BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC THĂNG LONG

KHÓA LUẬN TỐT NGHIỆP

ĐỀ TÀI:

ỨNG DỤNG DEEP LEARNING TRONG PHÂN TÍCH CẢM XÚC

SINH VIÊN THỰC HIỆN : TRẦN VĂN HÙNG

MÃ SINH VIÊN : A30037

CHUYÊN NGÀNH : TOÁN ỨNG DỤNG

HÀ NỘI – 2021

LỜI CẢM ƠN

Đầu tiên, em xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới đã tận tình chỉ bảo, hướng dẫn, động viên và giúp đỡ em trong quá trình thực hiện đề tài khóa luận này.

Em xin gửi lời cảm ơn sâu sắc tới quý thầy cô giáo trong Khoa Toán Tin nói riêng và trường Đại học Thăng Long nói chung đã truyền đạt kiến thức quý báu cho em trong những năm tháng ngồi trên ghế nhà trường.

Cuối cùng, tôi xin gửi lời cảm ơn tới bạn bè, đặc biệt là tập thể lớp TM29 đã ủng hộ, giúp đỡ tôi trong suốt quá trình học tập trên giảng đường đại học. Tôi xin chân thành cảm ơn!

Hà Nội, ngày tháng năm 2021

Sinh viên

HÙNG

TRẦN VĂN HÙNG

LỜI CAM ĐOAN

Tôi xin cam đoan các kỹ thuật sử dụng trong khoá luận này là do tôi thực hiện dưới sự hướng dẫn của

Tất cả những tài liệu tham khảo từ các nghiên cứu liên quan đều được trích dẫn nguồn gốc rõ ràng từ danh mục tài liệu tham khảo của khoá luận. Trong khoá luận này, không có việc sao chép tài liệu, các công trình nghiên cứu của người khác mà không ghi rõ trong tài liệu tham khảo.

Nếu phát hiện có bất kì sự gian lận nào, tôi xin hoàn toàn chịu trách nhiệm trước hội đồng cũng như kết quả khóa luận tốt nghiệp của mình..

Hà Nội, ngày tháng năm 2021

Sinh viên

Trần Văn Hùng

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Cụm từ tiếng Anh	Cụm từ tiếng Việt
1	CNN	Convolutional Neural Network	Mạng nơron tích chập
2	CRF	Conditional Random Fields	Trường điều kiện ngẫu nhiên
3	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
4	ReLU	Rectified Linear Unit	Đơn vị tinh chỉnh tuyến tính
5	RNN	Recurrent Neural Networks	Mạng nơron hồi quy
6	SVM	Support Vector Machine	Máy vectơ hỗ trợ

MỞ ĐẦU

Sentiment analysis (Phân tích cảm xúc) là công nghệ được sử dụng để đo lường cảm xúc trong thông điệp truyền tải dựa vào những đặc điểm được lập trình sẵn dựa trên thang điểm mặc định trong hệ thống, có sự tác động của một số yếu tố như ngữ cảnh, không gian, thời gian, ánh sáng, âm thanh,… Khi sử dụng Sentiment analysis cho phép người dùng đo lường được những đặc điểm, sự thay đổi về mặt cảm xúc, thái độ..từ đó có thể điều chỉnh những hành vi và lời nói nhằm đạt đến mục đích mong muốn.

Mặc dù phân tích quan điểm là một lĩnh vực quan trọng và đã có một loạt các ứng dụng, nhưng rõ ràng đây không phải là một nhiệm vụ đơn giản và có nhiều thách thức liên quan đến xử lý ngôn ngữ tự nhiên (natural language processing - NLP). Từ khoảng một thập kỷ trước, học sâu đã nổi lên như một kỹ thuật học máy mạnh mẽ và tạo ra kết quả tiên tiến trong nhiều lĩnh vực ứng dụng, từ thị giác máy và nhận dạng giọng nói đến NLP. Áp dụng học tập sâu để phân tích quan điểm cũng đã trở nên phổ biến hơn trong thời gian gần đây với một số mô hình nổi bật như mạng nơ ron tích chập xếp tầng của Gu và cộng sự, mạng nơ ron tích chập đa nhiệm của Ruder và cộng sự, tích hợp hai mô hình mạng nơ ron cho phân tích quan điểm mức khía cạnh của Toh và Su (2016).

Mục tiêu của khóa luận này là đưa ra một mô hình học sâu cho bài toán phân tích cảm xúc. Để tiếp cận mục tiêu này, khoán luận giới thiệu các phương pháp giải quyết bài toán phân tích quan điểm và một số mô hình trong phương pháp học sâu. Nội dung của khóa luận này được tổ chức thành 4 chương như sau:

Chương 1: Giới thiệu chung.

Chương này giới thiệu về phân tích quan điểm, học sâu, đồng thời trình bày về bài toán phân tích quan điểm của khóa luận.

Chương 2: Kỹ thuật học sâu cho phân tích cảm xúc.

Trong chương này, khóa luận giới thiệu về một số hướng tiếp cận và các kỹ thuật học sâu sử dụng trong phân tích quan điểm, đồng thời khảo sát một số mô hình học sâu và đưa ra phương pháp áp dụng cho bài toán khóa luận.

Chương 3: Mô hình học sâu cho phân tích cảm xúc.

Chương này trình bày về mô hình học sâu được sử dụng trong bài toán phân tích quan điểm của khóa luận.

Chương 4: Thực nghiệm và kết quả.

Ở chương này, khóa luận trình bày dữ liệu thực nghiệm và các kết quả thực nghiệm của khóa luận.Cuối cùng, phần kết luận tóm lược kết quả đạt được của khóa luận, chỉ ra những hạn chế và điểm chưa làm được, đồng thời đưa ra định hướng phát triển trong tương lai.

CHƯƠNG 1. GIỚI THIỆU CHUNG

Để hiểu và giải quyết được bài toán phân tích quan điểm, đầu tiên cần nắm vững các khái niệm liên quan đến bài toán. Chương này giới thiệu về khái niệm quan điểm, các ứng dụng, thách thức phân tích cảm xúc và các bài toán liên quan, đồng thời phát biểu bài toán của khóa luận. Bên cạnh đó, chương này cũng giới thiệu các kiến thức về phương pháp học sâu mà khóa luận sử dụng để giải quyết bài toán.

1.1. Giới thiệu bài toán phân tích quan điểm

1.1.1. Các khái niệm

Theo Bing Liu, một quan điểm (opinion) là một bộ *(g, s, h, t)*, trong đó *g* là đích của quan điểm, *s* là cảm xúc về đích *g*, *h* là chủ quan điểm (người hoặc tổ chức đưa ra quan điểm) và t là thời điểm quan điểm được bày tỏ.

Bốn thành phần ở đây đều rất cần thiết, không thể mất bất kỳ thành phần nào. Ví dụ:

- Thành phần thời gian thường rất quan trọng trong thực tế vì một quan điểm hai năm trước không giống một quan điểm bây giờ.
- Không có chủ quan điểm cũng có vấn đề.quan điểm từ một người rất quan trọng (như thủ tướng) có lẽ quan trọng hơn ý kiến của một người dân bình thường.Một ý kiến từ một tổ chức cũng thường quan trọng hơn một ý kiến từ một cá nhân riêng lẻ.

Cảm xúc (sentiment) cơ bản là cảm giác, thái độ, đánh giá hoặc tình cảm liên quan đến một quan điểm. Nó được biểu diễn dưới dạng bộ *(y, o, i)*, trong đó *y* là loại cảm xúc, *o* là định hướng của cảm xúc và *i* là cường độ của cảm xúc. Tuy nhiên, định nghĩa này có thể được đơn giản hóa. Trong nhiều ứng dụng, tích cực được ký hiệu là +1, tiêu cực ký hiệu là −1 và trung lập ký hiệu là 0. Trong hầu hết các ứng dụng, năm mức xếp hạng là đủ, ví dụ: 1-5 sao. Trong cả hai trường hợp, cảm xúc có thể được biểu diễn bằng một giá trị duy nhất. Hai thành phần khác trong bộ ba có thể được gấp lại thành giá trị này. Cảm xúc có thể là tích cực (positive), tiêu cực (negative) hoặc trung lập (neutral).

Một điều quan trọng về định nghĩa trên cần nhấn mạnh là quan điểm có mục tiêu. Mục tiêu của quan điểm là thực thể hoặc một phần, một thuộc tính của thực thể mà quan điểm đã được thể hiện. Điều này xác định một khái niệm mở rộng cho quan điểm: Một quan điểm là một bộ *(e, a, s, h, t)*. Trong đó *e* là thực thể đích, *a* là khía cạnh đích của thực thể *e* mà ý kiến đã được đưa ra, *s* là cảm xúc của quan điểm về khía cạnh *a* của thực thể *e*, *h* là chủ quan điểm và *t* là thời gian bày tỏ quan điểm. Khi một quan điểm chỉ trên toàn bộ thực thể, ta sử dụng khía cạnh đặc biệt chung để biểu thị nó. Ở đây *e* và *a* cùng đại diện cho mục tiêu quan điểm. Phân tích quan điểm dựa trên định nghĩa này thường được gọi là phân tích quan điểm dựa trên khía cạnh.

Quan điểm có thể được phân thành hai loại: quan điểm thông thường (regular opinion) và quan điểm so sánh (comparative opinion).

- Một quan điểm thông thường thể hiện một cảm xúc về một thực thể cụ thể hoặc một khía cạnh của thực thể, ví dụ, "Coca Cola có vị rất ngon." bày tỏ một cảm xúc hoặc quan điểm tích cực về khía cạnh hương vị của Coca Cola. Đây là loại quan điểm phổ biến nhất.
- Một quan điểm so sánh so sánh nhiều thực thể dựa trên một số khía cạnh được nêu ra, ví dụ, "Pepsi ngon hơn Coca Cola." so sánh Pepsi và Coca Cola và dựa trên thị hiếu của họ (một khía cạnh) và thể hiện sự ưa thích đối với Pepsi.

Ngoài ra, trên cơ sở tính khách quan, quan điểm cũng có thể được chia thành quan điểm chủ quan (subjective opinion) và quan điểm khách quan (fact-implied opinion), hay first-person opinion và non-first-person opinion dựa trên người giữ quan điểm.

Phân tích cảm xúc (sentiment analysis) hay khai phá quan điểm (opinion mining), là lĩnh vực nghiên cứu phân tích ý kiến, cảm xúc, đánh giá, thái độ và cảm xúc của con người đối với các thực thể và thuộc tính của chúng được thể hiện bằng văn bản. Các thực thể có thể là sản phẩm, dịch vụ, tổ chức, cá nhân, sự kiện, vấn đề hoặc chủ đề... Mục tiêu

là để xác định xem văn bản được tạo bởi người dùng có truyền đạt ý kiến tích cực, tiêu cực hoặc trung lập của họ không.

1.1.2. Ứng dụng

Sự khởi đầu và phát triển nhanh chóng của lĩnh vực này trùng khớp với sự phát triển của các phương tiện truyền thông xã hội trên Web, như các diễn đàn đánh giá, thảo luận, blog và mạng xã hội, vì lần đầu tiên trong lịch sử loài người, chúng ta có một khối lượng lớn dữ liệu quan điểm được ghi lại dưới dạng kỹ thuật số. Từ đầu năm 2000, phân tích quan điểm đã trở thành một trong những lĩnh vực nghiên cứu tích cực nhất trong xử lý ngôn ngữ tự nhiên. Nó cũng được nghiên cứu rộng rãi trong khai phá dữ liệu, khai phá Web, khai phá văn bản và truy xuất thông tin. Trên thực tế, nó đã lan rộng từ khoa học máy tính sang khoa học quản lý và khoa học xã hội như tiếp thị, tài chính, khoa học chính trị, truyền thông, khoa học sức khỏe và thậm chí cả lịch sử, do tầm quan trọng của nó đối với toàn bộ doanh nghiệp và xã hội. Sự phổ biến này là do thực tế, các ý kiến là trung tâm của hầu hết các hoạt động của con người và là nhân tố chính ảnh hưởng đến hành vi của chúng ta. Ở một mức độ đáng kể, niềm tin và nhận thức của chúng ta về thực tế và những lựa chọn chúng ta đưa ra dựa trên cách người khác nhìn và đánh giá thế giới. Vì lý do này, bất cứ khi nào chúng ta cần đưa ra quyết định, chúng ta thường tìm kiếm ý kiến của người khác. Điều này không chỉ đúng với cá nhân mà còn đúng với các tổ chức.

Ngày nay, nếu muốn mua một sản phẩm tiêu dùng, người ta không còn giới hạn trong việc hỏi ý kiến của bạn bè và gia đình vì có nhiều đánh giá và thảo luận của người dùng về sản phẩm này trên các diễn đàn, cộng đồng trên Web. Đối với một tổ chức, có thể không còn cần thiết phải tiến hành khảo sát, thăm dò ý kiến và các nhóm tập trung để thu thập ý kiến công chúng vì có rất nhiều thông tin như vậy có sẵn công khai.

Trong những năm gần đây, các bài đăng có ý kiến trong phương tiện truyền thông xã hội đã giúp định hình lại các doanh nghiệp, và làm lung lay quan điểm và cảm xúc của công chúng, những thứ đã tác động sâu sắc đến hệ thống chính trị xã hội của chúng ta.

Tuy nhiên, việc tìm kiếm và giám sát các trang web ý kiến và chắt lọc thông tin chứa trong đó vẫn là một nhiệm vụ khó khăn vì sự phổ biến và đa dạng của các trang web. Mỗi trang web thường chứa một khối lượng lớn văn bản ý kiến không phải lúc nào cũng dễ dàng được giải mã trong các blog và các bài đăng trên diễn đàn. Trung bình, người đọc sẽ gặp khó khăn trong việc xác định các trang web có liên quan và trích xuất, tóm tắt các ý kiến trong đó. Vì vậy, cần có hệ thống phân tích quan điểm tự động. Điều này khiến có nhiều start-up tập trung vào việc cung cấp dịch vụ phân tích quan điểm. Nhiều tập đoàn lớn cũng đã xây dựng hệ thống nội bộ của riêng họ. Những ứng dụng thực tế và lợi ích công nghiệp này đã cung cấp động lực mạnh mẽ cho nghiên cứu trong phân tích quan điểm.

1.1.3. Thách thức

Phân tích quan điểm cũng tồn tại nhiều thách thức khác nhau, trong đó có:

- Quan điểm ngầm: Có khả năng một câu có thể chứa cảm xúc ngầm mặc dù nó không có bất kỳ từ mang quan điểm nào.
- Phụ thuộc miền: Trong thách thức này, cảm xúc thay đổi từ một miền này sang một miền khác trong sự phụ thuộc miền.
- Vấn đề ngôn ngữ: Trong phân tích cảm xúc,tiếng Anh chủ yếu được sử dụng vì về tính sẵn có của tài nguyên, có nghĩa là từ vựng, từ điển và ngôn ngữ nhưng người dùng bị thu hút bằng cách sử dụng phân tích quan điểm với các ngôn ngữ khác ngoài tiếng Anh như tiếng Hindi, tiếng Pháp, tiếng Trung Quốc, tiếng Đức Ả Rập...
- Ý kiến giả mạo: Ý kiến giả mạo cũng được gọi là đánh giá giả và đề cập đến đánh giá không có thật hoặc giả. Ý kiến giả mạo đang gây hiểu lầm cho người dùng hoặc độc giả bằng cách cung cấp cho họ ý kiến tích cực hoặc tiêu cực không trung thực liên quan đến bất kỳ đối tượng nào. Đây là thách thức xã hội mà phân tích quan điểm phải đối mặt.

1.1.4. Các bài toán con

Nguồn dữ liệu quan trọng để phân tích quan điểm là phương tiện truyền thông xã hội trực tuyến khi người dùng tạo ra lượng thông tin ngày càng tăng. Mạng xã hội liên tục mở rộng, tạo ra nhiều thông tin phức tạp và liên quan đến nhau hơn nhiều. Do đó, các loại nguồn dữ liệu này phải được xem xét theo cách tiếp cận dữ liệu lớn, dữ liệu phải được xử lý, truy cập, lưu trữ hiệu quả để đảm bảo độ tin cậy của kết quả thu được. Ngoài ra, các kho dữ liệu này chứa cả các thông tin hữu ích được các cá nhân, tổ chức quan tâm và các thông tin rác không cần thiết. Điều này dẫn đến việc hình thành các bài toán con khác như phân lớp khía cạnh (aspect classification), phân tích tranh luận và bình luận (debate and comment analysis), khai phá ý định (intention mining), phân lớp chủ quan và khách quan (subjectivity classification), phân lớp ý kiến trái chiều (sentiment polarity classification), phát hiện ý kiến rác (spam opinion detection), tóm tắt và tổng hợp quan điểm (opinion summarization), phân tích tính đa diện của của một ý kiến (dual sentiment analysis)...

Nghiên cứu phân tích quan điểm chủ yếu được thực hiện ở ba mức độ chi tiết: mức độ tài liệu, mức độ câu và mức độ khía cạnh. Các bài toán con trong phân tích quan điểm cũng có thể được giải quyết ở nhiều mức độ khác nhau.

- **Mức độ tài liệu:** Nhiệm vụ ở mức độ tài liệu là phân loại xem toàn bộ tài liệu ý kiến thể hiện một cảm xúc tích cực hay tiêu cực. Ví dụ, khi đưa ra đánh giá sản phẩm, hệ thống sẽ xác định xem đánh giá thể hiện ý kiến tổng thể tích cực hay tiêu cực về sản phẩm. Mức phân tích này mặc nhiên giả định rằng mỗi tài liệu thể hiện ý kiến về một thực thể duy nhất (một sản phẩm hoặc dịch vụ). Do đó, nó không áp dụng cho các tài liệu đánh giá hoặc so sánh nhiều thực thể.
- **Mức độ câu:** Mức độ tiếp theo là xác định xem mỗi câu thể hiện ý kiến tích cực, tiêu cực hay trung lập. Mức độ phân tích này liên quan chặt chẽ đến phân lớp chủ quan và khách quan, trong đó phân biệt các câu diễn đạt thông tin thực tế (gọi là câu khách quan) với các câu thể hiện quan điểm và ý kiến chủ quan (gọi

là câu chủ quan). Tuy nhiên, tính chủ quan không tương đương với cảm xúc hoặc quan điểm bởi vì nhiều câu khách quan có thể ám chỉ cảm xúc hoặc quan điểm. Ví dụ: "Chúng tôi vừa mua chiếc xe vào tháng trước và cần gạt nước đã vỡ". Ngược lại, nhiều câu chủ quan có thể không thể hiện bất kỳ ý kiến hay cảm xúc nào, ví dụ: "Tôi nghĩ anh ấy về nhà sau bữa trưa".

- **Mức độ khía cạnh:** Các phân tích mức độ tài liệu và mức độ câu không khám phá chính xác những gì mọi người thích và không thích. Nói cách khác, họ không cho biết mỗi quan điểm nói về cái gì, nghĩa là mục tiêu của quan điểm. Ví dụ, nếu chúng ta chỉ biết rằng câu "Tôi thích iPhone 11" là tích cực thì nó bị hạn chế sử dụng trừ khi chúng ta biết rằng ý kiến tích cực là về iPhone 11. Có thể nói rằng, nếu chúng ta có thể phân loại một câu thành tích cực, mọi thứ trong câu có thể lấy quan điểm tích cực. Tuy nhiên, điều đó cũng không hiệu quả, bởi vì một câu có thể có nhiều quan điểm, ví dụ, "Apple đang làm rất tốt trong nền kinh tế nghèo nàn này". Không có ý nghĩa gì khi phân loại câu này là tích cực hay tiêu cực bởi vì nó tích cực về Apple nhưng tiêu cực về nền kinh tế. Để có được mức độ kết quả tốt này, chúng ta cần phải đi đến cấp độ khía cạnh. Thay vì nhìn vào các đơn vị ngôn ngữ (tài liệu, đoạn văn, câu, mệnh đề hoặc cụm từ), phân tích mức độ khía cạnh trực tiếp nhìn vào ý kiến và mục tiêu của nó (được gọi là mục tiêu của quan điểm). Tầm quan trọng của các mục tiêu của quan điểm giúp hiểu rõ hơn vấn đề phân tích quan điểm. Một ví dụ khác: "Mặc dù dịch vụ không tuyệt vời, tôi vẫn thích nhà hàng này". Câu này rõ ràng có một phần tích cực, nhưng chúng ta không thể nói rằng câu này là hoàn toàn tích cực.Ta chỉ có thể nói rằng câu đó là tích cực về nhà hàng (nhấn mạnh), nhưng nó vẫn tiêu cực về dịch vụ của nó (không nhấn mạnh). Nếu ai đó đọc quan điểm này quan tâm rất nhiều về dịch vụ, anh ta có thể sẽ không đi ăn ở nhà hàng. Trong các ứng dụng, mục tiêu quan điểm (ví dụ: nhà hàng và dịch vụ trong câu trước) thường được mô tả bởi các thực thể (ví dụ: nhà hàng) và các khía cạnh khác nhau của chúng (ví dụ: dịch vụ của nhà hàng). Do đó, mục tiêu của cấp độ phân tích này là khám phá quan về các thực thể và các khía cạnh của chúng. Trong trường hợp người dùng chỉ quan tâm

đến quan điểm về các thực thể, hệ thống có thể bỏ qua các khía cạnh của nó. Phân tích cấp độ khía cạnh là những gì cần thiết trong các ứng dụng và hầu hết tất cả các hệ thống phân tích quan điểm thực tế trong công nghiệp đều dựa trên mức độ phân tích này.

1.2. Giới thiệu về học sâu

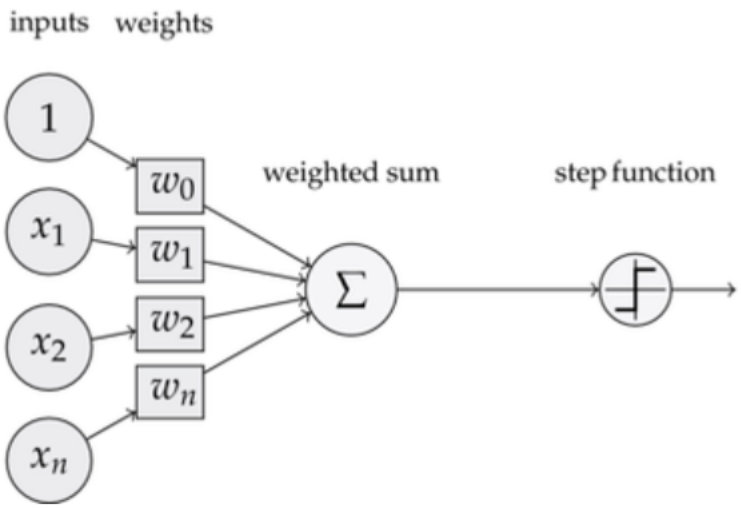
Học sâu (deep learning) là một nhánh của học máy (machine learning), sử dụng các mạng nơron - mô hình học lấy ý tưởng từ hệ thống kết nối các tế bào thần kinh trong bộ não người để xây dựng hệ thống học máy. Học sâu vượt trội hơn so với học máy truyền thống trong các vấn đề phức tạp như nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, phân loại hình ảnh...

Học sâu cho phép các mô hình tính toán gồm nhiều tầng xử lý để học biểu diễn dữ liệu với nhiều mức trừu tượng khác nhau. Học sâu có nhiều kiến trúc biến thể khác như: mạng nơron sâu (deep neural network), mạng niềm tin sâu (deep belief network), mạng nơron tích chập (convolutional neural network – CNN), mạng niềm tin sâu tích chập (convolutional deep belief network), mạng nơron lưu trữ và truy xuất bộ nhớ lớn (large memory storage and retrieval neural network), các máy Deep Boltzmann,…. Các mô hình học sâu có thể đạt được mức độ chính xác cao, đôi khi vượt quá hiệu suất ở cấp độ con người và thường được đào tạo bằng cách sử dụng một tập hợp lớn dữ liệu có nhãn và kiến trúc mạng nơron chứa nhiều lớp.

1.2.1. Mạng nơron nhân tạo

Mạng nơ ron là một hệ thống tính toán lấy cảm hứng từ sự hoạt động của các nơ ron sinh học trong hệ thần kinh. Một mạng nơ ron được cấu thành bởi các nơ ron đơn lẻ được gọi là các perceptron.

a. Perceptron



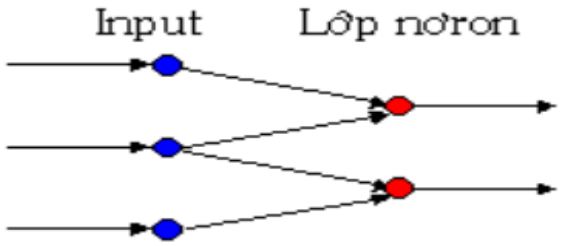
- Đầu vào : Tất cả các tính năng trở thành đầu vào cho một perseptron.Biểu thị đầu vào bằng [x1, x2, x3, .., xn], trong đó x đại diện cho giá trị tính năng và n đại diện cho tổng số tính năng.Có loại đầu vào đặc biệt gọi là sai lệch. Trong ảnh, mô tả giá trị của BIAS là w0.
- Trọng lượng : Các giá trị được tính toán theo thời gian đào tạo mô hình. Bắt đầu giá trị của trọng số với một số giá trị ban đầu và các giá trị này được cập nhật cho mỗi lỗi đào tạo. Các trọng số cho perceptron theo [w1, w2, w3, .. wn].
- Xu hướng : Một nơ ron thiên vị cho phép một bộ phân loại thay đổi ranh giới quyết định sang trái hoặc phải. Theo thuật ngữ đại số, nơ ron thiên vị cho phép một bộ phân loại dịch ranh giới quyết định của nó. Nó nhằm mục đích "di chuyển mọi điểm một khoảng cách không đổi theo một hướng xác định." Xu hướng giúp đào tạo mô hình nhanh hơn và với chất lượng tốt hơn.
- Tổng cộng có trọng số : Tổng cộng có trọng số là tổng của các giá trị mà nhận được sau khi nhân của mỗi trọng số [wn] được liên kết với từng giá trị tính năng [xn].Tổng kết có trọng số bằng $\sum w_i x_i$ cho tất cả i -> [1 đến n].
- Chức năng bước / kích hoạt : Vai trò của chức năng kích hoạt là làm cho các mạng thần kinh trở nên phi tuyến. Đối với phân loại tuyến tính, ví dụ, nó trở nên cần thiết để làm cho perceptron càng tuyến tính càng tốt.

- Đầu ra : Tổng kết có trọng số được chuyển đến hàm bước / kích hoạt và bất kỳ giá trị nào chúng tôi nhận được sau khi tính toán là đầu ra dự đoán của chúng tôi.

Một số hàm kích hoạt thường dùng như: hàm sigmoid, tanH, ReLU, Softmax...

b. Kiến trúc mạng nơ ron nhân tạo

Mạng nơ ron là sự kết hợp của các tầng perceptron như hình vẽ bên dưới:



Hình 1.1: Mạng nơ ron gồm nhiều perceptron

Một mạng nơ ron nhân tạo sẽ có 1 lớp đầu vào (input layer), 1 lớp đầu ra (output layer), có thể có hoặc không các lớp ẩn (hidden layer). Tất cả các nốt mạng (nơron) được kết hợp đôi một với nhau theo một chiều duy nhất từ tầng vào tới tầng ra, tức là mỗi nốt ở một tầng nào đó sẽ nhận đầu vào là tất cả các nốt ở tầng trước đó mà không cần suy luận ngược lại. Điều này gọi là truyền thẳng (feedforward).

c. Học với mạng nơ ron

Trong các mạng nơ ron, tương tự như các bài toán học máy khác, ta cần tìm một hàm mất mát (loss function) để đánh giá sai số giữa đầu ra mong muốn và đầu ra dự đoán. Để dự đoán chính xác, ta cần giảm thiểu sai số, tức là tối ưu hóa hàm mất mát. Với mạng nơ ron, ta sử dụng một giải thuật đặc biệt là lan truyền ngược. Phương pháp này tính đạo hàm của hàm mất mát, dựa trên quy tắc chuỗi đạo hàm của hàm hợp và phép tính ngược đạo hàm để thu được đạo hàm theo tất cả các tham số.

1.2.2. Mạng nơron tích chập

Mạng nơron tích chập (Convolutional Neural Network - CNN) được đề xuất bởi Yann LeCun và cộng sự vào năm 1998. Tên của mạng nơron tích chập được dựa trên phép tính quan trọng được sử dụng là toán tích chập. Trong mô hình mạng nơ ron, phép tích chập được thực hiện trên giá trị đầu vào của dữ liệu cùng ma trận nhân (kernel) để tạo ra một bản đồ đặc trưng (feature map). Công thức tích chập trên ma trận hai chiều như sau:

$$S(i,j)=(I\otimes K)(I,j)= \sum_m \sum_n \quad I(i+m,j+n)K(m,n)$$

Với I là ma trận đâu vào, K là ma trận nhân, $m\times n$ là kích thước của K.

Khi đó, trượt ma trận nhân theo dữ liệu đầu vào, tại mỗi vị trí, ta tiến hành phép nhân ma trận và tính tổng các giá trị để đưa vào bản đồ đặc trưng.

Mạng nơ ron tích chập gồm 2 phần:

- Phần trích rút đặc trưng: mạng sẽ tiến hành tính toán hàng loạt phép tích chập và hợp nhất (pooling) để phát hiện đặc trưng. Lớp hợp nhất (pooling layer) thường được dùng giữa các lớp tích chập để giảm kích thước dữ liệu nhưng vẫn giữ các thuộc tính quan trọng.
- Phần phân lớp: một lớp với các liên kết đầy đủ (fully connected layer) sẽ đóng vai trò như một bộ phân lớp các đặc trưng đã rút trích trước đó.

Trong mạng nơ ron tích chập, ta thực hiện phép tích chập trên đầu vào nhiều lần khác nhau, mỗi lần sử dụng ma trận nhân khác nhau. Kết quả, ta sẽ thu được những bản đồ đặc trưng khác nhau. Sau đó, ta kết hợp toàn bộ bản đồ đặc trưng này thành kết quả cuối cùng của tầng tích chập. Tương tự mạng nơ ron thông thường, ta sử dụng một hàm kích hoạt phi tuyến, ở đây ta sử dụng hàm ReLU (rectified linear unit). Trong quá trình trượt ma trận nhân trên dữ liệu đầu vào, ta sẽ quy định một bước nhảy (stride) với mỗi lần di chuyển, nếu bước nhảy tăng, ma trận nhân sẽ có ít ô trùng lặp. Vì kích thước đầu ra luôn nhỏ hơn đầu vào nên ta cần một phép xử lý đầu vào để đầu ra không bị co

dãn. Khi đó, ta chỉ cần thêm một lề nhỏ với giá trị 0 vào xung quanh đầu vào trước khi thực hiện phép tích chập. Sau mỗi tầng tích chập, ta cho kết quả qua một tầng hợp nhất để nhanh chóng giảm số chiều, giúp giảm thời gian học và hạn chế việc quá khớp.

Trong phần phân lớp, để xử lý kết quả của phần tích chập, ta sử dụng một vài tầng với các liên kết đầy đủ. Do đầu vào của tầng này là một chiều, ta cần làm phẳng đầu vào trước khi phân lớp. Tầng cuối cùng trong mạng nơron tích chập là một tầng với các liên kết đầy đủ hoạt động như mạng nơ ron thông thường. Kết quả thu được cuối cùng sẽ là một véc tơ với các giá trị xác suất cho việc dự đoán như mạng nơron thông thường.

1.2.3. Nhúng từ

Nhúng từ (word embedding) là tên gọi chung của các mô hình ngôn ngữ và các phương pháp học theo đặc trưng trong xử lý ngôn ngữ tự nhiên, ở đó một từ hoặc cụm từ được biểu diễn sang các vector số. Đây là công cụ đóng vai trò quan trọng đối với hầu hết các thuật toán, kiến trúc học máy, học sâu trong việc xử lý đầu vào dạng văn bản.

Nhúng từ được phân loại thành hai loại chính như sau:

a. Nhúng dựa trên tần suất

Các loại nhúng dựa trên tần suất (frequency-based embedding) dựa trên tần số xuất hiện của các từ để tạo ra các vector từ, trong đó có 3 loại phổ biến nhất:

- Count vector: là dạng đơn giản nhất, ở đó vector biểu diễn là một số nguyên có độ dài D, phần tử tại vị trí i là tần số xuất hiện của từ đó trong document d_i. tf-idf vector: quan tâm đến tần số xuất hiện của từ đó trong một document và cả toàn bộ tập dữ liệu D document, do đó có tính phân loại cao hơn count vector.
- Co-occurrence matrix: bảo tồn mối quan hệ ngữ nghĩa giữa các từ, được xây dựng dựa trên số lần xuất hiện của các cặp từ thay vì chỉ chú trọng đến tần số xuất hiện của một từ; thường là một ma trận vuông đối xứng, mỗi hàng, cột là véctơ biểu thị của từng từ tương ứng.

- GloVe (Global Vector) được xây dựng dựa trên co-occurrence matrix. Ý tưởng chính là độ tương tự ngữ nghĩa giữa hai từ *i, j* có thể được xác định thông qua độ tương tự ngữ nghĩa giữa chúng với từ *k*.

b. Nhúng dựa trên dự đoán

Các loại nhúng dựa trên dự đoán (prediction-based embedding) xây dựng các véctơ dựa vào mô hình dự đoán, tiêu biểu là Word2vec. Đây là sự kết hợp của 2 mô hình CBOW và Skip-gram. Cả 2 mô hình đều được xây dựng dựa trên mạng nơron gồm 3 tầng: 1 tầng vào, 1 tầng ẩn và 1 tầng ra với mục đích chính là học các trọng số biểu diễn véctơ từ.

- CBOW (Continois bag of words) dự đoán xác suất của một từ được đưa ra theo ngữ cảnh với đầu vào là một hoặc nhiều one-hot vector của các từ ngữ cảnh có chiều dài *V*, đầu ra là một véctơ xác suất cùng chiều dài *V* của từ liên quan hoặc còn thiếu, tầng ẩn có chiều dài *N*, *N* cũng là độ lớn của véctơ từ biểu thị.
- Skip-gram có cấu trúc tương tự CBOW nhưng mục đích là dự đoán ngữ cảnh của một từ đưa vào.

1.3. Phát biểu bài toán phân tích quan điểm của khóa luận

Khóa luận tập trung vào việc sử dụng phương pháp học sâu cho bài toán phân tích quan điểm mức khía cạnh. Trong bài toán phân tích quan điểm mức khía cạnh cần thực hiện hai bài toán con là phân lớp khía cạnh (aspect classification) và phân tích quan điểm (sentiment analysis). Cụ thể, phân lớp khía cạnh là bài toán phân lớp đa nhãn, cần trích xuất các khía cạnh có trong câu; sau đó thực hiện phân tích quan điểm (tích cực hay tiêu cực) cho quan điểm thể hiện trong câu, đây là bài toán phân lớp nhị phân.

- Khóa luận sẽ tiến hành thực hiện hai bài toán con này đồng thời bằng một mô hình sử dụng mạng nơron tích chập. Để thực nghiệm, tập dữ liệu tiếng Việt về nhận xét khách sạn đã được sử dụng để huấn luyện và đánh giá.
- Đầu vào: Tập dữ liệu tiếng Việt các bình luận của khách hàng.

- Đầu ra: Bình luận gồm những khía cạnh nào và nhãn cảm xúc của từng khía cạnh.

Tổng kết chương 1

Ở chương này, khóa luận đã giới thiệu các khái niệm cơ bản, phương pháp học sâu và trình bày về bài toán phân tích quan điểm của khóa luận. Trong chương tiếp theo, khóa luận sẽ khảo sát một số hướng tiếp cận trong giải quyết bài toán phân tích quan điểm.

CHƯƠNG 2. KỸ THUẬT HỌC SÂU CHO PHÂN TÍCH QUAN ĐIỂM

Ở chương này, khóa luận sẽ giới thiệu một số hướng tiếp cận cho bài toán phân tích quan điểm, đồng thời trình bày hướng tiếp cận bài toán phân tích quan điểm dựa trên học sâu và một số mô hình sử dụng mạng nơ ron tích chập trong bài toán phân tích quan điểm mức khía cạnh.

2.1. Một số hướng tiếp cận cho phân tích quan điểm

2.1.1. Tiếp cận dựa trên từ vựng

Dễ dàng nhận thấy rằng các từ và cụm từ truyền đạt quan điểm tích cực hoặc tiêu cực là công cụ để phân tích tình cảm. Các từ mang quan điểm tích cực như "đẹp, tuyệt vời" được sử dụng để thể hiện một số trạng thái mong muốn trong khi các từ mang tiêu cực như "xấu, khủng khiếp" được sử dụng để diễn tả một số trạng thái không mong muốn. Như vậy, từ vựng được coi như một chỉ số quan trọng cho quan điểm, được gọi là từ mang quan điểm. Cách tiếp cận này phụ thuộc vào quan điểm của từ vựng, được chia thành tiếp cận dựa trên từ điển (dictionary-based approach) và kho văn bản (corpus based approach).

Tiếp cận dựa trên từ điển được dựa trên cấu trúc từ đồng nghĩa và trái nghĩa của từ điển, trong khi với tiếp cận dựa trên văn bản, có hai ý tưởng chính được áp dụng để xác định các từ mang quan điểm trong một kho văn bản. Thứ nhất, đó là khai thác một số quy tắc hoặc quy ước ngôn ngữ trên các từ nối để xác định đồng thời các từ mang quan điểm và thể hiện của chúng. Thứ hai, sử dụng các quan hệ cú pháp của các quan điểm và mục tiêu để trích xuất các từ mang quan điểm. Tuy nhiên, chỉ sử dụng cách tiếp cận dựa trên kho văn bản không hiệu quả như cách tiếp cận dựa trên từ điển vì khó có thể chuẩn bị một khối lượng lớn để bao gồm tất cả các từ ngữ, nhưng cách tiếp cận này có một lợi thế, có thể giúp tìm các từ mang quan điểm cụ thể theo ngữ cảnh và định hướng của người viết bằng cách sử dụng một miền kho văn bản.

2.1.2. Tiếp cận dựa trên học máy

Các hướng tiếp cận học máy dựa vào các thuật toán học máy nổi tiếng để giải quyết phân tích quan điểm là một vấn đề phân lớp văn bản thông thường sử dụng các tính năng cú pháp và ngôn ngữ. Các phương pháp học máy bao gồm phân cụm cho học không giám sát và phân loại cây quyết định, phân loại tuyến tính, phân loại dựa trên luật, phân loại xác suất cho học có giám sát; trong đó có học sâu.

Có nhiều loại phân loại tuyến tính, trong số đó là máy vectơ hỗ trợ (Support Vector Machines - SVM) là một dạng phân loại cố gắng xác định các phân tách tuyến tính tốt giữa các lớp khác nhau. Nguyên tắc chính của các SVM là xác định các dấu phân cách tuyến tính trong không gian tìm kiếm có thể phân tách tốt nhất các lớp khác nhau. SVM được sử dụng trong nhiều ứng dụng, trong số đó là phân loại đánh giá theo chất lượng.

Một phương pháp khác là sử dụng mạng nơron nhân tạo. Mạng nơron bao gồm nhiều nơron trong đó nơron là đơn vị cơ bản của nó. Các đầu vào của các nơ-ron được biểu thị bằng vector tuyến tinh $\vec{X}_t$ là tần số trong tài liệu thứ i.Có một tập hợp các trọng số A được liên kết với mỗi nơron được sử dụng để tính toán một hàm của các đầu vào f(•). Hàm tuyến tính của mạng nơ ron là: p =A. $\vec{X}_t$. Trong bài toán phân loại nhị phân người ta cho rằng nhãn của lớp $\vec{X}_t$ được ký hiệu là y_i và dấu hiệu của hàm dự đoán p mang lại nhãn của lớp. Mạng nơron đa lớp được sử dụng cho các ranh giới phi tuyến tính. Nhiều lớp này được sử dụng để tạo ra nhiều ranh giới tuyến tính từng phần, được sử dụng để xấp xỉ các khu vực kín thuộc về một lớp cụ thể. Đầu ra của các nơron ở các lớp trước đưa vào các nơron ở các lớp sau. Quá trình đào tạo phức tạp hơn vì các lỗi cần phải được truyền lại trên các lớp khác nhau.

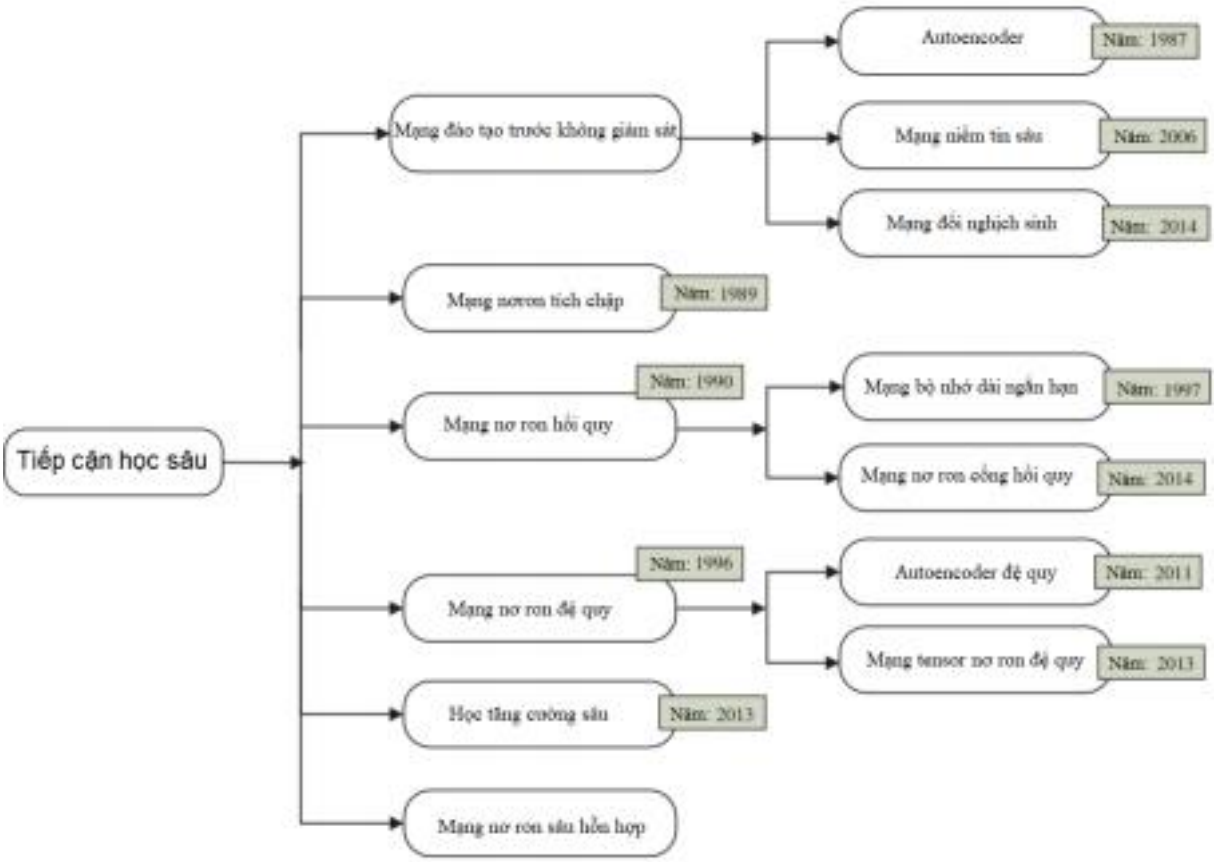
Phân loại xác suất sử dụng mô hình hỗn hợp để phân loại. Mô hình hỗn hợp giả định rằng mỗi lớp là một thành phần của hỗn hợp. Mỗi thành phần hỗn hợp là một mô hình tổng quát cung cấp xác suất lấy mẫu một kỳ hạn cụ thể cho thành phần đó. Phân loại Naïve Bayes là trình phân loại đơn giản nhất và được sử dụng phổ biến nhất. Mô hình phân loại của Naïve Bayes tính xác suất sau của một lớp, dựa trên sự phân phối các từ

trong tài liệu. Mô hình hoạt động với trích xuất tính năng BOW (bag of words) mà bỏ qua vị trí của từ trong tài liệu. Nó sử dụng Định lý Bayes để dự đoán xác suất một bộ đặc trưng nhất định thuộc về một nhãn cụ thể.

2.2. Học sâu cho phân tích quan điểm

Trong các hướng tiếp cận nêu trên với bài toán phân tích quan điểm, một nhược điểm của các phương pháp phân tích quan điểm dựa trên từ điển là tốn công sức, hạn chế về độ phủ và khả năng mở rộng. Với các phương pháp học máy dựa trên đặc trưng thông thường như SVM, phân loại cây quyết định... thì tốn công sức thiết kế tập đặc trưng phù hợp với dữ liệu. Học sâu có ưu điểm là cung cấp các cách học biểu diễn dữ liệu theo cách được giám sát và không giám sát với sự trợ giúp của hệ thống phân cấp các lớp, cho phép xử lý nhiều lần. Trước hết, việc áp dụng các phương pháp tiếp cận học sâu trong phân tích quan điểm đã được thúc đẩy bởi khả năng học tính năng tự động của chúng, nơi chúng có thể học tự động và khám phá các biểu diễn đầu vào từ chính dữ liệu. Hơn nữa, việc áp dụng chúng đã được thúc đẩy bởi sự gia tăng của dữ liệu huấn luyện với phân loại đa lớp và sự thành công của nhúng từ. Bên cạnh đó, sự sẵn có của các tài nguyên điện toán mạnh mẽ như bộ xử lý đồ họa (GPU) cho phép thao tác ma trận hiệu quả cũng trở thành động lực để nắm lấy các phương pháp học sâu.

Gần đây, các nhà nghiên cứu đã đề xuất một số lượng lớn các phương pháp học sâu, và hầu hết trong số chúng đã được áp dụng để phân tích quan điểm. Những phương pháp này đã được chứng minh là phương pháp hiệu quả trong phân tích quan điểm, được chứng minh bằng nhiều nghiên cứu đã được thực hiện thành công. Họ đã giải quyết các vấn đề phức tạp như thích ứng miền, có thể xử lý bối cảnh mà từ đó xuất hiện và mô hình hóa các phụ thuộc tầm xa có thể thay đổi tính phân cực của một nhận định trong một câu nhất định.



Hình 2.1: Các hướng tiếp cận học sâu cho phân tích quan điểm

Phân tích quan điểm dựa trên học sâu chia thành sáu loại: mạng đào tạo trước không giám sát (unsupervised pre-trained networks), mạng nơron tích chập, mạng nơ ron hồi quy (Recurrent neural networks - RNN), mạng nơ ron đệ quy (Recursivneural networks), học tăng cường sâu (Deep reinforcement learning), mạng nơron sâu hỗn hợp (hybrid deep neural network) [5]. Các mạng mạng đào tạo trước không giám sát bao gồm autoencoder, mạng niềm tin sâu và mạng đối nghịch sinh. Mạng nơron hồi quy cũng có các biến thể chính như mạng bộ nhớ dài ngắn hạn (long short-term memory network) và mạng nơron cổng hồi quy (gated recurrent neural network), trong khi mạng nơron đệ quy cũng có các biến thể như autoencoder đệ quy, mạng tensor nơron đệ quy.

Học sâu nói chung và mạng nơ ron tích chập nói riêng đã được ứng dụng thành công trong nhiều bài toán phân tích quan điểm. Lý do bởi mô hình CNN giả định rằng các từ khóa có thể chứa thuật ngữ khía cạnh và chỉ ra một danh mục hoặc xác định độ phân cực,

bất kể vị trí của chúng. CNN có khả năng học cách tìm các tính năng đó, do đó, có thể trích xuất các mẫu cục bộ từ dữ liệu bất kể vị trí của chúng. Điều này rất hữu ích để xác định các cụm từ có độ dài cố định. Một lợi thế khác là CNN là một mô hình phi tuyến tính, được kỳ vọng sẽ phù hợp với dữ liệu hơn các mô hình tuyến tính như CRF và không yêu cầu các tính năng thủ công mở rộng như quy tắc ngôn ngữ cố định.

Dưới đây, khóa luận sẽ trình bày một số mô hình mạng nơron tích chập nổi bật áp dụng trong phân tích quan điểm mức khía cạnh.

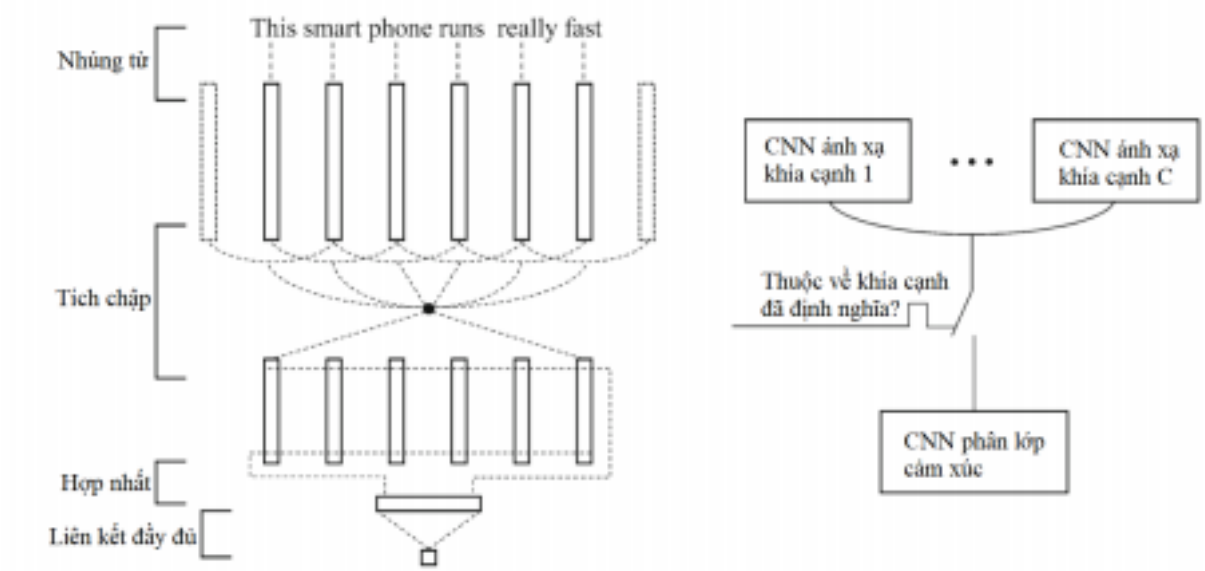
2.3. Một số mô hình mạng nơ ron tích chập trong phân tích quan điểm mức khía cạnh

2.3.1. Phân lớp quan điểm mức khía cạnh dựa trên mạng nơ ron tích chập xếp tầng

Đối với nhiệm vụ giải quyết hai bài toán phân lớp khía cạnh và phân tích quan điểm trong phân tích quan điểm mức khía cạnh, Gu và cộng sự (2017)đề xuất một mô hình xếp tầng với hai cấp độ CNN bao gồm các trình ánh xạ khía cạnh CNN và phân loại cảm xúc CNN với tên gọi Mạng nơ ron tích chập xếp tầng (Cascaded CNN). CNN ánh xạ khía cạnh và CNN phân loại cảm xúc được tổ chức theo cách xếp tầng. Mỗi trình ánh xạ xác định xem câu đầu vào có thuộc về khía cạnh tương ứng của nó hay không. Nếu có, phân loại cảm xúc dự đoán phân cực cảm xúc là tích cực hoặc tiêu cực.

Mạng CNN xếp tầng giải quyết hai cân nhắc về mạng xếp tầng. Thứ nhất, mạng chỉ chứa một phân lớp cảm xúc. Mọi người có thể nghĩ rằng nó có vấn đề khi chỉ câu (ví dụ: “This smart phone runs fast, but loses its charge too quickly!” (Điện thoại này chạy nhanh, nhưng nhanh hết pin!)” có thể chứa các khía cạnh khác nhau và cảm xúc đối với các khía cạnh này có thể trái ngược nhau. Mô hình này không đào tạo một trình phân lớp cảm xúc riêng biệt cho từng loại khía cạnh vì trong thực tế chỉ có một vài câu ngụ ý tình cảm trái ngược cho các khía cạnh khác nhau. Thứ hai, trình phân loại cảm xúc chỉ xử lý các câu thuộc ít nhất một loại khía cạnh được xác định trước vì các ứng dụng thực tế chỉ quan tâm đến cảm xúc của các câu liên quan đến khía cạnh. Ngoài ra, các câu không thuộc bất

kỳ khía cạnh nào được xác định trước có thể là câu khách quan. Nó không phù hợp để phân loại cảm xúc của câu khách quan là tích cực hay tiêu cực.



Hình 2.2: Mạng nơ ron tích chập xếp tầng cho ánh xạ khía cạnh và phân lớp cảm xúc

Mỗi CNN chứa một lớp nhúng từ, một lớp tích chập và hợp nhất, một lớp được kết nối đầy đủ.

- Lớp nhúng từ: Lớp này mã hóa từng từ trong câu đầu vào dưới dạng vectơ từ. Đặt độ dài câu, $|D| \in R$ là kích thước từ vựng và $W^{(1)} \in R^{k \times |D|}$ là ma trận nhúng của vectơ từ k chiều. Từ thứ i trong câu được chuyển thành vectơ k chiều bởi kết quả vectơ ma trận:

$$w_i = W^{(1)} x_i$$

Trong đó x_i đại diện cho x_i one-hot cho từ thứ i.

- Lớp tích chập: Sau khi mã hóa câu đầu vào bằng các vectơ từ, các phép toán tích chập được áp dụng trên đầu các vectơ này để tạo ra các đặc trưng mới. Phép toán tích chập liên quan đến bộ lọc $u \in R^{hk}$ được áp dụng cho cửa sổ $h = 2r + 1$ từ. Ví dụ: một tính năng f_i được tạo ra từ một cửa sổ các từ $w_{i-r:\, i+r}$ bởi $f_i = g\,(w_{i-r:\, i+r} \cdot u)$.

Ở đây g biểu thị một hàm kích hoạt phi tuyến. Bộ lọc này được áp dụng cho mọi cửa sổ có thể của câu đầu vào để tạo bản đồ đặc trưng.

$$f = [f_1, f_2, ..., f_l]$$

Phần trên mô tả quá trình một bản đồ đặc trưng được trích xuất từ một bộ lọc. Mạng sử dụng các bộ lọc m_i $(i = 1, 2, .., C)$ để tạo các bản đồ đặc trưng m_i cho bộ ánh xạ khía cạnh thứ i và các bộ lọc m_{C+1} cho trình phân loại cảm xúc. Trọng số bộ lọc cho trình ánh xạ khía cạnh thứ i được lưu trữ trong ma trận $hk \times m_i$ - chiều $W^{(2)} \in \mathbb{R}^{hk \cdot mi}$Đối với phân loại cảm xúc $W^{(2)} \in \mathbb{R}^{hk \cdot m2}$

- Lớp hợp nhất: Lớp này áp dụng hợp nhất tối đa theo thời gian (max-over-time pooling) cho mỗi bản đồ đặc trưng được tạo bởi các lớp tích chập:

$$f=\max(f1,f2,\ldots,fi)$$

Hợp nhất tối đa theo thời gian lấy yếu tố tối đa trong mỗi bản đồ đặc trưng và xử lý một cách tự nhiên với độ dài câu thay đổi. Nó tạo ra một vectơ đặc trưng có kích thước cố định $v_i \in \mathbb{R}^{m_i}$ cho tác vụ thứ i.

- Lớp kết nối đầy đủ: Các vectơ đặc trưng có kích thước cố định được tạo bởi các lớp hợp nhất được đưa vào các lớp kết nối đầy đủ. Cụ thể, v_i được chuyển đến một phân loại hồi quy logistic nhị phân.

$$a_i = \frac{1}{1+e^{-W_i^{(3)} v_i}},\ i = 1, 2, ..., C + 1$$

Ở đây $W^{(3)} \in \mathbb{R}^{n \cdot mi}$ là ma trận trọng số cho tác vụ thứ i và ai là vectơ đầu ra khía cạnh. Đối với trình ánh xạ khía cạnh, a_i $(i = 1, 2, .., C)$ là xác suất của câu đầu vào thuộc loại khía cạnh thứ i; đối với phân loại cảm xúc a_i $(i = C + 1)$ là xác suất cảm xúc tích cực.

2.3.2. Phân tích quan điểm mức khía cạnh dựa trên tích hợp hai mạng nơron

Toh và Su (2016) đã đạt được hiệu suất tốt nhất trong SemEval 2016 trong phát hiện danh mục khía cạnh với việc sử dụng hai mô hình học máy khác nhau. Nhóm tác giả coi phát hiện danh mục khía cạnh là một vấn đề phân loại nhiều lớp thành một cách tiếp cận nhị phân phù hợp. Đặc biệt, họ đã sử dụng nhiều phân loại nhị phân được đào tạo trên một mạng nơron truyền thẳng một lớp, sau đó kết hợp đầu ra xác suất từ một mạng nơ ron tích chập sâu để dự đoán xem văn bản có bao gồm một loại khía cạnh hay không.

Nhóm tác giả sử dụng các đặc trưng bổ sung từ hệ thống học sâu được xây dựng dựa trên mạng nơ ron tích chập sâu. Một ma trận câu $S \in R^{|s| \times d}$ được xây dựng cho mỗi câu đầu vào s, trong đó mỗi hàng i là một đại diện vectơ của từ i trong câu. Độ dài câu $|s|$ được cố định với độ dài câu tối đa của tập dữ liệu để tất cả các ma trận câu có cùng kích thước. Các câu ngắn hơn được đệm với các vectơ hàng 0 tương ứng. Mỗi vectơ hàng của ma trận câu được tạo thành từ các cột tương ứng với các đặc trưng đầu vào khác nhau được nối với nhau. Ma trận câu đầu vào S sau đó được chuyển qua một loạt các phép biến đổi lớp mạng trong hệ thống học sâu.

Mô hình tính toán tích chập giữa ma trận câu đầu vào S và ma trận lọc $F \in R^{m \times d}$ của cửa sổ ngữ cảnh kích thước m, dẫn đến một vectơ cột $c \in R^{|s|}$. Ma trận bộ lọc F sẽ trượt (với bước nhảy là 1) dọc theo kích thước hàng của S, tạo ra một giá trị cho mỗi từ trong câu. Thay vì một ma trận bộ lọc đơn, ma trận bộ lọc n được áp dụng cho ma trận câu S, dẫn đến ma trận tính năng tích chập $C \in R^{|s| \times n}$. Để tìm hiểu ranh giới quyết định phi tuyến, mỗi phần tử của C đều đi qua hàm kích hoạt tiếp tuyến hyperbol tanh. Ma trận đầu ra C sau đó được chuyển đến lớp lớp hợp nhất tối đa. Lớp này sẽ trả về giá trị tối đa của mỗi cột. Một lớp ẩn dày đặc với h đơn vị ẩn được áp dụng cho đầu ra của lớp gộp, sử dụng đơn vị tinh chỉnh tuyến tính (ReLU) làm chức năng kích hoạt. Một lớp softmax nhận đầu ra của lớp ẩn dày đặc trước đó và tính toán phân phối xác suất theo các danh mục có thể. Mô hình bao gồm một danh mục bổ sung "NIL" trực cho trường hợp câu không chứa danh mục khía cạnh. Vì một câu có thể chứa nhiều hơn một danh mục, mô hình xuất ra các danh mục có giá trị xác suất đầu ra lớn hơn ngưỡng t.

Nhóm tác giả coi việc trích xuất mục tiêu ý kiến là một nhiệm vụ ghi nhãn tuần tự. Các trình phân loại ghi nhãn tuần tự được đào tạo bằng trường ngẫu nhiên có điều kiện (Conditional Random Fields - CRF). Để tăng cường hệ thống CRF hiện tại, đầu ra của mạng nơ ron hồi quy hai chiều (Bidirectional Recurrent Neural Networks) được sử dụng làm các đặc trưng bổ sung. Một mô hình như vậy cho phép các phụ thuộc tầm xa từ tương lai cũng như từ quá khứ được nắm bắt, có lợi cho các nhiệm vụ ghi nhãn tuần tự. Lớp cuối cùng của mô hình là lớp softmax được kết nối đầy đủ để cho phép mô hình xuất ra xác suất.

2.3.3. Phân tích quan điểm mức khía cạnh dựa trên mạng nơ ron tích chập đa nhiệm

Ruder và cộng sự (2016) đề xuất một cách tiếp cận mạng nơron tích chập đa nhiệm để thực hiện cả hai bài toán phân lớp khía cạnh và phân tích quan điểm. Mô hình này coi trích xuất khía cạnh là một vấn đề phân loại đa nhãn nhưng tiếp cận vấn đề này thông qua ngưỡng phân phối xác suất.

Mô hình lấy làm đầu vào một văn bản, được đệm theo chiều dài n. Văn bản được biểu diễn dưới dạng nối của từ nhúng $x_{1:n}$ trong đó $x_1 \in \mathbb{R}^k$ là vectơ k chiều của từ thứ i trong văn bản.

Lớp tích chập trượt các bộ lọc có kích thước cửa sổ khác nhau trên các nhúng đầu vào. Mỗi bộ lọc có trọng số $w \in \mathbb{R}^{hk}$ tạo ra một đặc trưng mới c_i cho một cửa sổ của các từ h theo thao tác sau:

$$c_i = f(w \cdot x_{i:i+h-1} + b$$

Lưu ý rằng bias $b \in \mathbb{R}$ và f là hàm phi tuyến ReLU. Việc áp dụng bộ lọc qua từng cửa sổ có thể của từ h hoặc ký tự trong câu tạo ra bản đồ đặc trưng sau:

$$c = [c_1, c_2, \ldots, c_{n-h+1}]$$

Hợp nhất tối đa theo thời gian sẽ ngưng tụ vectơ tính năng này thành đặc trưng quan trọng nhất của nó bằng cách lấy giá trị tối đa của nó và xử lý một cách tự nhiên với độ dài đầu vào thay đổi.

Một lớp softmax cuối cùng lấy các giá trị tối đa của các bản đồ đặc trưng được tạo bởi tất cả các bộ lọc và đưa ra phân phối xác suất trên tất cả các lớp đầu ra.

Để trích xuất các khía cạnh, mô hình đã trích xuất khía cạnh như là một vấn đề phân loại đa nhãn và huấn luyện một mạng nơ ron tích chập để phân phối xác suất đầu ra trên các khía cạnh, giảm thiểu tổn thất chéo. Để mô hình hóa đầu ra đa nhãn dưới dạng phân phối xác suất, mô hình xác định xác suất *p* của một khía cạnh với một câu *s* là p(a|s)=1/n nếu *a* xuất hiện trong *s* và *s* chứa *n* khía cạnh, mặt khác p(a|s) = 0. Chúng tôi xác định ngưỡng *f* và loại bỏ tất cả các khía cạnh với p(a|s) < f. Sau khi huấn luyện, lựa chọn *f* tối đa hóa điểm F1 trên bộ xác nhận.

Ruder và cộng sự quan sát rằng các phân phối khía cạnh khác nhau đáng kể tùy thuộc vào tên miền và ngôn ngữ. Chẳng hạn, miền máy tính xách tay tiếng Anh chứa 82 khía cạnh, trong khi miền nhà hàng chỉ chứa 13 khía cạnh. Do đó, trong mọi miền, nhóm tác giả thay thế tất cả các khía cạnh xảy ra dưới 5 lần bằng một khía cạnh "OTHER" và thêm một khía cạnh "NONE" vào mỗi câu không chứa khía cạnh nào để cho phép CNN không đưa ra dự đoán về khía cạnh. Trong quá trình suy luận, mỗi khi mô hình dự đoán "OTHER", khía cạnh thường xuyên nhất được thay thế bởi "OTHER" cho mỗi tên miền là đầu ra thay thế. Cuối cùng, mô hình loại bỏ tất cả các khía cạnh được dự đoán là "NONE".

Để phân tích quan điểm dựa trên khía cạnh, vectơ khía cạnh được cung cấp cùng với các từ nhúng của câu đầu vào vào CNN. Để có được vectơ khía cạnh, khía cạnh được chia từng thành các mã thành phần (token) cấu thành nó, sau đó nhúng các mã thông báo của tất cả các khía cạnh trong một không gian nhúng. Sau đó, chúng tôi tìm kiếm sự nhúng của từng mã thông báo và lấy trung bình để lấy vectơ khía cạnh. Theo cách này,

mô hình học các khía cạnh chia sẻ cùng một thực thể. Qua nhiều thử nghiệm, nhóm tác giả thấy rằng việc ghép từng vectơ từ với vectơ khía cạnh trước khi tích chập mang lại kết quả tốt nhất.

Hoạt động của mô hình như sau: Đầu tiên mô hình đệm câu đầu vào, sau đó tìm kiếm các từ nhúng của từng từ đầu vào. Mô hình tạo ra vectơ khía cạnh bằng cách tìm kiếm các phần nhúng trong không gian nhúng khía cạnh và tính trung bình cả hai phần nhúng. Kết quả của vectơ khía cạnh sau đó được nối với mỗi vectơ từ, rồi được nối để tạo ra ma trận câu 100x600. Tích chập, hợp nhất tối đa và softmax được áp dụng cho ma trận này như được mô tả ở trên. Nhóm tác giả nhận thấy chỉ cần sử dụng không gian nhúng chiều thấp để nhúng các khía cạnh sẽ đạt được kết quả vượt trội trong một số trường hợp khi không sử dụng nhúng từ được đào tạo trước.

2.4. Ý tưởng áp dụng cho bài toán phân tích quan điểm của khóa luận

Các mô hình mạng nơron tích chập áp dụng trong phân tích quan điểm mức khía cạnh được đề cập ở trên cho thấy ba cách tiếp cận giải quyết bài toán phân lớp khía cạnh và phân tích quan điểm trong phân tích quan điểm mức khía cạnh. Thứ nhất, thực hiện lần lượt bài toán phân lớp quan điểm sau đó đến phân lớp quan điểm (mô hìnhpipeline). Thứ hai, thực hiện hai mô hình cho hai bài toán rồi kết hợp lại. Thứ ba, thực hiện một mô hình kết hợp cho cả hai bài toán.

Qua ý tưởng sử dụng mô hình mạng nơ ron tích chập đa nhiệm để thực hiện cả hai bài toán phân lớp khía cạnh và phân tích quan điểm do Ruder và cộng sự đề xuất như đã được trình bày ở Mục 2.5, khóa luận đề xuất một cách tiếp cận tương tự với việc xây dựng một mô hình thực hiện đồng thời hai bài toán trên trong phân tích quan điểm mức khía cạnh. Cụ thể, khóa luận sử dụng mô hình mạng nơron tích chập cho bài toán phân tích quan điểm mức khía cạnh, ở đó quy bài toán phân lớp khía cạnh là một vấn đề phân lớp đa nhãn thành nhiều bộ phân lớp đơn nhãn.

Tổng kết chương 2

Ở chương này, khóa luận đã giới thiệu các hướng tiếp cận và các kỹ thuật học sâu sử dụng trong phân tích quan điểm. Khóa luận cũng đã khảo sát một số mô hình mạng nơron tích chập cho bài toán phân tích quan điểm mức khía cạnh đồng thời đưa ra phương pháp áp dụng cho bài toán khóa luận. Tiếp theo, chương 3 sẽ trình bày mô hình học sâu giải quyết bài toán phân tích quan điểm của khóa luận.

CHƯƠNG 3. MÔ HÌNH HỌC SÂU CHO PHÂN TÍCH QUAN ĐIỂM

Với ý tưởng tiếp cận đã trình bày ở chương 2, trong chương này, khóa luận trình bày mô hình mô hình học sâu dựa trên mạng nơron tích chập và áp dụng trong bài toán toán phân tích quan điểm mức khía cạnh.

3.1. Mô hình đề xuất giải quyết bài toán học sâu cho phân tích quan điểm

Hình 3.1 Mô hình đề xuất giải quyết bài toán khóa luận.

Phát biểu bài toán: Đề xuất mô hình học sâu cho phân tích quan điểm kết hợp hai bài toán phân lớp khía cạnh và phân tích quan điểm trong phân tích quan điểm mức khía cạnh sử dụng mạng nơron tích chập.

- Đầu vào: Tập dữ liệu tiếng Việt các bình luận
- Đầu ra: Bình luận gồm những khía cạnh nào và nhãn cảm xúc của .

Ví dụ

- Các khía cạnh:
- Quan điểm của khía cạnh:

3.2. Tiền xử lý dữ liệu

Dữ liệu cần được tiền xử lý trước khi huấn luyện. Dữ liệu được đưa qua công cụ vi_spacy để tiến hành tách từ, loại bỏ từ dừng, chuẩn hóa và gán nhãn. Vi_spacy bao gồm các mô hình tiếng Việt cho spaCy – công cụ được phát triển nhằm mục đích hỗ trợ các bài toán xử lý ngôn ngữ tự nhiên.

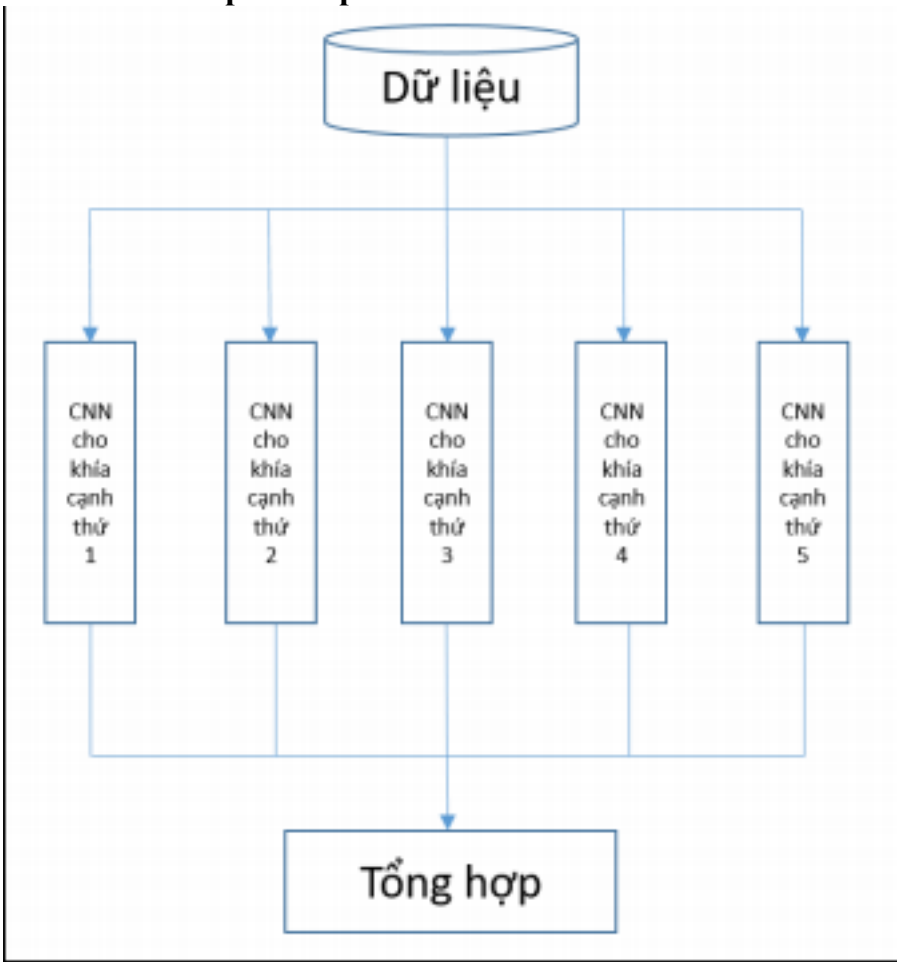
3.3. Biểu diễn dữ liệu

Tiếp theo, khóa luận sử dụng nhúng từ để biểu diễn các từ thành dạng vectơ. Ở đây, khóa luận sử dụng mô hình fastText đã huấn luyện trước trên kho dữ liệu của Wikipedia tiếng Việt, mỗi từ là một vec tơ 300 chiều. Khóa luận trích xuất toàn bộ từ có trong tập dữ liệu, sau đó so sánh với ma trận nhúng từ tiếng Việt của mô hình và lưu lại thành ma trận nhúng từ của riêng tập dữ liệu. Khóa luận cũng tạo một vec tơ 300 chiều đại diện cho các từ có trong dữ liệu nhưng không có trong ma trận nhúng. Mỗi từ trong câu sẽ được tìm kiếm trong ma trận nhúng để thu được về vectơ biểu diễn tương ứng. Để các câu có độ dài bằng nhau khi đưa vào lớp đầu vào, khóa luận lấy câu có số từ nhiều nhất làm chuẩn và thêm vào các câu còn lại các vec tơ 100 chiều, tương tự với các từ không có trong ma trận nhúng.

Đối với nhãn cho quan điểm của từng khía cạnh trong câu, khóa luận sử dụng phương pháp one-hot để mã hóa nhãn theo thứ tự quan điểm trung lập hoặc không đề cập, quan điểm tích cực và quan điểm tiêu cực. Đối với thể hiện cho quan điểm theo 5 khía cạnh trong câu, khóa luận mã hóa với các giá trị "0" đại diện cho quan điểm trung lập hoặc không đề cập, "1" cho quan điểm tích cực và "2" cho quan điểm tiêu cực, trong câu.

3.4. Mô hình mạng nơron tích chập

3.4.1. Mô hình phân lớp

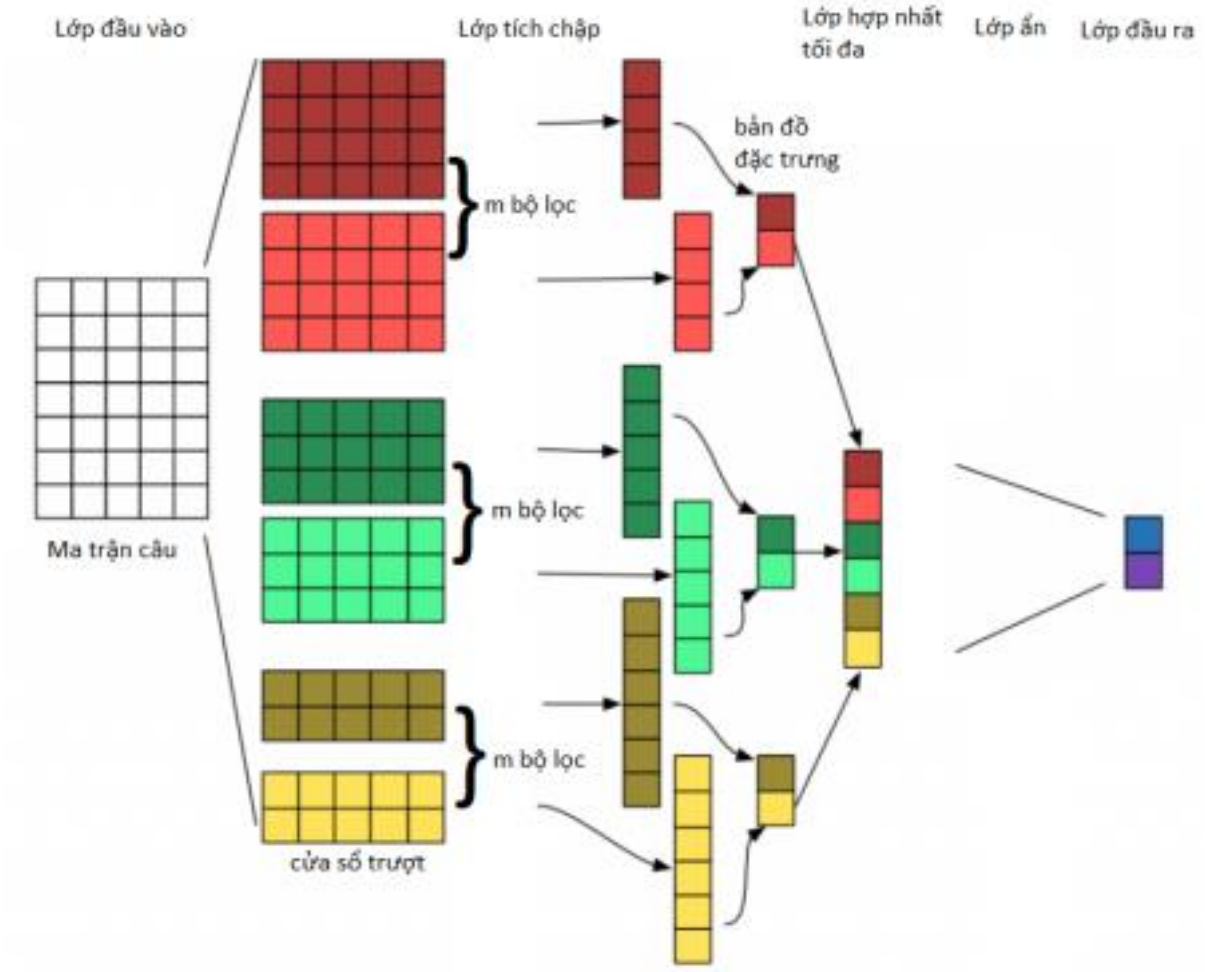


Hình 3.2: Mô hình phân lớp sử dụng mạng nơ ron tích chập

Khóa luận sử dụng mô hình phân lớp bao gồm 5 bộ phân lớp nhỏ, mỗi bộ phân lớp là mô hình mạng nơron tích chập thực hiện phân tích quan điểm theo từng khía cạnh với 3 nhãn tích cực, tiêu cực và không đề cập. Sau khi kết hợp 5 bộ phân lớp, mô hình xác định được những khía cạnh được đề cập trong câu và quan điểm của chúng, qua đó thu được kết quả cho hai bài toán phân lớp khía cạnh và phân tích cảm xúc.

3.4.2. Mô hình mạng nơ ron tích chập

Tổng quan mô hình mạng nơ ron tích chập bao gồm 5 lớp: lớp đầu vào, lớp tích chập, lớp hợp nhất tối đa (max pooling), lớp ẩn và cuối cùng là lớp đầu ra.



Hình 3.3: Mô hình mạng nơ ron tích chập

- **Lớp đầu vào:** Một câu bao gồm n từ cấu thành lên được biểu diễn dưới dạng:

$$x_{1:n} = x_1 \oplus x_2 \oplus \dots x_n$$

Trong đó $x_i \in \mathbb{R}_k$ là vectơ từ k chiều tương ứng với từ thứ i trong câu, $\oplus$ là toán tử ghép. Điều này mang lại một ma trận $S \in \mathbb{R}^{d \times n}$, được sử dụng làm đầu vào cho mạng nơron tích chập .

- **Lớp tích chập:** Trong lớp này, một bộ các bộ lọc *m* được áp dụng cho một cửa sổ trượt có độ dài *h* trên mỗi câu. Đặt $S_{[i,i+h]}$ biểu thị phép ghép của vecto từ s_i đến s_{i+h} .Một đặc trưng c_i được tạo cho bộ lọc F đã cho bằng cách:

$$c_i = \sum(S_{[i,i+h]})_{k,j} \cdot F_{k,j}$$

Việc ghép tất cả các vectơ trong câu tạo ra một vectơ đặc trưng $c \in \mathbb{R}^{n-h+1}$. Các vectơ *c* sau đó được tổng hợp trên tất cả các bộ lọc *m* tạo thành ma trân bản đồ đặc trưng $c \in \mathbb{R}^{m\times(n-h+1)}$.

- **Lớp hợp nhất tối đa:** Đầu ra của lớp tích chập được truyền qua một hàm kích hoạt phi tuyến trước khi vào lớp hợp nhất. Các phần tử sau này tổng hợp các phần tử vectơ bằng cách lấy tối đa trên một tập hợp các khoảng không chồng chéo cố định. Ma trận bản đồ đặc trưng được hợp nhất có dạng: $C_{pooled} \in R^{m.\frac{n-h+1}{s}}$, trong đó s là độ dài của mỗi khoảng. Trong trường hợp các khoảng chồng lấp với giá trị sải chân (stride) S_t, ma trận bản đồ đặc trưng được hợp nhất có dạng $C_{pooled} \in R^{m.\frac{n-h+1-s}{st}}$. Tùy thuộc vào việc bao gồm các đường viền hay không, kết quả của phân số được làm tròn lên hoặc xuống tương ứng.
- **Lớp ẩn:** Một lớp ẩn được kết nối đầy đủ sẽ tính toán phép biến đổi $\alpha(W*x+b)$, trong đó $W \in \mathbb{R}^{m.n}$ là ma trận trọng số, bias $b \in \mathbb{R}^m$ và α là hàm kích hoạt ReLU (Rectified linear unit). Vectơ đầu ra của lớp này, $x \in \mathbb{R}^m$, tương ứng với các nhúng câu cho mỗi câu.
- **Lớp đầu ra:** Cuối cùng, các đầu ra của lớp ẩn $x \in \mathbb{R}^m$ được kết nối hoàn toàn với một lớp hồi quy Soft-max, trả về lớp $y \in [1, K]$ với xác suất lớn nhất,

$$\hat{y} := \arg\max_j \frac{e^{x^T w_j + a_j}}{\sum_{k=1}^{K} e^{x^T w_k + a_j}},$$

trong đó w_j biểu thị vectơ trọng số của lớp *j* và bias

3.4.3. Huấn luyện mô hình

Sau bước tiền xử lý và biểu diễn dữ liệu, một phần dữ liệu được sử dụng để huấn luyện theo mô hình đã trình bày. Tập dữ liệu huấn luyện được chia thành các lô

(batch) nhỏ để xử lý theo lô. Mô hình sẽ tính giá trị mất mát (loss) và các độ đo theo lô, cuối mỗi lô mô hình sẽ cập nhật lại trọng số một lần. Để cập nhật lại các trọng số của mạng nơ ron sau mỗi lô xử lý, khóa luận sử dụng hàm tối ưu hóa Adaptive Moment Estimation (Adam). Gọi các trọng số của mô hình tại thời điểm lô thứ t là $W^{(t)}$ và mất mát (loss) tại thời điểm đó là $L^{(t)}$ .Thuật toán lan truyền ngược sử dụng hàm tối ưu hoá Adam như sau:

$$m_w^{(t+1)} \leftarrow \beta_1 m_w^{(t)} + (1 - \beta_1)\nabla_w L^{(t)}$$

$$v_w^{(t+1)} \leftarrow \beta_2 v_w^{(t)} + (1 - \beta_2)(\nabla_w L^{(t)})^2$$

$$\hat{m}_w = \frac{m_w^{(t+1)}}{1 - \beta_1^t}$$

$$\hat{v}_w = \frac{v_w^{(t+1)}}{1 - \beta_2^t}$$

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon}$$

trong đó, ϵ là một số rất nhỏ để tránh trường hợp chia cho 0, β_1 và β_2 là các tham số về độ giảm và mô men của độ dốc, η là tốc độ học của Adam.

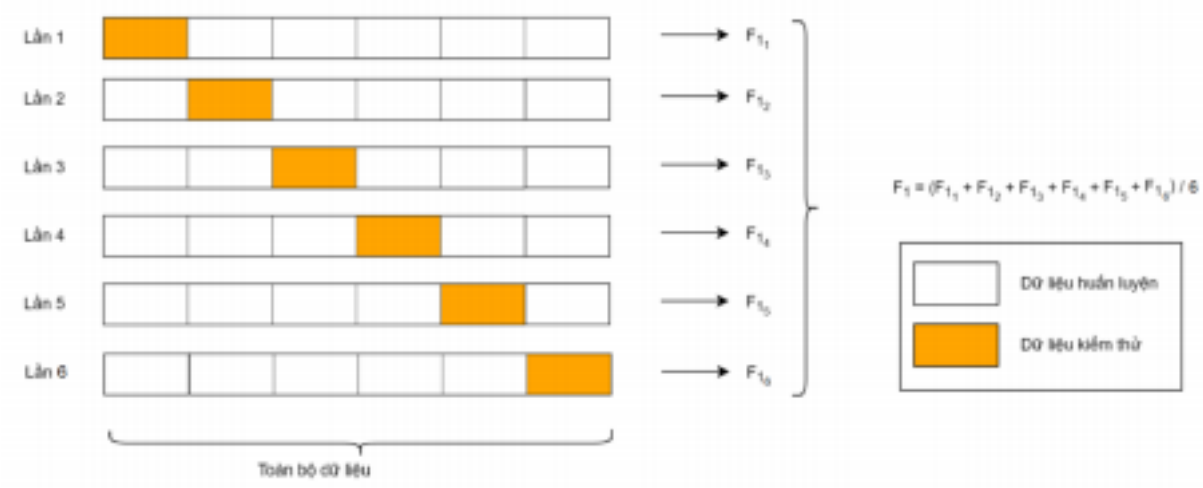
Để phòng tránh quá khớp, khóa luận sử dụng phương pháp dropout. Việc dropout ngăn chặn sự thích ứng đồng thời của các đơn vị ẩn bằng cách dropout ngẫu nhiên, tức là đặt thành 0 một tỷ lệ p của các đơn vị ẩn trong quá trình lan truyền ngược. Nghĩa là, cho lớp áp chót $z=[\sigma_1,..,\sigma_m]$ (lưu ý rằng có *m* bộ lọc), thay vì sử dụng $y = w \cdot z + b$, đối với đơn vị đầu ra *y* trong quá trình truyền thẳng, dropout sử dụng $y = w \cdot (z \otimes r) + b$, trong đó $\otimes$ là toán tử nhân element-wise và $r \in \mathbb{R}^m$ là một vectơ "mặt nạ" của các biến ngẫu nhiên Bernoulli với xác suất p là 1. Tóm lại, trong mỗi giai đoạn đào tạo, các

nút riêng lẻ được loại bỏ với xác suất p, mạng nơron suy giảm được cập nhật và sau đó các nút bị loại bỏ được xác nhận lại.

3.5. Đánh giá mô hình

Trong pha này, sau khi huấn luyện, một phần dữ liệu đã được tiền xử lý sẽ được dùng để đánh giá mô hình. Dữ liệu sẽ lần lượt đi qua 5 mô hình theo 5 khía cạnh để dự đoán quan điểm theo từng khía cạnh trong câu. Sau khi nhận được kết quả đầu ra là quan điểm theo từng khía cạnh, mô hình tổng hợp lại và so sánh với nhãn theo 5 khía cạnh trong câu để ra được kết quả trong câu gồm những khía cạnh nào và quan điểm cho từng khía cạnh, đồng thời đánh giá độ chính xác dựa trên độ đo và phương pháp được mô tả sau đây.

Khóa luận sử dụng phương pháp đánh giá chéo k lần (k-fold cross validation) với k = 6. Tập dữ liệu sẽ được chia ra k phần có kích thước xấp xỉ bằng nhau. Mô hình sẽ lặp k lần, tại lần lặp thứ i mô hình sẽ chọn phần dữ liệu thứ i làm tập đánh giá (testing) và (k−1) phần còn lại làm tập huấn luyện (training).



Hình 3.4: Phương pháp đánh giá chéo 6 lần

Tổng kết chương 3

Trong chương 3, khóa luận đã trình bày chi tiết về mô hình học sâu cho phân tích quan điểm dùng cho bài toán khóa luận và kiến trúc mạng nơ ron tích chập sử

dụng trong mô hình này cũng như các bước của phương pháp đề xuất. Chương tiếp theo, khóa luận sẽ trình bày về kết quả thực nghiệm và đánh giá mô hình này.

CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT QUẢ

Chương 3 đã đề cập đến mô hình học sâu cho phân tích quan điểm và các bước tiền hành. Tiếp theo, chương 4 sẽ chạy thực nghiệm cho mô hình nhằm làm rõ các bước thực hiện như đã giới thiệu. Mô hình được tiến hành trên tập dữ liệu tiếng Việt về nhận xét khách sạn.

4.1. Môi trường và các công cụ sử dụng thực nghiệm

4.1.1. Cấu hình phần cứng

Cấu hình phần cứng được khóa luận sử dụng để thực nghiệm được thể hiện như sau:

Bảng 4.1: Cấu hình phần cứng

Thành phần	**Thông số**
CPU	Intel Core i5 2.50 GHz
RAM	4.00 GB
Hệ điều hành	Windows 10 – 64 bit

4.1.2. Các phần mềm sử dụng

Bảng 4.2: Các phần mềm sử dụng

STT	**Tên phần mềm**	**Tác giả**	**Chức năng**	**Nguồn**
1	Pycharm Community 2020. 1.2		Môi trường phát triển	https://www.jetbrains .com/pycharm

2	vi_spacy 0.2.1	Trần Việt Trung	Tách từ	https://github.com/trungtv/vi_spacy
3	fastText 0.9.2		Bộ nhúng từ được huấn luyện sẵn	https://fasttext.cc/

4	Numpy 1.18.5		Thư viện Python để tính toán trên các ma trận	http://www.numpy.org/
5	Scikit-learn 0.23.1		Thư viện Python để chia dữ liệu và thực hiện đánh giá chéo	http://scikitlearn.org/stable/
6	TensorFlow 1.15.0		Thư viện Python để thiết kế mô hình học sâu	https://www.tensorflow.org/

4.2. Tiến hành thực nghiệm

4.2.1. Dữ liệu

Dữ liệu tiếng Việt về nhận xét khách sạn do Phòng Thí nghiệm DS&KTLab thu thập từ trang Web khachsan.chudu24.com và gán nhãn gồm 1493 câu đánh giá của người dùng về khách sạn với 5 khía cạnh bao gồm: chất lượng phục vụ, chất lượng phòng, chất lượng đồ ăn, vị trí và giá cả, trang thiết bị được mô tả như sau:

Bảng 4.3: Dữ liệu chi tiết theo từng khía cạnh trong tập dữ liệu nhẫn xét khách sạn

Số lượng Khía cạnh	**Tích cực**	**Tiêu cực**	**Trung lập hoặc không đề cập**

Chất lượng phục vụ	571	69	853
Chất lượng phòng	480	89	924
Chất lượng đồ ăn	282	121	1090
Vị trí và giá cả	320	56	1117
Trang thiết bị	154	144	1493

4.2.2. Cài đặt tham số

Như đã đề cập ở mục 3.3.1, khóa luận sử dụng nhúng từ với số chiều là 300. Do tập dữ liệu có kích thước nhỏ nên khóa luận chọn kích thước lô bằng 16. Đối với các tham số của hàm tối ưu hóa Adam, khóa luận để mặc định theo thư viện đang sử dụng. Đối với số epoch, dựa trên kinh nghiệm cá nhân, khóa luận chọn 30 epoch. Các tham số còn lại được mô tả như trong bảng mô tả dưới đây.

Bảng 4.4: Các danh sách tham số của mô hình

Tham số	**Giá trị**
Số chiều nhúng từ	300
Số bộ lọc CNN	16, 32, 32
Kích thước cửa sổ tích chập	2, 3, 4
Dropout	0,5

Kích thước lô		16
Số epoch		32
Adam	Tốc độ học ()	0,01
	��1	0,9
	��2	0,999
	∈	10^{-7}

4.2.3. Độ đo đánh giá

Trong mô hình này, khóa luận đánh giá hiệu năng hệ thống trên tập dữ liệu đã chuẩn bị qua độ đo trung bình mịn (micro average) và trung bình thô (macro average) dựa trên độ chính xác Precision, độ hồi tưởng Recall và độ đo điều hòa F1.

Độ chính xác (P) được tính bằng phần trăm các trường hợp được gán nhãn đúng (TP) trên tổng số nhãn dương của hệ thống (TP + FP).

Độ hồi tưởng (R) là phân trăm các kết quả đúng (TP) trên tất cả các mẫu dương hiện có trong dữ liệu (TP + FN).

Hình 4.1: Mô tả các độ đo chính xác, độ hồi tưởng và độ đo F1

4.3. Kết quả thực nghiệm và đánh giá

4.3.1. Kết quả thực nghiệm trong bài toán phân lớp khía cạnh

Bảng 4.5 thể hiện kết quả thực nghiệm chi tiết của bài toán phân lớp khía cạnh trên tập dữ liệu tại 6 lần lặp (fold) khác nhau của phương pháp đánh giá chéo. Trong đó, ký hiệu P (Precision) là độ chính xác, R (Recall) là độ hồi tưởng, F (F1-score) là độ đo F1.

Bảng 4.5: Kết quả thực nghiệm đánh giá chéo 6 lần cho bài toán phân lớp quan điểm (Đơn vị: %)

Lần lặp (fold) **Khía cạnh**		**1**	**2**	**3**	**4**	**5**	**6**
(1) Chất lượng phục vụ	P	96.81	100	93.64	100	100	85.11
	R	80.53	84.55	92.79	66.67	83.81	81.63
	F	87.92	91.63	93.21	80	91.19	83.33
(2) Chất lượng phòng	P	84.42	90.74	91.78	100	98.15	89.83

	R	63.11	52.13	73.63	72.73	56.38	53.54
	F	72.23	66.22	81.71	84.21	71.62	67.09
(3) Chất lượng đồ ăn	P	93.22	91.38	98.33	96.72	100	98.11
	R	79.71	88.33	89.39	85.51	85	88.14
	F	85.94	89.83	93.65	90.77	91.89	92.86
(4) Vị trí và giá cả	P	64.29	66.67	84.13	77.19	96.97	97.22
	R	35.29	56.67	76.81	68.75	44.44	58.33
	F	45.57	61.26	80.3	72.73	60.95	72.91
(5) Trang thiết bị	P	60	76.92	85.96	82.35	89.29	92.16
	R	64.71	50	71.01	65.62	69.44	78.33
	F	62.27	60.61	77.77	73.04	78.12	84.68

Bảng 4.6 thể hiện kết quả thực nghiệm chi tiết của bài toán phân lớp khía cạnh. Nhìn vào bảng, ta thu đươc kết quả với độ đo trung bình vĩ mô F1 đạt 78.18%, trung bình vi mô F1 đạt 79.93%.

Bảng 4.6: Kết quả thực nghiệm trên 5 khía cạnh cho bài toán phân lớp quan điểm (Đơn vị: %)

Độ đo **Khía cạnh**	**P**	**R**	**F**
(1) Chất lượng phục vụ	95.93	81.66	87.88
(2) Chất lượng phòng	92.49	61.92	73.85
(3) Chất lượng đồ ăn	96.29	86.01	90.82
(4) Vị trí và giá cả	81.08	56.72	65.62
(5) Trang thiết bị	81.11	66.52	72.75
Trung bình thô	89.38	70.57	78.18
Trung bình mịn	90.56	71.68	79.93

4.3.2. Kết quả thực nghiệm trong bài toán phân tích quan điểm

Tương tự, bảng 4.5 thể hiện kết quả thực nghiệm chi tiết của bài toán phân tích quan điểm trên tập dữ liệu tại 6 lần lặp (fold) khác nhau của phương pháp đánh giá chéo.

Bảng 4.7: Kết quả thực nghiệm đánh giá chéo 6 lần cho bài toán phân tích quan điểm (Đơn vị: %)

Lần lặp **(fold)** **Khía cạnh**		**1**	**2**	**3**	**4**	**5**	**6**
(1) Chất lượng phục vụ	P	86.17	93.27	84.4	95	94.32	78.72

	R	83.51	88.18	95.83	69.51	85.57	83.15
	F	84.82	90.65	89.75	80.28	89.73	80.87
(2) Chất lượng phòng	P	76.62	77.78	84.93	98.44	79.63	84.75
	R	66.29	55.26	76.54	81.82	56.58	61.73
	F	71.08	64.61	80.52	89.36	66.15	71.43
(3) Chất lượng đồ ăn	P	65.22	87.1	95.35	77.08	91.49	87.18
	R	69.77	69.23	80.39	74	72.88	85
	F	67.42	77.14	87.23	75.51	81.13	86.08
(4) Vị trí và giá cả	P	44.44	60.42	66.67	62.5	75	86.11
	R	30.77	52.73	61.82	67.31	37.5	56.36
	F	36.36	56.31	64.15	64.82	50	68.13
(5) Trang thiết bị	P	47.27	64.86	71.15	68.09	82.69	82.35
	R	66.67	43.64	67.27	61.54	67.19	76.36
	F	55.32	52.17	69.16	64.65	74.14	79.24

Bảng 4.6 thể hiện kết quả thực nghiệm chi tiết của bài toán phân tích quan điểm. Nhìn vào bảng, ta thu đươc kết quả với độ đo trung bình vĩ mô F1 đạt 72.28%, trung bình vi mô F1 đạt 74.72%.

Bảng 4.8: Kết quả thực nghiệm trên 5 khía cạnh cho bài toán phân tích quan điểm
(Đơn vị: %)

Độ đo **Khía cạnh**	**P**	**R**	**F**
(1) Chất lượng phục vụ	88.65	84.29	86.02
(2) Chất lượng phòng	83.69	66.37	73.86
(3) Chất lượng đồ ăn	83.9	75.21	79.09
(4) Vị trí và giá cả	65.86	51.08	56.63
(5) Trang thiết bị	69.4	63.78	65.78
Trung bình thô	78.3	68.15	72.28
Trung bình mịn	80.21	70.19	74.72

Bên cạnh việc đánh giá độc lập, khoá luận cũng so sánh kết quả thực nghiệm với một số mô hình khác. Cụ thể, khóa luận so sánh với hai mô hình sử dụng học máy bán giám sát cho bài toán phân lớp quan điểm trên cùng bộ dữ liệu về bình luận khách sạn gồm 1500 câu. Có thể thấy, khóa luận đã đạt được kết quả hết sức khả quan khi đạt xấp xỉ

ngưỡng 80%, chỉ kém hơn so với hai mô hình so sánh đạt kết quả lần lượt là 83.2% và 81%.

Bảng 4.9: So sánh kết quả thực nghiệm với một số mô hình khác
(Đơn vị: %)

Phương pháp	**Độ đo trung bình vi mô F1**
Học máy bán giám sát cho phân loại đa nhãn sử dụng BI (binary feature) và MI (mutual information) (kết quả tốt nhất) [12]	83.2
Học máy bán giám sát cho phân loại đa nhãn sử dụng thuật toán LIFT (Label specIfic FeaTures), TESC (TExt classification using Semi-supervised Clustering) và kNN (với k = 1) [13]	81
Học sâu của khóa luận	79.93

Kết luận chương 4

Trong chương này, khóa luận mô tả về tập dữ liệu và các tham số mà mô hình sử dụng. Bên cạnh đó, khóa luận cũng đã trình bày về quá trình thực nghiệm và kết quả thực nghiệm của khóa luận. Qua các kết quả thực nghiệm thu được, ta có thể thấy phương pháp đề xuất trong khóa luận có thể áp dụng được vào bài toán phân tích quan điểm trong thực tiễn.

KẾT LUẬN

Khóa luận đã tiếp cận được những phương pháp sử dụng trong bài toán phân tích quan điểm được nghiên cứu và công bố trên thế giới, đặc biệt là các mô hình học sâu. Dựa vào đó, khóa luận đã tiến hành phân tích và tiến hành xây dựng một mô hình học sâu cho phân tích quan điểm.

Khóa luận đã đạt được những kết quả sau:

1. Khảo sát, tìm hiểu về phân tích quan điểm, phương pháp học sâu cũng như các hướng tiếp cận trong phân tích quan điểm. Qua đó, khóa luận đưa ra mô hình sử dụng mạng nơ ron tích chập thực hiện đồng thời hai bài toán con là phân lớp khía cạnh và phân tích quan điểm trong bài toán phân tích quan điểm mức khía cạnh.

2. Khóa luận tiến hành thực nghiệm trên tập dữ liệu tiếng Việt về nhận xét khách sạn. Qua thực nghiệm đã thu được kết quả ban đầu khá khả quan với độ đo F1 trung bình thô đạt 78.18% cho bài toán phân lớp khía cạnh và 72.28% cho bài toán phân tích quan điểm; độ đo F1 trung bình mịn đạt 79.93% cho bài toán phân lớp khía cạnh và 74.72% cho bài toán phân tích quan điểm.

Do hạn chế về thời gian và kiến thức của cá nhân, khóa luận mới chỉ tập trung vào xây dựng mô hình chứ chưa xây dựng thành một hệ thống có ứng dụng cụ thể và trực quan. Bên cạnh đó, lượng dữ liệu khóa luận sử dụng để thực nghiệm cũng chưa nhiều.

Trong thời gian tới, khóa luận sẽ tiếp tục cải thiện hiệu suất học của mô hình bằng cách thu thập thêm dữ liệu và làm mịn dữ liệu, tinh chỉnh các tham số của mô hình, cũng như tìm hiểu và áp dụng thêm các phương pháp tránh quá khớp như dừng sớm (early stopping).

TÀI LIỆU THAM KHẢO

[1] Bing Liu. Sentiment analysis: mining opinions, sentiments, and emotions. The Cambridge University Press, 2015.

[2] Lei Zhang, Shuai Wang, Bing Liu. Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1253.

[3] Goodfellow I, Bengio Y, Courville A. Deep learning. The MIT Press. 2016.

[4] Walaa Medhat, Ahmed Hassan, Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

[5] Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu & Ge Yu. Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences 63.1 (2020): 1-36.

[6] Xiaodong Gu, Yiwei Gu, Haibing Wu. Cascaded convolutional neural networks for aspect-based opinion summary. Neural Processing Letters 46.2 (2017): 581-594.

[7] Zhiqiang Toh, Jian Su. NLANGP at SemEval-2016 task 5: Improving aspect based sentiment analysis using neural network features. Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). (2016)

[8] Sebastian Ruder, Parsa Ghaffari, John G. Breslin. Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis. arXiv preprint arXiv:1609.02748 (2016).

[9] Hai Ha Do, PWC Prasad, Angelika Maag, Abeer Alsadoon. Deep learning for aspect based sentiment analysis: a comparative review. Expert Systems with Applications 118 (2019): 272-299.

[10] Yoon Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).

[11] Nhan Cach Dang, María N. Moreno-García, Fernando De la Prieta. Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics 9.3 (2020): 483.

[12] Thi-Ngan Pham, Van-Quang Nguyen, Van-Hien Tran, Tri-Thanh Nguyen, Quang Thuy Ha. A semi-supervised multi-label classification framework with

feature reduction and enrichment. Journal of Information and Telecommunication 1.4 (2017): 305-318.

[13] Thi-Ngan Pham, Van-Quang Nguyen, Duc-Trong Dinh, Tri-Thanh Nguyen, Quang-Thuy Ha. MASS: A semi-supervised multi-label classification algorithm with specific features. Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2017.