ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM ĐỨC HỒNG

PHÁT TRIỂN CÁC MÔ HÌNH DỰA TRÊN MẠNG NƠ-RON CHO PHÂN TÍCH QUAN ĐIỂM THEO KHÍA CẠNH

onluyen.net

LUẬN ÁN TIẾN SĨ NGÀNH KHOA HỌC MÁY TÍNH

Hà Nội - 2019

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM ĐỨC HỒNG

PHÁT TRIỂN CÁC MÔ HÌNH DỰA TRÊN MẠNG NƠ-RON CHO PHÂN TÍCH QUAN ĐIỂM THEO KHÍA CẠNH

Chuyên ngành: Khoa học máy tính
Mã số: 9480101.01

onluyen.net

LUẬN ÁN TIẾN SĨ NGÀNH KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS. Lê Anh Cường

Hà Nội - 2019

LỜI CẢM ƠN

Luận án tiến sĩ ngành Khoa học Máy tính này được Chính phủ Việt Nam hỗ trợ một phần kinh phí thông qua Đề án 911 và được thực hiện tại Bộ môn Khoa học máy tính, Khoa Công nghệ thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội dưới sự hỗ trợ về mặt thủ tục của Phòng Tài chính kế toán, Trường Đại học Công nghệ. Bên cạnh đó còn có sự hỗ trợ kinh phí đối với các công bố có chỉ số SCIE, SCI của Trường Đại học Điện lực và của đề tài NAFOSTED, mã số 102.01-2014.22 thuộc Quỹ Phát triển khoa học và công nghệ Quốc gia. Tôi xin chân thành cảm ơn các đơn vị, tổ chức này đã giúp đỡ tôi trong thời gian nghiên cứu.

Luận án còn có sự hợp tác và hỗ trợ của các cá nhân, những người đã đóng góp rất nhiều trong quá trình hoàn thành các vấn đề nghiên cứu của luận án. Trước hết tôi xin chân thành cảm ơn PGS.TS. Lê Anh Cường đã trực tiếp hướng dẫn, giúp đỡ luôn sẵn lòng và tạo mọi điều kiện thuận lợi nhất cho tôi trong quá trình học tập và nghiên cứu. Tôi xin gửi lời cảm ơn chân thành tới PGS.TS. Hoàng Xuân Huấn, PGS.TS. Phan Xuân Hiếu, TS. Nguyễn Văn Vinh, TS. Lê Nguyên Khôi, TS. Nguyễn Bá Đạt, TS. Nguyễn Thị Ngọc Điệp (Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội), PGS.TS. Lê Thanh Hương (Trường Đại học Bách khoa Hà Nội), TS. Nguyễn Thị Minh Huyền (Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội), PGS. TS. Trần Đăng Hưng (Trường Đại học Sư Phạm Hà Nội), TS. Đặng Thị Thu Hiền (Trường Đại học Thủy lợi) vì sự góp ý rất chân thành và thẳng thắn giúp cho luận án của tôi được hoàn thiện tốt hơn.

Tôi biết ơn chân thành đối với PGS.TS. Nguyễn Lê Minh (Viện Khoa học công nghệ Tiên tiến Nhật Bản), TS. Trần Quốc Long (Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội). Hai thầy đã trực tiếp giảng dạy, chia sẻ cho tôi nhiều hiểu biết liên quan đến nội dung nghiên cứu.

Tôi xin gửi lời cảm ơn đến tất cả anh, chị, em và bạn bè đồng nghiệp, nghiên cứu sinh ở Bộ môn Khoa học máy tính, Khoa Công nghệ thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã giúp đỡ tôi hoàn thành các kế hoạch và thủ tục hành chính trong thời gian làm nghiên cứu sinh. Tôi cũng muốn cảm ơn đến anh/chị/em đồng nghiệp, giảng viên khoa Công nghệ Thông tin, Trường Đại học Điện lực đã luôn cổ vũ động viên và sát cánh bên tôi trong suốt quá trình nghiên cứu.

Cuối cùng, tôi muốn nói lời cảm ơn đặc biệt tới vợ tôi Lê Thị Kim Chung, và con trai tôi Phạm Công Phúc đã dành cho tôi tình yêu và sự cảm thông, cho phép tôi dành nhiều thời gian, tập trung cho công việc nghiên cứu. Tôi hết lòng biết ơn bố mẹ tôi về tình yêu và sự cống hiến to lớn để tôi trưởng thành như ngày hôm nay, cảm ơn các anh, chị, em của tôi về tình yêu gia đình và sự quan tâm giúp đỡ của họ cho công việc này.

onluyen.net

LỜI CAM ĐOAN

Tôi xin cam đoan luận án này là kết quả nghiên cứu của tôi, được thực hiện dưới sự hướng dẫn của PGS.TS. Lê Anh Cường. Các nội dung trích dẫn từ các nghiên cứu của các tác giả khác mà tôi trình bày trong luận án này đã được ghi rõ nguồn trong phần tài liệu tham khảo.

Phạm Đức Hồng

onluyen.net

Mục lục

Lời cảm ơn **i**

Lời cam đoan **ii**

Mục lục **iii**

Danh mục các chữ viết tắt **vii**

Danh mục các bảng **ix**

Danh mục các hình vẽ **xi**

Lời mở đầu **1**

1 Tổng quan vấn đề nghiên cứu **5**

1.1 Giới thiệu bài toán 5

1.2 Các bài toán trong phân tích quan điểm 6

1.2.1 Tổng quan một hệ thống phân tích quan điểm 6

1.2.2 Phân tích quan điểm cho toàn bộ văn bản 8

1.2.3 Phân tích quan điểm theo khía cạnh 8

1.2.4 Các bài toán trong phân tích quan điểm theo khía cạnh 8

1.3 Các nghiên cứu liên quan 10

1.3.1 Trích xuất từ thể hiện khía cạnh 10

1.3.2 Xác định khía cạnh 11

1.3.3 Phân đoạn khía cạnh 11

1.3.4 Phân loại quan điểm theo khía cạnh 12

onluyen.net

1.3.5 Xếp hạng khía cạnh 12
1.3.6 Xác định hạng và trọng số khía cạnh ẩn 13
1.4 Các tiếp cận giải quyết bài toán 14
1.5 Nghiên cứu trên thế giới và Việt nam 15
1.6 Thảo luận 16

2 Kiến thức cơ sở **18**
2.1 Các ký hiệu và khái niệm liên quan 18
2.2 Các mô hình học máy cơ sở cho phân tích quan điểm theo khía cạnh . . 21
2.2.1 Mô hình hồi quy đánh giá ẩn 21
2.2.2 Thuật toán xác suất xếp hạng khía cạnh 22
2.3 Các mô hình học biểu diễn mức từ, câu, đoạn/văn bản 25
2.3.1 Mô hình Word2Vec 25
2.3.2 Mô hình GloVe 26
2.3.3 Mô hình véc-tơ Paragraph 27
2.3.4 Mô hình mạng nơ-ron tích chập CNN 30
2.3.5 Mô hình véc-tơ kết hợp 33
2.4 Kết luận và thảo luận 34

3 Đề xuất mô hình dựa trên mạng nơ-ron xác định hạng và trọng số khía cạnh của thực thể **35**
3.1 Giới thiệu 35
3.2 Mô hình hóa bài toán 36
3.2.1 Bài toán xác định hạng và trọng số khía cạnh ẩn của thực thể . . 36
3.2.2 Bài toán xác định trọng số khía cạnh chung của các thực thể . . 37
3.3 Phương pháp đề xuất 38
3.3.1 Xác định hạng và trọng số khía cạnh ẩn của thực thể sử dụng mô hình mạng nơ-ron một lớp ẩn 38
3.3.2 Xác định hạng và trọng số khía cạnh ẩn của thực thể sử dụng mô hình học biểu diễn đa tầng 45

onluyen.net

3.3.3 Xác định trọng số khía cạnh chung của thực thể sử dụng mô hình mạng nơ-ron . . . 52
3.4 Thực nghiệm . . . 54
3.4.1 Các độ đo . . . 57
3.4.2 Cài đặt mô hình . . . 58
3.4.3 Kết quả thực nghiệm . . . 59
3.4.4 Đánh giá . . . 60
3.4.5 Hiệu quả của các tham số trong mô hình LRNN-ASR . . . 63
3.5 Kết luận . . . 66

4 Học véc-tơ biểu diễn từ cho phân tích quan điểm theo khía cạnh 68
4.1 Giới thiệu . . . 68
4.2 Mô hình hóa bài toán . . . 69
4.2.1 Bài toán tinh chỉnh véc-tơ biểu diễn từ . . . 70
4.2.2 Bài toán học véc-tơ biểu diễn từ . . . 71
4.3 Phương pháp đề xuất . . . 72
4.3.1 Mô hình tinh chỉnh véc-tơ biểu diễn từ . . . 72
4.3.2 Mô hình học véc-tơ biểu diễn từ SSCWE . . . 77
4.4 Thực nghiệm . . . 82
4.4.1 Dữ liệu thực nghiệm và các độ đo . . . 82
4.4.2 Các độ đo . . . 82
4.5 Cài đặt và đánh giá mô hình tinh chỉnh véc-tơ từ WEFT . . . 83
4.5.1 Cài đặt mô hình . . . 83
4.5.2 Đánh giá mô hình . . . 83
4.6 Cài đặt và đánh giá mô hình SSCWE . . . 85
4.6.1 Cài đặt mô hình . . . 85
4.6.2 Đánh giá mô hình . . . 85
4.6.3 So sánh hai mô hình WEFT và SSCWE . . . 88
4.7 Kết luận . . . 90

onluyen.net

5 Mô hình đa kênh dựa trên CNN nhằm khai thác đa véc-tơ biểu diễn từ và ký tự cho phân tích quan điểm theo khía cạnh 91

5.1 Giới thiệu . . . 92

5.2 Mô tả bài toán . . . 93

5.3 Phương pháp đề xuất . . . 94

5.3.1 Thành phần tích chập . . . 94

5.3.2 Mô hình mạng nơ-ron tích chập đa kênh cho phân tích quan điểm theo khía cạnh . . . 95

5.4 Thực nghiệm . . . 100

5.4.1 Dữ liệu thực nghiệm và cài đặt mô hình MCNN . . . 100

5.4.2 Môi trường và thời gian thực nghiệm . . . 101

5.4.3 Đánh giá . . . 101

5.4.4 Hiệu quả của các loại tham số . . . 105

5.5 Kết luận . . . 109

Kết luận 110

Danh mục công trình khoa học của tác giả liên quan đến luận án 112

Tài liệu tham khảo 113

onluyen.net

Danh mục các chữ viết tắt

LRNN	**L**atent **R**ating **N**eural **N**etwork (Mạng nơ-ron đánh giá ẩn)
LRR	Latent Rating Regression (Hồi quy đánh giá ẩn)
ASR	**A**spect **S**emantic **R**epresentation (Biểu diễn ngữ nghĩa khía cạnh)
NNAWs	**N**eural **N**etwork **A**spect **W**eights (Mạng nơ-ron trọng số khía cạnh)
CNN	**C**onvolutional **N**eural **N**etwork (Mạng nơ-ron tích chập)
MCNN	**M**ultichannel **C**onvolutional **N**eural **N**etwork (Mạng nơ-ron đa kênh tích chập)
NLP	**N**atural **L**anguage **P**rocessing (Xử lý ngôn ngữ tự nhiên)
POS	**P**art **O**f **S**peech (Nhãn từ loại)
SVM	**S**upport **V**ector **M**achine (Máy véc-tơ hỗ trợ)

onluyen.net

Danh sách bảng

3.1 Các từ hạt nhân được lựa chọn cho thuật toán phân đoạn khía cạnh . . . 55

3.2 Thống kế dữ liệu thực nghiệm 55

3.3 Kết quả dự đoán hạng của 5 khách sạn 59

3.4 Kết quả xác định trọng số khía cạnh của 5 khách sạn 59

3.5 So sánh mô hình LRNN với phương pháp LRR trong bốn trường hợp biểu diễn khía cạnh 61

3.6 Top 10 từ có trọng số tích cực và tiêu cực của từng khía cạnh 61

3.7 Các kết quả thực nghiệm và so sánh các mô hình trong việc xác định hạng khía cạnh 62

3.8 Kết quả so sánh chất lượng trọng số khía cạnh chung 63

3.9 Các kết quả thực nghiệm trên các trường hợp khởi tạo trọng số khía cạnh 64

3.10 Kết quả thực nghiệm mô hình đề xuất sử dụng trọng số khía cạnh chung so với sử dụng riêng 65

4.1 Thống kê tập dữ liệu thứ 2 82

4.2 Kết quả xác định khía cạnh 84

4.3 Kết quả phân loại quan điểm theo khía cạnh 84

4.4 Bốn từ gần ngữ nghĩa với từ đã cho đối với từng mô hình 85

4.5 Các kết quả xác định khía cạnh 87

4.6 Các kết quả phân loại quan điểm 87

4.7 Năm từ gần ngữ nghĩa với từ đã cho đối với từng mô hình 88

4.8 So sánh kết quả phân loại quan điểm giữa mô hình WEFT và SSCWE . 89

4.9 So sánh thời gian thực hiện giữa mô hình WEFT và SSCWE 90

5.1 Thống kê số lượng câu được sử dụng trong thực nghiệm 100

onluyen.net

5.2 Bảng từ điển các ký tự tiếng Anh 100
5.3 Kết quả xác định khía cạnh của mô hình MCNN và các mô hình cơ sở . 104
5.4 Kết quả dự đoán phân loại quan điểm theo khía canh của mô hình MCNN và các mô hình cơ sở 105
5.5 Các từ gần nhau được xác định bằng véc-tơ biểu diễn từ trong kênh CNN3 106
5.6 So sánh kết quả, thời gian sử dụng kỹ thuật *dropout* trong mô hình MCNN đối với cộng việc xác định khía cạnh 106
5.7 So sánh kết quả, thời gian sử dụng kỹ thuật *dropout* trong mô hình MCNN đối với công việc phân loại quan điểm khía cạnh 106
5.8 Kết quả xác định khía cạnh của mô hình MCNN với các mức lựa chọn số chiều véc-tơ ký tự nhúng khác nhau 107
5.9 Kết quả phân loại quan điểm của mô hình MCNN với các mức lựa chọn số chiều véc-tơ ký tự nhúng khác nhau 107
5.10 Kết quả xác định khía cạnh của mô hình MCNN khi sử dụng số lượng bộ lọc khác nhau 108
5.11 Kết quả phân loại quan điểm của mô hình MCNN khi sử dụng số lượng bộ lọc khác nhau 108

onluyen.net

Danh sách hình vẽ

1.1 Kiến trúc tổng quan của hệ thống phân tích quan điểm 7

2.1 Ví dụ một ý kiến khách hàng thể hiện quan điểm về dịch vụ khách sạn Vinpearl Phu Quoc Resort 19

2.2 Ví dụ về Hạng chung của sản phẩm iPhone X 64GB 20

2.3 Mô hình hồi quy đánh giá khía cạnh ẩn [1] 22

2.4 Mô hình Word2Vec với hai kiến trúc CBOW và Skip-gram 25

2.5 Mô hình véc-tơ Paragraph học biểu diễn câu, đoạn hoặc cả văn bản [2] . 28

2.6 Minh họa mô hình túi từ phân phối của các véc-tơ Paragraph [2] 29

2.7 Minh họa tích chập trong ma trận câu 31

2.8 Minh họa mô hình phân lớp câu sử dụng mạng CNN [3] 32

2.9 Minh họa việc sử dung mô hình véc-tơ kết hợp cho biểu diễn mức câu . 34

3.1 Ví dụ: đầu vào, đầu ra của toán xác định hạng và trọng số khía cạnh ẩn . 37

3.2 Ví dụ: đầu vào, đầu ra của toán xác định trọng số khía cạnh chung . . . 38

3.3 Các công việc cần giải quyết của bài toán xác định hạng và trọng số khía cạnh ẩn của thực thể 39

3.4 Minh họa mô hình mạng nơ-ron LRNN xếp hạng ẩn 41

3.5 Các công việc cần giải quyết của bài toán xác định hạng, trọng số khía cạnh ẩn sử dụng mô hình học biểu diễn đa tầng 45

3.6 Minh họa mô hình học biểu diễn đa tầng cho phân tích quan điểm theo khía cạnh 46

3.7 Minh họa mô hình xác định hạng khía cạnh chung 52

3.8 Quy trình thực nghiệm, đánh giá mô hình LRNN 56

3.9 Quy trình thực nghiệm, đánh giá mô hình LRNN-ASR 56

onluyen.net