiệu véc-tơ này là $\overline{\alpha}_{LRNN}$. Kết quả thực nghiệm đã cho thấy trọng số khía cạnh chung α_{NNAWs} có chất lượng nhất.

3.5 Kết luận

Trong chương này chúng tôi đã trình bày ba phương pháp sử dụng mô hình mạng nơ-ron cho việc xác định hạng và trọng số khía cạnh. Thứ nhất là một phương pháp xác định hạng và trọng số khía cạnh ẩn sử dụng mạng nơ-ron một lớp ẩn, thứ hai là phương pháp xác định hạng và trọng số khía cạnh ẩn sử dụng mạng nơ-ron đa lớp ẩn. Phương pháp thứ ba là một mô hình mạng nơ-ron xác định trọng số khía cạnh chung cho sản phẩm/dịch vụ.

Chương 4

Học véc-tơ biểu diễn từ cho phân tích quan điểm theo khía cạnh

4.1 Giới thiệu

Mặc dù các véc-tơ biểu diễn từ được học từ các mô hình dựa trên ngữ cảnh đã được sử dụng hiệu quả trong nhiều công việc xử lý ngôn ngữ tư nhiên (Collobert và các cộng sự, 2008), nhưng chúng được xem là biểu diễn thiếu thông tin khi được áp dụng vào các công việc cụ thể (Tang và các cộng sự, 2014) . Trong phân tích quan điểm dựa trên khía cạnh, các véc-tơ biểu diễn từ có thể trượt mất thông tin về khía cạnh và quan điểm. Ví dụ, cho một câu "*Rất thích dùng BIDV, nhân viên lúc nào cũng thân thiện và nhiệt tình*" được gán hai nhãn, nhãn khía cạnh "Dịch vụ" và nhãn quan điểm là "Tích cực", với câu này các mô hình học véc-tơ biểu diễn dựa trên các ngữ cảnh không bắt được thông tin khía cạnh "Dịch vụ" và thông tin quan điểm "Tích cực". Bài toán ở đây là làm thế nào chúng ta có thể mã hóa được các loại thông tin này vào trong các véc-tơ biểu diễn từ.

Trong chương này chúng tôi đề xuất hai mô hình sử dụng tập dữ liệu gồm các câu được gán nhãn và các câu không được gán nhãn để học các véc-tơ biểu diễn từ cho phân tích quan điểm dựa trên khía cạnh.

4.2 Các bài toán học biểu diễn từ cho phân tích quan điểm theo khía cạnh

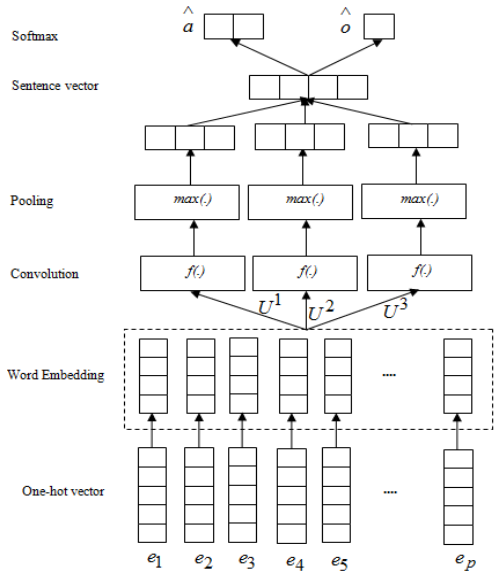
4.3 Phương pháp đề xuất

Luận án đề xuất hai mô hình học véc-tơ biểu diễn từ. Hai mô hình sử dụng một số ký hiệu cần thiết như sau:

Cho một tập các câu $D = \{d_1, d_2, ..., d_{|D|}\}$ được trích xuất từ một tập ý kiến đánh giá của một sản phẩm/dich vụ cụ thủ (ví dụ: dịch vụ nhà hàng). Ký hiệu k là số lượng nhãn khía cạnh và m là số lượng nhãn quan điểm khía cạnh. Ký hiệu $a_d \in \mathrm{R}^k$ là một véc-tơ nhị phân của các nhãn khía cạnh trong câu d. Từng giá trị trong a_d xác nhận câu d có thảo luận về một khía cạnh hay không. Ký hiệu $o_d \in \mathrm{R}^m$ là một véc-tơ nhị phân của các quan điểm trong câu d. Từng thành phần trong véc-tơ o_d xác nhận câu d có thảo luận một quan điểm khía cạnh hay không.

4.3.1 Mô hình tinh chỉnh véc-tơ biểu diễn từ

Mô hình tinh chỉnh véc-tơ biểu diễn từ được đề xuất là một mô hình mạng nơ-ron tích chập. Thực hiện tinh chỉnh các véc-tơ được học từ các mô hình không giam sát như Word2Vec hay Glove. Hình 4.1 là một minh họa mô hình tinh chỉnh véc-tơ biểu diễn từ *Word Embedding Fine-Tuning (WEFT)*.



Hình 4.1: Minh họa mô hình tinh chỉnh véc-tơ biểu diễn từ WEFT

Hàm mục tiêu của mô hình trên một tập các câu huấn luyện D như sau:

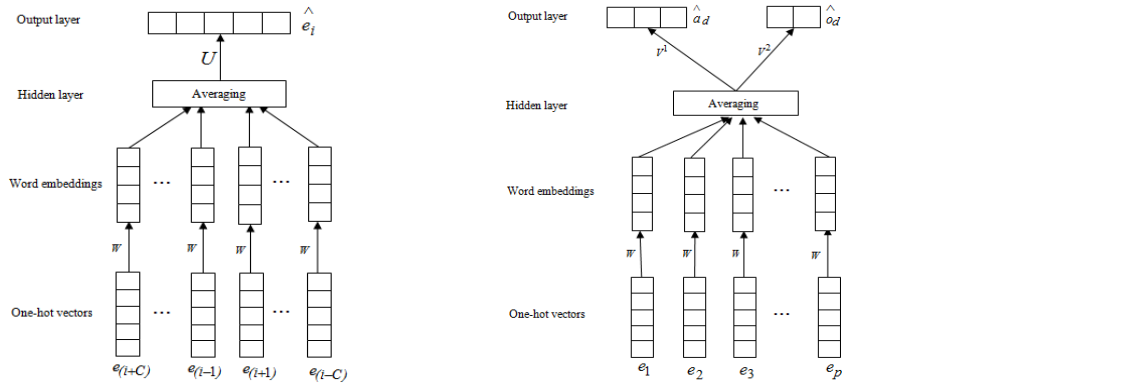
$$E(\theta) = -\sum_{d \in D}(\sum_{i=1}^{k} a_{di} \log \hat{a}_{di} + \sum_{i=1}^{m} o_{di} \log \hat{o}_{di}) + \frac{1}{2}\lambda_\theta \|\theta\|^2 \qquad (4.3.1)$$

với $\hat{a}_{di}$ và $\hat{o}_{di}$ được tính theo mô hình. $\theta = [U^1, U^2, U^3, W, V^1, V^2, u^1, u^2, u^3, b^1, b^2]$, λ_θ là hằng số chuẩn hóa.

Để học được mô hình WEFT, chúng tôi xây dựng 1 thuật toán lặp dựa trên thuật toán lan truyền ngược (backpropagation) để cực tiểu hóa hàm mục tiêu.

4.3.2 Mô hình học véc-tơ biểu diễn từ SSCWE

Mô hình học véc-tơ biểu diễn từ SSCWE gồm hai thành phần: Thành phần nhúng ngữ nghĩa của véc-tơ từ và thành phần nhúng khía cạnh và quan điểm. Hình 4.2 là một minh họa của hai thành phần này. Trong đó, thành phần nhúng ngữ nghĩa làm việc tương tự như mô hình CBOW của Word2Vec (Mikolov và các cộng sự, 2013), thành phần nhúng khía cạnh và quan điểm sử dụng thông tin khía cạnh và quan điểm khía cạnh làm giám sát trong đầu ra của mô hình.



(a) Thành phần nhúng ngữ nghĩa (b) Thành phần nhúng khía cạnh và quan điểm

Hình 4.2: Hai thành phần của mô hình học véc-tơ biểu diễn từ SSCWE

Hàm mục tiêu của mô hình trên một tập các câu huấn luyện D như sau:

$$E(\theta) = \sum_{d \in \mathrm{D}} \sum_{i=1}^{N_d} H(\hat{e_i}, e_i) + \sum_{d \in D'} H(\hat{a_d}, a_d, \hat{o_d}, o_d) + \frac{1}{2}\lambda_\theta \|\theta\|^2 \qquad (4.3.2)$$

với $\theta = [W, U, u_0, V^1, V^2, b^1, b^2]$, λ_θ là tham số và λ_θ là hằng số chuẩn hóa.

Để học được mô hình SSCWE, chúng tôi xây dựng 1 thuật toán lặp dựa trên thuật toán lan truyền ngược (backpropagation) để cực tiểu hóa hàm mục tiêu.

4.4 Thực nghiệm

4.4.1 Dữ liệu thực nghiệm và các độ đo

Luận án sử dụng hai tập dữ liệu trên miền dữ liệu của các sản phẩm/dịch vụ Nhà hàng cho công việc thực nghiệm. Tập dữ liệu thứ nhất gồm 3,111,239 câu không gán nhãn được trích xuất từ 229,907 ý kiến đánh giá[1]. Tập dữ liệu thứ hai

[1]https://www.yelp.com/datasetchallenge/

gồm 190,655 câu được trích xuất từ 52,574 ý kiến đánh giá[2]. Gồm các câu được gán 5 nhãn khía cạnh *Food, Price, Service, Ambience, Anecdotes*, và *Miscellaneous*. Và 4 nhãn quan điểm *Positive, Negative, Neutral* và *Conflict*. Từng câu được gán hai nhãn: khía cạnh và quan điểm khía cạnh. 75% số lượng câu gán nhãn được sử dụng để học véc-tơ từ, còn lại 25% được sử dụng để đánh giá chất lượng của mô hình WEFT và mô hình SSCWE.

4.5 Đánh giá mô hình WEFT

Mô hình WEFT được đánh giá thông qua các véc-tơ từ được học từ các mô hình: CBOW, skip-gram của Word2Vec và GloVe. Ký hiệu các phiên bản của mô hình WEFT như sau: WEFT-rand sử dụng các véc-tơ từ được khởi tạo ngẫu nhiên và sau đó sẽ được chỉnh sửa trong quá trình huấn luyện mô hình. Các mô hình WEFT-SG, WEFT-CB và WEFT-GV tinh chỉnh các véc-tơ từ được học từ các mô hình tương ứng skip-gram, CBOW và GloVe. Sử dụng các công cụ Word2Vec[3] và GloVe[4] để học các véc-tơ từ, kích thước véc tơ từ được cấu hình bằng 300 và kích thước cửa số ngữ cảnh bằng 4. Trong bảng 4.1 và 4.2 luận án thể hiện kết quả các véc-tơ tinh chỉnh (học từ mô hình WEFT) được sử dụng trong hai công việc của phân tích quan điểm theo khía cạnh: Xác định khía cạnh và phân tích quan điểm.

Bảng 4.1: Kết quả XĐKC

Method	F1 score
SG	77.87
CB	78.54
GV	79.19
WEFT-rand	81.43
WEFT-SG	81.50
WEFT-CB	81.76
WEFT-GV	**82.09**

Bảng 4.2: Kết quả phân loại quan điểm khía cạnh

Method	Pos-F1	Neg-F1	Neu-F1	Con-F1	Accuracy
SG	87.05	52.03	65.74	55.46	78.77
CB	86.93	52.25	66.60	55.93	79.22
GV	87.10	51.07	71.02	57.85	80.35
WEFT-rand	88.65	64.18	74.13	56.40	82.15
WEFT-SG	90.87	64.63	73.82	60.23	83.82
WEFT-CB	93.12	64.70	77.03	61.17	84.05
WEFT-GV	93.61	64.77	77.11	61.43	**84.23**

4.5.1 Đánh giá mô hình SSCWE

Mô hình của chúng tôi với các mô hình cơ sở như sau: **Word2Vec**, **GloVe**, **SCWE**: chỉ gồm thành phần SCWE của mô hình SSCWE, **SSCWE***: chỉ sử dụng các câu được gán nhãn làm đầu vào để học các véc-tơ từ. Số chiều véc-tơ trong tất cả các mô hình là 300. Ngoài ra, chúng tôi cũng sử dụng các véc-tơ từ đã được học

[2]http://spidr-ursa.rutgers.edu/datasets/
[3]https://github.com/piskvorky/gensim/
[4]https://nlp.stanford.edu/projects/glove/

từ những tập dữ liệu khác để so sánh với các mô hình này, gồm Pre-Word2Vec là các Word2Vec [5] và Pre-GloVe là các véc-tơ GloVe[6]. Trong bảng 4.3 và 4.4 luận án thể hiện kết quả các véc-tơ được sử dụng trong hai công việc của phân tích quan điểm theo khía cạnh: Xác định khía cạnh và phân tích quan điểm.

Bảng 4.3: Kết quả XĐKC

Method	F1 score
Word2Vec	78.54
GloVec	79.19
Pre-Word2Vec	77.24
Pre-GloVec	79.01
Our SCWE	80.04
Our SSCWE*	82.12
Our SSCWE	**82.77**

Bảng 4.4: Kết quả phân loại quan điểm

Phương pháp	Pos-F1	Neg-F1	Neu-F1	Con-F1	Accuracy
Word2Vec	86.93	52.25	66.60	55.93	79.22
GloVec	87.10	51.07	71.02	57.85	80.35
Pre-Word2Vec	82.04	49.53	68.44	53.16	79.01
Pre-GloVec	83.95	53.04	65.04	54.04	80.13
Our SCWE	89.54	64.00	74.01	56.30	81.41
Our SSCWE*	93.78	63.81	76.58	61.93	83.85
Our SSCWE	93.80	64.70	76.13	63.02	**84.69**

4.5.2 So sánh hai mô hình WEFT và SSCWE

Trong bảng 4.5 luận án thể hiện kết quả đạt được của mô hình WEFT so với mô hình SSCWE, trong hầu hết các trường hợp mô hình SSCWE cho kết quả nhỉnh hơn mô hình WEFT. Điều này chứng tỏ việc học liên hợp trong mô hình SSCWE tốt hơn là việc tinh chỉnh trong mô hình WEFT.

Bảng 4.5: So sánh kết quả phân loại quan điểm giữa mô hình WEFT và SSCWE

Phương pháp	Pos-F1	Neg-F1	Neu-F1	Con-F1	Accuracy
WEFT-SG	90.87	64.63	73.82	60.23	83.82
WEFT-CB	93.12	64.70	77.03	61.17	84.05
WEFT-GV	93.61	64.77	77.11	61.43	**84.23**
Our SSCWE	93.80	64.70	76.13	63.02	**84.69**

4.6 Kết luận

Trong chương này, luận án đã trình bày hai mô hình mới học véc-tơ biểu diễn từ cho phân tích quan điểm theo khía cạnh. Mô hình thứ nhất là mô hình mạng nơ-ron tích chập WEFT chỉnh sửa các véc-tơ được học từ các mô hình Word2Vec và Glove. Mô hình thứ hai là mô hình SSCWE, sử dụng sự kết hợp của kỹ thuật học giám sát và không giám sát để học véc-tơ từ cho các công việc của phân tích quan điểm dựa trên khía cạnh. Kết quả đạt được tốt hơn các mô hình cơ sở.

[5]https://code.google.com/archive/p/Word2Vec/
[6]http://nlp.stanford.edu/projects/glove/

Chương 5

Khai thác đa véc-tơ biểu diễn từ và ký tự cho phân tích quan điểm theo khía cạnh

5.1 Giới thiệu

Nhằm sử dụng và khai thác hiệu quả các phiên bản véc-tơ biểu diễn từ được công bố sẵn trên mạng internet kết hợp với các véc-tơ biểu diễn ký tự. Trong chương này luận án đề xuất mô hình mạng nơ-ron đa kênh tích chập khai thác đa véc-tơ biểu diễn từ và véc-tơ biểu diễn ký tự cho các công việc của phân tích quan điểm theo khía cạnh. Mục tiêu là giúp mô hình đề xuất bắt được thông tin kết hợp của các véc-tơ biểu diễn từ và các véc-tơ ký tự để sinh ra mức biểu diễn tốt nhất cho câu. Thông qua mục tiêu này các kết quả dự đoán trong phân tích quan điểm theo khía cạnh sẽ được cải thiện.

5.2 Các nghiên cứu liên quan

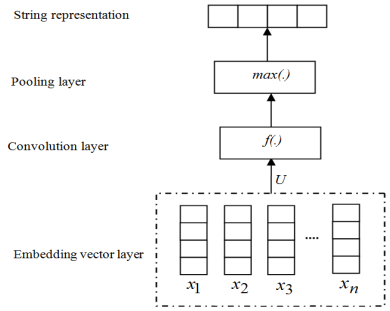
5.3 Mô tả đầu vào/ra của bài toán khai thác các mức biểu diễn cho phân tích quan điểm theo khía cạnh

5.4 Phương pháp đề xuất

Trong đoạn này, đầu tiên chúng tôi trình bày thành phần tích chập gồm hai tầng, tầng tích chập và tầng pooling. Sau đó chúng tôi áp dụng nó để trình bày một mô hình mạng nơ-ron tích chập đa tầng cho phân tích quan điểm theo khía cạnh.

5.4.1 Thành phần tích chập

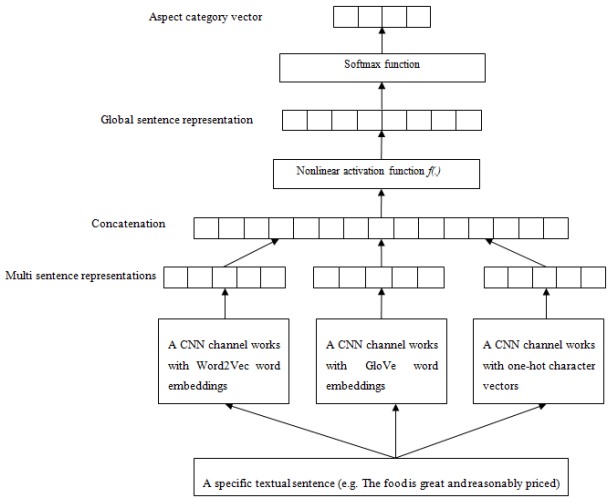
Một mô hình mạng nơ-ron tích chập truyền thống gồm có một tầng tích chập và một tầng thu thập đặc trưng (pooling layer). Chúng tôi sử dụng nó như là một thành phần trong mô hình đề xuất của chúng tôi, vì vậy mà chúng tôi gọi nó là thành phần tích chập.



Hình 5.1: Minh họa thành phần tích chập

5.4.2 Mô hình mạng nơ-ron tích chập đa kênh cho phân tích quan điểm theo khía cạnh

Mô hình đề xuất được minh họa như hình 5.2, gồm có ba kênh CNN, hai kênh đầu tiên, một kênh làm việc với véc-tơ biểu diễn từ WordVec, một kênh làm việc với véc-tơ biểu diễn từ Glove và kênh thứ ba sử dụng các véc-tơ biểu diễn ký tự làm đầu vào.



Hình 5.2: Mô hình mạng nơ-ron tích chập đa kênh MCNN (*Multichannel Convolutional Neural Network*) cho công việc xác định khía cạnh

5.5 Thực nghiệm

5.5.1 Dữ liệu thực nghiệm và cài đặt mô hình MCNN

Dữ liệu thực nghiệm[1] gồm 190,655 câu và được trích xuất từ 52,574 văn bản đánh giá, tập dữ liệu này đã được sử dụng trong nghiên cứu (Brody và các cộng sự). Gồm sáu nhãn khía cạnh *Price, Food, Service, Ambience, Anecdotes*, và *Miscellaneous*, và bốn nhãn quan điểm khía cạnh *Positive, Negative, Neutral* and *Conflict*. Từng câu được gán cả hai nhãn khía cạnh và quan điểm tương ứng của khía cạnh. 75% số lượng câu đã cho làm dữ liệu huấn luyện và 25% số lượng câu còn lại được sử dụng để đánh giá chất lượng mô hình.

Đối với các tập véc-tơ biểu diễn từ và bộ từ điển ký tự, chúng tôi sử dụng hai tập véc-tơ biểu diễn từ đang được sử dụng rộng rãi và công bố trên internet là Word2Vec tại địa chỉ[2] và GloVe tại địa chỉ[3]. Bộ từ điển ký tự được sử dụng gồm 52 ký tự thông thường của tiếng Anh.

5.5.2 Đánh giá

Mô hình đề xuất gồm ba nhóm:

- (1) Nhóm mô hình 1 bao gồm các mô hình chỉ sử dụng một kênh CNN đơn lẻ, CNN1 là mô hình với Kênh 1, sử dụng véc-tơ Word2Vec làm đầu vào; CNN2 chỉ gồm Kênh 2 với các véc-tơ Glove được sử dụng làm đầu vào; CNN3 là một mô hình chỉ gồm Kênh 3 với đầu vào là các véc-tơ one-hot ký tự.

- (2) Nhóm mô hình 2 bao gồm các mô hình lai được tạo từ sự kết hợp giữa các kênh. CNN1+CNN2 là mô hình lai giữa Kênh 1 và Kênh 2; CNN1+CNN2+CNN3 là mô hình lai giữa Kênh 1, Kênh 2 và Kênh 3.

- (3) Nhóm mô hình 3 bao gồm các mô hình sử dụng sự kết hợp giữa các kênh. CNN1+CNN2 là mô hình kết hợp giữa Kênh 1 và Kênh 2; MCNN là mô hình kết hợp giữa ba kênh CNN1, CNN2, và CNN3. Khác với các mô hình lai trong nhóm mô hình (2), quá trình học trong nhóm mô hình này được thực hiện bằng cách học liên hợp để cùng tạo ra một mô hình chung gồm các kênh kết hợp.

Các mô hình cơ sở khác được lựa chọn cho đánh giá như sau:

NLSE model (Astudillo và các cộng sự, 2015): Một mô hình mạng nơ-ron đơn giản sử dụng một tầng ẩn học véc-tơ biểu diễn từ cho phân tích quan điểm. Đầu vào của mô hình là các véc-tơ Word2Vec, sau đó các véc-tơ này được ánh xạ vào một không gian ngữ nghĩa với số chiều nhỏ hơn so với số chiều của Word2Vec trước

[1] http://spidr-ursa.rutgers.edu/datasets/

[2] https://code.google.com/archive/p/Word2Vec/

[3] http://nlp.stanford.edu/projects/glove/

đó. Quá trình học mô hình là quá trình học các véc-tơ biểu diễn từ và học phân loại quan điểm cho văn bản đầu vào.

Hybrid Feature Learning (Zhou và các cộng sự, 2015): Một mô hình mạng nơ-ron học lai đặc trưng cho công việc xác định khía cạnh, chúng tôi ký hiệu mô hình này là HFL. Mô hình gồm hai mô hình con, một mô hình học biểu diễn mức câu tại một tầng ẩn cho phân loại tất cả nhãn khía cạnh mức câu, một mô hình học biểu diễn mức câu tại một tầng ẩn cho từng nhãn khía cạnh. Đặc trưng lai đạt được là véc-tơ biểu diễn câu bằng cách nối hai véc-tơ biểu câu ở tầng ẩn của hai mô hình con với nhau.

CNN-non-static model (Kim và các cộng sự, 2014): Mô hình mạng nơ-ron chỉ gồm một tầng chập thực tinh chỉnh các véc-tơ biểu diễn từ cho phân loại mức câu. Đầu vào của mô hình là các véc-tơ Word2Vec, sau đó các véc-tơ Word2Vec được tính chỉnh trong quá trình học mô hình. Quá trình tinh chỉnh Word2Vec giúp mô hình dự đoán phân loại hơn và các véc-tơ Word2Vec cũng đạt được mức biểu diễn tốt hơn trong không gian ngữ nghĩa.

CNN-multichannel model (Kim và các cộng sự, 2014): Mô hình mạng nơ-ron sử dụng hai kênh tích chập sử dụng các véc-tơ Word2Vec làm đầu vào. Trong đó, các véc-tơ Word2Vec được giữ nguyên trong một kênh, kênh còn lại thực hiện tinh chỉnh Word2Vec trong qua trình huấn luyện mô hình.

CharSCNN model (dos Santos và các cộng sự, 2014): Mô hình mạng nơ-ron sử dụng hai tầng tích chập khai thác thông tin mức từ và mức ký tự cho phân tích quan điểm trong văn bản ngắn. Đối với thông tin mức họ sử các véc-tơ học từ mô hình Word2Vec, thông tin mức ký tự họ sử dụng các véc-tơ one-hot tương tự như mô hình MCNN.

Quan sát thông tin kết quả đạt được của các mô hình trong bảng 5.1 và bảng 5.2, chúng ta thấy rằng: Trong nhóm mô hình 1, hầu hết các mô hình thực hiện không tốt bằng các mô hình cơ sở. Điều này là do các mô hình cơ sở (ngoại trừ mô hình HFL) đều thực hiện tinh chỉnh véc-tơ biểu diễn từ trong quá trình huấn luyện. Các véc-tơ biểu diễn từ được cải thiện, dẫn đến mức biểu diễn câu đầu vào được tốt hơn nên kết quả dự đoán nhãn khía cạnh cũng được tốt hơn. Mô hình học lai HFL không thực hiện tinh chỉnh các véc-tơ biểu diễn từ nhưng nó cũng thực hiện nhỉnh hơn các mô hình CNN1, CNN2 và CNN3. Điều này cho chúng ta thấy rằng, đặc trưng lai sử dụng trong biểu diễn mức câu là quan trọng hơn so với việc chỉ sử dụng riêng lẻ đặc trưng biểu diễn câu trong các kênh CNN1, CNN2, CNN3.

Trong nhóm mô hình 2, mặc dù các mô hình không thực hiện tinh chỉnh các véc-tơ biểu diễn từ trong quá trình huấn luyện như các mô hình cơ sở, nhưng các mô hình thực hiện tốt hơn các mô hình cơ (trừ mô hình CharSCNN) và các mô hình trong nhóm 1. Ngoài ra, chúng ta thầy rằng mô hình lai CNN1+CNN2+CNN3

Bảng 5.1: Kết quả xác định khía canh của mô hình MCNN và các mô hình cơ sở

	Method	**Precision**	**Recall**	**F1 score**
Baselines	NLSE	77.82	81.53	79.63
	HFL	79.11	80.97	80.03
	CNN-non-static	79.08	81.23	80.14
	CNN-multichannel	80.18	81.41	80.79
	CharSCNN	82.30	80.17	81.22
Our	CNN1	78.45	76.87	77.65
	CNN2	80.02	78.57	79.29
	CNN3	77.78	73.18	75.41
	CNN1+CNN2 (hybrid)	81.88	79.79	80.82
	CNN1+CNN2+CNN3 (hybrid)	81.91	80.25	81.07
	CNN1+CNN2	83.40	81.27	82.32
	MCNN	**83.94**	**81.61**	**82.76**

thực hiện nhỉnh hơn mô hình lai CNN1+CNN2. Kết quả này xác nhận vai trò của thông tin mức ký tự sử dụng trong kỹ thuật lai đặc trưng cho mức câu.

Bảng 5.2: Kết quả dự đoán phân loại quan điểm theo khía canh của mô hình MCNN và các mô hình cơ sở

	Method	**Accuracy**
Baselines	NLSE	81.49
	HFL	81.71
	CNN-non-static	82.13
	CNN-multichannel	82.79
	CharSCNN	83.35
Our	CNN1	79.97
	CNN2	80.50
	CNN3	77.83
	CNN1+CNN2 (hybrid)	82.81
	CNN1+CNN2+CNN3 (hybrid)	83.02
	CNN1+CNN2	83.68
	MCNN	**84.16**

5.6 Kết luận

Trong chương 5, luận án đã trình bày mô hình mạng nơ-ron tích chập đa kênh để khai thác đa véc-tơ biểu diễn từ và các véc-tơ biểu diễn ký tự. Các kết quả thực nghiệm đã cho thấy tính hiệu quả của mô hình đề xuất. Đặc biệt thông tin mức ký tự cũng đã cho thấy vai trò quan trọng trong việc kết hợp với thông tin mức từ.

KẾT LUẬN

Các mô hình học biểu diễn đặc trưng mức từ, mức câu, mức khía cạnh có hiệu quả đối với các công việc phân tích quan điểm theo khía cạnh. Trong phần này, chúng tôi tóm lược lại các kết quả chính và những đóng góp của luận án. Ngoài ra, chúng tôi trình bày những định hướng phát triển cho các nghiên cứu tiếp theo trong tương lai.

Các đóng góp của luận án bao gồm:

- Đề xuất mô hình mạng nơ-ron xác định hạng và trọng số khía cạnh ẩn sản phẩm/dịch vụ. Sử dụng các véc-tơ biểu diễn khía cạnh được học từ mô hình véc-tơ Paragraph làm đầu vào.
- Đề xuất mô hình mạng nơ-ron xác định trọng số khía cạnh chung của sản phẩm/dịch vụ.
- Đề xuất mô hình mạng nơ-ron học đa tầng biểu diễn cho bài toán xác định hạng và trọng số khía cạnh ẩn.
- Đề xuất hai mô hình học véc-tơ biểu diễn từ: một mô hình thực hiện tinh chỉnh các véc-tơ được học từ mô hình Word2Vec và Glove; một mô hình học véc-tơ biểu diễn từ gồm hai thành phần: một thành phần được thiết kế dựa trên mô hình Word2Vec thực hiện bắt mối quan hệ ngữ nghĩa giữa các từ, một thành phần sử dụng các thông tin được giám sát để bắt lấy thông tin khía cạnh và quan điểm khía cạnh.
- Đề xuất mô hình mạng nơ-ron đa kênh tích chập khai thác đa phiên bản véc-tơ biểu diễn từ và véc-tơ biểu diễn ký tự.

Tất cả các mô hình đề xuất đã được thực nghiệm đánh giá chi tiết thông qua các tập dữ liệu tiếng Anh, trong miền dữ liệu là các sản phẩm/dịch vụ gồm các khía cạnh đã được các khách hàng thảo luận/đánh giá trong các ý kiến. Nhìn chung các kết quả đạt được trong các mô hình đề xuất đã nhỉnh hơn các phương pháp truyền thống. Đặc biệt với việc sử dụng mô hình mạng nơ-ron nhiều tầng học biểu diễn xác định hạng và trọng số khía cạnh ẩn đã chứng tỏ được sự hiệu quả vượt trội so với các phương pháp khác.

Trong tương lai chúng tôi tìm hiểu và thực hiện đánh giá các mô hình đề xuất trên các tập dữ liệu tiếng Anh khác. Chúng tôi cũng định hướng chú trọng việc áp dụng các mô hình đề xuất vào các hệ thống phân tích dữ liệu thực tế bằng tiếng Việt, như dữ liệu về Ngân hàng, Chứng khoán, Điện thoại di động.

Danh mục công trình khoa học của tác giả liên quan đến luận án

[1] Duc-Hong Pham, and Anh-Cuong Le, "*Exploiting Multiple Word Embeddings and One-hot Character Vectors for Aspect-Based Sentiment Analysis*", International Journal of Approximate Reasoning (IJAR), 103, 2018, pp. 1-10. (ISI-SCI)

[2] Duc-Hong Pham, and Anh-Cuong Le, "*Learning Multiple Layers of Knowledge Representation for Aspect Based Sentiment Analysis*", Journal: Data&Knowledge Engineering (DKE), 114, 2018, pp. 26-39. (ISI-SCIE)

[3] Duc-Hong Pham, Thi-Thanh-Tan Nguyen, and Anh-Cuong Le, "*Fine-Tuning Word Embeddings for Aspect-based Sentiment Analysis*", Proceedings of the 20th International Conference on Text, Speech and Dialogue (TSD), 2017, pp. 500-508. (Rank B1)

[4] Duc-Hong Pham, Anh-Cuong Le, and Thi-Kim-Chung Le, "*Learning Word Embeddings for Aspect-based Sentiment Analysis*", Proceedings of the 15th International Conference of the Pacific Association for Computational Linguistics (PACLING), 2017, pp. 28-40. (Rank B)

[5] Duc-Hong Pham, Anh-Cuong Le, and Thi-Thanh-Tan Nguyen, "*Determing Aspect Ratings and Aspect Weights from Textual Reviews by Using Neural Network with Paragraph Vector Model*", Proceedings of the 5th International Conference on Computational Social Networks (CSONet), 2016, pp. 309-320. (SCOPUS)

[6] Duc-Hong Pham, and Anh-Cuong Le, "*A Neural Network based Model for Determining Overall Aspect Weights in Opinion Mining and Sentiment Analysis*", Indian Journal of Science and Technology, 2016, pp. 1-6. (SCOPUS)