BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN VIẾT HÙNG

KHAI THÁC QUAN ĐIỂM CỦA CÁC BÌNH LUẬN TIẾNG ANH TRÊN MẠNG XÃ HỘI SỬ DỤNG PHƯƠNG PHÁP XỬ LÝ NGÔN NGỮ TỰ NHIÊN

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 4 năm 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN VIẾT HÙNG

KHAI THÁC QUAN ĐIỂM CỦA CÁC BÌNH LUẬN TIẾNG ANH TRÊN MẠNG XÃ HỘI SỬ DỤNG PHƯƠNG PHÁP XỬ LÝ NGÔN NGỮ TỰ NHIÊN

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC:

TS. NGÔ MINH VƯƠNG

TS. NGUYỄN THỊ THANH SANG

TP. HỒ CHÍ MINH, tháng 4 năm 2016

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : TS NGÔ MINH VƯƠNG

TS NGUYỄN THỊ THANH SANG

(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày … tháng … năm 2016

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	**Họ và tên**	**Chức danh Hội đồng**
1	PGS. TS. Võ Đình Bảy	Chủ tịch
2	TS. Lư Nhật Vinh	Phản biện 1
3	TS. Vũ Thanh Hiền	Phản biện 2
4	TS. Cao Tùng Anh	Ủy viên
5	TS. Nguyễn Thị Thúy Loan	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày..… tháng….. năm 2016

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: Nguyễn Viết Hùng — Giới tính: Nam

Ngày, tháng, năm sinh: 02/09/1990 — Nơi sinh: Hà Tĩnh

Chuyên ngành: Công nghệ thông tin — MSHV: 1441860011

I- Tên đề tài:

Khai thác quan điểm của các bình luận tiếng Anh trên mạng xã hội sử dụng phương pháp xử lý ngôn ngữ tự nhiên.

II- Nhiệm vụ và nội dung:

Xem xét, phân tích các ý kiến, quan điểm trong mạng xã hội hiện nay như Website, diễn đàn và mạng xã hội.

- Ý kiến tích cực (postitive)
- Ý kiến tiêu cực (negative)
- Ý kiến trung lập (neutral)

Tìm hiểu các công trình phân tích ý kiến hiện nay

Thiết kế và xây dựng mô hình phân tích ý kiến phù hợp với NLP.

Xây dựng chương trình và tiến hành đánh giá thực nghiệm mô hình đề xuất.

III- Ngày giao nhiệm vụ: 15/07/2015

IV- Ngày hoàn thành nhiệm vụ: 15/04/2016

V- Cán bộ hướng dẫn: TS. Ngô Minh Vương và TS. Nguyễn Thị Thanh Sang

CÁN BỘ HƯỚNG DẪN — **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) — (Họ tên và chữ ký)

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Học viên thực hiện Luận văn

(Ký và ghi rõ họ tên)

Nguyễn Viết Hùng

LỜI CẢM ƠN

Trước tiên, tôi xin được gửi lời cảm ơn đến Ban Giám Hiệu, toàn thể cán bộ nhân viên, giảng viên trường Đại Học HUTECH, Ban lãnh đạo Phòng Quản Lý Khoa Học và Đào Tạo Sau Đại Học, khoa Công Nghệ Thông Tin đã tạo điều kiện thuận lợi cho chúng tôi học tập và nghiên cứu trong suốt học trình cao học tại trường. Xin được gửi lời cảm ơn đến tất cả quý thầy cô đã giảng dạy trong chương trình Đào tạo thạc sĩ chuyên ngành Công nghệ thông tin, niên khóa 2014-2016, lớp 14SCT11 - Trường Đại học Công Nghệ TP HCM, những người đã truyền đạt cho tôi những kiến thức hữu ích để làm cơ sở cho tôi thực hiện tốt luận văn này.

Với lòng kính trọng và biết ơn, tôi xin bày tỏ lời cảm ơn đến TS Ngô Minh Vương và TS Nguyễn Thị Thanh Sang đã tận tình hướng dẫn cho tôi trong thời gian thực hiện luận văn. Mặc dù, trong quá trình thực hiện luận văn có giai đoạn không được thuận lợi, nhưng những gì thầy cô đã hướng dẫn, chỉ bảo đã cho tôi nhiều kinh nghiệm trong thời gian thực hiện luận văn.

Xin gửi lời cảm ơn đến Ths Đặng Thị Vân đã giúp đỡ và tư vấn cho tôi về ngôn ngữ tiếng Anh trong suốt quá trình tôi thực hiện luận văn.

Và đặc biệt, tôi xin gửi lời biết ơn sâu sắc đến bạn bè, gia đình, các anh chị trong tập thể lớp 14SCT11 đã luôn tạo điều kiện tốt nhất cho tôi trong suốt quá trình học cũng như thực hiện luận văn.

Sau cùng, tôi xin cảm ơn và ghi nhận tất cả những sự giúp đỡ kể trên. Với tất cả sự nỗ lực và cố gắng của bản thân trong hơn 9 tháng thực hiện, tôi đã hoàn thành được luận văn, và tất nhiên sẽ không tránh khỏi những thiếu sót cần phải hoàn thiện, rất mong nhận được sự góp ý của quý thầy cô và các bạn.

Nguyễn Viết Hùng

TÓM TẮT

Với sự đa dạng về ngôn ngữ, trong đó khai thác ngôn ngữ đang trở nên một ngành được chú tâm đối với nhiều nhà nghiên cứu khoa học hiện nay, đặc biệt khai thác quan điểm, ý kiến, tình cảm, cảm xúc đóng vài trò quan trọng trong phát triển mạng xã hội. Trong lĩnh vực rút trích thông tin, phân loại quan điểm có thể thực hiện bằng một loạt các ứng dụng trong việc khác thác theo phương pháp xử lý ngôn ngữ tự nhiên và học máy.

Khai thác quan điểm về các ngôn ngữ liên quan đến việc đánh giá, giải thích chính xác của việc sử dụng ngôn ngữ tự nhiên, và tất cả những điều này đã khai thác từ việc phân tích và đánh giá theo phương pháp xử lý ngôn ngữ tự nhiên. Mặc dù vậy, các công trình nghiên cứu cơ bản về chủ đề này đã thể hiện sự đáng ngạc nhiên và mang yếu tố quyết định trong quy trình khai thác theo phương pháp sử dụng các quy tắc dựa trên ngôn ngữ tự nhiên.

Phương pháp sử dụng ngôn ngữ tự nhiên được sử dụng thực hiện trong khai thác quan điểm trên mạng xã hội là chúng tôi đề ra các quy tắc khai phá quan điểm đánh giá từ những ứng dụng, kỹ thuật đã được nghiên cứu và phân tích cấu trúc ngữ pháp, và xây dựng bộ từ điển, cụm từ, xử lý kỹ thuật với các từ đặc biệt trong tiếng Anh.

Báo cáo này là sự nỗ lực nghiên cứu, khai thác quan điểm của khách hàng trên mạng xã hội bằng phương pháp xử lý ngôn ngữ tự nhiên. Ứng dụng kỹ thuật cho vấn đề này là việc phân loại, đánh giá các ý kiến, quan điểm trên mạng xã hội. Thông qua việc đánh giá này, chúng tôi đánh giá quan điểm của khác hàng về các chủ đề nhất định trên các trang mạng xã hội.

Kết hợp các phương pháp xử lý ngôn ngữ tự nhiên để phân loại cấu trúc câu và xử lý ngôn ngữ, có thể cải thiện hiệu suất khai thác ý kiến.

ABSTRACT

With the variety of languages, language mining sectors are being paid much attention by contemporary scientific researchers, especially the exploitation of views, ideas, emotions play an important role in social network development. Opinion classification through information extraction requires a series of natural language processing method and machine learning.

Exploiting opinions relate to the assessment and interpretation of natural language and all these things were fully exploited from the analysis and evaluation by the way of natural language processing. However, the basic research on this topic has shown surprising and brought the decisive factor in other processes according to the techniques based on natural language.

Natural language processing is implemented to exploit points of view on the social network, in that we have devised techniques to assess such as applied techniques to analyse grammatical structure, build dictionaries, phrases, technical terms.

This study has made an attempt to research, exploit customer opinions on the social network by the means of natural language processing. Applying this technique to this problem is to classify and evaluate the comments and views on the social network. By these reviews, we assess customers' views about certain topics on the social networking site.

Combining the method natural language processing techniques for sentence structure classification and language processing, can improve the performance of opinion mining.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Kí hiệu	English	Tiếng việt
NLP	Natural language processing	**Xử lý ngôn ngữ tự nhiên**
POS	Part – Of – Speech	**Từ loại**
SVM	Suppor vector Machine	**Phương pháp sử dụng Máy học**
PMI	Pointwise mutual information	**Chuẩn hóa thông tin lẫn nhau**

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

CHƯƠNG 1
MỞ ĐẦU

1.1 Giới thiệu

Ngày nay với nhu cầu công nghiệp hóa hiện đại hóa và phát triển đất nước, con người đã biết vận dụng khoa học thực tiễn vào đời sống của mình nhằm tăng hiệu quả trong sản xuất cũng như phát triển khoa học.

Dựa trên sự phát triển, có rất nhiều người muốn tham khảo ý kiến trên các phương tiện truyền thông nhằm quyết định mua các sản phẩm hoặc dịch vụ nào đó. Tuy nhiên, rất khó khăn trong việc phát hiện ý kiến rác bởi vì các nhận xét lừa đảo có thể được viết ra bởi các tổ chức cũng như cá nhân với nhiều mục đích khác nhau. Họ viết các nhận xét lừa đảo này nhằm mục đích đánh lừa người đọc hoặc hệ thống nhận diện tự động để đề cao sản phẩm của họ hoặc đánh giá thấp các sản phẩm đối thủ.

Thông qua các ý kiến từ việc đánh giá trên các phương tiện truyền thông, trên mạng xã hội ... Việc phát hiện các tình cảm, cảm xúc cũng là vấn đề khó khăn bởi tình cảm thể hiện tâm trạng của con người và tùy theo những điều kiện và cảm xúc của mỗi người. Do đó, sẽ có những đánh giá dựa trên sự cảm xúc yêu mến để đánh giá, dự trên những yếu tố đó dẫn đến sự đánh giá sai lệch về các sản phẩm. Qua đó mang đến yếu tố tiêu cực cho phần lớn các doanh nghiệp cũng như các cá nhân có thị hiếu mạng lại lợi ích cho khách hàng.

Chính vì những lý do trên, hệ thống phân tích quan điểm của nhận xét tiếng Anh của chúng tôi nhằm mục đích khai thác và tổng hợp lại những bày tỏ ý kiến, những bình luận về những sản phẩm đang thịnh hành trên các trang web thương mại điện tử như điện thoại di động và máy tính xách tay. Hệ thống của chúng tôi sẽ xác định chủ đề nào được đề cập đến trong nhận xét. Một nhận xét có thể có nhiều hơn một thực thể được nói đến, ví dụ như trong nhận xét kiểu so sánh. Một thực thể có thể được đề cập đến thông qua các đặc trưng, thuộc tính con của nó. Tiếp theo hệ thống sẽ tìm kiếm các

ý nghĩa mang tính tích cực, tiêu cực hoặc trung tính trong nhận xét. Từ đó hệ thống sẽ xác định được quan điểm của người viết nhận xét cho chủ đề được đề cập trong nhận xét này.

Với những vấn đề này, chúng tôi muốn nghiên cứu và tìm hiểu về những ý kiến, cảm xúc của khách hành qua đó đánh giá đúng những nhận xét tốt và xấu có hiệu quả tốt nhất phục vụ nhu cầu thị hiếu của khách hàng, và những yêu cầu đúng đắn của khách hàng.

1.2 Lý do chọn đề tài

Hiện nay, với nhu cầu cạnh tranh, đi sâu vào thiện ý của những khách hàng trên thị trường, các doanh nghiệp luôn tìm hiểu các quan điểm của khách hàng về sản phẩm, chất lượng, giá thành... Các doanh nghiệp ngày càng tăng sự cạnh tranh với nhau.

Với nhu cầu công nghiệp hóa hiện đại hóa đất nước, các doanh nghiệp trên thị trường luôn mong muốn có sự điều chỉnh hợp lý, mang lại hiểu quả cho khách hàng, và tạo sự uy tín của mình. Cho nên, các doanh nghiệp không ngừng thay đổi, đổi mới và muốn nắm bắt tâm tự nguyện vọng của người tiêu dùng.

Hầu hết các ứng dụng của phân tích ý kiến là nằm trong nhận xét của khách hàng về sản phẩm và dịch vụ [1]. Vì thế trong công trình này, chúng tôi sẽ khảo sát các ý kiến trên các website mạng xã hội. Mạng xã hội là một bước đột phá so với kinh doanh theo truyền thống (người bán và người mua trao đổi trực tiếp với nhau) và đang trở nên rất phổ biến. Số lượng khách hàng sử dụng dịch vụ mua bán qua mạng càng nhiều và sự quan tâm của họ với loại hình kinh doanh này cũng ngày càng tăng lên. Do đó, sẽ có rất nhiều thậm chí hàng trăm, hàng nghìn những lời bình luận, nhận xét cho những sản phẩm hoặc dịch vụ mà họ quan tâm hoặc những sản phẩm, dịch vụ đang phổ biến trên thị trường như máy ảnh, điện thoại di động, máy tính xách tay, phim điện ảnh chiếu rạp, sách, khách sạn và du lịch. Chính vì vậy, thật khó để cho những khách hàng tiềm năng có thể tìm đọc hết những lời bình luận, nhận xét của những khách hàng trước đó đã sử dụng để có thể đưa ra được những quyết định hợp lý. Và cũng thật khó để các

nhà sản xuất sản phẩm đó có thể theo dõi và quản lý những ý kiến của khách hàng để làm thỏa mãn khách hàng.

Vì vậy, đề tài nghiên cứu các vấn đề khai thác ngữ nghĩa, nhằm khai thác hiệu quả các ý kiến đánh giá của khách hàng. Thể hiện tính đúng đắn và mang lại hiệu quả tốt nhất cho khách hàng và tạo sự cạnh tranh lành mạnh của các doanh nghiệp trên thị trường.

1.3 Mục tiêu của đề tài

Đề xuất khai thác quan điểm của các bình luận tiếng Anh trên mạng xã hội bằng phương pháp xử lý ngôn ngữ tự nhiên.

Xây dựng hệ thống phân loại các bình luận.

1.4 Phương pháp luận và phương pháp nghiên cứu

Trích xuất và phân tích các appraisal groups (cụm đánh giá, ví dụ: rất đẹp, không quá mắc...) để phân loại tình cảm. Mỗi cụm đánh giá gồm một tính từ chính và các từ bổ nghĩa. Ví dụ: không thật sự hạnh phúc, "hạnh phúc" là tính từ chính, "không", "thật sự" là hai từ bổ nghĩa.

Các đặc điểm lấy từ việc phân tích appraisal group được kết hợp với bag-of-words giúp tăng độ chính xác của classifier. Tự động phát hiện các biểu hiện cảm xúc ẩn, dựa trên ngữ cảnh và những tri thức thông thường.

Xây dựng một cơ sở tri thức, gọi là EmotiNet. Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) để cải thiện phân loại tình cảm, ý kiến.

Ba kỹ thuật chính được sử dụng trong báo cáo này là:

- Gắn nhãn từ loại (POS tagging)
- Phân tích tính phụ thuộc bằng cách phân tách cấu trúc cây của câu.
- Xử lý các phép phủ định trong câu.

1.5 Cấu trúc luận văn

Nội dung báo cáo gồm những chương sau:

Chương 1: Mở đầu: Giới thiệu khái quát chung

Chương 2: Cơ sở kiến thức: Khái quát về ngôn ngữ tự nhiên và kỹ thuật xử lý ngôn ngữ tự nhiên Pos tangger.

Chương 3: Các công trình liên quan: Trình bày các công trình liên quan như Xử lý ngôn ngữ tự nhiên, Máy học và Ontology.

Chương 4: Mô hình đề xuất: Đề xuất mô hình, thực nghiệm cho quá trình nghiên cứu.

Chương 5: Kết luận và hướng phát triển: Khái quát lại những việc làm được và chưa làm được, định hướng phát triển của đề tài.

CHƯƠNG 2
CƠ SỞ LÝ THUYẾT

2.1 Khái quát về ngôn ngữ NLP

2.1.1 Khái niệm

NLP (Natural Language Processing) là khái niệm để chỉ các kĩ thuật, phương pháp thao tác trên ngôn ngữ tự nhiên bằng máy tính. Bạn cần phân biệt ngôn ngữ tự nhiên (ví dụ như tiếng Việt, tiếng Anh, tiếng Nhật… là những ngôn ngữ trong giao tiếp thường ngày) và ngôn ngữ nhân tạo (như ngôn ngữ lập trình, ngôn ngữ máy, …).

Ngoài ra, Xử lý ngôn ngữ tự nhiên (NLP) cũng là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp.

2.1.2 Khái quát chung

2.1.2.1 Ngôn ngữ tự nhiên

Ngôn ngữ là hệ thống để giao thiệp hay suy luận dùng một cách biểu diễn phép ẩn dụ và một loại ngữ pháp theo logic, mỗi cái đó bao hàm một tiêu chuẩn hay sự thật thuộc lịch sử và siêu việt. Nhiều ngôn ngữ sử dụng điệu bộ, âm thanh, ký hiệu, hay chữ viết, và cố gắng truyền khái niệm, ý nghĩa, và ý nghĩ, nhưng mà nhiều khi những khía cạnh này nằm sát quá, cho nên khó phân biệt nó.

2.1.2.2 Trí tuệ nhân tạo

Trí tuệ nhân tạo hay trí thông minh nhân tạo (tiếng Anh: artificial intelligence hay machine intelligence, thường được viết tắt là AI) là trí tuệ được biểu diễn bởi bất cứ một hệ thống nhân tạo nào. Thuật ngữ này thường dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và ứng dụng của trí tuệ nhân tạo.

2.1.2.3 Nhập nhằng trong ngôn ngữ

Nhập nhằng trong ngôn ngữ học là hiện tượng thường gặp, trong giao tiếp hàng ngày con người ít để ý đến nó bởi vì họ xử lý tốt hiện tượng này. Nhưng trong các ứng dụng liên quan đến xử lý ngôn ngữ tự nhiên khi phải thao tác với ý nghĩa từ vựng mà điển hình là dịch tự động thì nhập nhằng trở thành vấn đề nghiêm trọng . Ví dụ trong một câu cần dịch có xuất hiện từ "đường" như trong câu "ra chợ mua cho mẹ ít đường" vấn đề nảy sinh là cần dịch từ này là đường (sử dụng trong thức ăn của con người) hay đường (sử dụng trong giao thông), con người xác định chúng khá dễ dàng căn cứ vào văn cảnh và các dấu hiệu nhận biết khác nhưng với máy thì không. Một số hiện tượng nhập nhằng: Nhập nhằng ranh giới từ, Nhập nhằng từ đa nghĩa, Nhập nhằng từ đồng âm (đồng tự), Nhập nhằng từ loại.

2.1.2.4 Dịch máy

Dịch máy là một trong những ứng dụng chính của xử lý ngôn ngữ tự nhiên, dùng máy tính để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác. Mặc dù dịch máy đã được nghiên cứu và phát triển hơn 50 năm qua, xong vẫn tồn tại nhiều vấn đề cần nghiên cứu. Ở Việt Nam, dịch máy đã được nghiên cứu hơn 20 năm, nhưng các sản phẩm dịch máy hiện tại cho chất lượng dịch còn nhiều hạn chế. Hiện nay, dịch máy được phân chia thành một số phương pháp như: dịch máy trên cơ sở luật, dịch máy thống kê và dịch máy trên cơ sở ví dụ.

2.2. Khái quát về POS Tagger NLP

2.1 Khái niệm

A Part-Of-Speech Tagger (POS Tagger) là một phần mềm được sử dụng nhiều nhằm nghiên cứu, khai thác ngôn ngữ như tiếng Anh, tiếng Nhật,… Với việc gán nhãn các từ loại, phần mềm hỗ trợ phân tích các danh từ, động từ, tính từ, vv...

2.2 Khái quát chung

Pos tagger NLP được xem là một nền tảng thông dụng trong việc ứng dụng xử lý theo ngôn ngữ tự nhiên, việc khai thác các quan điển trên mạng xã hội và ứng dụng

chúng. Sử dụng Pos tagger chúng ta đã phần nào hạn chế sự nhập nhằn của các quan điểm khi đánh giá về một vấn đề bằng việc phân chia các từ, cụm từ trong câu, với ứng dụng này chúng ta khai thác quan điểm tối ưu hơn. Với Bài luận này đã khai thác nó rất hiểu quả trong quy trình nghiên cứu và đánh giá các quan điểm theo mô hình tối ưu hơn.

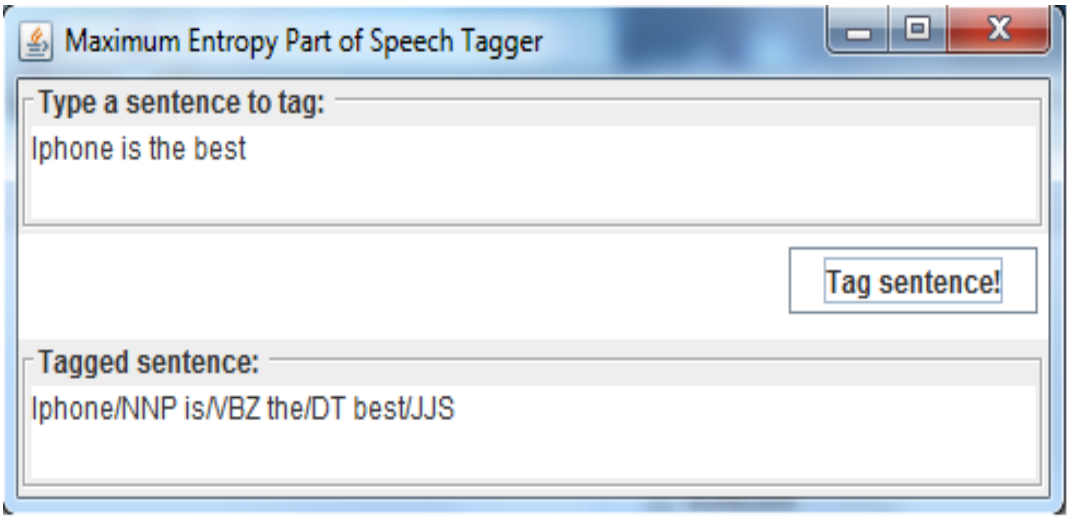
Stanford POS Tagger là một ứng dụng miễn phí thuộc lĩnh vực xử lý ngôn ngữ tự nhiên do đại học Stanford xây dựng nhằm gán nhãn từ loại cho văn bản với độ chính xác rất cao. Ứng dụng được thiết kế rất đa dạng: người sử dụng có thể chạy ứng dụng trực tiếp từ dòng lệnh hoặc có thể sử dụng các API có sẵn để xây dựng nên một chương trình cho riêng mình. Stanford POS Tagger hiện đã xử lý được rất nhiều ngôn ngữ khác nhau như: tiếng Anh, tiếng Ả Rập, tiếng Trung Quốc, … Tính từ lúc chương trình đầu tiên được giới thiệu vào năm 2004 cho tới thời điểm hiện tại, đã có hơn 25 phiên bản Stanford POS Tagger ra đời với sự cải tiến không ngừng về hiệu suất cũng như giải thuật. Trong đề tài này, chúng tôi chọn một phiên bản mới cập nhật là phiên bản 3.6.0 (được công bố vào ngày 09/12/2015) để được vào sử dụng trong luận văn.

Với đề tài Khai thác quan điểm của các bình luận tiếng Anh trên mạng xã hội sử dụng phương pháp xử lý ngôn ngữ tự nhiên, chúng tôi đặc biệt quan tâm đến các mô hình gán nhãn từ loại cho tiếng Anh. Với việc gán nhãn các từ loại chúng tôi đánh giá tối ưu quy trình khác thác tốt hơn. Chẳng hạn:

Khi chúng tôi phân tích câu bằng tiếng anh sau “Iphone is the best”

Mô hình phân tích câu được thể hiện trong chương trình chạy là:

Iphone/NNP is/VBZ the/DT best/JJS



Hình 2.1: Mô hình Pos tagger

Với việc sử dụng Pos tagger, chúng ta sẽ phân tích được danh từ, động từ và tính từ, qua đó đánh giá và suy xét các quan điểm của các khách hàng khi khác thác các quan điểm.

2.3 Phân lớp quan điểm

2.3.1 Giới thiệu phân lớp quan điểm

2.3.1.1 Khái niệm phân lớp quan điểm

Theo Huifeng Tang và cộng sự [5], phân lớp quan điểm bao gồm hai dạng phân lớp: phân lớp quan điểm nhị phân và phân lớp quan điểm đa lớp. Cho một tập văn bản cần đánh giá D = {d_1,d_2, ... ,d_n}và một tập đánh giá được xác định trước C = {tích cực (positive), tiêu cực (negative)}. Phân lớp quan điểm nhị phân là phân loại mỗi tài liệu $d_i \subset D$ vào một trong hai lớp: tích cực và tiêu cực. Nếu d thuộc lớp tích cực có nghĩa là tài liệu d thể hiện quan điểm tích cực. Ngược lại, d thuộc tiêu cực có nghĩa tài liệu d thể hiện quan điểm tiêu cực.

Ví dụ: Đưa ra một vài nhận xét về một bộ phim, hệ thống sẽ phân loại các nhận xét thành hai loại: nhận xét tích cực, nhận xét tiêu cực.

Để chuyển sang phân lớp quan điểm đa lớp, thiết lập tập C* = {tích cực mạnh (strong positive), tích cực (positive), trung lập (neutral), tiêu cực (negative), tiêu cực mạnh (negative strong)} và phân loại mỗi $d_i \subset D$ vào một trong các lớp trong C* .

2.3.1.2 Một số phương pháp phân lớp quan điểm

Bing Liu đưa ra ba phương pháp chính để phân lớp quan điểm.

- Phân lớp dựa vào cụm từ thể hiện quan điểm.
- Phân lớp dựa vào phương pháp phân lớp văn bản.
- Phân lớp dựa hàm tính điểm số.

a) Phân lớp dựa vào cụm từ thể hiện quan điểm

Phương pháp phân lớp dựa vào từ thể hiện quan điểm tích cực hay tiêu cực trong mỗi văn bản đánh giá. Thuật toán mô tả dựa trên nghiên cứu của Turney [3], được thiết kế để phân loại đánh giá của khách hàng. Thuật toán này sử dụng kỹ thuật xử lí ngôn ngữ tự nhiên gọi là gán nhãn từ loại (part-of-speech). Đánh dấu cho một từ được xác định bởi cú pháp ngữ nghĩa của nó. Các loại nhãn chung cho ngữ pháp tiếng Anh là: danh từ, động từ, tính từ, trạng từ, đại từ, giới từ, từ chuyển tiếp và thán từ. Có nhiều loại từ sinh ra từ các kiểu khác nhau của các loại này. Ở đây, ta có thể sử dụng bảng sau:

Bảng 2.1. Bảng các nhãn từ loại của Pennn Treebank

Thẻ	**Mô tả**	**Thẻ**	**Mô tả**
CC	Từ nối	PRP$	Đại từ sở hữu
CD	Số đếm	RB	Trạng từ
DT	Từ hạn định	RBR	Trạng từ, so sánh hơn
EX	Từ chỉ sự tồn tại	RBS	Trạng từ, so sánh hơn nhất

FW	Từ nước ngoài	RP	Tiền tố, hậu tố
IN	Giới từ, hoặc từ nối ngoài	SYM	Từ đại diện
JJ	Tính từ	TO	Trong
JJR	Tính từ, so sánh hơn	UH	Thán từ
JJS	Tính từ, so sánh hơn nhất	VB	Động từ, từ nguyên thể
LS	Danh sách mục đánh dâu	VBD	Động từ thì quá khứ
MD	Từ chỉ cách thức	VBG	Động từ, danh động từ, hiện tại hoàn thành
NNS	Danh từ, từ số nhiều	VBP	Động từ chia ở thời hiện tại thứ ba số nhiều
NNP	Đại từ, từ số ít	VBZ	Động từ chia ở thời hiện tại thứ ba số ít
NNPS	Đại từ số nhiều	WDT	Từ hạn định bắt đầu bằng Wh
PDT	Từ hạn định	WP	Đại từ bắt đầu bằng Wh
POS	Từ sở hữu	WP$	Đại từ sở hữu bắt đầu bằng Wh
PRP	Đại từ chỉ người	WRB	Trạng từ bắt đầu bằng Wh
NN	Danh từ, từ số ít	VBN	Động từ, từ quá khứ

Thuật toán mô tả dựa trên nghiên cứu của Turney [3] được trình bày chia làm ba bước sau:

Bước 1:

Trích chọn ra các cụm từ chứa tính từ hay trạng từ. Bởi vì, theo các nghiên cứu đã chỉ ra thì tính từ và trạng từ tốt để chỉ ra quan điểm, đánh giá chủ quan. Tuy nhiên, với một tính từ cô lập thể hiện chủ quan nhưng không đầy đủ ngữ cảnh thì khó xác định được hướng ngữ nghĩa của cụm từ đó.

Ví dụ: "không đoán trước được".

Trong câu "anh ta không đoán trước được cơ hội" thì mang hướng tiêu cực. Trong câu "hắn ta không đoán trước được âm mưu" thì mang hướng tích cực.

Do đó, thuật toán trích chọn hai từ liên tiếp trong đó một từ là tính từ hoặc trạng từ, từ kia thể hiện ngữ cảnh. Hai từ được trích chọn nếu nhãn của chúng phù hợp với bất kì các mẫu nào trong sau:

Bảng 2.2 Nhãn của mẫu cho trích chọn với cụm có hai từ

STT	Từ thứ nhất	Từ thứ hai	Từ thứ ba
1	JJ	NN hay NNS	Bất kỳ
2	RB, RBR, hay RBS	JJ	Không phải NN và NNS
3	JJ	J	Không phải NN và NNS
4	NN hoặc NNS	JJ	Không phải NN hoặc NNS
5	RB, RBR hoặc RBS	VB, VBD, VBN , VBG	Bất kỳ

Ví dụ: Xét câu: "This camera produces beautiful pictures" thì cụm từ "beautiful pictures" sẽ được trích chọn do khớp với mẫu 1

Bước 2:

Xác định xu hướng quan điểm của cụm từ thu được dựa trên độ đo pointwise mutual information (PMI).

Độ tương đồng ngữ nghĩa giữa hai cụm từ tính theo công thức sau:

$$PMI\ (term_1, term_2) = Log_2\ (\frac{\Pr\ (term_1 \wedge\ term_2)}{\Pr\ (term_1)\Pr\ (term_2)}) \quad (2.1)$$

Trong đó:

$Pr(term_1 \wedge term_2)$: xác suất đồng xuất hiện của $term_1$ và $term_2$.

$Pr(term_1)$, $Pr(term_2)$: xác suất mà $term_1$, $term_2$ xuất hiện khi thống kê chúng riêng rẽ.

Log của tỉ lệ trên là lượng thông tin mà ta có được về sự hiện diện của một term khi ta quan sát term kia. Xu hướng ngữ nghĩa hay quan điểm (SO) của một từ/cụm từ được tính dựa trên việc tính toán độ đo PMI của từ/cụm đó với 2 từ "excellent" và "poor" theo công thức sau:

$$SO(phrase) = PMI(phrase,"excellent") - PMI(phrase,"poor") \quad (2.2)$$

Sử dụng máy tìm kiếm để tính toán PMI như công thức (2.3)

• Pr(term): số kết quả trả về (hits) của máy tìm kiếm khi truy vấn là term. Thêm 0,01 vào hits tránh trường hợp chia cho 0.

• $Pr(term_1 \wedge term_2)$: Số kết quả trả về khi sử dụng máy tìm kiếm Alta Vista sử thêm toán tử NEAR.

$$SO(phrase) = \log_2(\frac{hits(pharseNEAR"excellent")hits("poor")}{hits(phraseNEAR"poor")hits("excellent")}) \quad (2.3)$$

Bước 3:

Với mỗi một đánh giá, hệ thống sẽ tính trung bình các chỉ số SO của tất các cụm từ trích chọn được và phân lớp chúng:

- Nếu chỉ số này dương thì xếp vào lớp pos.
- Nếu chỉ số này âm thì xếp vào lớp neg.

b) Phân lớp dựa vào phân lớp văn bản

Đây là phương pháp đơn giản nhất để giải quyết các bài toán phân lớp quan điểm dựa vào chủ đề. Sau đó, có thể áp dụng bất kì kỹ thuật học máy nào để phân lớp như Bayesian, SVM, KNN…

Cách tiếp cận này được thử nghiệm với Bo Pang và các cộng sự [4] áp dụng để đánh giá xem phim thành hai lớp tích cực hay tiêu cực. Bài toán chỉ ra việc sử dụng unigram trong phân lớp cho kết quả thực nghiệm tốt khi sử dụng Bayesian hoặc SVM.

Kết quả thực nghiệm qua sử dụng 700 đánh giá tiêu cực và 700 đánh giá tích cực cho thấy các thuật toán phân lớp đạt độ chính xác là 81% và 82% tương ứng với hai thuật toán Baysian hoặc SVM. Tuy nhiên, việc đánh giá trung lập không được đề cập trong bài báo.

c) Phân lớp dựa vào hàm tính điểm số

Phương pháp phân lớp dựa vào tính điểm được Kushal Dave và cộng sự [5] đưa ra gồm hai bước sau:

Bước 1:

Tính điểm các từ trong văn bản của tập dữ liệu học theo công thức (2.4)

$$score(t_i) = \frac{\Pr(t_i|C) - \Pr(t_i|C')}{\Pr(t_i|C) + \Pr(t_i|C')} \quad (2.4)$$

Trong đó:

t_i là từ cần được tính điểm.

C là một lớp quan điểm; C’ là lớp phần bù của C (not C).

Pr(t|C): xác suất t xuất hiện ở lớp C, được tính bằng số lần xuất hiện của t trong lớp C.

Điểm số được chuẩn hóa trong khoảng [-1, 1].

Bước 2:

Một văn bản mới $d_i = t_1 \ldots t_n$ sẽ được phân lớp theo công thức (2.5) sau:

$$class(d_i) = \begin{cases} C & eval(d_i) > 0 \\ C' & otherwise, \end{cases} \quad (2.5)$$

Với

$$eval(d_i) = \sum_j Score\ (t_i)$$

Kết quả:

- Kiểm thử trên 13000 đánh giá của 7 sản phẩm
- Bigrams và trigrams cho kết quả chính xác cao nhất, từ 84.6% tới 88.3%

2.3.1.3 Phân lớp dựa vào kỹ thuật học máy

Có rất nhiều phương pháp có thể xác định tính phân cực của một tài liệu. Trong [2], Huifeng Tang và cộng sự đề cập phương pháp sử dung kỹ thuật học máy để xác định tích cực hay tiêu cực của bình luận với việc chuẩn bị dữ liệu học bằng tay. Các tác giả cũng đưa ra hai vấn đề quan trọng khi phân lớp quan điểm dựa vào kỹ thuật học máy: trích chọn đặc trưng và huấn luyện bộ phân lớp.

a) Trích chọn đặc trưng

Với tập dữ liệu thô (tập trang web để phân lớp quan điểm), thực hiện tách các thẻ HTML, chia văn bản thành các câu. Các câu này được chạy qua bộ phân tích trước khi chia nhỏ thành các từ đơn. Có một số phương pháp để trích chọn đặc trưng, chẳng hạn như dựa vào từ vựng. Có hai phương pháp dựa vào từ vựng được sử dụng như sau:

Thứ nhất là dựa trên từ điển WordNet. WordNet thay thế các từ trong đánh giá bởi một tập từ đồng nghĩa chung với nó có trong WordNet. Bởi vì các từ trong các bình luận có thể không phải là từ phổ biến trong đánh giá. Có rất nhiều nghiên cứu sử dụng kỹ thuật này, nghiên cứu gần đây nhất là của Taboada [6]. Nghiên cứu mô tả và so sánh các phương pháp để tạo bộ từ điển tương ứng với ngữ nghĩa định hướng của nó (SO). Qua thực nghiệm, tác giả cho thấy hiệu quả của việc sử dụng từ điển để xác định hướng ngữ nghĩa cho văn bản. Để trích chọn riêng mỗi từ, tác giả sử dụng

một phương pháp dựa trên độ đo thông tin lẫn nhau (PMI). Thông tin lẫn nhau giữa một tập từ mồi và tập từ mục tiêu được tính toán bằng cách sử dụng hai phương pháp khác nhau: một là tìm kiếm NEAR dựa vào kỹ thuật tìm kiếm Altavista, hai là AND tìm kiếm trên Google. Hai tập từ điển được thử nghiệm với từ điển được gán nhãn tích cực, tiêu cực bằng tay. Kết quả của ba phương pháp khá gần nhau và không một phương pháp nào trong số họ thực sự có hiệu quả đặc biệt. Bài báo cũng chỉ ra hướng nghiên cứu tiềm năng trong việc tính toán độ đo thông tin lẫn nhau PMI bằng cách sử dụng Google.

Thứ hai là sử dụng bộ gán nhãn nhằm phát hiện ra từ và các cụm từ thể hiện quan điểm. Dựa vào tập từ gán nhãn cho các từ thể hiện quan điểm, ta loại được tập các từ không thể hiện quan điểm. Đây là tập dữ liệu nhiễu. Bộ gán nhãn được phát triển để giảm bớt dữ liệu nhiễu.

Đánh giá tính từ: Phương pháp đánh giá tính từ tập trung vào trích chọn và phân tích nhóm đánh giá tính từ bởi tính từ chính như đẹp, buồn… và tùy chọn thay đổi bởi một chuỗi các từ sửa đổi như rất, không, hơi… Phương pháp phân tích chi tiết hơn về ngữ nghĩa của câu thể hiện quan điểm, đánh giá với các nhóm tính từ đặc biệt như "vô cùng nhàm chán", "không thực sự tốt", … Phương pháp này gồm 4 bước:

- Xây dựng bộ tập từ vựng sử dụng kỹ thuật bán tự động, thu về, phân loại tính từ và tập từ bổ nghĩa để phân loại một số các thuộc tính đánh giá.
- Trích xuất nhóm tính từ đánh giá từ văn bản và tính toán các giá trị thuộc tính theo từ vựng đó.
- Coi biểu diễn của một văn bản như là vector đặc trưng tần suất tương đối sử dụng các nhóm này.
- Sử dụng máy hỗ trợ vector SVM để phân biệt hướng tích cực hay tiêu cực của tài liệu.

Beineke và cộng sự [7] mở rộng phương pháp này bằng cách trích chọn tập các

đặc trưng được kết hợp tuyến tính để xác định hướng quan điểm. Với phương pháp này có thể nâng cao kết quả so với phương pháp gốc, chủ yếu thông qua hai chiến lược: tích hợp các đặc trưng bổ sung vào mô hình và có thể sử dụng gán nhãn dữ liệu để đánh giá ảnh hưởng của chúng trong ngữ cảnh. Đặc biệt, khóa luận tập trung vào phương pháp của Takamura và cộng sự [8] sử dụng kỹ thuật trích chọn đặc trưng chuỗi từ con và cây con phụ thuộc của câu dựa trên thuật toán tính tần suất khai phá mẫu và kết hợp sử dụng máy hỗ trợ vector.

b) Huấn luyện bộ phận lớp SVM nhị phân

Bài toán gốc của phân lớp quan điểm là bài toán phân lớp văn bản. Có thể coi phân lớp quan điểm là bài toán phân lớp văn bản theo hai lớp tích cực và tiêu cực. Do đó một số kỹ thuật phân lớp văn bản như K người láng giềng gần nhất, Naïve Bayes, Maximum entropy và SVM có thể sử dụng trong phương pháp học máy phân lớp quan điểm.

Mặt khác, trong số các công cụ trên , SVM được chứng minh là công cụ phân lớp mạnh, hiệu quả hơn phân lớp văn bản truyền thống như Naïve Bayes [9]. Thêm vào đó, B.Pang và các cộng sự [4] áp dụng kỹ thuật Naïve Bayes, maximum entropy và SVM để xác định hướng quan điểm phân cực trong bình luận về phim. Kết quả phân lớp sử dụng mô hình unigram và phân lớp SVM đạt kết quả cao nhất 82.9%. Điều đó cho ta thấy rằng SVM vẫn là một công cụ hiệu quả cho phân lớp quan điểm.

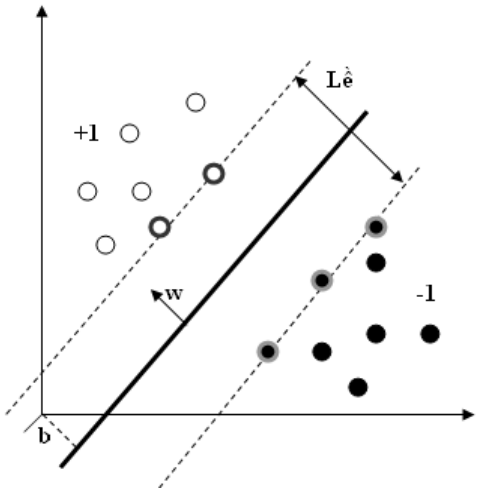
SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis [10] xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dựng trong thực tiễn. Tư tưởng chung của học máy SVM như sau:

- Giai đoạn xây dựng mô hình: Cho một tập mẫu dữ liệu huấn luyện đã được gán nhãn lớp, như vậy có một tập nhãn lớp tương ứng xác định tên tập mẫu. Mỗi mẫu dữ liệu được biểu diễn dưới dạng một vector đặc trưng. Dựa vào vector đặc trưng của các mẫu dữ liệu huấn luyện, mô hình máy vector hỗ trợ sẽ được xây dựng để phân tách các mẫu học. Trong trường hợp khả tách tuyến tính, nó là

một siêu phẳng (hyperplane) trong không gian dùng để phân tách tuyến tính các mẫu thuộc các nhãn lớp khác nhau với khoảng cách lớn nhất có thể. Trong trường hợp không khả tách tuyến tính, chúng ta có thể sử dụng lề mềm (soft margin) để phân tách mẫu học, hay sử dụng ánh xạ phi tuyến để chuyển không gian ban đầu sang không gian mới có số chiều lớn hơn mà ở đó các mẫu học có khả năng phân tách tuyến tính.

- Giai đoạn sử dụng mô hình: Mô hình đã xây dựng sẽ được sử dụng để gán nhãn lớp cho các mẫu dữ liệu mới.

b.1) Trường hợp khả tách tuyến tính



Hình 2.2 Mô hình máy vector hỗ trợ khả tách tuyến tính

Đầu vào của thuật toán là một tập dữ liệu huấn luyện, mỗi mẫu được đánh dấu rơi vào một trong hai lớp gọi chung là lớp mẫu âm (negative) và lớp mẫu dương (positive). Đầu ra của mô hình là một mặt siêu phẳng phân tách các mẫu dương và mẫu âm với khoảng cách lề cực đại.

Thuật toán SVM được mô tả cụ thể như sau: Cho 1 tập huấn luyện các cặp (x_i, y_i), với i = 1, …, l; trong đó $x_i \in R^n$ là không gian vector đặc trưng n chiều; $y_i \in \{-1, +1\}$, các mẫu dương là các mẫu x_i thuộc lĩnh vực quan tâm và được gán nhãn y_i =

+1, các mẫu âm là các mẫu x_i không thuộc lĩnh vực quan tâm và được gán nhãn y_i = −1.

Trong trường hợp này, bộ phân lớp SVM là một siêu phẳng phân tách tập mẫu dương khỏi tập mẫu âm với độ chênh lệnh cực đại. Độ chênh lệch cực đại này còn gọi là lề của siêu phẳng (margin). Lề xác định khoảng cách giữa các mẫu dương với mẫu âm gần mặt siêu phẳng nhất (chính là khoảng cách giữa các mẫu nằm trên 2 đường nét đứt tới đường nét đậm). Các mặt siêu phẳng trong không gian đối tượng có phương trình là $w^Tx + b = 0$, trong đó w là vector pháp tuyến, b là tham số mô hình phân lớp (bộ phân lớp). Khi thay đổi w và b, hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi. Các giá trị khác nhau của lề cho ta các họ mặt siêu phẳng khác nhau, và lề càng lớn thì lỗi tổng quát hóa của bộ phân lớp càng giảm. Thuật toán SVM ước lượng các tham số w và b nhằm tìm ra mặt siêu phẳng phân tách lớp mẫu dương khỏi lớp mẫu âm với lề cực đại. Mặt siêu phẳng này còn được gọi là mặt siêu phẳng lề tối ưu hay ranh giới quyết định (decision boundary), hoặc là lề cứng (hard margin).

Bộ phân lớp SVM được định nghĩa như sau:

$$f(x) = \text{sign}(w^Tx + b) \qquad (2.6)$$

Trong đó :

$$\text{sign}(z) = +1 \text{ nếu } z \geq 0,$$

$$\text{sign}(z) = -1 \text{ nếu } z < 0.$$

Nếu f(x) = +1 thì x thuộc về lớp dương, và ngược lại, nếu f(x) = −1 thì x thuộc về lớp âm.

Tập dữ liệu huấn luyện là khả tách tuyến tính, ta có các ràng buộc sau :

$$w^T x_i + b \geq +1 \text{ nếu } y_i = +1 \qquad (2.7)$$

$$w^T x_i + b \leq -1 \text{ nếu } y_i = -1 \qquad (2.8)$$

Hai mặt siêu phẳng có phương trình là $w^Tx + b = \pm 1$ được gọi là các mặt siêu phẳng hỗ trợ (các đường nét đứt trên hình).

Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán cực đại hóa như sau:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Với các ràng buộc

$$a_i \geq 0$$

$$và \sum_{i=1}^{N} \alpha_i y_j = 0$$

Trong đó các hệ số Lagrange α_i, i = 1, 2, ..., N, là các biến cần được tối ưu hóa.

b.2 Trường hợp không khả tách tuyến tính

Có thể giải quyết theo 2 phương pháp sau:

Cách thứ nhất sử dụng một mặt siêu phẳng lề mềm, nghĩa là cho phép một số mẫu huấn luyện nằm về phía sai của mặt siêu phẳng phân tách hoặc vẫn ở vị trí đúng nhưng rơi vào vùng giữa mặt siêu phẳng phân tách và mặt siêu phẳng hỗ trợ tương ứng. Trong trường hợp này, các hệ số Lagrange của bài toán quy hoạch toàn phương có thêm một cận trên C dương – tham số do người sử dụng lựa chọn. Tham số này tương ứng với giá trị phạt đối với các mẫu bị phân loại sai.

Cụ thể, tập dữ liệu huấn luyện là khả tách tuyến tính, ta có các ràng buộc sau:

$$w^T x_j + b \geq +1 - \varepsilon \; nếu \; y_j = +1 \qquad (2.9)$$

$$w^T x_j + b \leq -1 + \varepsilon \; nếu \; y_j = -1 \qquad (2.10)$$

$$\varepsilon \geq 0$$

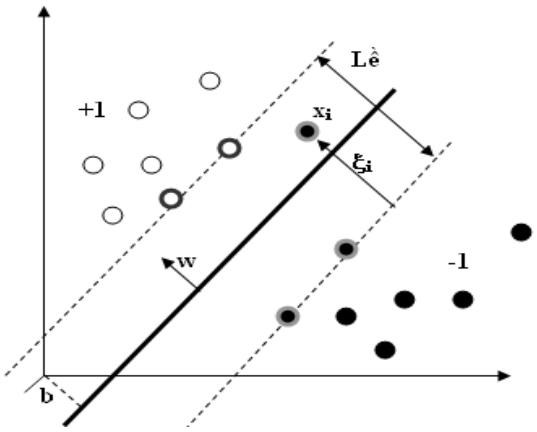
Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán cực đại hóa như sau:

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

với các ràng buộc

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C$$



Hình 2.3 Phương pháp lề mềm

Cách thứ hai sử dụng một ánh xạ phi tuyến Φ để ánh xạ các điểm dữ liệu đầu vào sang một không gian mới có số chiều cao hơn.

$$\emptyset: R^d \to R^D (D \gg d) x \to \emptyset(x)$$

Trong không gian này, các điểm dữ liệu trở thành khả tách tuyến tính, hoặc có thể phân tách với ít lỗi hơn so với trường hợp sử dụng không gian ban đầu. Một mặt quyết định tuyến tính trong không gian mới sẽ tương ứng với một mặt quyết định phi tuyến trong không gian ban đầu. Khi đó, bài toán ban đầu sẽ trở thành cực đại hóa như sau:

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j)$$

với các ràng buộc:

$$0 \leq \alpha_i \leq C$$
$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Trong đó k là một hàm nhân thoản mãn:

$$k(x_i, x_j) = \emptyset(x_i)^T . \emptyset(x)$$

Với việc dùng một hàm nhân, ta không cần biết rõ về ánh xạ Φ. Hơn nữa, bằng cách chọn một nhân phù hợp, ta có thể xây dựng được nhiều bộ phân lớp khác nhau.

2.3.2 Thuật toán tính tần suất mẫu

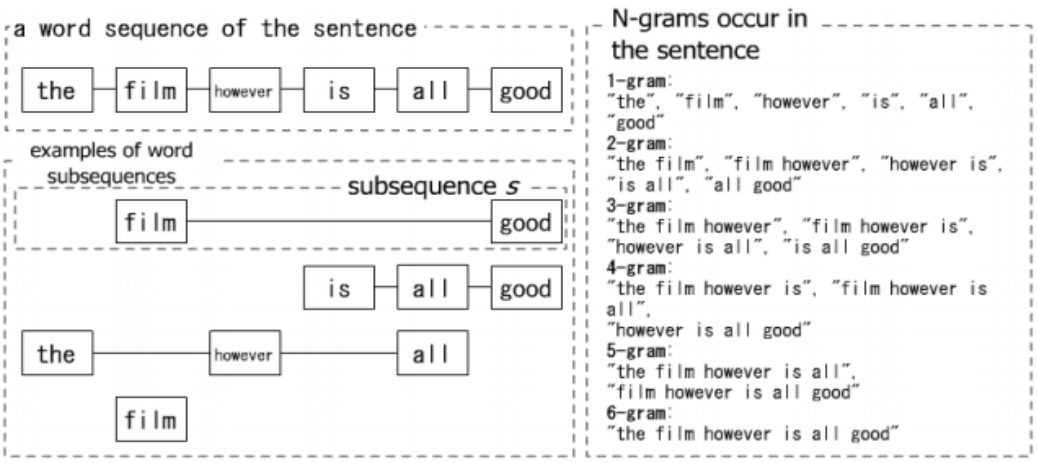
2.3.2.1 Chuỗi từ con

Chuỗi từ con (word subsequence): là một thể hiện cấu trúc của một câu. Từ chuỗi từ, ta có thể xác định được thứ tự của từ trong một câu.

Định nghĩa: Một chuỗi từ con của một chuỗi từ coi như một chuỗi thu được bởi không loại bỏ hay loại bỏ một hoặc nhiều từ trong một câu gốc. Trong chuỗi từ con, thứ tự các từ vẫn được giữ nguyên như trong câu gốc.

Trong khi n-grams chỉ thể hiện sự đồng xuất hiện của n từ liên tục trong một câu, chuỗi từ con thể hiện sự đồng xuất hiện của một số lượng bất kì các từ không liên tục cũng như liên tục. Do đó, sự kết hợp của các chuỗi con vào phân lớp là hiệu quả.

Ví dụ: N-grams không thể hiện được sự đồng xuất hiện của “film” và “good”, khi một từ khác xuất hiện giữa hai từ như trong hình (2.4). Ngược lại, với chuỗi con luôn chứa mẫu “film-good”, được chú ý bởi s trong hình.



Hình 2.4 Một ví dụ chuỗi con trong câu "The film however is all good"

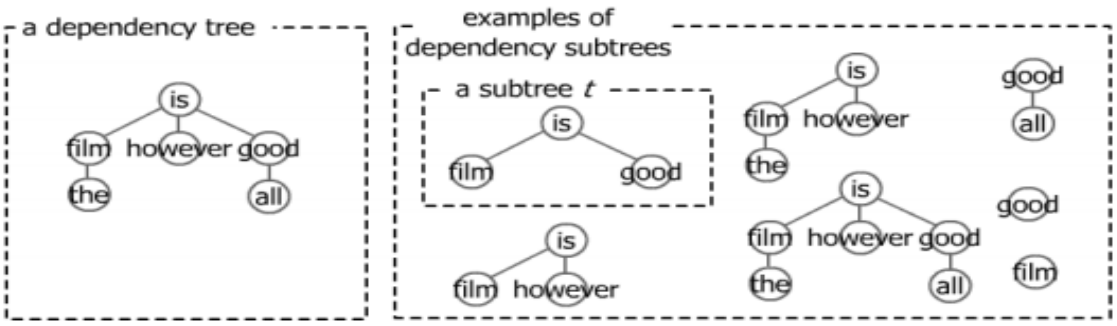
2.3.2.2 Cây con phụ thuộc

Một cây phụ thuộc là một thể hiện cấu trúc của tài liệu. Cây phụ thuộc thể hiện sự phụ thuộc của các từ trong một câu bởi quan hệ cha con giữa các nút.

Định nghĩa: Cây con phụ thuộc coi như một cây thu được bởi không loại bỏ hoặc loại bỏ một hay nhiều nút và nhánh từ cây gốc.

Các cây con phụ thuộc giữ được sự phụ thuộc giữa các từ trong một câu gốc. Vì mỗi nút tương ứng với một từ được kết nối bởi một nhánh, cây con phụ thuộc cung cấp thông tin giàu ngữ nghĩa hơn n-gram và một chuỗi từ.

Ví dụ: Trong hình (2.5), thể hiện quan hệ giữa các từ "good" và "film", sự phụ thuộc cây con t (được chú ý như là is((film)(good))) không chỉ thấy được sự đồng xuất hiện của từ "good" và "film", mà còn bảo đảm "good" và "film" được kết nối cú pháp với nhau qua từ "is".



Hình 2.5 Một ví dụ cây con phụ thuộc trong câu “The film however is all good”

2.3.2.3 Thuật toán tính tần suất mẫu

Vì số lượng của các sub-patterns của câu trong tài liệu là lớn. Vì vậy, ở đây ta không quan tâm đến tất cả sub-patterns nhưng mà chỉ quan tâm đến tần suất của các sub- patterns. Một câu chứa một mẫu khi và chỉ khi mẫu đó là một chuỗi con hoặc một cây con trong câu.

Định nghĩa: Độ hỗ trợ của một mẫu con (support of sub-pattern) là số lượng các câu chứa mẫu con đó. Nếu độ hỗ trợ của một mẫu con đạt đến ngưỡng hỗ trợ (support threshold) hoặc lớn hơn thì mẫu con đó là thường xuyên (frequent).

a) Tần suất khai phá chuỗi con

Giả sử có tất cả các items của một thành phần được sắp xếp theo thứ tự a,b,c. Với một chuỗi $\alpha = e1e2e3.....en$ và một chuỗi $\beta = e'1e'2e'3...e'm$ $(m \leq n)$ là tiền tố (prefix) của α khi và chỉ khi:

- $ei' = ei$ for$(i \leq m-1)$
- $em' \subseteq em$
- Tất cả các items trong $(em - e'm)$ được sắp xếp sau $e'm$

Với một chuỗi con α và β như thế, β là chuỗi con của α kí hiệu β ô α. Một chuỗi con α’ của chuỗi α gọi là hình chiếu (projection) của α tương ứng với tiền tố của β khi và chỉ khi:

- α' có tiền tố β.

- Không tồn tại chuỗi α’’ nào là tiền tố của β mà lớn hơn α’.

Với a’=e1e2e3.....en là hình chiếu (projection) của α tương ứng với tiền tố β = e1e2e3.....em-1e'm.

Chuỗi con γ =(e"mem + 1...en) gọi là hậu tố của α tương ứng với tiền tố của β khi $\gamma = \alpha / \beta$ với e"m =$(em - e'm)^2$.

b) Tần suất khai phá cây con

Thuật toán freqt tính tần suất của tất cả các cây con trong một cây được Kenji Abe và cộng sự. Đầu tiên, thuật toán bắt đầu với một tập hợp tần suất của các cây con gồm các từ đơn (single node). Sau đó, thuật toán được mở rộng, với mỗi cây con có kích thước k gắn thêm một từ mới để tính được tần suất của cây con có kích thước k+1. Thuật toán tính được tất cả tần suất của chuỗi con thông qua lặp đệ quy.

Tuy nhiên, việc mở rộng cây con bằng cách thêm một nút mới vào bất kì vị trí của lá có thể dẫn đến tình trạng trùng lặp các cây con mới được sinh ra. Để tránh điều này, thuật toán hạn chế vị trí đính kèm một nút mới vào cuối cây con mới theo ưu tiên độ sâu.

CHƯƠNG 3
CÁC CÔNG TRÌNH LIÊU QUAN

3.1 Khái quát chung

Trong thời đại bùng nổ công nghệ thông tin như hiện nay, việc các công ty hay một tổ chức tự xây dựng cho mình một trang web là rất phổ biến, trang Web này như là một kênh thông tin riêng hoặc thể hiện tiềm năng của họ. Ở đó, họ có thể giới thiệu những gì mà họ đang làm như là các sản phẩm, các dịch vụ, các cuộc thảo luận về vấn đề xã hội, buôn bán, thu thập ý kiến của khách hàng [11].

Ngoài ra, trong tài liệu nghiên cứu [11] tác giả đề cập tới một ý kiến khác với những nghiên cứu trên mà tác giả gọi là ý kiến dự đoán. Ta đã nghe khá nhiều về những ý kiến dự đoán về tương lai của một chủ đề như thị trường bất động sản, kết quả của các trận đấu bóng đá hay là các cuộc bầu cử. Những dự đoán này thường dựa trên niềm tin và kiến thức của người đưa ra dự đoán là chính (thường là các chuyên gia). Ví dụ trong câu: "Giá của bất động sản sẽ được giảm xuống trong vài tháng tới", như ta thấy đây là một câu dự đoán ở tương lai về thị trường bất động sản và trong câu này nó cũng thể hiện mặt tích cực trong vế "giá bất động sản sẽ giảm".

Do đó, để đưa ra được các dự đoán về thời tiết, động đất, sóng thần, thì các tác giả cần một cơ sở dữ liệu số (số liệu) rất lớn, tiến hành phân tích trên những con số đó. Tuy nhiên, trong bài [11] không đề cập về những phân tích dự đoán dựa trên việc phân tích số liệu mà tác giả muốn đề cập đến việc phân tích những câu phát biểu dự đoán trong các đoạn văn bản không có cấu trúc xác định dựa trên kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), từ đó đưa ra kết luận về mức độ tích cực của từng dự đoán như ví dụ ở trên.

Cụ thể, trong bài [11] tác giả đã thí nghiệm trên tập các ý kiến dự đoán trên Web của một cuộc bầu cử. Mục đích của tác giả là phân tích tự động tập ý kiến rất lớn từ những người dùng đã bày tỏ trên web, từ đó rút ra được Đảng nào đang chiếm ưu thế

về niềm tin của người dân và đưa ra được con số phần trăm thắng cử của mỗi Đảng tham gia vào cuộc bầu cử này. Vì thế tác giả đã đưa ra mô hình đánh giá như sau:

ElectionPredictionOpinion = (Party, Valence).

- ElectionPredictionOpinion**:** Kết quả đánh giá cho mỗi Đảng.
- Party**:** là Đảng mà người bình luận đang muốn nói đến.
- Valence**:** là tỷ lệ phần trăm thắng cử mà hệ thống tính toán được.

Tác giả cũng xây dựng hệ thống này gồm có ba bước chính được huấn luyện bởi học máy sử dụng chức năng n-gram và SVM (Support Vector Machine) để lượng giá:

1. Chức năng tổng quát hóa: chức năng này làm công việc xác định những ai liên quan đến Đảng đang nói đến trong câu thì qui về tên của Đảng mà người đó đang làm việc.

2. Phân loại và đánh giá từng câu ý kiến dự đoán sử dụng kỹ thuật SVM dựa trên mô hình (Party, Valence).

3. Tổng hợp lại tất cả các phân loại mà bước 2 đã làm và đưa ra được kết quả là phần trăm của Đảng được dự đoán thắng cuộc.

Như vậy, ưu điểm của hệ thống trên là việc tổng quát hóa các đối tượng liên quan lại thành một đối tượng để đánh giá, giúp cho việc đánh giá được chính xác hơn.

Khai thác ý kiến và phân tích tình cảm là nhánh con của NLP và khai thác văn bản (text mining) nhằm mục đích khám phá và khai thác tự động tri thức về tình cảm con người, sự đánh giá những ý kiến từ những dữ liệu văn bản gốc như là trang nhật ký của một người nào đó, những nhận xét trên website và trong những bản phản hồi từ khách hàng.

Ngoài ra, trong các bài nghiên cứu [12], [13], [14], [15], [16], [17], [18], các tác giả thường phân tích và khai thác các ý kiến đánh giá của khách hàng trên các sản phẩm như là điện thoại, máy ảnh, máy tính, sách, phim ảnh. Những ý kiến này thường thể hiện là thích hay không thích các sản phẩm đó, người ta gọi những ý kiến này là

những ý kiến nhận định hoặc phán xét. Tuy nhiên những ý kiến này không phải là tiếng Việt.

Những tác giả trong công trình [19] đã đề xuất một phương pháp xây dựng một miền Ontology tự động từ một mạng ngữ nghĩa ConceptNet. ConceptNet này được xây dựng bởi các tình nguyện viên trên thế giới. Và kết quả của công trình trên có thể sử dụng như nguồn từ vựng hoặc thực hiện xác định các mục tiêu để phân tích tình cảm trong thời gian đó.

Khác với những công trình trước (được đề cập đến trong 5.1), trong công trình nghiên cứu [19] các tác giả đã đưa ra được giải pháp cải tiến hơn so với những giải pháp truyền thống. Sự khác biệt trong giải pháp của tác giả so với những giải pháp khác là tác giả đã đề xuất một mạng ngữ nghĩa của common-sense knowledge-base (ConceptNet) để tạo tự động ra một miền ontology của những tính năng của sản phẩm và các thuộc tính của nó. Những công trình trước thường đánh giá chung những danh sách tính năng của sản phẩm, trong khi đó tác giả tạo ra một ontology ở đó những tính năng của sản phẩm như một khái niệm hoặc là những nút trên mạng ngữ nghĩa được kết nối với những nút khác sử dụng nhiều kiểu của quan hệ ngữ nghĩa (semantic relationship - theo nhiều kiểu khác nhau hoặc nhiều mối quan hệ khác nhau). Vì thế mà sản phẩm sẽ xuất hiện trong miền ontology và từ vựng đã tạo ra từ cách này mang ngữ nghĩa phong phú hơn là từ vựng.

Tác giả tận dụng ConceptNet để xây dựng miền ontology. Các nút thể hiện các khái niệm, các cạnh thể hiện các tính chất, thuộc tính, quan hệ. Tác giả không những đưa ra những mối quan hệ ngữ nghĩa giống như IsA, HasA mà còn nhiều mối quan hệ khác CreatedBy, MadeOf, PartOf, DesireOf và DefineAs. Phạm vi của ConceptNet là những tri thức chung và không giới hạn bởi một miền cụ thể nào, rất là hữu dụng trong việc khai thác những câu bình luận về những tính năng của sản phẩm.

Với mong muốn đáp ứng được tính ưu biệt khi phân tích tình cảm, cảm xúc của khác hàng, với việc sử dụng (NLP) kỹ thuật trong một nỗ lực để cải thiện độ chính xác

phân loại và để phân loại một tập kiểm tra 100 văn bản ngôn ngữ tự nhiên và các kết quả so với một bộ phân loại cơ sở, để xác định xem việc sử dụng ngôn ngữ cơ bản cấu trúc có thể cung cấp một lợi thế trong phân loại tình cảm.

Với việc sử dụng NLP đã cải thiện đáng kể, trong việc phân loại ý kiến, bằng việc sử dụng các kỹ thuật khác nhau như: Gắm nhẫn từ loại (POS tagging), phân tích tính phụ thuộc bằng cách phân tách cấu trúc cây, xử lý các phép phủ định trong câu. Với những kỹ thuật mới mang tính ưu biệt đã phân giải và tăng độ chính xác của việc phân loại ngôn ngữ trong câu.

Ngoài ra với việc sử dụng công cụ POS tagging đã mang tới việc phân tích có độ chính xác cao hơn, bằng việc sử dụng các thuật toán Pos tagging mà kết hợp từ ngữ rời rạc, cũng như các phần ẩn của bài phát biểu, phù hợp với một bộ các thẻ mô tả. Qua đó dễ dàng đánh giá các ý kiến có tính ưu biệt cao hơn.

Với việc sử dụng các tập huấn luyện từ các thí nghiệm trên tập dữ liệu gồm 100 câu bình luận mà chúng tôi đã thực hiện được. Với việc thu thập các dataset trên các trang mạng xã hội khác nhau, chúng tôi xây dựng các dataset với các chủ đề khác nhau và sau đó thực thi với tập được huấn luyện, tiến hành thử nghiệm trên tập huấn luyện này để rút ra điểm số đáng tin cậy nhất cho việc xác định điểm số trung tính cho thực thể. Với các thực hiện này chúng tôi xây dựng mô hình thực thể, đánh giá và đưa ra kết luận là tích cực (positive), tiêu cực (negative), trung tính (neutral).

Phương pháp sử dụng Ontology kết hợp từ điển để xác định mức độ thể hiện tình cảm trong nhận xét của khách hàng. Với phương pháp này xác định được những thực thể xuất hiện trong miền Ontology. Sau đó, dựa vào tập từ điển để xác định các từ thể hiện tình cảm và điểm số của chúng (tính từ, động từ, trợ từ, phủ định từ). Từ đó, chúng tôi tiến hành đánh giá điểm số, tổng hợp và đưa ra được kết luận tích cực, tiêu cực hay trung tính cho từng thực thể tương ứng. Qua việc sử dụng phương pháp đánh giá theo Ontolory của một thực thể ở mỗi bình luận cao hơn hai tập dữ liệu còn lại dựa trên đánh giá điểm số cho từng thực thể tương ứng ở cả hai độ đo được đề cập ở trên.

Việc khai thác các quan điểm của chúng tôi, chủ yếu tập trung vào xác định và đánh giá những ý kiến của khách hàng thông qua các trang mạng xã hội. Trong tương lai, nếu có điều kiện chúng tôi muốn mở rộng ra nhiều trang khác nhau, có các bình luận trên các sản phẩm khác.

Ngoài ra, chúng tôi chỉ đánh giá cho các bình luận thuộc thể loại câu khẳng định, phủ định và so sánh, mà chưa đề cập đến những mẫu câu khác như câu nhân hóa, câu điều kiện, và câu cảm thán. Vì thế đây là một vấn đề cần nghiên cứu thêm để xác định được quan điểm đúng của các câu thuộc các thể loại này.

3.2 Các công trình liên quan

3.2.1 Các công trình sử dụng NLP

Xử lý ngôn ngữ chính là xử lý thông tin khi đầu vào là "dữ liệu ngôn ngữ" (dữ liệu cần biến đổi), tức dữ liệu "văn bản". Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) đang dần trở thành kiểu dữ liệu chính con người có và lưu trữ dưới dạng điện tử. Đặc điểm chính của các kiểu dữ liệu này là không có cấu trúc hoặc nửa cấu trúc và chúng không thể lưu trữ trong các khuôn dạng cố định như các bảng biểu. Theo đánh giá của công ty Oracle, hiện có đến 80% dữ liệu không cấu trúc trong lượng dữ liệu của loài người đang có [20]. Với sự ra đời và phổ biến của Internet, của sách báo điện tử, của máy tính cá nhân, của viễn thông, của thiết bị âm thanh,… Người người ai cũng có thể tạo ra dữ liệu văn bản. Vấn đề là làm sao ta có thể xử lý chúng, tức chuyển chúng từ các dạng ta chưa hiểu được thành các dạng ta có thể hiểu và giải thích được, tức là ta có thể tìm ra thông tin, tri thức hữu ích cho mình.

Tuy nhiên, một văn bản thật sự (một bài báo khoa học chẳng hạn) có thể có đến hàng nghìn câu và ta không phải có một mà có hàng triệu văn bản. Web là một nguồn dữ liệu văn bản khổng lồ, và cùng với các thư viện điện tử − khi trong một tương lai gần các sách báo xưa nay và các nguồn dữ liệu được chuyển hết vào máy tính (chẳng hạn bằng các chương trình nhận dạng chữ, thu nhập âm thanh, hoặc gõ thẳng vào máy tính) - sẽ sớm chứa hầu như toàn bộ kiến thức của nhân loại. Vấn đề là làm sao "xử lý"

(chuyển đổi) được khối dữ liệu văn bản khổng lồ này qua dạng khác để mỗi người có được thông tin và tri thức mà bản thân cần chúng.

Xử lý ngôn ngữ tự nhiên đã được ứng dụng trong thực tế để giải quyết các bài toán như: nhận dạng chữ viết, nhận dạng tiếng nói, tổng hợp tiếng nói, dịch tự động, tìm kiếm thông tin, tóm tắt văn bản và khai phá dữ liệu và phát hiện tri thức.

Với các công trình nghiên cứu về xử lý ngôn ngữ tự nhiên con người đã phân tích và đưa ra yêu cầu cho máy tính hiểu, Với những kỹ thuật và sự nghiên cứu không giới hạn từ các nhà khoa học xưa và nay, đã phần nào giúp máy tính chúng ta gần với con người hơn. Với việc sử dụng khía cạnh khai thác ngôn ngữ, đánh giám dự đoán theo phương pháp xử lý ngôn ngữ tự nhiên đã phần nào xây dựng nên những trang sử về khai thác các ngôn ngữ từ các nguồn khác nhau như: tiếng Anh, tiếng Pháp, tiếng Trung,...

Một công trình nghiên cứu được đánh giá là thể hiện tính ưu biệt trong việc xử lý ngôn ngữ tự nhiên, đó là công trình nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng việt. Đã phần nào cho chúng ta nhìn nhận được cách thức cũng như các kỹ thuật đã được sử dụng một cách hiểu quả trong quá trình dùng phương pháp xử lý ngôn ngữ tự nhiên. [21] với công trình này đã mang lại giá trị khả thi trong việc xử lý ngôn ngữ tự nhiên.

Các phương pháp tổng hợp quan điểm đã nhận được nhiều sự quan tâm của các nhà nghiên cứu trên thế giới. Trong [22], AurélienBossard và cộng sự đã tiếp cận hướng tổng hợp quan điểm trên nhiều tài liệu sử dụng học máy SVM, với dữ liệu là các bài viết trên blog. Trong [12], Hu và Liu đề xuất phương pháp tổng hợp đánh giá người dùng về sản phẩm bằng cách biểu diễn những quan điểm tích cực/tiêu cực của người dùng về những đặc trưng của sản phẩm. Trong [19], Amitava Das và cộng sự đã đưa ra phương pháp tổng hợp quan điểm dựa trên chủ đề sử dụng từ điển Bengali Senti WordNet.

Nhiều kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được áp dụng, các phương pháp học máy như phân lớp Naïve Bayes, cực đại hóa entropy và SVM được nghiên cứu và áp dụng thử nghiệm. Tuy nhiên, trong [4], Lang, Pee và Vaithyanathan đã chứng minh rằng các kỹ thuật NLP không thực hiện tốt như giải thuật phân lớp dựa trên chủ đề truyền thống. Trong [23, 24], các tác giả đã đưa ra mô hình và phương pháp tổng hợp quan điểm dựa trên truy vấn dưới hai góc độ tiếp cận và học máy SVM. Kết quả nghiên cứu của hai nhóm này là cơ sở quan trọng để chúng tôi phân tích và đưa ra được mô hình áp dụng phù hợp.

3.2.2 Sử dụng máy học

Các thông tin văn bản trên thế giới hiện nay được phân làm hai loại chính: Những thông tin về sự kiện (hoặc thực thể) và những ý kiến. Ý kiến của một người là thể hiện tình cảm hoặc đánh giá của người đó về các thực thể, sự kiện hoặc các thuộc tính của chúng.

Hệ thống phân tích ý kiến là một phương pháp quan trọng để xác định tự động các quan điểm của những bình luận trên internet như website, diễn đàn và mạng xã hội. Việc phân tích này rất quan trọng vì giúp các công ty có thể hiểu khách hàng nghĩ gì về chất lượng của các sản phẩm và sự phục vụ của họ, cũng như của các đối thủ cạnh tranh. Ở khía cạch khác, thông qua hệ thống, các công ty cũng có thể để biết được sở thích và nhu cầu của khách hàng.

Có rất nhiều các phương pháp, thuật toán được sử dụng trong các hệ thống phân tích ý kiến hiện nay tiêu biểu: Phương pháp Naïve Bayes, Phương pháp Học máy Suppor vector Machine (SVM), phương pháp sử dụng bộ từ vựng (Lexicon-Base), Formal Concept Analysis (FCA)... Trong đó phương pháp học máy (SVM) và phương pháp sử dụng bộ từ vựng được sử dụng rất nhiều trong các hệ thống phân tích ý kiến và đã được chứng minh là có rất nhiều ưu điểm. Tuy nhiên, hai phương pháp trên vẫn còn tồn tại khá nhiều giới hạn.

Phương pháp học máy (SVM) được giới thiệu bởi Cortes và Vapnik vào năm 1995 [25] đã được ứng dụng vào rất nhiều các lĩnh vực sử dụng tin học như phân tích dữ liệu sinh học, văn bản và nhận diện hình ảnh. SVM được xây dựng cho bài toán phân loại nhị phân, cụ thể dữ liệu sẽ được phân làm 2 nhãn: tích cực (Positive) và tiêu cực (Negative). Ngoài ra phương pháp này cũng có thể phân thêm một nhãn nữa là Trung lập (Neutral). Phương pháp này được ứng dụng cho các hệ thống phân tích ý kiến [26], [27], [28]. Nhược điểm của phương pháp này là rút trích các đặc trưng của văn bản là từ nên không gian đặc trưng là rất lớn, bao gồm mọi từ của ngôn ngữ hoặc tập dữ liệu về ngôn ngữ. Chiều dài của không gian đặc trưng càng lớn càng làm gia tăng khả năng nhiễu và phân loại không chính xác.

Phương pháp sử dụng bộ từ vựng (Lexicon-Base) thực hiện việc phân loại dựa trên các cụm từ cú pháp bày tỏ ý kiến trong tiếng anh bằng cách trích lọc các cụm từ chứa tính từ, trạng từ. Phương pháp này thực hiện việc phân loại xem là câu chủ quan hay câu khách quan và xu hướng của câu chủ quan về hai hướng (tích cực hoặc tiêu cực). Đối với phương pháp này chúng ta có thể thu thập bộ từ vựng bằng 3 phương pháp chính: Thủ công, dựa vào từ điển, dựa trên văn bản. Với phương pháp thủ công thường sẽ tốn rất nhiều thời gian do đó chúng ta có thể sử dụng bộ từ điển trực tuyến dựa trên bootstrapping như WordNet [29] với việc thu thập một lượng nhỏ các từ có ý kiến và kết hợp từ bộ từ này bằng cách tìm kiếm trên WordNet.

3.2.3 Sử dụng Ontology

Trong những năm gần đây, thuật ngữ "Ontology" không chỉ được sử dụng ở trong các phòng thì nghiệm trên lĩnh vực trí tuệ nhân tạo mà đã trở nên phổ biến đối với nhiều miền lĩnh vực trong đời sống. Đứng trên quan điểm của ngành trí tuệ nhân tạo, một Ontology là sự môt tả về những khái niệm và những quan hệ của các khái niệm đó nhằm mục đích thể hiện một góc nhìn về thế giới. Trên miền ứng dụng khác của khoa học, một Ontology bao gồm tập các từ vựng cơ bản hay một tài nguyên trên

một miền lĩnh vực cụ thể, nhờ đó những nhà nghiên cứu có thể lưu trữ, quản lý và trao đổi tri thức cho nhau theo một cách tiện lợi nhất.

Ontology được chia làm 3 thành phần bản chất, bao gồm: lớp (Class), thuộc tính (Property), cá thể (Individual).

Lớp (class) là một bộ những thực thể, các thực thể được mô tả logic đề định nghĩa các đối tượng của lớp; lớp được xây dựng theo cấu trúc phân cấp cha con như là một sự phân loại các đối tượng..

Cá thể (individual) hay còn gọi là "thể hiện". Đây được xem là thể hiện của một lớp, làm rõ hơn về lớp đó và có thể được hiểu là một đối tương nào đó trong tự nhiên.

Một vài nguyên do cần xây dựng Ontology và để lý giải cho việc trả về kết quả tìm kiếm không hợp lý được đề cập tới như sau:

Thứ nhất, để chia sẻ những hiểu hiểu biết chung về các khái niệm, cấu trúc thông tin giữa con người hoặc giữa các hệ thống phần mềm: đây là vai trò quan trọng nhất của một Ontology, không những trong lĩnh vực Web ngữ nghĩa mà còn trong nhiều ngành và lĩnh vực khác. Về phương diện này, có thể hình dung Ontology giống như một cuốn từ điển chuyên ngành, cung cấp và giải thích các thuật ngữ cho người không có cùng chuyên môn khi được yêu cầu. Không chỉ được sử dụng bởi con người, Ontology còn hữu ích khi cần sự hợp tác giữa các hệ thống phần mềm. Ví dụ, Open Biological là bộ Ontology nổi tiếng được phát triển bởi trường đại học Stanford nhằm cung cấp các thuật ngữ một cách đầy đủ trong ngành sinh vật học. Ontology này hiện đã được tích hợp vào một số ứng dụng Web trên Internet. Sau đó, một phần mềm tra cứu hoặc dạy sinh học trên máy tính có thể kết nối với các ứng dụng Web trên để lấy thông tin cho mục tiêu chú giải.

Thứ hai, cho phép tái sử dụng tri thức: đây là một vấn đề khó và là mục tiêu nghiên cứu quan trọng trong những năm gần đây. Nó liên quan đến bài toán trộn hai hay nhiều Ontology thành một Ontology lớn và đầy đủ hơn. Nhưng vấn đề ở đây là tên

các khái niệm được định nghĩa trong các Ontology này có thể giống nhau trong khi chúng được dùng để mô tả các loại vật hoàn toàn khác nhau. Tuy nhiên cũng có thể có trường hợp ngược lại, khi tên các khái niệm khác nhau nhưng cùng mô tả một sự vật. Ngoài ra, làm thế nào để bổ sung các quan hệ, thuộc tính có sẵn vào một hệ thống mới càng làm cho vấn đề trở nên phức tạp.

Thứ ba, do sự mơ hồ, không rõ ràng của ngôn ngữ tự nhiên, người tìm kiếm có thể nhập vào những từ khóa đồng nghĩa, có nghĩa là từ khóa này đại diện cho nhiều đối tượng khác nhau, ví dụ khi người dùng nhập vào "Vietnam", công cụ tìm kiếm không thể biết được người này đang tìm kiếm tên của một người, hay tên của một tên một đất nước... Ngoài ra, Ontology có thể sẽ trở thành hướng đi mới cho một lĩnh vực đã quen thuộc là dịch tài liệu tự động. Có thể nói như vậy, bởi ngữ nghĩa các từ vựng trong văn bản sẽ được dịch chính xác hơn khi được ánh xạ vào đúng ngữ cảnh của nó.

Thứ tư, cho phép tri thức trở nên nhất quán và tường minh: các khái niệm khác nhau trong một hay nhiều lĩnh vực cụ thể có thể cùng tên và gây nhập nhằng về ngữ nghĩa, tuy nhiên khi được đưa vào một hệ thống Ontology thì tên mỗi khái niệm là duy nhất. Một giải pháp cho vấn đề này là Ontology sẽ sử dụng các tham khảo URI làm định danh thật sự cho khái niệm trong khi vẫn sử dụng các nhãn gợi nhớ bên trên để thuận tiện cho người dùng.

Thứ năm, cung cấp một phương tiện cho công việc mô hình hóa và suy luận: Ontology là một tập các khái niệm phân cấp được liên kết với nhau bởi các quan hệ. Cơ bản mỗi khái niệm có thể xem như là một lớp, mà đối tượng của lớp đó cùng các quan hệ đã góp phần tạo nên cấu trúc logic hay vấn đề cần giải quyết. Hơn nữa hiện nay, một số ngôn ngữ Ontology đã tích hợp lớp Ontology suy luận (Ontology Inference Layer) bên trong cho mục đích suy luận logic trên tập quan hệ giữa các đối tượng trong hệ thống.

Cuối cùng: do những khái niệm cấp cao, mơ hồ hay trừu tượng. Những khái niệm dạng này thường không được nhắc tới tường minh trong các tài liệu. Công cụ tìm

kiếm khó có thể nhận diện được những kết quả phù hợp cho những truy vấn thuộc loại này.

Để giải quyết những vấn đề nêu trên, tác giả thêm vào quy trình truy hồi thông tin truyền thống của một công cụ tìm kiếm những cơ sở tri thức được lưu trong một Ontology, còn gọi là hệ thống truy hồi thông tin dựa trên Ontology. Công cụ tìm kiếm sẽ tham khảo cơ sở tri thức từ Ontology để đưa ra các kết quả tìm kiếm tối ưu hơn, dựa trên ngữ nghĩa, quan hệ giữa các từ khóa.

Một nghiên cứu khác về Ontology được mô tả trong [30]. Điện thoại di động nói chung và smartphone nói riêng trong năm gần đây đã và đang đạt được những bước tiến đáng kể về mặt công nghệ. Sự phát triển của công nghệ cho ra đời nhiều loại điện thoại khác nhau, với nhiều đặc điểm và công nghệ đi kèm khác nhau, kéo theo sự mở rộng đáng kể về số lượng các khái niệm trong tri thức về mobile. Đó chính là một khó khăn lớn đối với những ai cần nghiên cứu về mobile, đặc biệt là các nhà nghiên cứu, lập trình hay thậm chí cả khách hàng cần mua sản phẩm mobilephone và tham khảo thông tin sản phẩm của họ trên các website. Các website hiện nay giới thiệu sản phẩm với một vài thuộc tính sơ lược, nhiều vấn đề sẽ phát sinh nếu như xuất hiện những thuộc tính trùng tên, hay những thuộc tính không được đề cập rõ ràng chi tiết gây nhập nhằng cho người sử dụng.

CHƯƠNG 4
MÔ HÌNH ĐỀ XUẤT

4.1 Mô hình hệ thống

4.1.1 Giới thiệu

Khai thác quan điểm là một qui trình nghiên cứu rất phức tạp, được tìm hiểu trên nhiều khía cạnh khác nhau. Ở đây việc khai phá quan điểm trên mạng xã hội là một lĩnh vực mới, nhận được nhiều sự quan tâm trong những năm gần đây, và đánh dấu một bước phát triển trong khai phá văn bản (text mining). Khai phá văn bản hướng tới việc phân tích ngữ nghĩa, giúp máy móc thực sự "hiểu" nội dung văn bản nói và quan điểm của người viết như thế nào (ví dụ: khen/chê) trong văn bản đó.

Nhu cầu một máy tìm kiếm quan điểm được đặt ra đáp ứng nhu cầu tìm kiếm quan điểm người dùng. Máy tìm kiếm quan điểm nhận đầu vào là một truy vấn từ người dùng và kết quả trả về là những quan điểm về vấn đề mà người dùng quan tâm, thay vì trả về tập các văn bản liên quan tới truy vấn của người dùng như các máy tìm kiếm thông thường.

Khóa luận tập trung nghiên cứu phương pháp và xây dựng mô hình thống kê cho tổng hợp khai thác quan điểm trên mạng xã hội sử dụng phương pháp xử lý ngôn ngữ tự nhiên bằng tiếng Anh nhằm ứng dụng vào máy tính tìm kiếm, khai thác quan điểm trên mạng xã hội. Với đầu vào là một câu quan điểm mà người dùng quan tâm, hệ thống tiến hành tìm kiếm, đánh giá và cho ra kết quả tích cực, tiêu cực hay trung tính của quan điểm đó.

Với mô hình đề xuất, khóa luận tiến hành xây dựng thử nghiệm áp dụng mô hình khai thác quan điểm của các bình luận bằng tiếng Anh trên mạng xã hội. Dự vào nguồn dữ liệu được lấy [31] của Giáo sư Bing liu. Ông đã đưa ra những phương pháp và kỹ thuật khai thác quan điểm, đánh giá các bình luận và đưa kết quả đáng ngạc nhiên.

Mặt khác ở Việt Nam, việc khai phá quan điểm được coi là một lĩnh vực mới, dành được nhiều quan tâm trong thời gian gần đây và chỉ mới đạt được một số kết quả bước đầu, do đó còn rất nhiều vấn đề trong khai phá quan điểm chưa được giải quyết trên nhiều khía cạnh.

Khai thác quan điểm có vai trò rất quan trọng, bởi khi chúng ta cần quyết định một vấn đề gì chúng ta thường đặt ra câu hỏi "Người khác nghĩ về vấn đề đó như thế nào?". Chẳng hạn khi bạn muốn mua một chiếc laptop Dell bạn sẽ muốn hỏi bạn bè và người thân "Máy Dell có tốt không? Dòng Dell như thế nào? Pin dùng có lâu không?..v.v". Như vậy quan điểm của người khác giúp các cá nhân có thêm thông tin trước khi quyết định một vấn đề. Ngoài ra khai phá quan điểm giúp các công ty, tổ chức biết được ý kiến, quan điểm của một bộ phận người quan tâm về vấn đề của công ty, tổ chức. Dựa trên những vấn đề đó chúng tôi tiến hành hiện thực và kiểm tra mô hình hệ thống khái quát.

4.1.2 Mô hình hệ thống

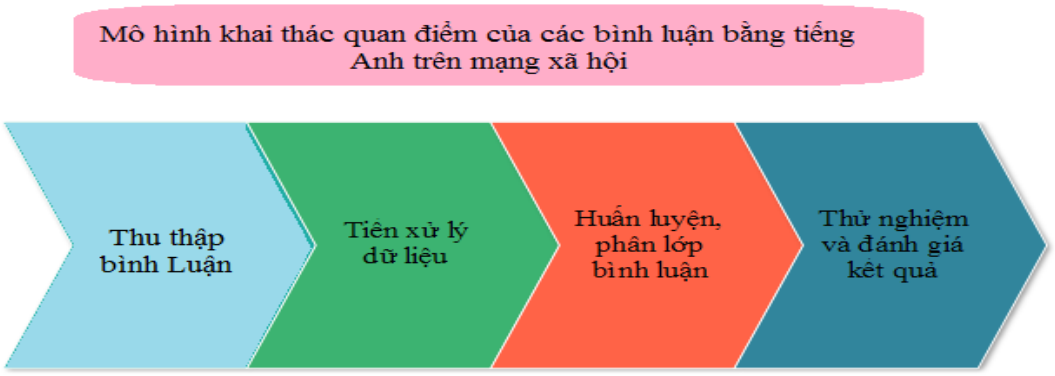
Mô hình hiện thực và kiểm tra hệ thống như sau:

Bước 1: Thu thập bình luận

Bước 2: Tiền xử lý dữ liệu

Bước 3: Huấn luyện và phân lớp câu bình luận

Bước 4: Thử nghiệm và đánh giá kết quả



Hình 4.1: Mô hình khác thác quan điểm của các bình luận bằng tiếng Anh trên mạng xã hội sử dụng phương pháp xử lý ngôn ngữ tự nhiên

4.1.2.1 Thu thập bình luận

Module thực hiện hai quá trình thu thập dữ liệu và rút trích bình luận. Với việc thu thập bình luận, chúng tôi tiến hành lọc và thu thập các bình luận trên mạng xã hội như diễn đàn, facebook,... theo từng chủ đề. Chúng tôi thu thập khoảng 1000 câu bình luận trên các trang mạng xã hội, tạo cơ sở dữ liệu cho quá trình nghiên cứu.

Rút trích bình luận, chúng tôi dựa vào nhóm phát triển trên công nghệ Java và cơ sở dữ liệu lưu trữ Postgres. Cơ sở dữ liệu Postgres có vai trò lưu thông tin các liên kết cần phải duyệt qua và kết quả thu thập được. Khởi đầu, crawler (crawler là phần mềm có khả năng tự động lấy dữ liệu như ảnh, text,… trên WWW) sẽ được cung cấp một số liên kết (URL) khởi đầu và tiến hành thu thập toàn bộ nội dung HTML chứa thông tin. Mã HTML được phân tích cấu trúc DOM3, theo các luật quy định sẵn, crawler sẽ xác định vùng dữ liệu cần bóc tách: liên kết tương tự, thông tin bình luận cần thu thập. Các liên kết được chọn lọc và lưu trữ trong một hàng đợi URL. Để rút trích đúng mục tiêu bài viết, các liên kết hoặc tiêu đề bài viết được lọc lại theo từ khóa ứng với sản phẩm cần thu thập. Quá trình này đi lặp lại, cho tới khi không còn liên kết nào trong hàng đợi hoặc đủ số lượng cần thiết.

4.1.2.2 Tiền xử lý dữ liệu

Dữ liệu sau khi rút trích được tiền xử lý để có được một tập dữ liệu rõ ràng, không trùng lấp, loại bỏ các liên kết, trích dẫn trong bình luận. Mudule tiền xử lý này là rất quan trọng, bởi lẽ làm giảm sự nhập nhằng cho chương trình, cũng như quá trình thực thi, thực nghiệm chương trình.

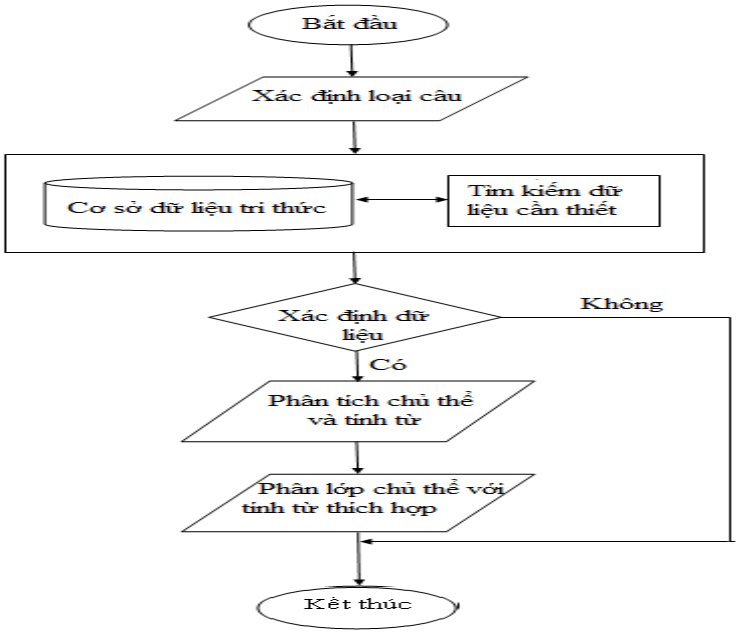
Từng dữ liệu sẽ được tách từ, đối với tiếng Anh, các từ được phân cách bởi dấu cách hoặc các dấu câu khác. Chẳng hạn có câu bình luận sau bằng tiếng Anh "Dell is more beautiful than Hp." ở đây các từ được cách nhau bằng một khoảng trắng và được tách ra thành những từ riêng lẻ như sau: "Dell" "is" "more" beautiful" "than" "Hp". Thông qua việc phân tích và tách các từ trong câu chúng ta có thể tìm các động từ, tính từ, danh từ,… trong câu và từ đó, phân tích câu đó theo các khía cạnh quan điểm thích hợp là tích cực, tiêu cực hay trung tính.

4.1.2.3 Phân lớp phản hồi, bình luận

Phản hồi sau khi được thu thập sẽ được phân thành các lớp khác nhau để phục vụ việc thống kê, tạo báo cáo.

Phân lớp văn bản (Text Classification) là quá trình gán nhãn các văn bản ngôn ngữ tự nhiên một cách tự động vào một hoặc nhiều lớp cho trước, "nhóm" các đối tượng "giống" nhau vào "một lớp" dựa trên các đặc trưng dữ liệu của chúng. Hệ thống đánh giá phân lớp các bình luận rút trích được thành 3 nhóm: "Tích cực", "Tiêu cực" và "Trung lập" tương ứng là: "Positive", "Negative", "Neutral".

Luận văn trình bày bộ phân lớp dựa trên hai giải thuật là Naïve Bayes, SVM và một số quy tắc của giáo sư Bing Liu. Qua quá trình nghiên cứu chúng tôi đề xuất mô hình giải thuật cho bài luận này như sau:



Hình 4.2: Sơ đồ giải quyết bài toán đề xuất

Với mô hình thuật toán trên chúng tôi xây dựng theo các Module sau:

Module 1: Xác định loại câu.

Module 2: Tìm kiếm dữ liệu cần thiết.

Module 3: Cơ sở dữ liệu tri thức.

Module 4: Phân tích chủ thể, tính từ.

Module 5: Phân lớp chủ thể với tính từ thích hợp.

Trong đó:

Module 1: Xác định loại câu là sau khi người dùng nhập vào một câu bằng tiếng Anh bất kỳ theo quan điểm đánh giá của chính người nhập, chẳng hạn như câu tích cực, tiêu cực hay trung tính thì module này sẽ xác định được loại câu như câu phủ định, câu khẳng định, câu so sánh bằng hay câu so sánh hơn. Ví dụ khi người dùng nhập câu sau: “Dell is the best“ thì mudule này sẽ xác định đây là loại câu khẳng định. Với module này chúng tôi thực hiện phương pháp phân tích xử lý theo ngôn ngữ tự nhiên.

Module 2: Tìm kiếm dữ liệu cần thiết đây là bước rất quan trọng ở mô hình giải thuật trên.

Sau khi xác định được loại câu, việc tìm kiếm dữ liệu sẽ phụ thuộc vào loại câu gì để truy vấn vào cơ sở dữ liệu tri thức. Cở sở dữ liệu tri thức ở đây là tập dữ liệu được chọn lọc mà chúng tôi thu thập được từ các trang web trên mạng xã hội … (dữ liệu có sẵn và được chọn lọc). Chẳng hạn khi chúng tôi xét câu sau: "Hp is not so good as Dell" thì chúng tôi đánh giá loại câu này là câu so sánh bằng, cho nên thuật toán sẽ phân tích và đánh giá theo phương pháp xử lý ngôn ngữ tự nhiên như sau: Hp và Dell là hai chủ thể riêng biệt, với cụm từ "not so good as" có nghĩa là "không tốt bằng". Vậy câu được phân tích như sau: "Hp" +not so + adj +as + "Dell", dẫn tới "Hp" là tiêu cực còn "Dell" là tích cực. Ở module này, chúng tôi sử dụng phương pháp xử lý ngôn ngữ tự nhiên là tìm kiếm các chủ thể, các cụm từ cần thiết để các module tiếp theo phân tích và đưa ra kết qủa đánh giá.

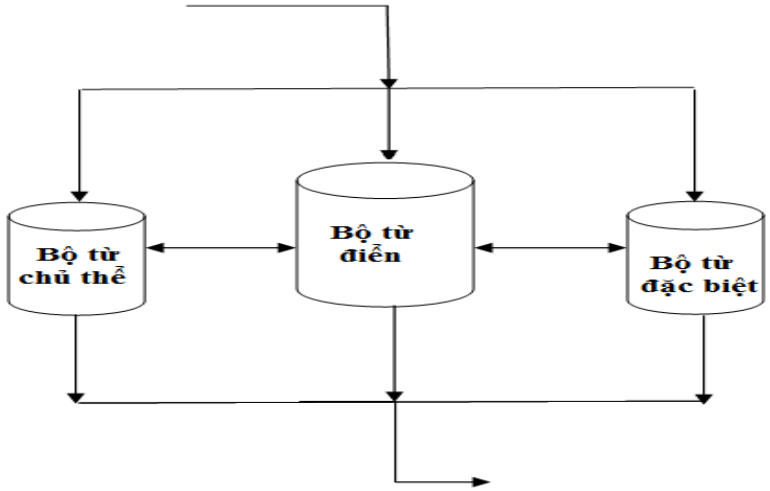
Thực ra, với module này chỉ là sử dụng phương pháp xử lý ngô ngữ tự nhiên bằng hai giải thuật SVM, NaiveBayes và 10 quy tắc của Bing Liu để loại bỏ những dữ liệu không cần thiết và chọn lọc những dữ liệu cần thiết để các module sau thực hiện đánh giá quan điểm.

Module 3: Cơ sở dữ liệu tri thức là tệp dữ liệu chúng tôi được chọn lọc, gồm bộ từ điển khoảng 6800 từ bao gồm tiêu cực và tích cực được thu thập từ [31], ngoài ra chúng tôi cũng tạo thêm khoảng 20 từ loại đặc biệt để xử lý trong ngôn ngữ tiếng Anh (chúng tôi gọi là bộ từ đặc biệt), chẳng hạn như từ "but", "nobody", "none", not, … và xây dựng thêm bộ từ bằng ontology làm chủ thể (chúng tôi gọi là bộ từ chủ thể) liên quan đến máy tính khoảng 15 từ là tên các hãng máy tính nổi tiếng hiện nay, chẳng hạn như các loại máy tính "Dell", "Hp",… để thực hiện quá trình đánh giá, khai thác quan điểm trong các loại câu so sánh, câu đơn, câu khẳng định,...

Như đã phân tích ở module 2 nêu trên, chúng ta sẽ dễ hình dung khi cơ sở dữ liệu không tồn tại thì ở các module tiếp theo sẽ không thực hiện mà kết thúc và đây

cũng là một vấn đề rất quan trọng khi mà dữ liệu không có hoặc thiếu thì kết quả sẽ không chính xác và có thể dẫn đến sai lệch. Chẳng hạn như xét câu "Dell is not good" thì giả sử từ "good" không tồn tại trong kho dữ liệu thì câu này sẽ cho kết quả là trung tính là sai với thực tế do con người đánh giá,… Trường hợp khi xét có chứa đầy đủ cơ sở dữ liệu thì câu được phân tích quan điểm như sau: "Hp is not so good as Dell" thì Hp và Dell sẽ được chọn lọc từ bộ từ chủ thể để phân tích, từ "good" được tìm thấy trong bộ từ điển và "not" là từ loại đặc biệt thuộc bộ từ loại đặc biệt,…

Từ những phân tích trên chúng tôi xây dựng quy trình thực hiện chuyển đổi giữa các hệ cơ sở dự liệu ở module này như sau:



Hình 4.3: Mô hình cơ sở dữ liệu

Nhìn chung, cở sở dữ liệu là rất quan trọng trong việc đánh giá quan điểm. Chính vì vậy, việc cơ sở dữ liệu không tồn tại thì việc đánh giá có thể dẫn tới sai lệch, nhầm lẫn, không đáng tin cậy.

Module 4: Phân tích chủ thể và tính từ là thực hiện sau khi nhận dữ liệu từ module 2. Dựa vào dữ liệu này module thực hiện việc phân tích chủ thể và tính từ đi kèm với nhau. Ví dụ với câu bình luận là câu đơn như sau: "Dell is the best" thì

module thực hiện phân tích chủ thể là "Dell" và tính từ đi kèm là "best". Như vậy, với module này sẽ xác định chủ thể nào đi với tính từ nào hay động từ nào tương ứng.

Để làm được điều này, chúng tôi cần kết hợp hai giải thuật SVM, NaiveBayes và 10 quy tắc của Bing Liu để phân tích một câu bình luận bằng tiếng Anh sử dụng phương pháp xử lý ngôn ngữ tự nhiên.

Module 5: Phân lớp chủ thể cùng với tính từ thích hợp là module phân tích các chủ thể, các tính từ có cùng tính chất đi vào một nhóm và xác định nhóm theo phân lớp. Điều này là thế mạnh khi sử dụng hai giải thuật là SVM và NaiveBayes, Ví dụ khi xét câu so sánh tiếng Anh như sau: "Dell is more beautiful than Acer" thì SVM sẽ phân làm hai lớp là Dell và Acer. Tuy nhiên việc phân lớp chủ thể Dell cùng với tính từ "beautiful" là lớp tích cực (+1) còn Acer là lớp khác thể hiện lớp tiêu cực (-1). Ở đây chúng tôi dùng giải thuật NaiveBayes.

Nói chung, ở module này chúng tôi kết hợp hai giải thuật SVM và NaiveBayes, với SVM thực hiện phân lớp trong câu còn NaiveBayes thực hiện việc gán nhãn từ loại với chủ thể tương ứng.

Dưới đây là hai giải thuật chúng tôi áp dụng trong quá trình nghiên cứu.

a) Naïve Bayes

Đây là kỹ thuật phân lớp giám sát được đề xuất bởi Thomas Bayes [32]. Phương pháp của Naïve Bayes được sử dụng khá phổ biến trong các lĩnh vực tìm kiếm, lọc mail, phân lớp, ... Kỹ thuật này sử dụng xác suất có điều kiện giữa từ và chủ đề để xác định chủ đề của văn bản. Các xác suất này dựa trên việc thống kê sự xuất hiện của từ và chủ đề trong tập huấn luyện. Tập huấn luyện lớp có thể mang lại kết quả khả quan cho Naïve Bayes. Ưu điểm của phương pháp này là đơn giản, tốc độ nhanh, cài đặt không quá phức tạp phù hợp với thời gian cho phép.

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

a.1 <u>Huấn luyện:</u>

Tính $P(C_i)$ và $P(X_k|C_i)$

X_k là được biểu diễn là vector k chiều: (X_1, X_2, X_3,… , X_n)

C_i là một tập xác định các nhãn lớp: (C_1, C_2, C_3,… , C_m).

$P(C_i)$ là giá trị xác xuất của các nhãn được xác định theo công thức sau:

$$P(C_i) = \frac{|docs_i| + 1}{|total\ docs| + m}$$

$P(X_k|C_i)$ là tỷ lệ giá trị xác suất của độ dài vertor với giá trị xác xuất nhãn phân lớp.

Đầu vào:

- Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
- Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

Đầu ra:

- Các giá trị xác suất $P(C_i)$ và $P(X_k|C_i)$.

a.2 Phân lớp:

Đầu vào:

- Vector đặc trưng của văn bản cần phân lớp.
- Các giá trị xác suất $P(C_i)$ và $P(X_k|C_i)$.

Đầu ra:

- Nhãn/lớp của văn bản cần phân loại.

b) Support Vector Machines (SVM)

Support Vector Machines là một phương pháp máy học do Vladimir Vapnik và các cộng sự xây dựng nên từ những năm 70 của thế kỉ 20. SVMs là bộ phân lớp nhị phân, để áp dụng trong bài toán phân loại đa lớp, một số chiến thuật phân lớp đã được đề xuất, như One-Against-One (OAO), One-Against-Rest (OAR), dựa trên cấu trúc đồ thị (DDAG, ADAG), Half-Against-Half (HAH) và phương pháp phân loại nhiều lớp

mờ. Thuật toán phân lớp SVM nhóm chọn là OAO do có nhiều thực nghiệm cho kết quả tương đối tối ưu, được triển khai trên thư viện Mlib của Apache Spark7.

Phương pháp One-Against-One (OAO)

b.1 Huấn luyện:

Đầu vào:

- Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
- Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.
- Các tham số cho mô hình SVM: C, γ (tham số của hàm kernel, nhóm sử dụng hàm Gauss)

Đầu ra:

- Mô hình SVM (Các Support Vector, nhân tử Lagrange a, tham số b).

b.2 Phân lớp:

Đầu vào:

- Vector đặc trưng của văn bản cần phân lớp.
- Mô hình SVM

Đầu ra:

- Nhãn/lớp của văn bản cần phân loại

c) Các quy tắc phân loại quan điểm [33]

Theo Binh Liu [33] thì có 32 quy tắc được xây dựng và trình bày ở dạng chuẩn Boy-Code (tương tự như định nghĩa cú pháp trong ngôn ngữ lập trình) nhận diện quan điểm.

Ở đây chúng tôi sử dụng 10 quy tắc của Bing Liu như sau.

1	Positive	::= P	Tích cực	:: P	
2		| PO			| PO
3		| đổi_hướng N			

4 | đổi_hướng NE

5 Negative ::= N Tiêu cực :: N

6 | NE | NE

7 | đổi_hướng P

8 | đổi_hướng PO

Trong 8 quy tắc này thì

P/PO là hai biểu diễn quan điểm tích cực.

N/NE là hai biểu diễn quan điểm tiêu cực.

P là biểu diễn quan điểm tích cực (từ/cụm từ quan điểm tích cực).

PO là biểu diễn hợp thành của hai quan điểm tích cực.

N là biểu diễn quan điểm tiêu cực (từ/cụm từ quan điểm tiêu cực).

NE là biểu diễn hợp thành của hai quan điểm tiêu cực.

Positive và Negative là biểu diễn quan điểm kết thúc xác định quan điểm cho đối tượng trên khía cạnh đó.

9. P ::= một từ/cụm từ quan điểm tích cực

10. N ::= một từ/cụm từ quan điểm tiêu cực

Quy tắc 9 và 10 là hai quy tắc biểu diễn đơn giản nhất: từ/cụm từ; từ/cụm từ tự chúng biểu diễn quan điểm tích cực/tiêu cực.

4.2 Thử nghiệm và đánh giá kết quả

Thu thập dữ liệu: Dữ liệu mà đề tài chuẩn bị thu thập [31] gồm khoảng 6800 từ (từ quan điểm tích cực và tiêu cực) [31], tập này được dùng cho mô hình, giải thuật, theo chúng tôi dữ liệu này là phù hợp. Ngoài ra, trong quá trình xây dựng ontology, chúng tôi nghiên cứu về máy tính nên chọn tập dữ liệu về máy tính và [31] chúng tôi chọn bộ dữ liệu gồm 531 câu về computer và 879 câu về Wirless router để thử nghiệm mô hình, giải thuật.

Về thử nghiệm, chúng tôi chọn 705 câu ngẫu nhiên từ dữ liệu trên, với mong muốn mang lại độ tin cậy cho mô hình của chúng tôi. Dùng tập huấn luyện 1410 câu, kiểm thử 705 câu, tỷ lệ kiểm thử trên tập huấn luyện đạt ~50%, với chúng tôi là hợp lý.

Xử lý dữ liệu: Các bài viết, bình luận sau khi thu thập được tiền xử lý và chuẩn hóa. Lọc bỏ các liên kết, lọc bỏ trích dẫn (quote) trong bình luận, gán nhãn dữ liệu. Mỗi bình luận được gán nhãn bằng tay, gồm nhãn: tích cực (1), tiêu cực (-1), trung lập (0), không liên quan (-2).

Chọn lựa thuật toán: Với mô hình chúng tôi đưa ra, chúng tôi kết hợp hai giải thuật SVM, NaiveBayes và các 10 quy tắc đầu tiên [33].

- B.Pang và các cộng sự [4] áp dụng giải thuật Naïve Bayes và SVM để xác định hướng quan điểm phân cực trong bình luận.
- Khi Sử dụng unigram trong phân lớp cho kết quả thực nghiệm tốt khi sử dụng Bayesian hoặc SVM.
- Naïve Bayes và SVM có thể sử dụng trong phương pháp học máy phân lớp quan điểm.
- Nhiều kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được áp dụng, các phương pháp học máy như phân lớp Naïve Bayes, cực đại hóa entropy và SVM được nghiên cứu và áp dụng thử nghiệm.
- SVM được xây dựng cho bài toán phân loại nhị phân, cụ thể dữ liệu sẽ được phân làm 2 nhãn: tích cực (Positive) và tiêu cực (Negative). Ngoài ra phương pháp này cũng có thể phân thêm một nhãn nữa là Trung lập (Neutral). Nhược điểm của phương pháp này là rút trích các đặc trưng của văn bản là từ nên không gian đặc trưng là rất lớn, bao gồm mọi từ của ngôn ngữ hoặc tập dữ liệu về ngôn ngữ. Chiều dài của không gian đặc trưng càng lớn càng làm gia tăng khả năng nhiễu và phân loại không chính xác.

- Phương pháp của Naïve Bayes được sử dụng khá phổ biến trong các lãnh vực tìm kiếm, lọc mail, phân lớp, ... Kỹ thuật này sử dụng xác suất có điều kiện giữa từ và chủ đề để xác định chủ đề của văn bản. Ưu điểm của phương pháp này là đơn giản, tốc độ nhanh, cài đặt không quá phức tạp phù hợp với thời gian cho phép.
- Kết hợp hai giải thuật trên, chúng tôi vận dụng 10 quy tắc đầu tiên của Bing Liu mà không phải tất cả 32 quy tắc trong [31], nhằm xây dựng và phân tích những trường hợp chuyển hướng quan điểm tích cực, tiêu cực hoặc trung tính trong ngôn ngữ tiếng Anh, điển hình như từ nhưng "but"...

Sau đây là kết quả đánh giá và thử nghiệm của mô hình thuật toán đề xuất như sau:

Bảng 4.1: Bảng đánh giá kết quả mô hình áp dụng (đơn vị: bình luận)

Stt	**Mô hình**	**Huấn luyện**	**Kiểm thử**	**Kết quả**		
				Độ chính xác	**Độ sai lệch**	**Không xác định**
1	Mô hình đề xuất	1410	750	87.23 %	8.37 %	4.4 %

Kết quả 87.23% là độ chính xác với các loại câu phủ định, khẳng định, câu ghép cùng khía cạnh, chẳng hạn xét một số câu sau:

- "I really love this netbook" cho kết quả là tích cực, đối với loại câu này là một câu khẳng định cho nên việc đánh giá chỉ dựa vào tính từ đề đánh giá cho câu. Dễ thấy nhìn vào câu chương trình sẽ tìm thấy tính từ love thể hiện tình cảm yêu mến, tốt, điều này thể hiện tích cực.

- “This monitor is much, much better but still not great” cho kết quả tiêu cực, đối với câu này chứa nhiều khía cạnh, ở đây là hai khía cạnh tích cực và tiêu cực, với loại câu này chương trình thực hiện tìm kiếm đánh giá vế sau từ “but” trước từ nhưng ở cây này bỏ qua, ta xét vế sau từ nhưng “but”, đây là từ đặc biệt trong khai thác quan điểm bằng ngôn ngữ tiếng Anh. Việc đánh giá câu này chương trình thực hiện việc tìm kiếm tính từ sau từ nhưng và nhận thấy từ “great” là tích cực nhưng trong trường hợp câu này trước từ great là một phủ định, cho nên từ tích cực chuyển về tiêu cực là hoàn toàn đúng.
- “My other monitor is a 23 inch ACER” ở câu này cho kết quả trung tính là hoàn toàn đúng, với nhưng loại câu này chương trình thực hiện việc tìm kiếm tính từ nhưng không thấy các từ thể hiện yếu tố tịch cực hay tiêu cực, do vậy, loại câu này là trung tính.

Kết quả 4.4% không xác định vì đây là trường hợp chương trình gặp phải loại câu so sánh cụ thể ở hai câu sau, cụm từ so sánh được gạch chân, đây là những cụm từ chương trình không thể xác định.

- It 's very light-weight , which is why I use this more than my Alienware .
- It is slightly slower than the dell, but it is hard to notice with the very nice screen quality and larger hard drive.

Sau đây là một số câu thể hiện độ sai lệch trong mô hình đề xuất chiếm 8.37 % đối với nhưng câu phức tạp như sau:

- “The operation with the increased memory option was flawless with the Windows 7 Home Professional instaled otherwise other netbooks with Windows 7 Starter doesnt allow, and nobody mention this” ở câu này cho kết quả tiêu cực là sai lệch, bởi chương trình cần truy xét và tìm tới cụm từ tích cực là “increased memory option” nhưng ở đây việc xét câu này chương trình

bị nhầm lần với cụm từ nobody chuyển hướng quan điểm nên theo chương trình thì tiêu cực là không đúng.

- "2 weeks later, The monitor start having other issues, only half screen turns on, or it become a full green screen and doesnt change, you need to turn it off and on again until it works" cho kết quả đúng là tiêu cực nhưng chương trình đề xuất cho kết quả là tích cực. Thực tế, đối với những loại câu này, câu chứa nhiều khía cạnh thì chương trình rất khó để tìm thấy tính từ phù hợp, do chúng tôi đề xuất chỉ xét một đoạn của câu đó được chia cắt bởi từ "or", có thể dẫn tới sai lệch khi không chọn đúng đoạn chính của câu và đây cũng là những vấn đề của chương trình gặp sự nhập nhằng khi đánh giá những loại câu phức tạp, điều này dẫn tới sư sai lệch khi đánh giá câu và hiện thực kết quả.

Ngoài ra, để kiểm tra và so sánh đánh giá với các mô hình khác chúng tôi cùng chọn tệp dữ liệu về máy tính khoảng 100 câu bình luận [31] với mô hình cơ sở sử dụng hai giải thuật SVM và Naïve Bayes kết quả hiện thị như sau:

Bảng 4.2: Bảng đánh giá kết quả so sánh với mô hình cơ sở (đơn vị: bình luận)

Stt	Mô hình	Kết quả độ chính xác
1	Mô hình cơ sở	79%
2	Mô hình đề xuất	86%

Nhìn vào bảng đánh giá kết quả so sánh khi sử dụng giải thuật toán SVM và Naïve Bayes với mô hình đề xuất của chúng tôi gồm hai giải thuật toán SVM và Naïve Bayes và 10 quy tắc của Bing Liu cho thấy mô hình đề xuất của chúng tôi đạt kết quả tốt hơn so với mô hình cơ sở.

a) Về ưu điểm của mô hình áp dụng:

+ Xây dựng được mô hình đơn giản, thân thiện, rõ ràng và không phức tạp cho người dùng.

+ Xác định đánh giá được một số loại câu như: phủ định, khẳng định, so sánh.

b) Về nhược điểm của mô hình áp dụng:

+ Việc phải nhập chủ thể đối với câu so sánh là mất thời gian cho người sử dụng.

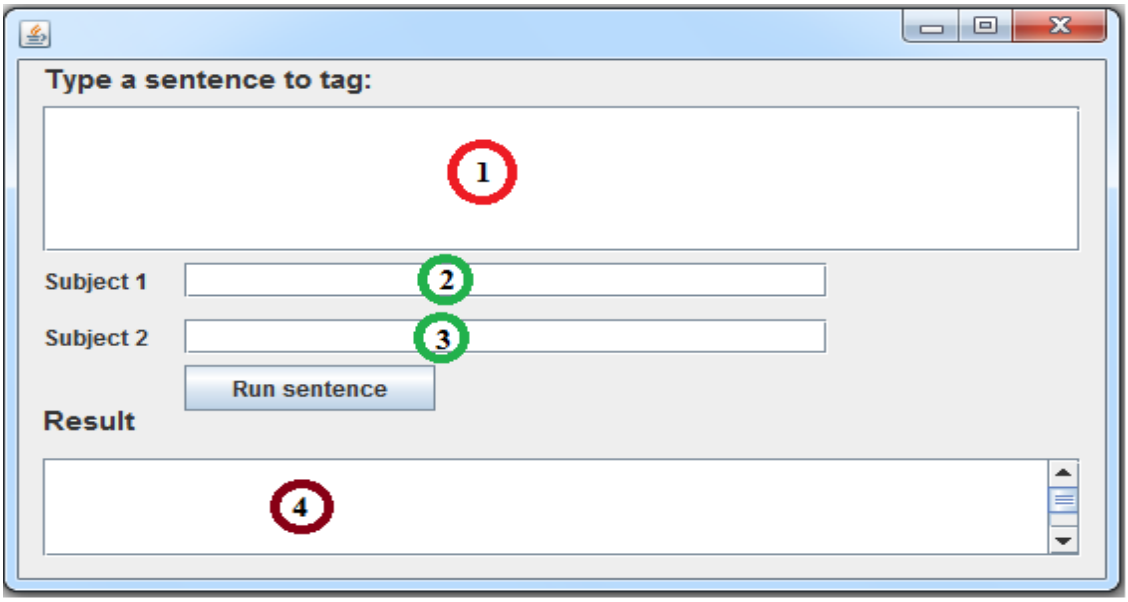
+ Chương trình mới dừng lại ở những câu đơn giản như câu phủ định, khẳng định, đối với câu so sánh mới dừng lại ở hai loại câu so sánh là so sánh hơn và so sánh bằng.

+ Vốn từ vựng còn hạn chế nên chương trình có thể bắt gặp nhưng kết quả chưa chuẩn xác.

c) Mô tả chương trình ứng dụng.

Như ta đã đề cập ở trên, người sử dụng luôn quan tâm đến các quan điểm đánh giá về sản phẩm thương mại, mà vấn đề là xem họ muốn biết được sản phẩm mà họ quan tâm là như thế nào thông qua các bình luận trên các trang mạng xã hội, để từ đó người dùng có những lựa chọn hợp lý, còn nhà cung cấp có thể dựa vào đó để có những quyết định chính xác.

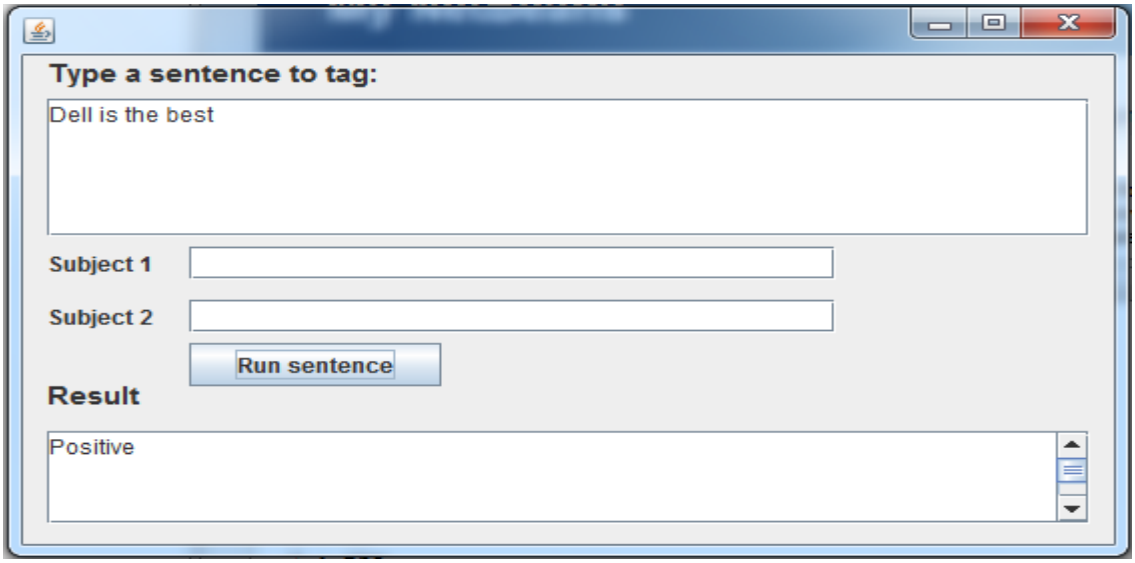
Vì vậy, để người dùng có thể sử dụng cũng như người nghiên cứu có thể kiểm thử những giải thuật mà chúng tôi đã nêu ở trên, do đó chúng tôi xây dựng một chương trình demo nhỏ. Chương trình này được phát triển trên ngôn ngữ java.



Hình 4.4: Giao diện chính của chương trình chính

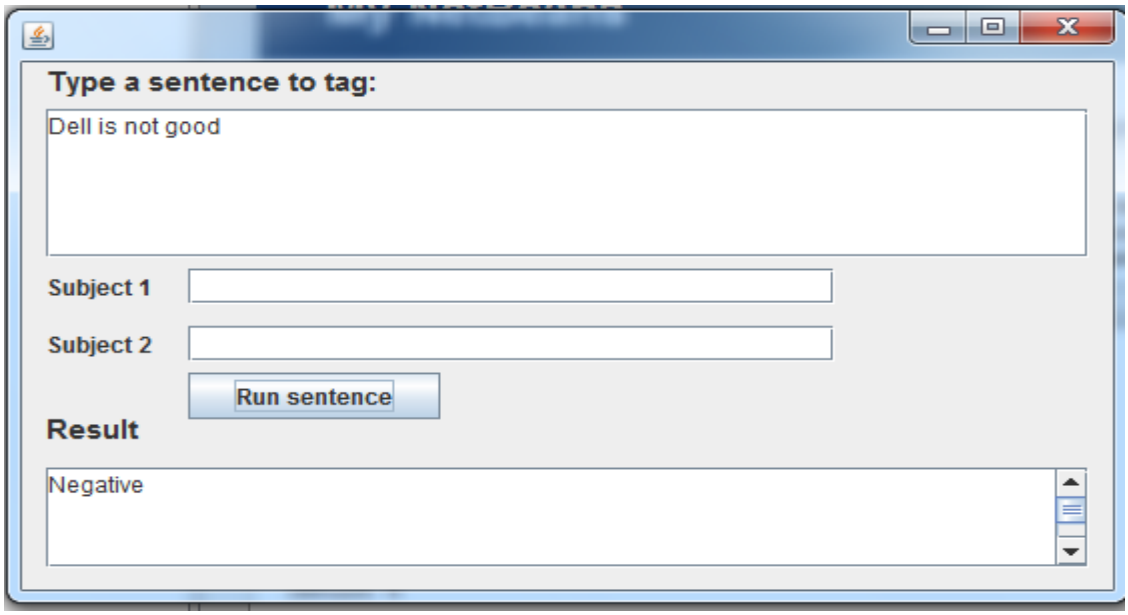
Để thực thi chương trình người dùng cần nhập vào vùng khung vòng tròn số 1 ở trên tức phần khung phía dưới “Type a sentence to tag” một câu bình luận bất kỳ bằng tiếng Anh. Hai hàng ngang “Subject 1” và “Subject 2” tương ứng vùng số 2 và 3 là người dùng nhập chủ thể nếu cần thiết hoặc không, nếu không quan tâm đến chủ thể là gì. Và cuối cùng, click vào “Run sentence” để chạy chương trình và hiện thị kết quả ở vùng số 4. Sau đây là một số giao diện thể hiện câu bình luận bằng tiếng Anh.

- Câu khẳng định



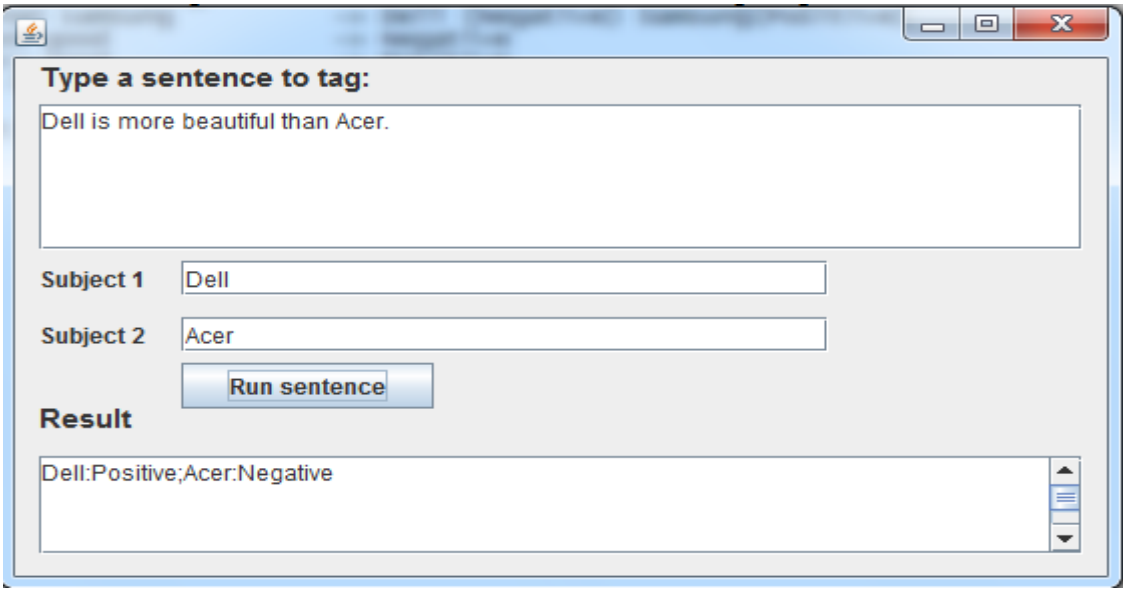
Hình 4.5: Kết quả câu đơn tích cực (Positive).

- Câu phủ định



Hình 4.6: Kết quả câu đơn tiêu cực (Negative).

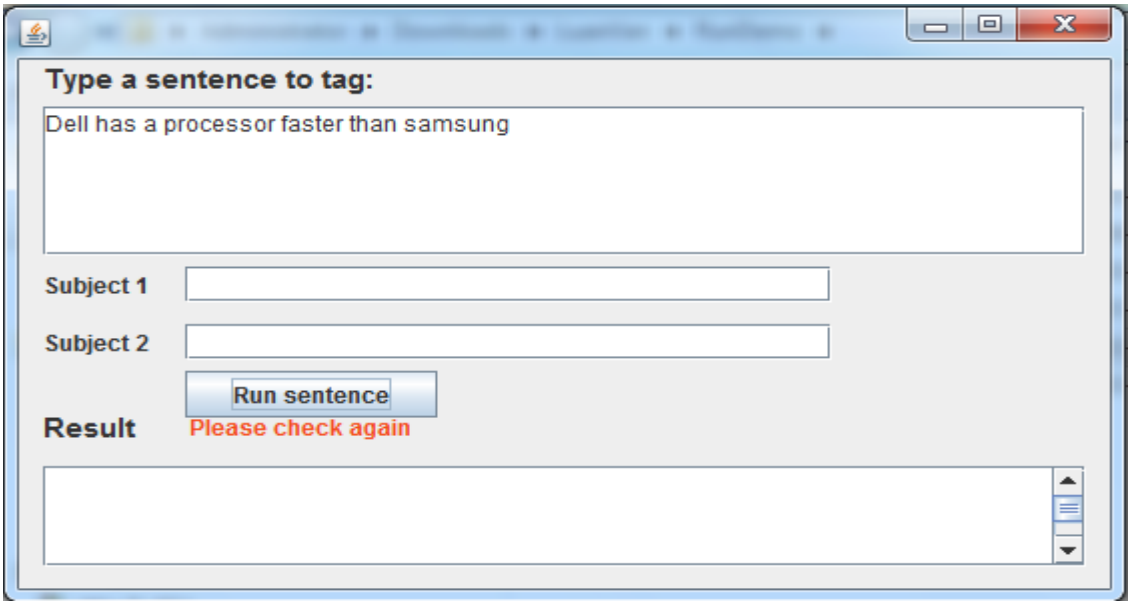
- Câu so sánh



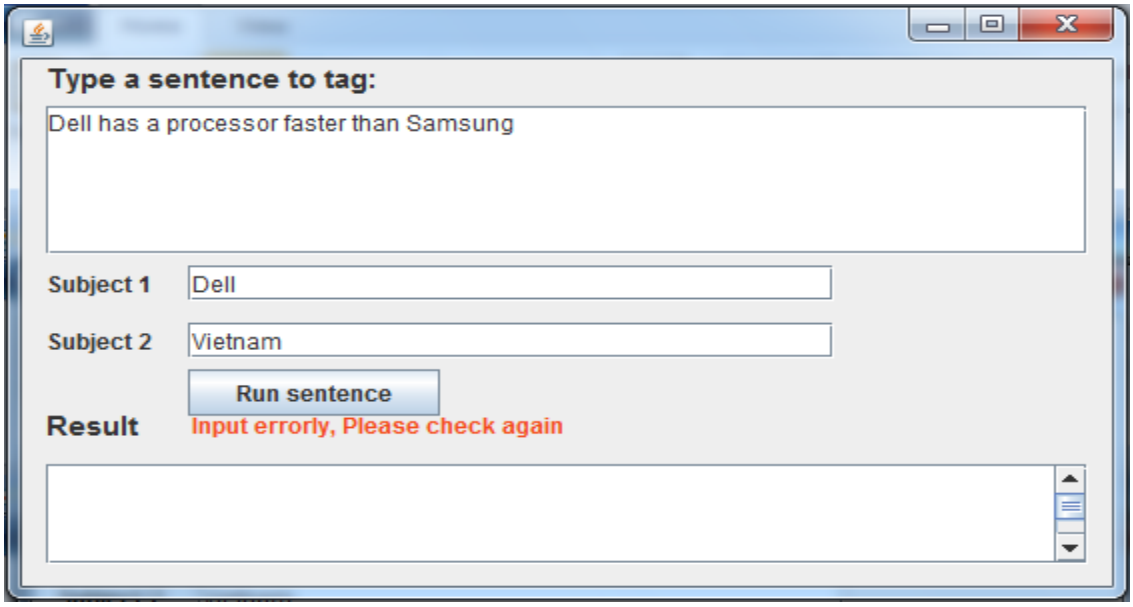
Hình 4.7: Kết quả câu so sánh hơn.

Đối với loại câu so sánh người dùng cần nhập hai chủ thể để thực thi chương trình, trong trường hợp người dùng không nhập, nhập sai hay nhập một chủ thể nào đó

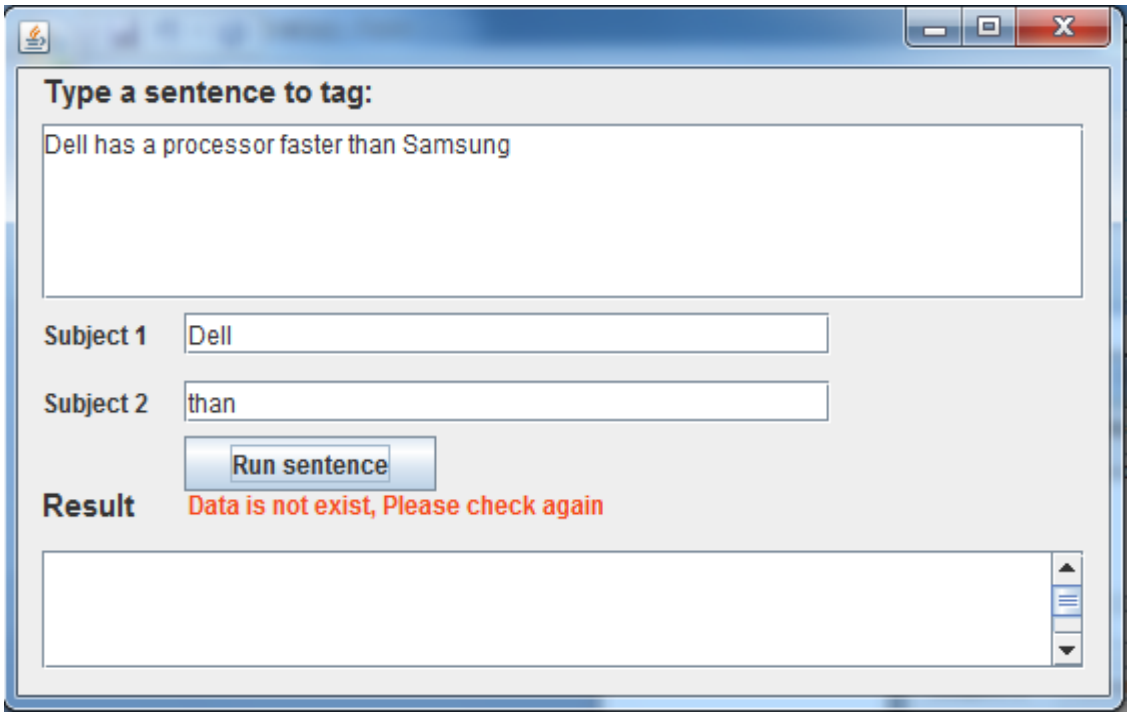
thì chương trình không hiện thị kết quả mà hiện thị thông báo. Sau đây là một số mô hình không hiện thị kết quả.



Hình 4.8: Lỗi không nhập chủ thể hoặc nhập thiếu chủ thể.



Hình 4.9: Lỗi nhập chủ thể không đúng.



Hình 4.10: Lỗi nhập chủ thể không đúng (tiếp).

Để hiểu rõ hơn chúng tôi xây dựng bảng kết quả của 10 câu bình luận đơn giản như sau:

Bảng 4.3: Bảng kết quả phân lớp các câu bình luận (đơn vị: bình luận)

Stt	**Câu bình luận bất kỳ bằng tiếng Anh**	**Kết quả hiện thị đánh giá đúng theo**				
		Mô hình đề xuất			**Con người**	
		Positive	**Negative**	**Neutral**	**Yes**	**No**
1	It's good	x			đ	
2	It's bad		x		đ	
3	It's so so			x	đ	
4	Dell is as bad as Samsung		(Dell, Hp)x		đ	
5	Dell is more beautiful than Acer	x(Dell)	x(Acer)		đ	

6	Dell is good but I don't like it		x		đ	
7	Hp is better than Dell	x(Hp)	x(Dell)		đ	
8	Dell has a processor faster than Hp	x(Desll)	x(Hp)		đ	
9	Dell has a battery faster than Hp	x(Hp)	x(Dell)		đ	
10	It’s good and I love them	x			đ	

Chúng tôi xét 10 câu bình luận khác nhau và kết quả nhận được tương ứng là x (chủ thể tương ứng) hoặc x. Với x có thể là “Positive”, “Negative” và “Neutral” theo bảng phân tích trên. Ngoài ra, chúng tôi cũng xây dựng thêm hai cột để con người có thể đánh giá, so sánh, đối chiếu với kết quả của chính họ với mô hình đề xuất của chúng tôi, trong đó kết quả đánh giá “Yes” là đúng và “No” là không đúng, được đánh dấu “đ”.

CHƯƠNG 5
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Chúng tôi đã xây dựng được mô hình và giải thuật đánh giá được một số câu đơn giản như câu đơn, câu so sánh và câu ghép là đáng tin cây.

Luận văn đã xây dựng được mô hình đánh giá tự động về khai thác quan điểm với giao diện thân thiện, rõ ràng và các chức năng giúp người sử dụng thao tác thuận tiện.

Ngoài ra, chúng tôi đã nghiên cứu các công trình liên quan để giải quyết bài toán phân tích mức độ tình cảm thể hiện trong câu nhận xét, ý kiến.

5.2 Hướng phát triển

Hệ thống của chúng tôi chủ yếu tập trung vào xác định các ý kiến đánh giá của khách hàng chủ yếu là về máy tính, đây là mặt hàng phổ biến hiện nay. Trong tương lai nếu có điều kiện chúng tôi sẽ mở rộng hơn, với nhiều chủ đề khác nhau trên mạng xã hội.

Hệ thống của chúng tôi mới chỉ xác định được mốt số loại câu cơ đơn giản như phủ định, khẳng định, so sánh. Trong tương lai chúng tôi muốn mở rộng thêm các loại câu phức tạp hơn trong tiếng Anh.

Ngoài ra mô hình ontology trong quá trình nghiên cứu mới dừng lại ở mức làm quen, bị giới hạn bởi phạm vi sản phẩm, một hướng phát triển đó là bổ sung thêm vào ontology các tri thức của các sản phẩm như điện thoại, sách, quần áo,... và đây cũng là vấn đề chúng tôi sẽ nghiên cứu tiếp trong tương lai không xa.

TÀI LIỆU THAM KHẢO

[1]. Feldman, R. - Techniques and Applications for Sentiment Analysis. In Communications of the ACM, pp.82-89, 2013.

[2]. Huifeng Tang, Songbo Tan, Xueqi Cheng, A survey on sentiment detection of reviews, Journal Expert Systems with Applications: An International Journal archive, pp.10760- 10773, 2009.

[3]. Peter Turney, Thumbs up or thumbs down, semantic orientation applied to unsupervised classification of reviews, Proc. of the 40th ACL, pp.417-424, 2002.

[4]. B. Pang, L. Lee Thumbs up Sentiment classification using machine learning techniques, pp.1-8, 2002.

[5]. Kushal Dave, Steve Lawrence, and DavidM. Pennock, Mining the peanut gallery Opinion extraction and semantic classification of product reviews, In Proceedings of WWW, pp. 519–528, 2003.

[6]. Taboada, M., Caroline A, & Kimberly V, Creating semantic orientation Dictionaries, in Proceedings of 5th international conference on language resources and evaluation, Italy, 2006.

[7]. Beineke, P.Hastie, T.Vaithyanathan, & S. The sentimental factor: Improving review classification via human-provided information. In Proceedings of the, 42nd ACL conference, 2004.

[8]. Shotaro Matsumoto, Hiroya Takamura, Manabu Okumura, Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees, pp.301-311, 2005.

[9]. Thorsten Joachims. Making large-scale SVM learning practical. In Bernhard Scholkopf and Alexander Smola, editors, Advances in Kernel Methods Support Vector Learning, pp.44–56, 1999.

[10]. Corinna Cortes, Vladimir Vapnik, Support-Vector Networks, Machine Learning, pp.273-297, 1995.

[11]. Kim S. and Eduard H. - Crystal: Analyzing Predictive Opinions on the Web. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.1056-1064, 2007.

[12]. Hu, M. and Liu, B. - Mining and Summarizing Customer Reviews. In Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.166-177, 2004.

[13]. Hu, M. and Liu, B. - Mining Opinion Features in Customer Reviews. In Proceedings of 19th National Conference on Artificial Intelligence, pp.755-761, 2004.

[14]. Alexander O. - Sentiment Mining for Natural Language Documents. In COMP3006 PROJECT REPORT, Computer Science Research Project, Department of Computer Science Australian National University, 2009.

[15]. Casey W., Navendu G. and Shlomo A. - Using Appraisal Groups for Sentiment Analysis. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp.625-631, 2005.

[16]. Ramanathan, N. Bing, L. and Alok, C. - Sentiment Analysis of Conditional Sentences. In Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing, pp.180-189, 2009.

[17]. Khin, S. - Ontology Based Combined Approach for Sentiment Classification. In Proceedings of 3th International Conference on Communications and Information Technology, pp.112-115, 2009.

[18]. Ginsca L., et al. - Sentimatrix – Multilingual Sentiment Analysis Service. In Proceedings of 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL pp.189-195, 2011.

[19]. Ashish S., Vikram G., Denzil C. and Anirban M. - Generating Domain-Specific Ontology from Common-Sense Semantic Network for Target-Specific Sentiment Analysis. In Proceedings of 5th International Conference of the Global WordNet Association, 2010.

[20]http://www.oracle.com/technetwork/database/enterprise-edition/index-098492.html

[20]http://vlsp.vietlp.org:8080/demo/?page=seg_pos_chunk

[22]. AurélienBossard, Michel Généreux and ThierryPoibeau. CBSEAS, a Summarization System Integration of Opinion Mining Techniques to Summarize Blogs, 2008.

[23]. Sushant Kumar and Diptesh Chatterjee. Statistical Model for Opinion Summarization, 2008.

[24]. Jack G. Conrad, Jochen L. Leidner, Frank Schilder, Ravi Kondadadi. Querybased Opinion Summarizationfor Legal Blog Entries, 2008.

[25]. Cortes, C. and Vapnik, V. - Support-Vector Networks. In Journal Machine Learning, pp.273-297, 1995.

[26] Trần Thị Ngọc Thảo, Nguyễn Ngọc Kim Liên, Ngô Minh Vương - Phân Tích Ý Kiến Của Nhận Xét Tiếng Anh Dựa Trên Phương Pháp Học Máy, pp.1-13, 2014.

[27] Walaa Medhata, Ahmed Hassan, Hoda Korashy - Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, pp.1-21, 2014.

[28] G.Vinodhini, RM.Chandrasekaran - Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, pp.1-11, 2012.

[29] C. Fellbaum, ed., Wordnet: An Electronic Lexical Database. MIT Press, 1998.

[30]. C. P. Cheng, G. T. Lau, J. Pan and K. H. Law. Domain-Specific Ontology Mapping by Corpus-Based Semantic Similarity. Proceedings of 2008 NSF CMMI Engineering Research and Innovation Conference, pp.7-10, 2008.

[31] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

[32] https://bayesian.org/bayes

[33] Bing Liu. Chapter 5. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.