ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thị Xuân Hương

NGHIÊN CỨU HỌC MÁY THỐNG KÊ CHO PHÂN TÍCH QUAN ĐIỂM

TÓM TẮT LUẬN ÁN TIẾN SỸ CÔNG NGHỆ THÔNG TIN

Hà Nội - 2018

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thị Xuân Hương

NGHIÊN CỨU HỌC MÁY THỐNG KÊ CHO PHÂN TÍCH QUAN ĐIỂM

Chuyên ngành: Khoa học máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SỸ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TS Lê Anh Cường
2. PGS.TS Nguyễn Lê Minh

Hà Nội - 2018

Mục lục

1 GIỚI THIỆU **1**

1.1 Đặt vấn đề . . . 1

1.2 Các kết quả chính của luận án . . . 2

1.3 Bố cục của luận án . . . 2

2 TỔNG QUAN **3**

2.1 Phân tích quan điểm . . . 3

2.1.1 Phân tích tình cảm (Sentiment Analysis) hay khai thác quan điểm (Opinion Mining) . . . 3

2.2 Phát biểu bài toán . . . 3

2.2.1 Bài toán phân tích quan điểm . . . 3

2.2.2 Phân loại tính chủ quan (Subjectivity Classification . . . 3

2.2.3 Phân loại quan điểm (Setiment classification) . . . 4

2.2.4 Phân loại quan điểm theo khía cạnh (Aspect based sentiment classification) . . . 4

2.2.5 Đặc trưng cho toán phân tích quan điểm . . . 4

2.2.6 Các miền dữ liệu và dữ liệu Microblog . . . 4

2.3 Các thảo luận và mục tiêu nghiên cứu của đề tài . . . 5

2.3.1 Bài toán Phân loại tính chủ quan . . . 5

2.3.2 Bài toán phân loại quan điểm theo khía cạnh . . . 5

2.3.3 Phân tích quan điểm tiếng Việt và dữ liệu dạng Microblog . . . 5

3 PHÂN LOẠI TÍNH CHỦ QUAN **6**

3.1 Giới thiệu . . . 6

3.2 Phương pháp đề xuất sử dụng các đặc trưng ngôn ngữ cho phân lớp khách quan . . . 6

3.2.1 Trích các đặc trưng . . . 7

3.2.2 Thực nghiệm và đánh giá. . . . 8

3.3 Phương pháp đề xuất học tự động các mẫu cho bài toán xác định câu chủ quan tiếng Việt . . . 9

3.3.1 Dữ liệu huấn luyện . . . 9

3.3.2 Định nghĩa các khuôn dạng . . . 9

3.3.3 Trích xuất và đánh giá các mẫu . . . 10

3.3.4 Kết quả thực nghiệm và thảo luận . . . 11

3.3.5 Đánh giá các mẫu học được . . . 12

3.3.6 Kết luận . . . 12

4 PHÂN TÍCH QUAN ĐIỂM THEO KHÍA CẠNH **13**

4.1 Giới thiệu . . . 13

4.2 Mô tả bài toán . . . 13

4.3 Mô hình đề xuất . . . 14

4.3.1 Mô hình CNN hai pha cho phân tích quan điểm theo khía cạnh (A two-phase CNN model for Aspect based Sentiment Analysis) . 14
4.3.2 Mô hình CNN với các đặc trưng ngoài (The CNN Model with External Features) . . . 16
4.4 Thực nghiệm . . . 16
4.4.1 Dữ liệu . . . 16
4.4.2 Tiền xử lý dữ liệu . . . 16
4.4.3 Các mô hình và các kết quả . . . 17
4.4.4 Các kết quả . . . 17
4.5 Kết luận . . . 17

5 PHÂN TÍCH QUAN ĐIỂM TIẾNG VIỆT 18
5.1 Giới thiệu . . . 18
5.2 Phương pháp kiểm tra chính tả cho dữ liệu MicroBlogs sử dụng n-gram lớn 18
5.2.1 Một số lỗi chính tả thường gặp . . . 18
5.2.2 Mô hình kiểm tra chính tả đề xuất . . . 18
5.2.3 Tiền xử lý dữ liệu . . . 19
5.2.4 Thuật toán kiểm tra chính tả mở rộng ngữ cảnh ở cả hai bên . . . 19
5.2.5 Mô hình N-gram lớn và nén N-gram . . . 20
5.2.6 Thực nghiệm của chúng tôi . . . 20
5.3 Phương pháp tách từ cho dữ liệu Micro-blogs tiếng Việt . . . 22
5.3.1 Tiếp cận của chúng tôi cho bài toán tách từ dữ liệu Micro-blogs . 22
5.3.2 Hệ thống tách từ có sử dụng kiểm tra chính tả (Adaption to word segmentation by spell-checking system) . . . 23
5.3.3 Các thực nghiệm . . . 23
5.4 Kết luận . . . 24

Danh mục các công trình khoa học 26

Chương 1

GIỚI THIỆU

1.1 Đặt vấn đề

Phân tích quan điểm người dùng là lĩnh vực đã và đang thu hút được sự quan tâm của cộng đồng các nhà nghiên cứu cũng như các nhà phát triển các ứng dụng trong công nghiệp. Trong những năm gần đây, do sự phát bùng nổ lượng dữ liệu đánh giá của người dùng trên các trang mạng xã hội, các diễn đàn, các trang đánh giá sản phẩm, việc phát triển các phương pháp mới và công cụ nhằm phân tích và rút trích ra quan điểm giúp có thể hiểu được xu thế mọi người đang bình luận hay đánh giá về một thực thể mục tiêu. Kết quả của những nghiên cứu này là hữu ích cho các cá nhân và doanh nghiệp khi họ cần tham khảo các thông tin đánh giá về thực thể mục tiêu mà họ quan tâm.

Đã có nhiều tiếp cận nghiên cứu khác nhau được đề xuất cho các nhiệm vụ trong phân tích quan điểm. Các tiếp cận này thường dựa trên việc trích chọn các đặc trưng thể hiện quan điểm, nhận xét, đánh giá, tình cảm hay cảm xúc người dùng về thực thể được đánh giá cho các bài toán mục tiêu. Trong luận án này, chúng tôi tập trung nghiên cứu về việc trích chọn các đặc trưng ngữ pháp hữu ích cho một số nhiệm vụ trong bài toán phân tích quan điểm với cả hai loại dữ liệu tiếng Anh và tiếng Việt.

Bài toán thứ nhất được chúng tôi đề cập là phân loại chủ quan. Đây là bài toán quan trọng đầu tiên trong phân tích quan điểm nhằm phân loại câu hay tài liệu chủ quan chứa quan điểm và câu hay tài liệu khách quan không chứa quan điểm. Đối với bài toán này, chúng tôi đề xuất hai phương pháp, một là trích các đặc trưng ngôn ngữ dựa trên các mẫu cú pháp cho dữ liệu tiếng Anh, hai là đề xuất phương pháp học tự động dựa theo thống kê mẫu ngữ pháp để phân loại câu chủ quan tiếng Việt.

Bài toán thứ hai là phân loại quan điểm theo khía cạnh với các tài liệu chứa quan điểm. Chúng tôi đề xuất một mô hình tích hợp các đặc trưng giàu thông tin bên ngoài vào mô hình mạng nơ ron tích chập để tăng hệu suất thực hiện cho mô hình.

Trong quá trình phát triển phương pháp phân tích quan điểm trên đối tượng dữ liệu tiếng Việt, chúng tôi nhận thấy các dữ liệu bình luận tiếng Việt trên các diễn đàn thường là những câu ngắn và được viết không theo chuẩn ngữ pháp, ngoài ra còn chứa rất nhiều lỗi và từ viết tắt hay ngôn ngữ ký hiệu riêng của giới trẻ. Loại dữ liệu này được gọi là dữ liệu dạng Microblog. Một số phương pháp tiền xử lý dữ liệu tiếng Việt hầu hết được phát triển cho dữ liệu chính thống, nên khi áp dụng cho dữ liệu dạng Microblog là không hiệu quả. Do đó, để xử lý dữ liệu phục vụ cho bào toán nghiên cứu, chứng tôi đề xuất phương pháp kiểm tra chính tả cho dữ liệu Microbog tieensg Việt sử dụng n-ram được huấn luyện từ kho ngữ liệu lớn. Chúng tôi cũng đề xuất một mô hình sử dụng hệ thống kiểm tra từ viết tắt và kiểm tra chính tả trong tách từ tiếng Việt để phù hợp với dữ liệu dạng Microblog.

Mục tiêu của luận án ***"Nghiên cứu học máy thống kê cho phân tích quan điểm"*** tập trung vào nhiệm vụ *"Đề xuất các phương pháp cho phân loại khách quan và phân loại quan điểm theo khía cạnh"*.

Phương pháp tiếp cận của luận án là xây dựng các mẫu để trích chọn các thông tin ngữ pháp hữu ích cho các mô hình học phân loại. Đối tượng nghiên cứu của luận án là các dữ liệu bình luận tiếng Anh và tiếng Việt. Trong quá trình xây dựng ngữ liệu bình luận tiếng Việt, chúng tôi cũng thực hiện các nghiên cứu nhằm cải thiện chất lượng của dữ liệu bình luận dạng Microblog với hai đề xuất xây dựng mô hình kiểm tra chính tả và tách từ thích ứng với dữ liệu Microblog

1.2 Các kết quả chính của luận án

Các kết quả nghiên cứu của luận án góp phần bổ sung và hoàn thiện cho các phương pháp phân tích quan điểm. Cụ thể, luận án đã có một số đóng góp chính như sau:

— Đề xuất một số phương pháp xác định văn bản chứa quan điểm. Chúng tôi đề xuất một phương pháp phân loại câu khách quan và câu chủ quan cho dữ liệu tiếng Anh được công bố tại kỷ yếu hội nghị IALP năm 2012. Chúng tôi cũng đã đề xuất một phương pháp thống kê tự động trích mẫu cho phân loại chủ quan tiếng Việt. Đóng góp này được công bố trong kỷ yếu hội thảo quốc tế Knowledge and Systems Engineering (KSE) năm 2014.

— Đề xuất phương pháp thêm các đặc trưng ngoài cho mạng nơ ron phân tích quan điểm theo khía cạnh. Đóng góp này được công bố trong kỷ yếu hội thảo quốc tế NAFOSTED Conference on Information and Computer Science (NICS) năm 2018.

— Đề xuất một số phương pháp để tiền xử lý cho dữ liệu Microblog tiếng Việt. Chúng tôi đề xuất hai phương pháp: Phương pháp thứ nhất dùng để kiểm tra chính tả cho dữ liệu tiếng Việt. Đóng góp này được công bố ở kỷ yếu hội thảo quốc tế Knowledge and Systems Engineering (KSE) năm 2014. Phương pháp thứ hai dùng để tách từ cho dữ liệu Microblog tiếng Việt được công bố trong kỷ yếu hội thảo quốc tế Asian Conference on Information Systems (ACIS) năm 2014.

1.3 Bố cục của luận án

Ngoài phần mở đầu và kết luận, luận án được tổ chức thành 5 chương, với bố cục như sau: Chương 1: Giới thiệu. Chương 2: Tổng quan. Trong chương này, chúng tôi trình bày tổng quan về các vấn đề nghiên cứu trong luận án. Chúng tôi phân tích, đánh giá 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 nội dung nghiên cứu của luận án. Chương 3: Phân loại khách quan. Chúng tôi trình bày nội dung, kết quả nghiên cứu cho nhiệm vụ xác định văn bản chứa quan điểm. Chương 4: Phân tích quan điểm theo khía cạnh. Trong đó, chúng tôi trình bày nội dung, kết quả nghiên cứu về phân tích quan điểm theo khía cạnh. Chương 5: Phân tích về phân tích quan điểm tiếng Việt dạng bài nhật ký trực tuyến ngắn (Microblog). Chúng tôi trình bày nội dung, kết quả nghiên cứu cho một số bước chuẩn hóa dữ liệu Microblog tiếng Viêt. Kết luận. Chúng tôi trình bày về các nhận xét và kết luận về kết quả đã thực hiện được trong luận án và hướng nghiên cứu tiếp theo.

Chương 2

TỔNG QUAN

2.1 Phân tích quan điểm

2.1.1 Phân tích tình cảm (Sentiment Analysis) hay khai thác quan điểm (Opinion Mining)

Phân tích tình cảm (Sentiment Analysis - SA) hay khai thác quan điểm (Opinion Mining - OM) là lĩnh vực nghiên cứu phân tích các quan điểm, tình cảm, đánh giá, thái độ và cảm xúc của con người cho các thực thể và các thuộc tính của chúng được thể hiện trong văn bản. Các thực thể có thể là các sản phẩm, dịch vụ, tổ chức, cá nhân, sự kiện, các vấn đề hoặc các chủ đề. Một số các tên gọi khác liên quan như, phân tích tình cảm(sentiment analysis), khai thác quan điểm opinion mining, phân tích quan điểm (opinion analysis), trích quan điểm (opinion extraction), khai thác tình cảm (sentiment mining), phân tích chủ quan (subjectivity analysis), phân tích khía cạnh (affect analysis), phân tích cảm xúc (emotion analysis), và phân tích đánh giá (review mining) đều nằm trong phạm vi của phân tích tình cảm. Trong luận án này, chúng tôi sử dụng hai thuật ngữ cho bài toán này là "Phân tích tình cảm"(Sentiment Analysis (SA)) hay "Phân tích quan điểm" (Opinion Analysis (OA)).

2.2 Phát biểu bài toán

2.2.1 Bài toán phân tích quan điểm

Đã có rất nhiều nhiệm vụ trong phân tích quan điểm đã và đang được nghiên cứu và ứng dụng trong thực tế. Có nhiều quan điểm khác nhau về việc phân chia các nhiệm vụ trong phân tích quan điểm. Tuy nhiên, chúng tôi đề cập đến 5 loại bài toán chính như sau:
1. Phân loại chủ quan - Subjectivity classification
2. Phân loại quan điểm - Sentiment classification
3. Phân loại quan điểm theo khía cạnh - Aspect-based Sentiment Classification
4. Tóm tắt quan điểm - Opinion Summarization
5. Phát hiện quan điểm giả mạo hay lừa đảo - Detecting Fake or Deceptive Opinions

2.2.2 Phân loại tính chủ quan (Subjectivity Classification

Phân loại tính chủ quan là xác định một câu/tài liệu là chủ quan hay khách quan. Phân loại tính chủ quan là bài toán đầu tiên quan trọng trong phân tích quan điểm. Kết quả của bài toán này được sử dụng là đầu vào cho nhiều bước phân tích tiếp theo trong lĩnh vực này.

2.2.3 Phân loại quan điểm (Setiment classification)

Phân loại quan điểm là phân chia một câu/tài liệu chứa quan điểm vào một trong các mức độ phân cực là tích cực, tiêu cực hay trung lập (hay các thứ hạng, 1*, 2*, 3*, 4*, 5*).
Phân loại quan điểm mức tài liệu có thể có một số hạn chế vì một tài liệu bình luận có thể có nhiều đánh giá về nhiều thực thể và định hướng quan điểm trên các thực thể khác nhau có thể khác nhau. Người đưa ra đánh giá có thể có ý kiến tích cực về một số thực thể và tiêu cực về những người khác. Trong trường hợp này, nhiệm vụ phân loại quan điểm mức tài liệu trở nên ít có ý nghĩa hơn vì nó sẽ không phù hợp để gán một quan điểm cho toàn bộ tài liệu. Do đó, phân loại quan điểm mức tài liệu là phân loại thô cho các ứng dụng thực tế. Để làm mịn hơn, ta phân loại quan điểm mức câu từ đó giúp trích xuất được các khía cạnh thể hiện quan điểm và làm mịn dần cho đến tiếp cận mức khía cạnh.
Phân loại quan điểm mức câu về cơ bản giống như phân loại cấp tài liệu vì câu có thể được coi là văn bản ngắn.

2.2.4 Phân loại quan điểm theo khía cạnh (Aspect based sentiment classification)

Phân loại quan điểm khía cạnh gồm hai nhiệm vụ chính: trích khía cạnh (Aspect extraction) và phân loại quan điểm với khía cạnh được trích (Aspect sentiment classification).
Trích khía cạnh: là nhiệm vụ trích các khía cạnh và các thực thể được đánh giá.
Phân loại quan điểm theo khía cạnh: là nhiệm vụ xác định các quan điểm về một khía cạnh được trích là tích cực, tiêu cực hay trung lập.

2.2.5 Đặc trưng cho toán phân tích quan điểm

Một số đặc trưng sử dụng trong các bài toán phân tích quan điểm:
1.Tần suất xuất hiện - Term presence and Frequency
2. Mô hình ngôn ngữ - n-gram
3. Thông tin nhãn từ loại - Parts of Speech
3. Thông tin phân tích cú pháp - Syntax Passer
4. Biểu diễn véc tơ từ - Word Embedding
5. Véc tơ biểu diễn ký tự - Character Embedding

2.2.6 Các miền dữ liệu và dữ liệu Microblog

— **Dữ liệu quan điểm và miền dữ liệu**
Có hai loại văn bản trong truyền thông xã hội: các bài đăng độc lập là các bài đánh giá, nhận xét hay bình luận về một thực thể mục tiêu xác định nào đó, và các đối thoại trực tuyến có tính tương tác và thường liên quan sự trao đổi tương tác của hai hoặc nhiều người tham gia.Trong nghiên cứu của chúng tôi, chúng tôi giới hạn chỉ xét đến các tài liệu/câu quan điểm độc lập.

— **Dữ liệu Microblog** Các bài viết đánh giá trang mạng xã hội, diễn đàn, blog... thường gồm những câu ngắn, không theo chuẩn quy tắc ngữ pháp thông thường, có từ được viết tắt hay sử dụng ký hiệu và từ lóng. do đó cần thiết phải chuẩn hóa dữ liệu này trước khi thực hiện phân tích quan điểm.

2.3 Các thảo luận và mục tiêu nghiên cứu của đề tài

Bằng công việc khảo cứu các nhiệm vụ và các nghiên cứu với những vấn đề được nêu ra trong phân tích quan điểm, chúng tôi đã xác định một số nhiệm vụ nghiên cứu cho đề tài như sau:

2.3.1 Bài toán Phân loại tính chủ quan

Qua những khảo cứu và phân tích từ các nghiên cứu trước đây, chúng tôi nhận thấy việc nâng cao hiệu suất của nhiệm vụ Phân loại tính chủ quan là cần thiết cho các bước nghiên cứu tiếp theo của phân tích quan điểm. Do đó, chúng tôi đã nghiên cứu việc trích chọn các thông tin hữu ích giúp phân lớp câu là khách quan hay chủ quan cho cả dữ liệu tiếng Anh và tiếng Việt. Các nghiên cứu và ứng dụng của phân tích quan điểm tiếng Việt đã và đang thu hút được sự quan tâm của cộng động trong một vài năm gần đây. Do đó, việc phát triển các nghiên cứu cho bài toán này thực sự có ý nghĩa cả trong lĩnh vực nghiên cứu lẫn trong công nghiệp.

2.3.2 Bài toán phân loại quan điểm theo khía cạnh

Phân loại quan điểm theo đặc trưng là mức phân loại chi tiết hơn cho các phân loại quan điểm mức câu hoặc tài liệu. Khi trong một bình luận chứa nhiều đánh giá cho từng khía cạnh khác nhau của một mục tiêu đánh giá, việc xác định các tình cảm nào gắn với khía cạnh nào sẽ giúp cho việc tổng hợp và tóm tát quan điểm trở lên dễ thực hiện hơn và giúp cho người dùng có cái nhìn tổng quan hơn về các nhận xét cho đối tượng họ quan tâm. Các nghiên cứu áp dụng các thuật toán sử dụng mạng nơ ron và học sâu đã và đang có nhiều kết quả hứa hẹn. Do đó chúng tôi đã chọn nghiên cứu việc tích hợp các đặc trưng giàu thông tin để làm tăng hiệu suất thực hiện cho các mô hình này.

2.3.3 Phân tích quan điểm tiếng Việt và dữ liệu dạng Microblog

Dữ liệu là vấn đề then chốt đối với việc xây dựng các thuật toán xử lý trên đó. Đối với bài toán phân tích quan điểm cho tiếng Việt, các dữ liệu của người dùng bình luận, đánh giá mà chúng ta thu thập được từ từ các trang mạng xã hội, diễn đàn, blog... thường là không chuẩn và chứa nhiều lỗi chính tả và viết tắt, gọi là dữ liệu kiểu Microblog. Trong khi đó, các công cụ hiện có chủ yếu là phát triển cho các văn bản chính thống và chưa xử lý hiệu quả cho việc chữa các lỗi này. Kết quả của việc sửa lỗi từ cũng sẽ ảnh hưởng đến bài toán tách từ cho loại dữ liệu Microblog. Do đó, chúng tôi nhận thấy cần có bước thực hiện hai nhiệm vụ kiểm tra chính tả và tách từ để phù hợp cho dữ liệu Microblog tiếng Việt.

Chương 3

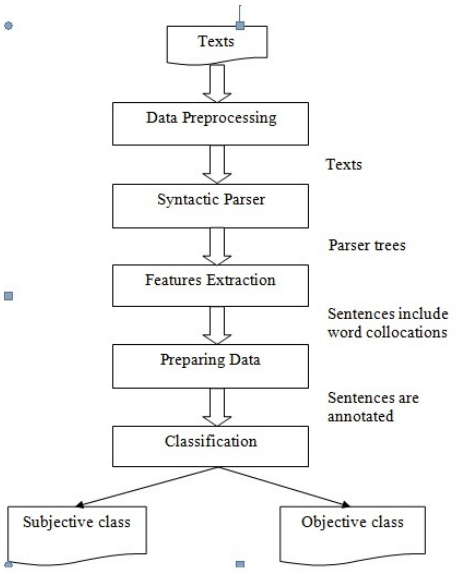
PHÂN LOẠI TÍNH CHỦ QUAN

3.1 Giới thiệu

Trong chương này, chúng tôi trình bày về nhiệm vụ phân loại chủ quan và một số phương pháp đề xuất của chúng tôi cho bài toán này. Chúng tôi đề xuất một phương pháp xác định câu chủ quan dựa trên các thông tin trích từ các mẫu ngữ pháp cho dữ liệu tiếng Anh. Chúng tôi giới thiệu một phương pháp thống kê để giúp hệ thống học các mẫu cú pháp và tự động đánh giá chúng từ dữ liệu huấn luyện có gắn nhãn tiếng Việt.

3.2 Phương pháp đề xuất sử dụng các đặc trưng ngôn ngữ cho phân lớp khách quan

Trong phần này chúng tôi giới thiệu mô hình phân loại khách quan của chúng tôi sử dụng các đặc trưng ngôn ngữ được trích dựa trên các mẫu được xác định trước được mô tả như sau:



— Bước 1: Tiền xử lý dữ liệu.
— Bước 2: Sử dụng công cụ phân tích cú pháp cho ngữ liệu ban đầu.
— Bước 3: Trích các đặc trưng dựa trên các mẫu cú pháp.
— Bước 4: Chuẩn bị dữ liệu cho phương pháp phân loại
— Bước 5: Sử dụng phương pháp phân loại Maximum Entropy để phân loại ngữ liệu văn bản câu thành hai lớp, khách quan và chủ quan.

3.2.1 Trích các đặc trưng

Thông tin ngữ pháp Để trích các đặc trưng ngôn ngữ từ các câu, chúng tôi sử dụng các thông tin cú pháp bằng cách sử dụng công cụ phân tích cú pháp Stanford Parser phân tích cho ngữ liệu vào.

Trích các đặc trưng khách quan

Chúng tôi tập trung phân tích bốn từ loại bao gồm, các tính từ, các trạng từ, các động từ và một số dạng mở rộng của động từ và các danh từ để tạo ra các mẫu dựa trên cú pháp. Chúng tôi trích các đặc trưng ngôn ngữ bằng cách sử dụng các mẫu dựa trên thông tin ngữ pháp để phân biệt các câu khách quan và câu chủ quan. Các thông tin được trích không chỉ phản ánh thể hiện chủ quan mà nó cũng giúp phân biệt câu khách quan. Sự khác nhau sẽ được phân loại nhờ việc áp dụng phương pháp học phân loại có giám sát Maximum entropy.

1. Trích các mẫu dựa trên cú pháp chứa các tính từ

Mẫu	Mô tả
[ADJP] [TO] [VB]	Tính từ trong mẫu thể h khả năng hoặc đánh giá của người dùng khi thực hiện một việc nào đó
[ADJP] [CC] [ADJP]	Hai cụm tính từ được liên kết bởi từ nối
[VP contains VBZ/VBG] [ADJ]	Tính từ dạng động từ mô tả thông tin hoặc đánh giá về một thực thể hay đối tượng
[ADJP contains only JJ]	Tính từ dạng cảm thán trong câu dùng để thể hiện tính chủ quan

Bảng 3.1: Các mẫu ngữ pháp chứa các tính từ

2. Trích các mẫu dựa trên cú pháp chứa các trạng từ

Mẫu	Mô tả
[VP contains[ADVP] /[VB/VBN/VBG/VBZ/VBD]	Trạng từ bổ nghĩa cho động từ.
[VP][with[PRT] or not] [ADVP][with[PP] or not]	Cụm trạng từ bổ nghĩa cho động từ chứa cụm giới từ.
[VP][ADVP][ADJP]	Trạng từ bổ nghĩa cho tính từ.
[ADJP][ADVP][JJ]	Cụm tính từ chưa các cụm trạng từ bổ nghĩa cho tính từ
[ADVP][VP]	Cụm trạng từ bổ nghĩa cho nội động từ.
[PP contain RB] [NP]	Cụm giới từ chứa trạng từ đứng trước cụm danh từ.
[ADVP]	Trạng từ bổ nghĩa cho động từ đặt ở cuối câu.

Bảng 3.2: Các mẫu ngữ pháp chứa các trạng từ

3.Trích các mẫu ngữ pháp chứa các động từ.

4. Trích các mẫu dựa trên ngữ pháp chứa danh từ.

Mẫu	Mô tả
[VP][TO] [VP]	Cụm động từ diễn tả mục đích
[MD][VP][TO][VP]	Động từ khuyết thiếu thường dùng để diễn tả khả năng hoặc giả định có thể diễn đạt quan điểm.
[VP][VBN]/[VBG]/[NN]	Mô tả về thực thể hoặc sự kiện cùng với trạng từ để nhấn mạnh
[VP][PP/PRT]	Cụm động từ có thể diễn tả hành động hoặc trạng thái của thực thể hoặc đối tượng.

Bảng 3.3: Các mẫu ngữ pháp chứa các động từ

Mẫu	Mô tả
$[NPcontainsJJ/JJR/JJS/$ $RB/RBR/RBS/VBG/VBN]$	Cụm danh từ bao gồm các nhãn từ loại là so sánh của các tính từ và trạng từ, tính từ dạng danh động từ và tính từ dạng quá khứ phân từ
$[NPcontainsJJ/VBN/VBG][CC][NP]$ $or[NP]CC[NPcontainsJJ/VBN/VBG]$	Hai cụm danh từ được kết nối bởi từ nối dùng để mô tả về tính chất hay đánh giá về một đối tượng.
$[NPNN/NNSCCNN/NS]$	Hai cụm danh từ được kết nối bởi từ nối mô tả về một đối tượng.
$[NP][POS][NP]$	Hai cụm danh từ thể hiện sở hữu cách. Sự kết hợp này có thể thể hiện quan điểm hoặc không
$[NP][IN][NP]$	Hai cụm danh từ được liên kết với nhau bởi giới từ hoặc quan hệ kết nối phụ thuộc.
$[NPcontains[QP][NN]]$	Cụm danh từ bao gồm cụm chỉ số lượng có thể diễn đạt quan điểm
$[NPcontainsNN/NNS]$	Cụm danh từ chứa hai danh từ có thể phản ánh quan điểm của người dùng.

Bảng 3.4: Các mẫu ngữ pháp chứa các danh từ

3.2.2 Thực nghiệm và đánh giá.

3.2.2.1 Thực nghiệm.

Thực nghiệm của chúng tôi được thực hiện trên dữ liệu Movie Review được giới thiệu bởi Pang và Lee (Pang and Lee, 2002). Dữ liệu này chứa 5.000 câu chủ quan và 5.000 câu khách quan.
Subjectivity dataset 1.0 [1].
Trước tiên, chúng tôi tiền xử lý dữ liệu và sử dụng bộ phân tích cú pháp Stanford Parser để lấy các thông tin ngữ pháp. Chúng tôi trích xuất các đặc trưng ngôn ngữ theo các mẫu đã đề xuất để phân biệt câu chủ quan và khách quan. Sử dụng mô hình Maximum Entropy để phân loại câu thành hai lớp chủ quan và khách quan. Dữ liệu được chia thành 10 folds và chúng tôi sử dụng 8 folds cho huấn luyện phương pháp và 2 folds cho dữ liệu đánh giá.

1. link: http://www.cs.cornell.edu/people/pabo/movie-review-data/

3.2.2.2 Đánh giá thực nghiệm.

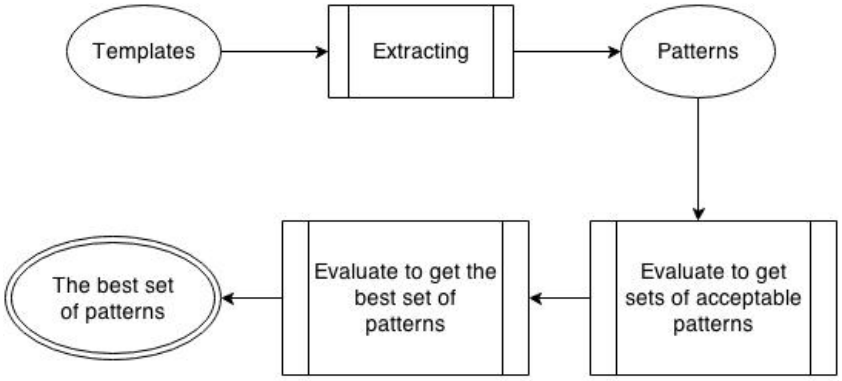
Chúng tôi có một so sánh giữa kết quả của chúng tôi với baseline (Pang và Lee, 2004) sử dụng các nhận xét dữ liệu phim, họ thực hiện 10 folds kiểm tra chéo và so sánh với nghiên cứu của Riloff và các cộng sự, 2006 sử dụng một hệ thống phân cấp trích các đặc trưng sau đó sử dụng SVM để phân loại dữ liệu chủ quan. Phương pháp của chúng tôi sử dụng ME để phân loại các đặc trưng ngôn ngữ đã trích dựa trên mẫu cú pháp. Chúng tôi giữ lại 66% dữ liệu từ các đánh giá từ đánh giá ban đầu và kết quả của chúng tôi đạt được độ chính xác 92,1% cho xác định câu chủ quan. Việc so sánh các phương pháp của chúng tôi với một số nghiên cứu trước được trình bày trong bảng 3.5

Phương pháp	Độ chính xác
Phương pháp của chúng tôi	**92.1%**
NB+Prox (Pang and Lee, 2004)	86.4%
SVM+Prox (Pang and Lee, 2004)	86.15%
Riloff06	82.7 %

Bảng 3.5: Bảng so sánh độ chính xác của các phương pháp .

3.3 Phương pháp đề xuất học tự động các mẫu cho bài toán xác định câu chủ quan tiếng Việt

Quá trình học tự động các mẫu cho phân loại chủ quan:



Hình 3.1: Quá trình học các mẫu từ loại

3.3.1 Dữ liệu huấn luyện

Trong dữ liệu huấn luyện, mỗi bình luận của chúng tôi được gán nhãn theo hai loại là <sub>(chủ quan) và <obj> (khách quan)và số các bình luận chủ quan bằng với số các bình luận khách quan. Dữ liệu được thực hiện các bước tiền xử lý (sử dụng công cụ kiểm tra chính tả và tách từ được đề xuất trong chương 5) trước khi gán nhãn từ loại.

3.3.2 Định nghĩa các khuôn dạng

Chúng tôi tập trung sử dụng các tính từ và các động từ là các đặc trưng cho phân loại, do đó các khuôn dạng được tạo ra bởi chúng và các nhãn từ loại xung quanh chúng.

Các nhãn từ loại xung quanh chúng có thể là danh từ (N), Danh từ riêng (Np), các tính từ khác (A), hoặc động từ (V), trạng từ (R), các từ nối (Cc), các trợ động từ (T). Chúng tôi giới thiệu hai kiểu khuôn dạng để học như sau:

- **Kiểu 1:** Khuôn dạng được xây dựng để trích các mẫu chỉ bao gồm các nhãn từ loại. Chúng tôi xem xét động từ và tình từ với các nhãn từ loại xung quanh chúng về bên trái, bên phải hoặc cả hai phía.
- **Kiểu 2:** Kiểu 2 cũng tương tự như kiểu 1 nhưng đặc biệt hơn kiểu 1 vì chúng tôi xây dựng khuôn dạng này để trích các mẫu bao gồm các từ (các tính từ và động từ) và các nhãn từ loại xung quanh chúng.

Chúng tôi có thể sử dụng các khuôn dạng của kiểu 1 và kiểu 2, trong phần này chúng tôi đưa ra các kết quả thực nghiệm khi áp dụng cả hai kiểu của khuôn dạng để trích xuất các mẫu. Bảng 3.6 và bảng 3.7 liệt kê các ví dụ của các khuôn dạng cho cả hai kiểu.

Bảng 3.6: Các khuôn dạng của kiểu 1

Khuôn dạng	Mô tả
tag-tag[+1]	Nếu nhãn hiện tại là tính từ hoặc động từ, xem xét khuôn dạng bao gồm nhãn hiện tại và nhãn tiếp theo.
tag-tag[-1]	Nếu nhãn hiện tại là tính từ hoặc động từ, xem xét khuôn dạng bao gồm nhãn hiện tại và một nhãn trước đó.
tag-tag[-1] & tag[+1]	Nếu nhãn hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm nhãn hiện tại và một nhãn trước đó và một nhãn tiếp theo (các nhãn theo cả hai hướng của nhãn hiện tại)
tag-tag[+2]	Nếu nhãn hiện tại là tính từ hoặc động từ, xem xét các khuôn dạng gồm nhãn hiện tại và hai nhãn tiếp theo.
tag-tag[-2]	Nếu nhãn hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm nhãn hiện tại và hai nhãn trước đó.

Bảng 3.7: Các khuôn dạng của kiểu 2

Khuôn dạng	Mô tả
word-tag[+1]	Nếu thẻ hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm từ hiện tại và một nhãn tiếp theo.
word-tag[-1]	Nếu thẻ hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm từ hiện tại và một nhãn trước đó.
word-tag[-1] & tag[+1]	Nếu thẻ hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm từ hiện tại và một nhãn trước đó và một nhãn tiếp theo (các nhãn theo cả hai phía của từ hiện tại).
word-tag[+2]	Nếu thẻ hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm từ hiện tại và hai nhãn tiếp theo.
word-tag[-2]	Nếu thẻ hiện tại là tính từ hoặc động từ, xem xét khuôn dạng gồm từ hiện tại và hai nhãn trước đó.

3.3.3 Trích xuất và đánh giá các mẫu

Các mẫu đã được định nghĩa trước ở phần trên được áp dụng trên dữ liệu huấn luyện đã được gán nhãn từ loại để trích ra tất cả các mẫu có thể. Sau đó chúng tôi đánh giá chúng để lấy ra tập các mẫu tốt nhất.
Đánh giá các mẫu chấp nhận được: chúng tôi chỉ xem xét các mẫu với hai ràng buộc:

- Một mẫu được cho là thể hiện chủ quan khi và chỉ khi:
  P (<sub> | $pattern_i$) > P (<obj> | $pattern_i$)

Công thức sau được giới thiệu để lấy ra các tập các mẫu chấp nhận được:

$$\frac{P(<sub>|pattern_i)}{P(<sub>|pattern_i)+P(<obj>|pattern_i)} > \text{threshold}$$

Để thỏa mãn ràng buộc trên, ngưỡng (threshold) đặt ra là lớn hơn 0.5. Ngưỡng này có thể được tăng để lấy ra các tập khác nhau cho các mẫu chấp nhận được. Dãy các ngưỡng này nằm trong khoảng từ [0.5, 1.0) ($0.5 \leq \text{threshold} < 1.0$).
Đánh giá tập các mẫu tốt nhất: Từ mỗi tập các mẫu chấp nhận được, trích các các đặc trưng tương ứng để phân loại trên dữ liệu huấn luyện và đánh giá bởi 10 phần qua đánh giá chéo. và chọn một tập có giá trị cao nhất.

3.3.4 Kết quả thực nghiệm và thảo luận

3.3.4.1 Dữ liệu thực nghiệm

Thực nghiệm của chúng tôi được thực hiện trên dữ liệu đánh giá sản phẩm kỹ thuật thu thập từ một số diễn đàn kỹ thuật Việt như textit tinhte.vn, voz.vn, thegioididong.com. Chúng tôi gán nhãn bằng tay 9000 bình luận tiếng Việt thu thập được với hai loại nhãn "< sub >"(chủ quan) và "< obj >" (khách quan). Dữ liệu huấn luyện gồm: 3000 ý kiến chủ quan và 3000 ý kiến khách quan. Dữ liệu đánh giá gồm: 3000 bình luận còn lại chia đều cho hai nhãn. Các dữ liệu huấn luyện và đánh giá được tách ra thành từ và được gán nhãn từ loại.Chúng tôi sử dụng một số công cụ trong Weka [2] để đánh giá quá trình học và chất lượng của các mẫu được học trong dữ liệu đánh giá.

3.3.4.2 Các kết quả thực nghiệm

Quá trình học:
Bước 1: chúng tôi đã học các N-gram (unigram, Bigram) của các từ trong dữ liệu huấn luyện với ngưỡng trong khoảng [0,5; 0,6; 0,7; 0,8; 0,9]. Chúng tôi sử dụng thư viện liblinear trong WEKA để thực hiện phân loại SVM. Chúng tôi thực hiện trong 10 lần bằng cách đánh giá chéo, sau đó đánh giá kết quả thực hiện phân loại.
Bước 2: chúng tôi đã học được những mẫu từ loại của hai loại khuôn dạng. Chúng tôi

Bảng 3.8: Các kết quả phân loại của unigram and bigram

Ngưỡng	Unigram	Bigram
0.5	82.59%	72.93%
0.6	83.14%	73.52%
0.7	83.47%	75.52%
0.8	83.29%	77.47%
0.9	81.82%	79.27%

cũng sử dụng ngưỡng trong khoảng [0,5; 0,6; 0,7; 0,8; 0.9]. Các mẫu trong mỗi tập được áp dụng trên dữ liệu huấn luyện để trích xuất các cụm từ, tính từ, động từ là các đặc trưng cho phân loại chủ quan. Chúng tôi cũng sử dụng liblinear trong WEKA để đánh giá tập các mẫu.

2. http://www.cs.waikato.ac.nz/ml/weka/

Bảng 3.9: Các kết quả phân loại của học các mẫu kiểu 1

ngưỡng	tag-tag[+1]	tag-tag[-1]	tag-tag[+2]	tag-tag[-1] & tag[+1]	tag-tag[-2]
0.5	82.81%	82.82%	**83.29%**	82.66%	82.62%
0.6	81.16%	79.54%	81.41%	81.46%	81.67%
0.7	75.47%	79.41%	80.56%	78.27%	80.99%
0.8	67.92%	50.03%	76.19%	76.17%	76.81%
0.9	50.03%	50.03%	71.32%	72.74%	68.95%

Bảng 3.10: Các kết quả phân loại của học các mẫu loại 2

ngưỡng	word-tag[+1]	word-tag[-1]	word-tag[+2]	word-tag[-1] & tag[+1]	word-tag[-2]
0.5	82.95%	82.87%	82.69%	82.89%	82.76%
0.6	82.77%	83.06%	82.91%	82.89%	82.77%
0.7	82.66%	83.02%	82.91%	82.96%	82.57%
0.8	82.82%	83.21%	82.97%	82.86%	82.69%
0.9	82.42%	**83.37%**	83.06%	82.82%	82.71%

3.3.5 Đánh giá các mẫu học được

Chúng tôi phân tích hiệu quả của tập các đặc trưng: n-gram, các từ và các cụm từ được trích từ các mẫu được học. Các kết quả được đưa ra trong bảng 3.11.

Bảng 3.11: Các kết quả phân lớp trên dữ liệu đánh giá

	# Các đặc trưng	SVM	Naive Bayes
unigram	3894	82.29%	79.28%
bigram	3210	64.25%	59.24%
unigram + bigram	7104	82.54%	79.67%
words of patterns (type 1)	2972	82.56%	77.46%
words of patterns (type 2)	1655	82.68%	68.27%
words and phrases of patterns (type 1)	4472	82.56%	77.46%
words and phrases of patterns (type 2)	3062	82.68%	68.27%
words (type 1) + unigram + bigram	9847	83.47%	78.22%
words (type 2) + unigram + bigram	8421	84.03%	76.72%

3.3.6 Kết luận

Chúng tôi đã giới thiệu hai phương pháp cho phân tích chủ quan, trong đó phương pháp trích đặc trưng ngôn ngữ dựa vào các mẫu cú pháp và sử dụng bộ phân loại MEM cho độ chính xác cao (92.1%) cho dữ liệu đánh giá phim ảnh tiếng Anh. Phương pháp thống kê mới thứ 2 được chúng tôi giới thiệu để làm giàu các đặc trưng dựa trên các mẫu từ loại cho phân lớp khách quan tiếng Việt được phát triển dựa trên ý tưởng của phương pháp thứ nhất. Sử dụng SVM và NB để phân loại chủ quan, bằng cách kết hợp unigram, Bigram và từ trích xuất từ mẫu từ loại, hệ thống đạt được độ chính xác là 84,04 % với trường hợp tốt nhất trên phân loại SVM. Trong tương lai, chúng tôi sẽ mở rộng các đặc trưng bằng cách sử dụng các nhãn từ loại khác và khai thác thêm các khuôn mẫu mới.

Chương 4

PHÂN TÍCH QUAN ĐIỂM THEO KHÍA CẠNH

4.1 Giới thiệu

Phân tích quan điểm theo khía cạnh đều có hai giai đoạn riêng biệt, đầu tiên là phát hiện khía cạnh và tiếp theo là phân loại quan điểm tương ứng với khía cạnh được phát hiện.
Trong công việc này, chúng tôi đề xuất một mô hình dựa trên CNN được xây dựng để phân loại nhiều nhãn và được thiết kế cho cả hai nhiệm vụ: phát hiện khía cạnh và quan điểm được gán với phân loại khía cạnh. Mô hình hai pha của chúng tôi có đầu ra của giai đoạn 1 (để phát hiện khía cạnh) sẽ được sử dụng trong giai đoạn 2, và sau đó mô hình cuối cùng trong giai đoạn 2 sẽ tạo ra quan điểm gắn liền với khía cạnh tương ứng. Chúng tôi đề xuất một mô hình CNN mở rộng nhằm tích hợp các đặc trưng bên ngoài vào mô hình CNN thông thường. Các đặc trưng bên ngoài này phải từ các nguồn giàu thông tin để bổ sung thêm thông tin cho mô hình này. Chúng tôi sử dụng tiêu chí TF-IDF để chọn các mẫu ngôn ngữ thông tin là các đặc trưng ngoài.

4.2 Mô tả bài toán

Example 4.1. *Mô tả bài toán phân tích quan điểm theo khía cạnh.*
Cho một văn bản sau::

```
<sentences>
This place has got to be the best
Japanese restaurant
the New York area. I had a great
experience.
Food is great. Service is top
notch.
I have been going back again and
again.
<sentences>
```

Nhiệm vụ là xác định các đặc trưng được đề cập trong văn bản này cũng như quan điểm tương ứng được gắn với từng đặc trưng.
Từ ví dụ trên, chúng tôi cần nhận nhận được:

```
<sentiment>
RESTAURANT-GENERAL: positive
FOOD-QUALITY: positive
SERVICE-GENERAL: positive
</sentiment>
```

Trong đó, RESTAURANT-GENERAL, FOOD-QUALITY and SERVICE-GENERAL là các khía cạnh, mức độ gán nhãn của đánh giá là tích cực (positive)(trong trường hợp

này phân cực của quan điểm gồm: positive, negative, and neutral).
Nhiệm vụ của chúng tôi là xây dựng một mô hình thực hiện nhiệm vụ trên bằng cách sử dụng tập dữ liệu huấn luyện, trong đó mỗi văn bản được gắn nhãn với các khía cạnh và các lớp quan điểm tương ứng.

Sau đây là mô hình hoá của chúng tôi cho bài toán này:
Giả sử, chúng tôi cần thực hiện trên các nhận xét/đánh giá của một đối tượng, chúng tôi định nghĩa các tập sau: A là một tập các khía cạnh của đối tượng bao gồm s khía cạnh được biểu diễn như sau: $A = \{a_1, a_2, \ldots, a_s\}$.
C là một tập gồm k các lớp mức độ phân cực quan điểm được biểu diễn như sau: $C = \{c_1, c_2, \ldots, c_k\}$
Một tập dữ liệu huấn luyện D bao gồm N tài liệu được biểu diễn như sau: $D = \{d_1, d_2, \ldots, d_N\}$.
Trong đó: mỗi tài liệu d_i là một bình luận hoặc một đánh giá về một sản phẩm hoặc dịch vụ.

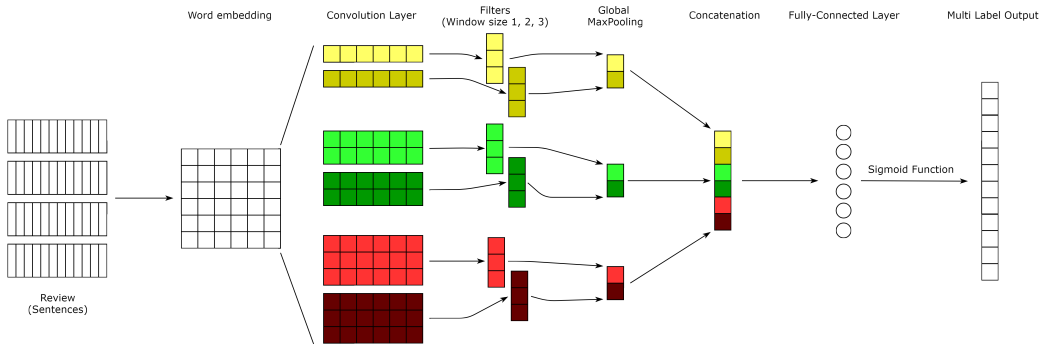
Lưu ý rằng các khía cạnh và quan điểm của chúng được gán bên ngoài toàn bộ văn bản d_i, Điều này làm cho công việc trở nên khó khăn hơn (so với trường hợp các đặc trưng và quan điểm của chúng được gán cho mỗi câu).
Ví dụ sau đây là mẫu cho một tài liệu được gán nhãn từ Task 5 - subtask 2, SemEval 2016.

```
<sentences>
$d_$
<sentences>
<sentiment>
$a_$:$c_$
. . .
</sentiment>
```

Nhiệm vụ chúng tôi cần thực hiện là làm thế nào tạo ra được một mô hình từ tập dữ liệu huấn luyện, và nó giúp sinh ra ra các đặc trưng và các lớp quan điểm tương ứng từ một tài liệu đầu vào mới.

4.3 Mô hình đề xuất

4.3.1 Mô hình CNN hai pha cho phân tích quan điểm theo khía cạnh (A two-phase CNN model for Aspect based Sentiment Analysis)



Hình 4.1: Mô hình CNN cho phân tích quan điểm theo khía cạnh.

Một mạng nơ-ron tích chập (CNN) bao gồm một tầng co giãn và một tầng gộp. Kiến trúc chung của một thành phần tích chập được thể hiện trong hình 4.1.
Giả sử rằng cho một câu vào có chứa một danh sách các từ $e_1, e_2, ..., e_n$, bằng cách sử dụng công cụ Word2vec, chúng ta có $x_1, x_2, ..., x_n$ tương ứng vectơ biểu diễn cho các từ này.

Lớp tích chập (Convolution layer)

Lớp này nhận $x_1, x_2, ..., x_n$ làm đầu vào và sử dụng các phép tính tích chập để thu được các vectơ đại diện mới. Ví dụ, bằng cách áp dụng một phép tính chập, chúng ta thu được các vectơ mới $y_1^1, y_2^1, ..., y_n^1$ theo phương trình sau:

$$y_i^1 = f(U.x_{i:i+h-1} + b), \tag{4.1}$$

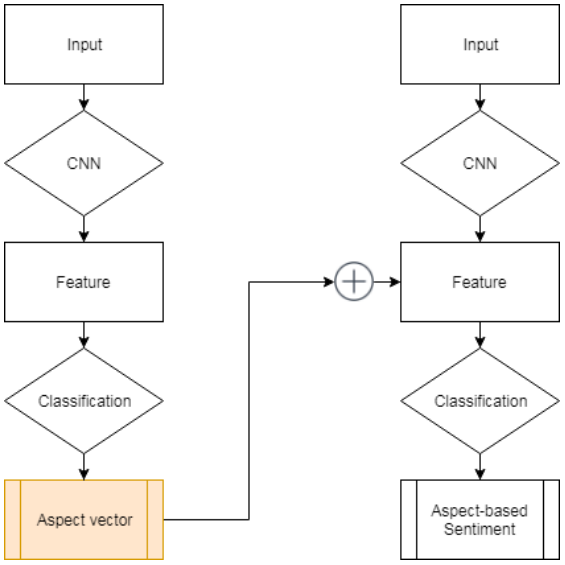
trong đó: $x_{i:i+h-1}$ biểu thị liên kết của các vectơ nhúng $x_i, x_{i+1}, ..., x_{i+h-1}$, h là kích thước cửa sổ cho các vectơ nhúng cần được kết hợp, $f(.)$ là một hàm phi tuyến kích hoạt phần tử, chúng tôi sử dụng hàm $f(t) = tanh(t) = frac e^t - e^{-t} e^t + e^{-t})$, $U\ trong\ rmR^{C\ lnm}$ và $b\ trong\ rmR^C$ là các tham số thành phần đã học trong giai đoạn huấn luyện và C là chiều đầu ra.

Lớp kết nối (Pooling layer)

Chúng tôi áp dụng phép toán tổng hợp lớn nhất để kết hợp các đặc trưng từ lớp tích chập thành một vector có kích thước cố định:

$$y^2 = [max(y_{i1}^1), max(y_{i2}^1), ..., max(y_{iC}^1)], \tag{4.2}$$

trong đó y_{ij}^1 biểu thị chiều thứ j của y_i^1,
$y^2\ inR^C$ là vectơ đầu ra của thành phần convolution .



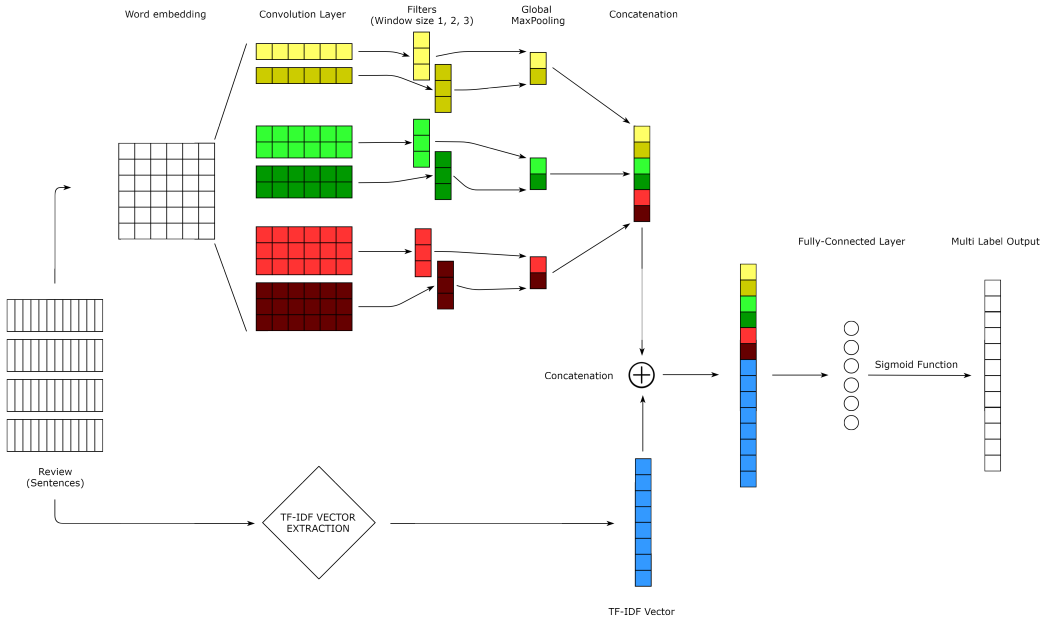
Hình 4.2: Hai pha của phân tích quan điểm theo khía cạnh.

Lớp được kết nối đầy đủ với nhiều lớp (Fully-connected Layer with multiple layer)

Nhận dữ liệu đầu vào là véc tơ số thực từ được biểu diễn của đầu vào dùng cho nhiệm vụ dự đoán nhãn.
Để dự đoán nhiều nhãn (cho nhiều khía cạnh), sử dụng hàm Sigmoid để lấy xác suất cho mỗi phần tử của mảng đầu ra. Mọi nhãn khía cạnh có xác suất cao hơn ngưỡng sẽ được xác định là một khía cạnh được dự đoán.

4.3.2 Mô hình CNN với các đặc trưng ngoài (The CNN Model with External Features)

Chúng tôi đề xuất mô hình CNN hai pha được tích hợp các đặc trưng ngoài để cải thiện chất lượng của mô hình khi dữ liệu thưa.



Hình 4.3: CNN with External Features.

4.4 Thực nghiệm

4.4.1 Dữ liệu

Trong công việc này, chúng tôi thực hiện phương pháp đã đề xuất trên tập dữ liệu ABSA 2016 [1]. Chúng tôi sử dụng dữ liệu từ Task 5, Subtask 2 (mức văn bản) gồm một tập hợp các đánh giá của khách hàng về dữ liệu nhà hàng. Mục tiêu là xác định một tập các bộ{khía cạnh (aspect), Mức độ phân cực (polarity)} để tóm tắt các ý kiến được thể hiện trong mỗi bài đánh giá.

4.4.2 Tiền xử lý dữ liệu

1. Biểu diễn từ cho mô hình CNN (Word Embeddings for CNN)
2. Trích các đặc trưng ngoài (Extracting external features)
Chúng tôi chọn 150 từ đặc trưng ngoài cho xác định khía cạnh và 300 từ đặc trưng cho phân loại quan điểm chi khía cạnh. Với mỗi văn bản đầu vào, chúng tôi chọn tập các đặc trưng ngoài và sau đó tạo ra một one-hot vectơ để tạo thành vectơ đặc trưng bên ngoài.

1. http://alt.qcri.org/semeval2016/task5/

4.4.3 Các mô hình và các kết quả

Chúng tôi thực hiện hai mô hình: mô hình đầu tiên là mô hình CNN cơ bản để phân loại nhiều nhãn; mô hình thứ hai là mô hình CNN cơ bản với các đặc trưng ngoài. Hai mô hình này được mô tả trong phần 4.3. Trong phần dưới đây, chúng tôi sẽ mô tả mô hình 2 (sử dụng các đặc trưng bên ngoài). Lưu ý rằng đối với mô hình 1, chúng tôi chỉ loại bỏ các đặc trưng bên ngoài khỏi mô hình 2.
Phương pháp này được thực hiện theo hai giai đoạn:
Giai đoạn 1: Phát hiện khía cạnh.
Giai đoạn 2: Phân loại phân cực quan điểm
Đối với nhiệm vụ phân loại phân cực quan điểm, chúng tôi nối đầu ra của CNN với vectơ ngoài và vectơ thu được từ pha 1, sau đó chuyển tiếp chúng qua lớp được kết nối đầy đủ. Chúng tôi huấn luyện ba mô hình tương ứng với ba lớp quan điểm (tích cực, tiêu cực và trung tính). Sau đó, chúng tôi tóm tắt ba kết quả đầu ra này cho kết quả cuối cùng.

4.4.4 Các kết quả

Bảng 4.1: Các kết quả đánh giá

Model	**Nghiên cứu của Soufian**		
	Precision	***Recall***	***F1***
Word2Vec+POS	65.90	**71.00**	68.40
Word2Vec+POS+Sentics	66.30	69.70	67.90
Word2Vec-Retro+POS	65.10	70.80	67.80
Word2Vec-Retro+POS+Sentics	65.50	70.60	67.90
	Nghiên cứu của Sebastian		
INSIGHT-1	-	-	68.11
	Các mô hình của chúng tôi		
CNN+multi-class+two-phases	**91.07**	61.70	73.53
CNN+multi-class+two-phases +External-Features	90.23	68.45	**77.84**

4.5 Kết luận

Trong chương này, chúng tôi đã đề xuất các mô hình dựa trên CNN cho bài toán phân tích quan điểm theo khía cạnh. Chúng tôi đã thêm các đặc trưng bên ngoài vào mô hình này để cải thiện hiệu suất của hệ thống. Thực nghiệm cho thấy các mô hình của chúng tôi thực hiện rất hiệu quả và đạt được kết quả tốt hơn nhiều so với các nghiên cứu trước đây được so sánh trên cùng một tập dữ liệu. Mô hình dựa trên CNN với các đặc trưng bên ngoài là mô hình hiệu quả nhất. Trong nghiên cứu tiếp theo, chúng tôi sẽ tiếp tục xem xét tích hợp các đặc trưng hữu ích khác vào mô hình nhằm cải thiện hiệu suất của hệ thống.

Chương 5

PHÂN TÍCH QUAN ĐIỂM TIẾNG VIỆT

5.1 Giới thiệu

Trong thời gian gần đây, nhiều nghiên cứu về phân tích cảm tính và khai thác ý kiến dữ liệu tiếng Việt được áp dụng vào thực tế để khai thác dữ liệu từ các trang mạng xã hội, diễn đàn, blog... Dữ liệu đã xử lý trong những ứng dụng này thường là không chuẩn và chứa nhiều lỗi chính tả và viết tắt, chúng ta gọi là dữ liệu kiểu Microblog.

Để có thể thực hiện các nghiên cứu trên loại dữ liệu này, nhiệm vụ đầu tiên cần thực hiện là các bước tiền xử lý văn bản như kiểm tra chính tả và tách từ. Các nghiên cứu và công cụ sử dụng để kiểm tra chính tả cho văn bản tiếng Việt còn hạn chế và mục tiêu ban đầu là thực hiện trên các dữ liệu chính thống. Đối với nhiệm vụ tách từ, hầu hết các công trình nghiên cứu và công cụ hiện có được thực hiện cho dữ liệu chuẩn tiếng Việt đạt kết quả cao. Tuy nhiên, độ chính xác sẽ giảm khi áp dụng cho kiểu dữ liệu Microblog. Do đó, việc phát triển các phương pháp tiền xử lý văn bản tiếng Việt cho dữ liệu Microblog là thực sự cần thiết.

Trong chương này, chúng tôi giới thiệu hai tiếp cận cho vấn đề này là phương pháp sử dụng n-gram lớn cho kiểm tra lỗi chính tả tiếng Việt và tách từ tiếng Việt cho dữ liệu Micoblog.

5.2 Phương pháp kiểm tra chính tả cho dữ liệu MicroBlogs sử dụng n-gram lớn

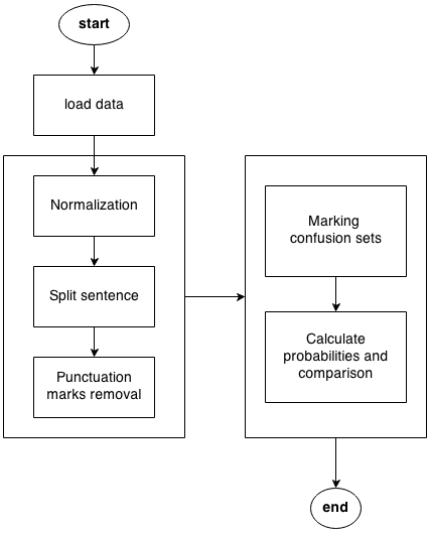
5.2.1 Một số lỗi chính tả thường gặp

Trong các ngôn ngữ nói chung, các lỗi chính tả được xem xét trong các nghiên cứu trước đây bao gồm: 1. Lỗi chính tả không phải từ (non-word error) và lỗi chính tả từ thực (real-word error). Lỗi không phải là từ: Lỗi do soạn thảo: Example: “bof” -> “bò” Lỗi từ thực:

- Lỗi do phát âm: (Tones making error) Ví dụ: “hõi” -> “hỏi”
- Lỗi phụ âm đầu (Initial consonant error) Ví dụ: “bức chanh” -> “bức tranh”
- Lỗi phụ âm cuối (End consonant error) Ví dụ: “bắt buột” -> “bắt buộc”.
- Lỗi vùng miền (Region error): Việt Nam có nhiều vùng có phương ngữ khác nhau, do đó cần phải thay đổi sang ngôn ngữ phổ thông) Ví dụ: “kím” -> “kiếm”

5.2.2 Mô hình kiểm tra chính tả đề xuất

Chúng tôi sử dụng tiếp cận dựa trên ngữ cảnh cho hệ thống kiểm tra chính tả. Trong đó, chúng tôi thực hiện tính toán độ đo mối quan hệ giữa các âm tiết và láng giềng của chúng và đánh giá kết quả đó để chọn âm tiết nhiều khả năng là đúng nhất. Chúng tôi đã mở rộng ngữ cảnh về cả hai phía của âm tiết và sử dụng kho ngữ liệu lớn để huấn luyện n-gram và nén để tối ưu hóa bộ nhớ. Kiến trúc của hệ thống được minh họa trong hình:



Hình 5.1: Kiến trúc của hệ thống kiểm tra chính tả

5.2.3 Tiền xử lý dữ liệu

Giai đoạn tiền xử lý có ba bước:

- **Bước 1:** Nhận biết các âm tiết đặc biệt như địa chỉ web, email, số :. và thay thế chúng bằng ký hiệu đặc biệt.
- **Bước 2:** Tách tài liệu thành các câu vì hai âm tiết trong các câu khác nhau không có mối quan hệ với câu khác.
- **Bước 3:** Xóa tất cả các dấu ngắt câu trong các câu bởi vì chúng không có mối quan hệ của ý nghĩa với các từ.

5.2.4 Thuật toán kiểm tra chính tả mở rộng ngữ cảnh ở cả hai bên

Thành phần chính của hệ thống kiểm tra chính tả của chúng tôi bao gồm hai bước:

- **Bước 1:** Xây dựng tập hợp lỗi cho mỗi âm tiết dựa trên khoảng cách soạn thảo vàcác đặc điểm ngôn ngữ tiếng Việt được chọn.
- **Bước 2:** Tính toán độ đo mối quan hệ giữa một âm tiết với các láng giềng của nó dựa trên mô hình N-gram để quyết định xem âm tiết hiện tại có đúng hay không sau đó chọn ứng cử viên có khả năng nhất để sửa nó.

Hệ thống của chúng tôi sử dụng ngữ cảnh là cửa sổ trượt bán kính 2 của các âm tiết xung quanh nó. Có nghĩa là nếu chúng ta biểu thị âm tiết hiện tại là w_0 và ngữ cảnh của nó là w_{-1}, w_{-2}, w_1, w_2. Chúng ta có thể mô hình hóa sự phụ thuộc của w_0 vào các âm tiết hàng xóm của nó bằng xác suất có điều kiện sau:

$$\mathrm{P}(w_0 \mid w_{-2}\ ,\ w_{-1},\ w_1,\ w_2)$$

Xác suất náy có thể ước lượng bằng hàm sau:

$$\mathrm{P}(w_0|w_{-2},w_{-1},w_1,w_2) = \mathrm{f}(\mathrm{P}(w_0|w_{-2},w_{-1}),\ \mathrm{P}(w_0|w_{-1},w_1),\ \mathrm{P}(w_0|w_1,w_2))$$

Trong đó f là một hàm trung bình nhân (geometric mean function). Để tính xác suất này, chúng ta cần 5-gram và 4-gram. Điều này là không thể thực hiện được vì số lượng kết họp quá lớn và dữ liệu quá rải rác. Thay vào đó, chúng ta tính xác suất: P(w_0 | w_{-2}, w_{-1}, w_1, w_2) và ước lượng các xác suất 3: P($w_0|w_{-2}$, w_{-1}), P($w_0|w_{-1}$, w_1), P($w_0|$

w_1, w_2). Các n-gram có xác suất p là các logarit trung bình nhân của ba xác suất 3. Chúng tôi chọn hàm trung bình hình học vì tên thực thể là tên người hoặc tổ chức có thể làm yếu đi liên kết của âm tiết với ngữ cảnh của nó. Các lỗi xuất hiện khi một âm tiết được xác định là lỗi, nhưng đây thực sự là lỗi chính tả. Để giảm số lỗi này, chúng tôi đã sử dụng các hệ số heuristic gọi là "ngưỡng lỗi" (error threshold) và "ngưỡng chênh lệch" (difference threshold), được viết tắt là *e_thresh* và *d_thresh*. Giả sử rằng âm tiết hiện tại là w_0 có giá trị N-gram là p và một âm tiết từ tập lỗi là w_0' có giá trị N-gram là p', w_0' được xem là "tốt hơn" w_0 khi và chỉ khi nó thỏa mãn hai bất đẳng thức sau:

- $p' > e\_thresh$
- $p' > p + d\_thresh$

e_thresh là một hằng số được xác định dựa trên dữ liệu phát triển, nó đảm bảo rằng nếu một âm tiết sẽ được sử dụng để sửa âm tiết hiện tại thì xác suất của nó phải cao hơn một ngưỡng nhất định; điều này giúp chúng ta giảm sai số do thực thể tên.

5.2.5 Mô hình N-gram lớn và nén N-gram

Để tính các xác suất 3, do đó phải xác định tần suất của các 2-gram (bigrams) và 3-gram (trigrams). **Nén n-gram** Trong quá trình mã hóa, chúng tôi thu thập từ điển âm tiết tiếng Việt bao gồm khoảng 6800 âm tiết. Mỗi âm tiết được biểu diễn bằng một số bắt đầu từ 0. chúng ta cần từ 0 đến 6800 để biểu diễn, mỗi số cần hai byte để lưu trữ. Đối với 2-gram (bigram), thể được lưu trữ bởi 4 byte (một số nguyên) và cần 6 byte để mã hóa 3-gram (trigram).

5.2.6 Thực nghiệm của chúng tôi

5.2.6.1 Dữ liệu thực nghiệm

- **Training data** Để xây dựng mô hình N-gram, chúng tôi đã thu thập dữ liệu từ nhiều nguồn khác nhau như *Wikipedia.org, dantri.com.vn, vnExpress.net*. Dữ liệu gồm nhiều chủ đề như: toán học, vật lý, khoa học, văn học, triết học, lịch sử, kinh tế, thể thao, luật, tin tức, giải trí ... Kích thước của kho ngữ liệu của chúng tôi là khoảng 2GB. Chúng tôi đã tính tần suất của unigram, bigram, trigram sau đó loại bỏ n-gram có tần suất nhỏ hơn 5.
- **Testing data** Chúng tôi đã tạo hai bộ thử nghiệm để đánh giá hệ thống này. Trước tiên, chúng tôi đã thu thập văn bản từ Internet. Trong tập đầu tiên, chúng tôi đã kiểm tra thủ công để đảm bảo rằng không có lỗi chính tả trong đó. Sau đó, chúng tôi tạo ra lỗi chính tả giả trong bộ kiểm tra và đánh dấu các lỗi này để đánh giá hiệu suất của hệ thống. Trong tập thứ hai, chúng tôi cũng tìm và đánh dấu lỗi chính tả. Tập đánh giá đầu tiên được sử dụng trong thử nghiệm 1 và 2, tập thứ hai trong thử nghiệm 3, tương ứng. Tập đánh giá đầu tiên chứa 2500 câu và câu thứ hai chứa 632 câu.

5.2.6.2 Các kết quả thực nghiệm

Trước khi đánh giá hiệu suất của hệ thống kiểm tra chính tả, chúng tôi đã áp dụng phương pháp nén n-gram được xây dựng từ dữ liệu đào tạo. Các kết quả nén được minh họa trong bảng:

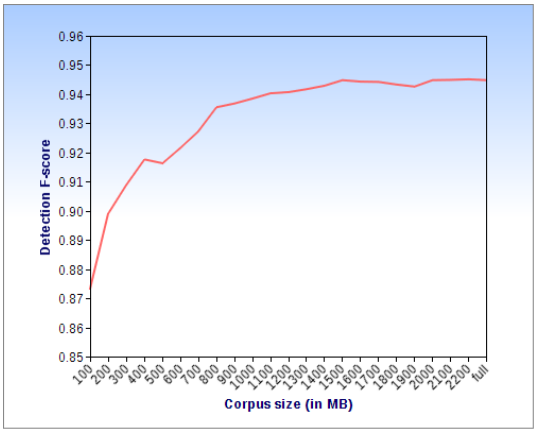
5.2.6.3 Thực nghiệm 1:

Phân tích ảnh hưởng của kích thước ngữ liệu huấn luyện n-gram đối với hiệu suất của hệ thống. Chia ngữ liệu thành các đoạn khoảng 100 MB. Một ngữ liệu nhỏ được tạo

Bảng 5.1: Các kết quả nén N-gram

	# số n-gram	kích thước dữ liệu trước khi nén	kích thước dữ liệu sau khi nén
unigram	6776	77.9 KB	13.55 KB
bigram	1208943	15.6 MB	4.6 MB
trigram	4886364	84 MB	28 MB

ra bằng cách ghép các đoạn lại với nhau. Chúng tôi đánh giá F-score của hệ thống với từng ngữ liệu nhỏ.



Hình 5.2: Ảnh hưởng của kích thước ngữ liệu đến hiệu suất của hệ thống

5.2.6.4 Thực nghiệm 2:

Chúng tôi đánh giá ảnh hưởng của ngữ cảnh đến độ chính xác của hệ thống. Bảng 2 đưa ra các kết quả đánh giá của mỗi ngữ cảnh.

Bảng 5.2: Ảnh hưởng của ngữ cảnh đến hiệu suất của hệ thống

Context	DP	DR	CP	DF	FPR
w_{-2}, w_{-1}	89.42%	52.22%	97.31%	65.93%	0.12%
w_{-1}, w_1	94.04%	91.53%	98.26%	92.76%	0.11%
w_1, w_2	93.83%	73.63%	96.79%	82.51%	0.09%
w_{-2}, w_{-1}, w_1, w_2	94.68%	94.26%	99.32%	94.46%	0.1%

5.2.6.5 Thực nghiệm 3:

Chúng tôi đã so sánh hệ thống của mình với một hệ thống kiểm tra chính tả khác cho tiếng Việt: copcon 5.0.3 beta footnote link: http://chinhta.vn. Chúng tôi sẽ so sánh độ chính xác của việc phát hiện và kiểm tra lỗi của chúng trên bộ thử nghiệm thứ hai. Kết quả được thể hiện trong bảng 5.3.

Bảng 5.3: So sánh độ chính xác hệ thống của chúng tôi và hệ thống kiểm tra chính tả Copcon

	DP	DR	CP	DF	FPR
Our system	92.62%	91.12%	95.45%	91.86%	0.2%
Copcon 5.0.3 beta	80.8%	77.6%	87.5%	79.2%	0%

5.3 Phương pháp tách từ cho dữ liệu Micro-blogs tiếng Việt

Trong phần này, chúng tôi giới thiệu phương pháp tách từ cho dữ liệu Micro-blogs tiếng Việt. Trong ngôn ngữ Tiếng Việt, các từ không được phân tách bằng các khoảng trắng. Trên thực tế, một từ trong tiếng Việt có thể chứa một hoặc nhiều âm tiết. Do đó có sự nhập nhằng của ranh giới từ làm cho nhiệm vụ tách từ trở lên khó khăn hơn. Có hai loại nhập nhằng chính được đề cập đến gồm: "Nhập nhằng chồng chéo" và "Nhập nhằng liên kết". Chúng tôi dùng hệ thống kiểm tra chính tả đã được giới thiệu để chuẩn hóa dữ liệu Microblogs trước khi áp dụng thuật toán tách từ.

5.3.1 Tiếp cận của chúng tôi cho bài toán tách từ dữ liệu Micro-blogs

Chúng tôi giới thiệu một số cải tiến để cải thiện nhược điểm của phương pháp kết hợp dài nhất. Hệ thống này bao gồm 3 bước, (1) sử dụng phương pháp nhận dạng tên riêng (name entity recognitio - NER),(2) Phát hiện nhập nhằng, (3) Lựa chọn khả năng thích hợp nhất. Để phát hiện nhập nhằng: thay vì dùng thuật toán kết hợp dài nhất chúng tôi phát hiện sự tồn tại của nhập nhằng trong câu nhập vào để đưa ra các trường hợp được tách từ và dựa vào mô hình N-gram để tính toán xác suất của mỗi ứng cử được chia tách để lựa chọn khả năng thích hợp nhất. - Để xử lý nhập nhằng chồng chéo chúng tôi tìm các từ có trong từ điển về cả hai phía để phát hiện hai từ liên tiếp có các âm tiết chung. Nếu các ứng cử được tách là như nhau thì không có nhầm lẫn chồng chéo, trong trường hợp ngược lại chúng tôi tính toán xác suất của mỗi ứng cử để chọn cái thích hợp nhất. Chúng ta theo dõi bảng dưới đây.

Bảng 5.4: Phát hiện các nhập nhằng chồng chéo

	segmented candidate
w_1 (matching from left to right)	*Tốc_độ truyền_thông tin ngày_càng tăng*
w_2 (matching from right to left)	*Tốc_độ truyền thông_tin ngày_càng tăng*

Trong ví dụ trên, có một nhầm lẫn chồng chéo vì các ứng cử được tách là khác nhau. Sau đó, chúng tôi tính $P(w_1)$ và $P(w_2)$ và chọn w_2 nếu $P(w_2) > P(w_1)$ hoặc chọn w_1 trong trường hợp: $P(w_1) > P(w_2)$.

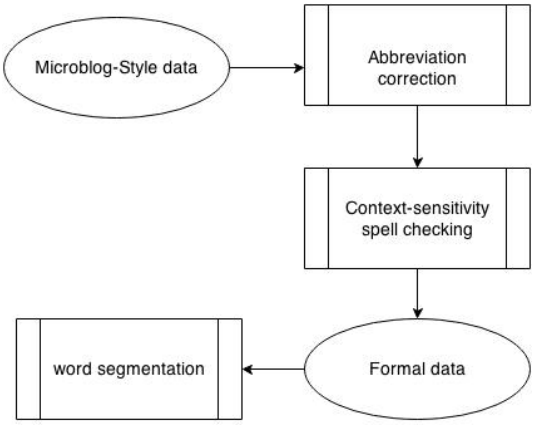
- Để phát hiện nhập nhằng liên kết, chúng tôi tách mỗi từ dài thành các từ ngắn hơn. Ví dụ, trong câu: "*Bàn là công cụ học tập*". Sử dụng thuật toán ghép cặp lớn nhất, chúng tôi thu được một số các từ ghép như sau: "*bàn_là*", "*công_cụ*", "*học_tập*", và tách từng từ ghép từ trái sang phải tương ứng. Sau đó, chúng rôi tính xác suất của các ứng cử được tạo ra và chọn ra cái tốt nhất trước khi chuyển sang từ dài tiếp theo. Công việc này giúp tránh tạo ra một số lượng lớn các kết hợp. Chúng tôi cũng sử dụng mô hình N-gram để tính toán xác suất của từng ứng cử. Ví dụ, sau khi sử dụng thuật toán ghép cặp dài nhất, câu *"Bàn là công cụ học tập"* được tách như sau: *Bàn_là công_cụ học_tập* Sau khi tách từ ghép thứ nhất "*Bàn_là*" chúng tôi có các ứng cử trong bảng 5.5 Nếu $P(w_2) > P(w_1)$, thì $w2$ sẽ được chọn làm câu có khả năng phân đoạn nhất tại thời điểm này, ngược lại, chúng tôi sẽ chọn w_1. Sau đó, chúng tôi thực hiệnlàm tương tự cho các từ tiếp theo "*công_cụ*", "*học_tập*" trên câu phân đoạn được chọn sau khi tách "*Bàn_là*"

Bảng 5.5: Phát hiện các nhập nhằng liên kết

	Ứng cử được tách
w_1	*Bàn_là công_cụ học_tập*
w_2	*Bàn là công_cụ học_tập*

để có được kết quả cuối cùng.

Trong hệ thống của chúng tôi, nhập nhằng chồng chéo được phát hiện trước nhập nhằng liên kêt.



Hình 5.3: Hệ thống tách từ có sử dụng sửa lỗi chính tả

5.3.2 Hệ thống tách từ có sử dụng kiểm tra chính tả (Adaption to word segmentation by spell-checking system)

Trong phần này, chúng tôi đề xuất một cách sử dụng kiểm tra chính tả được đề xuất ở trên để chuẩn hóa cho các văn bản Micro-blogs để cải thiện hiệu suất tách từ. (1) Kiểm tra viết tắt: Để phát hiện các từ viết tắt, sau đó thay thế chúng bằng các từ hoặc cụm từ chính xác.(2) Kiểm tra chính tả dựa trên ngữ cảnh: Để phát hiện và sửa lỗi chính tả thông thường.

5.3.3 Các thực nghiệm

Bảng 5.6: Dữ liệu huấn luyện của hệ thống kiểm tra chính tả

	# Số phần tử	Kích thước trước khi mã hóa	Kích thước sau khi mã hóa
Unigram	6776	77.9 KB	13.55 KB
Bigram	1208943	15.6 MB	4.6 MB
Trigram	4886364	84 MB	28 MB

Bảng 5.7: Dữ liệu test

	# Số câu	kích thước	# số lỗi
Fomal data	2000	285 KB	0
Informal data	2000	322 KB	4754

5.3.3.1 Chuẩn bị dữ liệu

Chúng tôi sử dụng từ điển tiếng Việt VCL_SP7.2 (SP7.2) [1] bao gồm khoảng 35000 từ. Tập dữ liệu huấn luyện gồm 77000 câu đã được tách từ, dùng để xử lý nhập nhằng bằng tay. Chúng tôi đã trích xuất thêm 5000 từ mới từ tập dữ liệu huấn luyện để làm phong phú thêm từ điển VCL_SP7.2. Thu thập hai loại dữ liệu văn bản, đó là dữ liệu văn bản chính thống gồm nhiều nguồn khác nhau (huấn luyện n-gram cho hệ thống kiểm tra chính tả cảm ngữ cảnh - 2 GB)và dữ liệu Micro-blogs từ các trang diễn đàn kỹ thuật. Từ điển các từ viết tắt được xây dựng từ dữ liệu Microblog, và thu được 281 từ viết tắt. Dữ liệu test gồm 2 bộ: 2000 câu từ dữ liệu văn bản chính thức và 2000 câu từ dữ liệu Microblog. cho hai bộ thử nghiệm. Bộ kiểm tra Microblog được kiểm tra chính tả theo cách thủ công.

5.3.3.2 Các kết quả thực nghiệm và thảo luận

Hiệu suất của tách từ thử nghiệm trên dữ liệu Microblog. Hiệu suất của tách từ trên

Bảng 5.8: Word segmentation on formal data and Microblog-Style data

Data	Precsion	Recall	F-measure
Formal data	97.48%	98.41%	97.94%
Microblog data	94.35%	95.21%	94.78%

dữ liệu Microblog sau khi sử dụng tính năng kiểm tra chính tả theo ngữ cảnh và kiểm tra từ viết tắt.

Bảng 5.9: Tách từ trên dữ liệu Microblog-Style sau khi sử dụng kiểm tra chính tả

Dữ liêu	Precsion	Recall	F-measure
Original Microblog data	94.35%	95.21%	94.78%
Context-sensitive checking	95.13%	95.26%	95.2%
Abbreviation correction	96.17%	97.05%	96.61%
Abbreviation correction and context-sensitive checking	97%	97.12%	97.06%

5.4 Kết luận

Chúng tôi đã trình bày hai đề xuất là xây dựng hệ thống kiểm tra chính tả cho dữ liệu Microblogs-Style và hệ thống tách từ cho dữ liệu này trong đó thống kiểm tra chính tả cảm ngữ cảnh và kiểm tra viết tắt. Việc sử dụng kiểm tra chính tả trước khi tách đã giúp dữ liệu Microblog thích ứng với quá trình phân đoạn từ và hiệu suất đã tăng lên đáng kể. Kết quả này cũng đóng góp vào việc làm tăng hiệu suất của phân tích quan điểm tiếng Việt trên dữ liệu này.

1. http://vlsp.vietlp.org:8080/demo/ ?&lang=en

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Tóm lược các kết quả và đóng góp của luận án

Luận án đã tập trung vào nghiên cứu việc trích chọn các đặc trưng ngữ pháp hữu ích áp dụng cho phân tích quan điểm trên dữ liệu tiếng Anh và tiếng Việt.

Chúng tôi đề xuất hai phương pháp cho phân loại chủ quan cho dữ liệu tiếng Việt và tiếng Anh. Với dữ liệu tiếng anh, chúng tôi giới thiệu mô hình trích các đặc trưng ngôn ngữ dựa trên các mẫu cú pháp cho để phân loại câu chủ quan. Chúng tôi đã thử nghiệm phương pháp trên bộ dữ liệu đánh giá phim ảnh.

Với bài toán phân loại quan điểm theo khía cạnh chúng tôi cũng đề xuất một mô hình tích hợp các đặc trưng giàu thông tin bên ngoài vào mô hình mạng nơ ron tích chập để tăng hệu suất thực hiện cho mô hình.

Trong quá trình phát triển phương pháp phân tích quan điểm trên đối tượng dữ liệu tiếng Việt, chúng tôi cũng đề xuất mô hình kiểm tra chính tả cho dữ liệu Microblog tiếng Việt và mô hình tách từ sử dụng hệ thống kiểm tra từ viết tắt và kiểm tra chính tả trong tách từ tiếng Việt để phù hợp với dữ liệu dạng Microblog nhằm tăng hiệu xuất thực hiện cho phân tích quan điểm tiếng Việt.

2. Những hạn chế và hướng nghiên cứu tiếp theo của luận

Bài toán phân loại chủ quan: chúng tôi sẽ tiếp tục xem xét việc tích hợp các đặc trưng hữu ích cho các mô hình học sâu mới nhằm nâng cao hiệu suất thực hiện. Đối với phân tích tính chủ quan dữ liệu tiếng Việt, chúng tôi sẽ tiếp tục mở rộng các đặc trưng bằng cách sử dụng các nhãn từ loại khác và khai thác thêm các khuôn mẫu mới.

Bài toán phân loại quan điểm theo khía cạnh, chúng tôi sẽ tiếp tục khai thác các mô hình học sâu mạnh và xem xét tích hợp các đặc trưng hữu ích khác vào mô hình nhằm cải thiện hiệu suất của hệ thống. Chúng tôi cũng tiếp tục nghiên cứu bài toán này cho cả hai đối tượng dữ liệu tiếng Anh và tiếng Việt.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC

1 . **Huong Nguyen Thi Xuan**, Vo Cong Hieu, and Anh-Cuong Le (2018). "Adding External Features to Convolutional Neural Network for Aspect-based Sentiment Analysis", In In Proc. The 5th NAFOSTED Conference on Information and Computer Science (NICS), pp. 53-59.

2 . **Nguyen Thi Xuan Huong**, Tran-Thai Dang, Anh-Cuong Le (2014), "Adapting Vietnamese Word Segmentation for Microblog-Style Data", In In Proc., The Third Asian Conference on Information Systems, pp. 164-171.

3 . Tran-Thai Dang, **Nguyen Thi Xuan Huong** and Anh-Cuong Le and Van-Nam Huynh (2014), "Automatically Learning Patterns in Subjectivity Classification for Vietnamese", In Proc. The Sixth International Conference on Knowledge and Systems Engineering (KSE 2014), pp. 675-690.

4 . **Nguyen Thi Xuan Huong**, Tran-Thai Dang, The-Tung Nguyen, Anh-Cuong Le (2014), "Using Large N-gram for Vietnamese Spell Checking", In Proc. The Sixth International Conference on Knowledge and Systems Engineering (KSE 2014), pp. 655-674.

5 . **Huong Nguyen Thi Xuan**, Aanh-Cuong Le and Le Minh Nguyen, (2012), "Linguistic Features for Subjectivity Classification.", In Proc. of the 6th International Conf. The International Conference on Asian Language Processing (IALP 2012), pp. 17-20.

Danh mục này gồm 05 công trình.