

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

NGUYỄN THÀNH LƯU

VÕ LÊ MINH

KHÓA LUẬN TỐT NGHIỆP
PHÂN TÍCH CẢM XÚC DỰA VÀO BÌNH LUẬN
TRÊN MẠNG XÃ HỘI

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2015

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

NGUYỄN THÀNH LƯU - 11520225

VÕ LÊ MINH - 11520229

KHÓA LUẬN TỐT NGHIỆP
PHÂN TÍCH CẢM XÚC DỰA VÀO BÌNH LUẬN
TRÊN MẠNG XÃ HỘI

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

GIẢNG VIÊN HƯỚNG DẪN

PGS. TS. ĐỖ PHÚC

TP. HỒ CHÍ MINH, 2015

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số ngày của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

1. TS. Phạm Văn Hậu – Chủ tịch.
2. ThS. Ngô Quốc Hưng – Thư ký.
3. TS. Ngô Thanh Hùng – Ủy viên.

[illegible]

This image shows a full page of blank handwriting practice paper. It features multiple sets of horizontal lines across the entire page. Each set consists of three lines: a solid top line, a dashed middle line, and a solid bottom line. These sets are repeated vertically down the page, providing a guide for letter height and placement. The background is plain white, and there are no margins or additional markings.

LỜI CẢM ƠN

Để hoàn thành khóa luận này, chúng tôi xin chân thành xin gửi những lời cảm ơn chân thành nhất đến quý thầy cô của trường Đại học Công nghệ thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh, bạn bè trong và ngoài trường,... đã chỉ bảo, quan tâm, giúp đỡ tận tình trong suốt quá trình thực hiện đề tài. Nhờ đó chúng tôi đã có thêm nhiều kinh nghiệm để xử lý những khó khăn gặp phải và hoàn thành tốt đề tài này.

Chúng tôi xin gửi đến lời cảm ơn sâu sắc nhất đến Phó Giáo sư Tiến sĩ Đỗ Phúc đã trực tiếp hướng dẫn, định hướng chuyên môn, quan tâm giúp đỡ tận tình và tạo mọi điều kiện thuận lợi trong quá trình thực hiện khóa luận.

Trong thời gian làm khóa luận, chúng tôi đã có những trải nghiệm bổ ích. Chúng tôi đã được học tập, tìm hiểu nhiều kiến thức mới mẻ. Hơn hết, chúng tôi được tiếp cận với những thành tựu nghiên cứu liên quan đến đề tài cả trong và ngoài nước. Đồng thời, được sự hướng dẫn tận tình và cách làm việc chuyên nghiệp của người hướng dẫn – PGS. TS. Đỗ Phúc chúng tôi đã tích lũy được nhiều kinh nghiệm cho bản thân, phục vụ cho công việc và những đề tài nghiên cứu sau này.

Mặc dù chúng tôi đã cố gắng và nỗ lực để hoàn thành tốt khóa luận của mình, nhưng khó tránh khỏi sai sót, rất mong nhận được sự góp ý và chỉ bảo của quý Thầy, Cô để đề tài được tốt hơn.

Lời cuối cùng, chúng tôi muốn nói cảm ơn tất cả mọi người, những người đã giúp cho chúng tôi có được ngày hôm nay.

TP. Hồ Chí Minh, tháng 07 năm 2015

Sinh viên thực hiện

Nguyễn Thành Lưu – Võ Lê Minh

MỤC LỤC

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI	4
1.1. Tổng quan về mạng xã hội	4
1.2. Tổng quan đề tài	7
1.2.1. Phát biểu bài toán	7
1.2.2. Mục tiêu của đề tài	8
1.2.3. Mô hình tổng quan	12
1.3. Tổng quan tình hình nghiên cứu trong nước và trên thế giới	13
1.3.1. Trong nước	13
1.3.2. Nước ngoài	14
1.4. Bố cục khóa luận	16
Chương 2. CƠ SỞ LÝ THUYẾT	17
2.1. Bộ từ điển cảm xúc SO-CAL tiếng Anh	17
2.2. Phương pháp phân loại chủ quan	21
2.2.1. Câu có từ hàm chứa cảm xúc	21
2.2.2. Các trường hợp ngoại lệ	21
2.3. Phương pháp phân loại cảm xúc	22
2.3.1. Giá trị cảm xúc của câu phụ thuộc vào từ hàm chứa cảm xúc ..	23
2.3.2. Giá trị cảm xúc của câu phụ thuộc vào từ tăng cường	24
2.3.3. Giá trị cảm xúc của câu phụ thuộc vào từ phủ định	24
2.3.4. Giá trị cảm xúc của câu phụ thuộc vào từ khiếm khuyết	25
2.3.5. Giá trị cảm xúc của câu có xu hướng tích cực	25
2.4. Phương pháp phân lớp Support Vector Machine (SVM)	26
Chương 3. XÂY DỰNG HỆ THỐNG THỬ NGHIỆM	28

3.1.	Giới thiệu	28
3.2.	Bộ từ điển cảm xúc SO-CAL tiếng Việt.....	29
3.3.	Thu thập dữ liệu	35
3.4.	Đặc trưng dữ liệu từ mạng xã hội	39
3.5.	Tiền xử lý dữ liệu	40
3.6.	Bộ dữ liệu huấn luyện	42
3.6.1.	Gán nhãn câu bằng tay.....	42
3.6.2.	Mô tả bộ dữ liệu huấn luyện	45
3.7.	Phương pháp phân loại chủ quan.....	46
3.8.	Phương pháp phân loại cảm xúc	48
3.9.	Giao diện hệ thống thực nghiệm.....	56
Chương 4.	KẾT QUẢ THỬ NGHIỆM	63
4.1.	Bộ dữ liệu thử nghiệm.....	63
4.2.	Kết quả đánh giá phương pháp phân loại chủ quan.....	64
4.3.	Kết quả đánh giá phương pháp phân loại cảm xúc	64
Chương 5.	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	66
5.1.	Kết quả đạt được	66
5.2.	Hướng phát triển	66

DANH MỤC HÌNH VẼ

Hình 1-1 Mô hình tổng quan hệ thống phân tích cảm xúc dựa vào bình luận trên mạng xã hội	12
Hình 3-1 Mô hình hệ thống thực nghiệm	28
Hình 3-2 Mô hình Graph API.....	35
Hình 3-3 Ví dụ về dữ liệu thô chưa xử lý.....	37
Hình 3-4 Những bình luận của trang VnExpress.net trên mạng xã hội Facebook.....	38
Hình 3-5 Nội dung các bình luận được lấy về thông qua thư viện Facebook Graph API ..	39
Hình 3-6 Ví dụ dữ liệu dạng mã UNICODETH.....	41
Hình 3-7 Ví dụ dữ liệu sau khi chuyển mã.....	41
Hình 3-8 Giao diện đánh giá độ chính xác của dữ liệu thử nghiệm.....	57
Hình 3-9 Giao diện phân loại cảm xúc, đánh giá mức độ quan tâm của người dùng	58
Hình 3-10 Giao diện thu thập dữ liệu	59
Hình 3-11 Giao diện màn hình bắt đầu	60
Hình 3-12 Giao diện hiển thị từ điển.....	61
Hình 3-13 Giao diện thông tin tác giả	62
Hình I-1 Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp “+” và “-” với khoảng cách biên lớn nhất.....	71
Hình I-2 Minh hoạ bài toán phân hai lớp với phương pháp SVM	74

DANH MỤC BẢNG

Bảng 1-1 Thống kê về số người dùng của một số mạng xã hội phổ biến trên thế giới tính tới tháng 1/2015 (đơn vị: triệu người)	6
Bảng 2-1 Một số từ tăng cường	18
Bảng 2-2 Bảng so sánh hiệu suất của các bộ từ điển khác nhau với từ điển SO-CAL	20
Bảng 3-1 Một số từ trong bộ từ điển danh từ	31
Bảng 3-2 Một số từ trong bộ từ điển động từ	32
Bảng 3-3 Một số từ trong bộ từ điển tính từ	33
Bảng 3-4 Một số từ trong bộ từ điển trạng từ	33
Bảng 3-5 Một số từ trong bộ từ điển từ tăng cường	34
Bảng 3-6 Kết quả phân loại chủ quan bằng tay	45
Bảng 3-7 Kết quả phân loại cảm xúc bằng tay	45
Bảng 4-1 Kết quả phân loại bằng tay bộ dữ liệu thử nghiệm	64
Bảng 4-2 Kết quả đánh giá độ chính xác phương pháp phân loại chủ quan	64
Bảng 4-3 Kết quả đánh giá độ chính xác phương pháp phân loại cảm xúc	65

DANH MỤC TỪ VIẾT TẮT

AAC (Adverb-Adjective Combinations)

API (Application Programming Interface)

cURL (Client for Uniform Resource Locator)

HTTP (The Hypertext Transfer Protocol)

ID (Identification)

PGS. TS (Phó Giáo sư Tiến sĩ)

SO (Semantic Orientation)

SO-CAL (the Semantic Orientation CALculator)

SRM (Structural Risk Minimization)

SVM (Support Vector Machines)

STT (Số thứ tự)

url lib (Uniform Resource Locator Library)

VC (Vapnik-Chervonenkis)

Wifi (Wireless Fidelity)

TÓM TẮT

Mạng xã hội là một công cụ giúp mọi người có thể kết nối với nhau dễ dàng hơn thông qua những những chia sẻ, thông điệp, bình luận hay ý kiến cá nhân về tất cả sự vật, sự việc diễn ra hằng ngày. Từ đó, mọi người sẽ có cái nhìn tổng quan, những thông tin cần thiết hỗ trợ việc đánh giá và đưa ra quyết định đối với mọi vấn đề diễn ra trong cuộc sống. Điều này không chỉ đúng trên phương diện cá nhân mà còn được các tổ chức sử dụng rộng rãi.

Việc thu thập, tổng hợp và phân tích những thông tin trên nếu làm theo phương pháp thủ công sẽ tốn rất nhiều thời gian cũng như chi phí để thực hiện. Nếu có một hệ thống tự động thu thập các thông tin trên mạng xã hội sau đó xử lý, phân loại chúng dựa trên cảm xúc của người dùng sẽ giúp tiết kiệm về nhiều mặt.

Đã có nhiều bài báo, công trình nghiên cứu có liên quan đến đề tài này nhưng hầu hết chúng được sử dụng cho tiếng Anh. Hầu hết các công trình này đều thu được nhiều kết quả khả quan. Tuy nhiên, bài báo và công trình nghiên cứu tương tự bằng tiếng Việt còn rất hạn chế. Hiện nay, với sự phát triển mạnh mẽ của truyền thông mạng xã hội và nhu cầu thu thập ý kiến về các sự vật, sự việc diễn ra xung quanh chúng ta, hướng nghiên cứu này dần được chú ý nhiều hơn ở Việt Nam.

Trong khoá luận này, chúng tôi nghiên cứu các lý thuyết, giải thuật giúp phân loại cảm xúc và tìm hiểu những đặc tính cơ bản của văn phạm tiếng Việt. Từ đó xây dựng mô hình phân tích cảm xúc tiếng Việt và áp dụng trực tiếp trên các bình luận của mạng xã hội. Ngoài ra chúng tôi còn sử dụng một số kỹ thuật xử lý ngôn ngữ tự nhiên hỗ trợ cho việc phân tích dữ liệu hiệu quả và nhanh chóng.

Cuối cùng, chúng tôi tổng hợp kết quả đã đạt được và đưa ra những đánh giá về mô hình phân tích cảm xúc tiếng Việt dựa vào bình luận trên mạng xã hội. Sau đó, đề ra hướng phát triển của đề tài trong tương lai.

MỞ ĐẦU

Với sự phát triển mạnh mẽ của mạng Internet cộng với sự bùng nổ thông tin trên toàn cầu, mạng xã hội đã được sử dụng rộng rãi và dần trở thành một phần không thể thiếu trong cuộc sống con người đặc biệt là giới trẻ - những người luôn quan tâm và cập nhật tin tức thường xuyên. Những tin tức, bình luận, đánh giá về nhiều lĩnh vực được chia sẻ nhanh chóng từ lúc sự việc đang diễn ra và ngay lập tức được lan truyền đến mọi nơi.

Phân loại tâm lý, cảm xúc và khai thác ý kiến trên mạng xã hội sẽ hỗ trợ cho việc nghiên cứu, phân tích cảm xúc, đánh giá thái độ của người dùng mạng xã hội đối với những chủ đề thời sự được chia sẻ. Đây là một trong những lĩnh vực được nghiên cứu rộng rãi trong khai thác dữ liệu Big Data, đồng thời có ý nghĩa quan trọng trong ngành khoa học xử lý ngôn ngữ tự nhiên. Trong thực tế, mức độ ảnh hưởng của nó ngày càng được coi trọng và tỷ lệ thuận với sự bùng nổ thông tin trên mạng Internet.

Chẳng hạn như khi muốn mua một sản phẩm nào đó, chúng ta muốn biết nó tốt hay không? Những lời quảng cáo hoa mỹ của nhà sản xuất chưa đủ thuyết phục, chúng ta muốn nghe những lời đánh giá chân thực từ những người đã sử dụng hoặc có hiểu biết về sản phẩm đó. Hay đối với các tổ chức, những con số khô khan về doanh thu sản phẩm không đủ để họ hài lòng. Họ muốn biết những đánh giá của khách hàng và người dùng về sản phẩm của họ. Những khía cạnh tốt sẽ được duy trì, phát huy và những mặt xấu, không tốt sẽ được họ cải thiện để dần hoàn thiện chất lượng sản phẩm về mọi mặt.

Từ lý do này, chúng tôi lựa chọn đề tài: “Phân tích cảm xúc dựa vào bình luận trên mạng xã hội” nhằm phát triển một phương pháp nghiên cứu phân tích cảm xúc trên ngôn ngữ tiếng Việt dựa trên đặc trưng nguồn dữ liệu từ mạng xã hội ở Việt Nam. Và xây dựng một chương trình thử nghiệm nhằm đánh giá độ đúng đắn của phương pháp bên trên, đồng thời có thể tự động đánh giá những cảm xúc của người dùng trang mạng xã hội đối với những thông tin được đăng tải, chia sẻ.

Đối tượng nghiên cứu của chúng tôi là những bình luận tiếng Việt của người dùng mạng xã hội. Phạm vi của đề tài là xây dựng mô hình phân tích cảm xúc dựa vào bình luận được thu thập trên mạng xã hội Facebook.

Quá trình thực hiện đề tài còn nhiều hạn chế và thiếu sót. Chúng tôi mong nhận được sự đóng góp ý kiến chân thành từ Thầy, Cô và các bạn. Chúng tôi xin cảm ơn.

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Tổng quan về mạng xã hội

Theo Mrutyunjaya Panda và Satchidananda Dehuri: “Một mạng xã hội là một chiếc ô cùng với các nút là các cá nhân, nhóm, tổ chức và các hệ thống liên quan được liên kết với nhau trong một hay nhiều loại phụ thuộc.” [15].

Mạng xã hội có những tính năng như nhắn tin, e-mail, phim ảnh, gọi thoại, chia sẻ dữ liệu, blog và xã luận. Chúng thay đổi hoàn toàn cách người dùng mạng Internet liên kết với nhau và dần trở thành một phần không thể thiếu của hàng trăm triệu thành viên sử dụng mạng xã hội trên khắp thế giới. Các dịch vụ này có nhiều cách để các thành viên tìm kiếm bạn bè, đối tác: dựa theo nhóm (ví dụ như tên trường hoặc tên thành phố), dựa trên thông tin cá nhân (như địa chỉ e-mail hoặc tên hiển thị), hoặc dựa trên sở thích cá nhân (như thể thao, phim ảnh, sách báo, hoặc ca nhạc), lĩnh vực quan tâm: kinh doanh, mua bán, v.v...

Hiện nay thế giới có hàng trăm trang mạng xã hội khác nhau, với MySpace và Facebook nổi tiếng nhất trong thị trường Bắc Mỹ và Tây Âu; Orkut và Hi5 tại Nam Mỹ; Friendster tại Châu Á và các đảo quốc Thái Bình Dương. Mạng xã hội khác gặt hái được thành công đáng kể tại quốc gia mà nó ra đời như Bebo tại Anh Quốc, CyWorld tại Hàn Quốc, Mixi tại Nhật Bản và tại Việt Nam xuất hiện rất nhiều các mạng xã hội như: Zing Me, YuMe, Tamtay...

Lịch sử mạng xã hội:

- Mạng xã hội xuất hiện lần đầu tiên năm 1995 với sự ra đời của trang Classmate với mục đích kết nối bạn học, tiếp theo là sự xuất hiện của SixDegrees vào năm 1997 với mục đích giao lưu kết bạn dựa theo sở thích.
- Năm 2002, Friendster trở thành một trào lưu mới tại Hoa Kỳ với hàng triệu thành viên ghi danh. Tuy nhiên sự phát triển quá nhanh này cũng là con dao hai lưỡi: server của Friendster thường bị quá tải mỗi ngày, gây bất bình cho rất nhiều thành viên.
- Năm 2004, MySpace ra đời với các tính năng nổi bật như phim ảnh (embedded video) và nhanh chóng thu hút hàng chục ngàn thành viên mới mỗi ngày. Các

thành viên cũ của Friendster đã đồng loạt chuyển qua sử dụng MySpace và trong vòng một năm, MySpace trở thành mạng xã hội đầu tiên có nhiều lượt xem hơn cả Google và được tập đoàn News Corporation mua lại với giá 580 triệu USD.

- Năm 2006, sự ra đời của Facebook đánh dấu bước ngoặt mới cho hệ thống mạng xã hội trực tuyến với nền tảng lập trình "Facebook Platform" cho phép thành viên tạo ra những công cụ mới cho cá nhân mình cũng như các thành viên khác dùng. Facebook Platform nhanh chóng gặt hái được thành công vượt bậc, mang lại hàng trăm tính năng mới cho Facebook và đóng góp không nhỏ cho con số trung bình 19 phút mà các thành viên bỏ ra trên trang này mỗi ngày.

Về cơ bản, mạng xã hội giống như một trang web mở với nhiều ứng dụng khác nhau. Mạng xã hội khác với trang web thông thường ở cách truyền tải thông tin và tích hợp ứng dụng. Trang web thông thường cũng giống như truyền hình, cung cấp càng nhiều thông tin, thông tin càng hấp dẫn càng tốt, còn mạng xã hội tạo ra các ứng dụng mở, các công cụ tương tác để mọi người tự tương tác và tạo ra bản tin rồi cùng lan truyền bản tin đó.

Theo thống kê 01-2015 của Wearesocial [16], dân số thế giới là 7.21 tỷ người và số người sử dụng Internet là 3.01 tỷ người. Số người có tài khoản mạng xã hội là 2.078 tỷ người, chiếm gần 1/3 dân số thế giới và chiếm 2/3 người sử dụng Internet. So với 12 tháng trước thì số người có tài khoản mạng xã hội đã tăng thêm 12% (222 triệu tài khoản). Bảng 1-1 là bảng thống kê số người dùng của một số Mạng xã hội phổ biến trên thế giới.

Bảng 1-1 Thống kê về số người dùng của một số mạng xã hội phổ biến trên thế giới tính tới tháng 1/2015 (đơn vị: triệu người)

Tên Mạng xã hội	Số tài khoản người dùng
Facebook	1,366
Qzone	629
Google+	343
Instagram	300
Twitter	284

Tính đến tháng 01-2015, dân số Việt Nam là 90.7 triệu người. Số lượng người dùng Internet và số tài khoản Mạng xã hội lần lượt là 39.8 triệu và 28 triệu. So với năm 2014, số lượng người dùng Mạng xã hội đã tăng 40%. Trung bình mỗi ngày, người dân Việt Nam dành 3 giờ 4 phút cho Mạng xã hội. Mạng xã hội phổ biến nhất Việt Nam là Facebook chiếm 21% người sử dụng Mạng xã hội. Kế tiếp là Goolge+ với 13%, Twitter với 8%, Pinterest với 5%, Linkedin với 5%, Instagram với 5%,...

Những ưu điểm của mạng xã hội

- Kết nối toàn cầu. Mạng xã hội giúp mọi người kết nối, tương tác với nhau một cách nhanh chóng và dễ dàng. Chỉ cần có Internet và tham gia một Mạng xã hội bất kì như Facebook, Twitter, LinkedIn,.. thì mọi người trên khắp thế giới có thể cập nhật tin tức của nhau hàng ngày, trao đổi về nhiều vấn đề trong cuộc sống, chia sẻ với nhau nhiều điều thú vị,...
- Quyền lợi cộng đồng. Người dùng mạng xã hội có thể tham gia một cộng đồng mạng xã hội và chọn lựa những người bạn thích, không thích. Ví dụ nếu thích cờ vua, người dùng có thể tham gia một cộng đồng về chủ đề này. Đây là nơi có thể tìm thấy những người cùng đam mê và chia sẻ những trải nghiệm về cờ vua với họ.
- Chia sẻ thông tin thời gian thực. Ngay sau khi đăng một tin hay nhấn một tin nhấn đến ai đó, nó sẽ được gửi đi ngay. Vì vậy chỉ cần có mạng và một thiết bị có thể truy cập vào Mạng xã hội, bạn có thể thoải mái trao đổi hay cập nhật

tức thời mọi tin tức, thông tin của mọi người cũng như của những cộng đồng với chủ đề mà bạn quan tâm.

- Truyền thông Mạng xã hội. Chỉ cần một vài thao tác đơn giản, tên công ty, tên sản phẩm hoặc một hoạt động mới của bạn sẽ được gửi tới hàng triệu người dùng Mạng xã hội trên khắp thế giới.

Bên cạnh những ưu điểm, Mạng xã hội vẫn tồn tại một số hạn chế:

- Gặp gỡ trực tiếp ngày càng ít và đang ở mức báo động. Bằng việc cung cấp những công cụ vô cùng tiện lợi, người dùng mạng xã hội thay vì gặp gỡ trao đổi trực tiếp, mọi người lại lựa chọn giao tiếp thông qua mạng xã hội. Nơi mà họ không cần nhìn vào ánh mắt, quan tâm đến khuôn mặt hay bất cứ ngôn ngữ cơ thể nào của người đối diện. Các kỹ năng giao tiếp xã hội theo đó bị tê liệt và ngày càng giảm sút.
- Lãng phí quá nhiều thời gian cho Mạng xã hội. Thay vì sử dụng Mạng xã hội như một công cụ hỗ trợ tích cực cho cuộc sống hàng ngày, nhiều người lãng phí quá nhiều thời gian cho các hoạt động không cần thiết trên đó. Gần đây hội chứng “nghiện mạng xã hội” đang ngày một gia tăng trong giới trẻ. Nếu không được khắc phục kịp thời sẽ dẫn đến những tác động tiêu cực đến đời sống cá nhân.
- Vấn đề bí mật thông tin cá nhân. Mạng xã hội mời các tập đoàn lớn xâm nhập quyền riêng tư và bán thông tin cá nhân của bạn. Đã bao giờ bạn thấy một nội dung quảng cáo xuất hiện có nội dung liên quan đến một tin hay bình luận bạn vừa chia sẻ trên mạng xã hội? Và nếu các thông tin của bạn không chỉ được sử dụng cho việc quảng cáo mà còn bị kẻ xấu lợi dụng để sử dụng vào các hành vi phi pháp hoặc gây ảnh hưởng xấu tới bạn?

1.2. Tổng quan đề tài

1.2.1. Phát biểu bài toán

Kể từ năm 2000, cùng với sự lớn mạnh của truyền thông xã hội trên mạng Internet như diễn đàn, blog và đặc biệt là mạng xã hội (Facebook, Google plus, Twitter, Instagram,...), phân tích cảm xúc (Sentiment Analysis) đã phát triển nhanh

chóng và trở thành lĩnh vực nghiên cứu sôi động nhất trong chuyên ngành xử lý ngôn ngữ tự nhiên. Mạng xã hội ngày càng có tầm ảnh hưởng không chỉ với doanh nghiệp mà còn với toàn xã hội.

Ý kiến là trung tâm của hầu hết các hoạt động và có ảnh hưởng lớn đến hành vi của con người. Thông thường khi cần phải đưa ra quyết định, chúng ta thường tham khảo ý kiến của người khác. Đối với cá nhân, họ thường tham khảo người thân, bạn bè hay mọi người xung quanh. Còn với tổ chức, họ tham khảo ý kiến của các hội đồng, của nhân viên, khách hàng,...

Chẳng hạn, có một cô gái trẻ đang đọc các tin tức trên bảng tin của một mạng xã hội nào đó. Bất chợt cô ấy thấy một chiếc điện thoại mới được giới thiệu kèm theo nhiều tính năng hiện đại với mức giá cực kỳ hấp dẫn. Nhưng ngay lập tức, cô ấy liền đặt câu hỏi: “Chiếc điện thoại này có tốt như những gì nhà sản xuất quảng cáo không?”. Bình thường thì cô ấy phải vất vả đọc thủ công từng bình luận chia sẻ của người dùng. Sau đó tổng hợp lại và đưa ra đánh giá cuối cùng. Công việc vô cùng đơn giản với năm hay mười bình luận. Nhưng nếu số bình luận lên đến năm mươi, một trăm hay vài trăm thì công việc này trở nên phức tạp hơn nhiều. Và liệu cô gái trẻ có nhớ và tổng hợp hết những bình luận đó một cách chính xác trong thời gian ngắn hay không?

Một ví dụ khác: Mỗi năm một công ty kinh doanh hàng hóa bỏ ra một lượng lớn thời gian, công sức và tiền bạc để khảo sát xu hướng thị trường hay nói cách khác là họ tìm hiểu xem thị hiếu của người dùng hiện tại là gì? Việc khai thác các hoạt động của người dùng trên mạng xã hội hay cụ thể hơn là các bình luận, đánh giá của họ sẽ giúp cho công ty thực hiện việc khảo sát một cách dễ dàng cũng như tiết kiệm được nhiều chi phí.

1.2.2. Mục tiêu của đề tài

Mục tiêu của đề tài là phát hiện những cảm xúc của người dùng mạng xã hội thông qua việc phân tích những bình luận của họ đối với thông tin được đăng tải hay chia sẻ. Để thực hiện được mục tiêu này, chúng tôi đã chia thành mục tiêu thành 3 mục tiêu thành phần. Đó là:

- Lấy dữ liệu từ mạng xã hội.
- Xây dựng bộ từ điển cảm xúc.
- Phân tích, đánh giá cảm xúc.

a) Lấy dữ liệu từ mạng xã hội

Mạng xã hội có mức độ bùng nổ thông tin rất cao. Những mạng xã hội nổi tiếng như Facebook, Twitter, Google Plus, Youtube,... có lượt người truy cập rất lớn.

- Twitter: dữ liệu cập nhật ngày 11/07/2014 [17].
 - Tổng số người dùng: 646 triệu người dùng.
 - Số người dùng mới đăng ký hàng ngày: 135 nghìn người dùng.
 - Số người dùng truy cập hàng tháng: 190 triệu người dùng.
 - Số tweets mỗi ngày: 58 triệu.
- Facebook: dữ liệu cập nhật ngày 01/07/2014 [18].
 - Số người dùng hoạt động hàng tháng: 1,3 tỷ người dùng.
 - Cứ mỗi 20 phút, Facebook sẽ:
 - Chia sẻ 1 triệu liên kết.
 - Gởi 3 triệu tin nhắn.
 - Gởi 2 triệu yêu cầu kết bạn.

Từ đó chúng ta có thể nhận thấy khối lượng dữ liệu trên các trang mạng xã hội là vô cùng lớn. Đi kèm với đó là lượng thông tin mà các trang mạng xã hội này cung cấp cho người dùng cũng rất nhiều. Tuy nhiên, không phải tất cả thông tin đó đều có ích và theo dạng chuẩn của ngôn ngữ tiếng Việt. Do đó bài toán đặt ra cho chúng tôi ba khó khăn cần giải quyết, đó là:

- Lấy dữ liệu lớn từ mạng xã hội.
- Chuẩn hóa dữ liệu cho phù hợp với phương pháp.
- Phân tích cảm xúc dựa vào nguồn dữ liệu được chuẩn hóa trên.

b) Xây dựng bộ từ điển cảm xúc

Hiện nay, có hai phương pháp tiếp cận chính để giải quyết vấn đề trích xuất cảm xúc tự động. Cách đầu tiên dựa vào các từ vựng thông qua việc tính toán giá trị

ngữ nghĩa (semantic orientation) của các từ hay cụm từ trong tài liệu. Cách tiếp cận thứ hai sử dụng một phương pháp thống kê hoặc máy học để giải quyết vấn đề.

Sau nhiều thời gian tìm tòi nghiên cứu, chúng tôi quyết định sử dụng phương pháp đầu tiên. Và bước đầu tiên để tính toán được các giá trị ngữ nghĩa cần dựa trên một tập hợp các từ và giá trị ngữ nghĩa của chúng hay còn gọi là từ điển cảm xúc.

Theo khả năng hiểu biết của chúng tôi, hiện tại chưa có một bộ từ điển cảm xúc cho tiếng Việt nào được công bố chính thức. Việc xây dựng một bộ từ điển cần đầu tư nhiều thời gian, kinh phí và nhất là cần có sự hợp tác của những chuyên gia về ngôn ngữ học. Vì vậy, chúng tôi đã sử dụng bộ từ điển tiếng Anh có tên từ điển SO-CAL [19] (Dictionaries for the Semantic Orientation CALculator) của nhóm tác giả Maite Taboada [4] và dịch bộ từ điển này sang tiếng Việt. Từ điển cảm xúc SO-CAL có khoảng 6600 từ chia thành năm từ điển nhỏ gồm có: từ điển danh từ, từ điển động từ, từ điển tính từ, từ điển động từ và từ điển từ tăng cường (intensifier). Mỗi từ điển bao gồm một danh sách các từ cảm xúc và các giá trị SO kèm theo.

c) Phân tích, đánh giá cảm xúc

Sau khi xây dựng xong từ điển cảm xúc, mục tiêu cuối cùng sẽ là phân tích đánh giá cảm xúc dựa vào những bình luận được thu thập trên mạng xã hội. Để phân tích cảm xúc có hai vấn đề cần giải quyết là phân loại câu có cảm xúc hay không có cảm xúc và phân loại câu có cảm xúc tích cực hay tiêu cực.

Theo Bing Luu [5], phân tích cảm xúc hiện được tập trung nghiên cứu chủ yếu ở 3 mức độ:

- Phân tích cảm xúc mức văn bản (document level):

Mục tiêu ở mức độ này là phân loại xem quan điểm tổng thể của văn bản diễn tả một cảm xúc tiêu cực hay tích cực. Phân tích cảm xúc mức văn bản giả định rằng mỗi văn bản thể hiện quan điểm về một thực thể duy nhất.

Có nhiều phương pháp đã và đang được nghiên cứu ở mức này như phương pháp giám sát, phương pháp không giám sát, phương pháp học máy (Support Vector

Machine, Maximum Entropy, K-Nearest Neighbors, Naïve Bayes, Centroid Classification), v.v...

- **Phân tích cảm xúc mức câu (sentence level):**

Ở mức độ này sẽ tập trung vào các câu và xác định xem chúng bày tỏ một quan điểm tích cực, tiêu cực hay trung tính. Theo Wiebe, Bruce và O'Hara [6], khi phân loại chủ quan một câu được chia làm hai loại là câu chủ quan (câu có cảm xúc) và câu khách quan (câu không có cảm xúc). Câu khách quan thể hiện một số thông tin thực tế còn câu chủ quan thường mang đến góc nhìn hay ý kiến cá nhân. Trong thực tế, câu chủ quan có thể diễn tả nhiều loại thông tin như ý kiến, đánh giá, cảm xúc, niềm tin, suy đoán, phán đoán, cáo buộc,... Để đánh giá trạng thái cảm xúc của câu chủ quan, người ta chia nó thành hai loại là câu có cảm xúc tích cực (như vui, thích, yêu, hưng phấn, tự tin) và câu có cảm xúc tiêu cực (như chán, ghét, hận, tức giận, sợ hãi).

- **Phân tích cảm xúc mức thực thể và khía cạnh của thực thể (Entity and Aspect level):**

Cả hai mức độ văn bản và câu đều không phát hiện được chính xác những quan điểm của người viết. Mức thực thể và khía cạnh của thực thể thực hiện phân tích sâu và chi tiết hơn. Thay vì nhìn vào cấu trúc ngôn ngữ (như văn bản, đoạn văn, câu văn, mệnh đề hay cụm từ), mức này tập trung trực tiếp vào ý kiến, quan điểm của người viết. Nó dựa trên ý tưởng rằng một ý kiến, quan điểm bao gồm một cảm xúc (tích cực hoặc tiêu cực) và một mục tiêu cụ thể. Mục tiêu này giúp chúng ta phân tích cảm xúc tốt hơn. Trong nhiều ứng dụng, mục tiêu của quan điểm, ý kiến được mô tả dựa vào các thực thể và các khía cạnh của chúng. Như vậy, mục tiêu của mức phân tích này là xác định cảm xúc về các thực thể và/hoặc các khía cạnh của các thực thể đó.

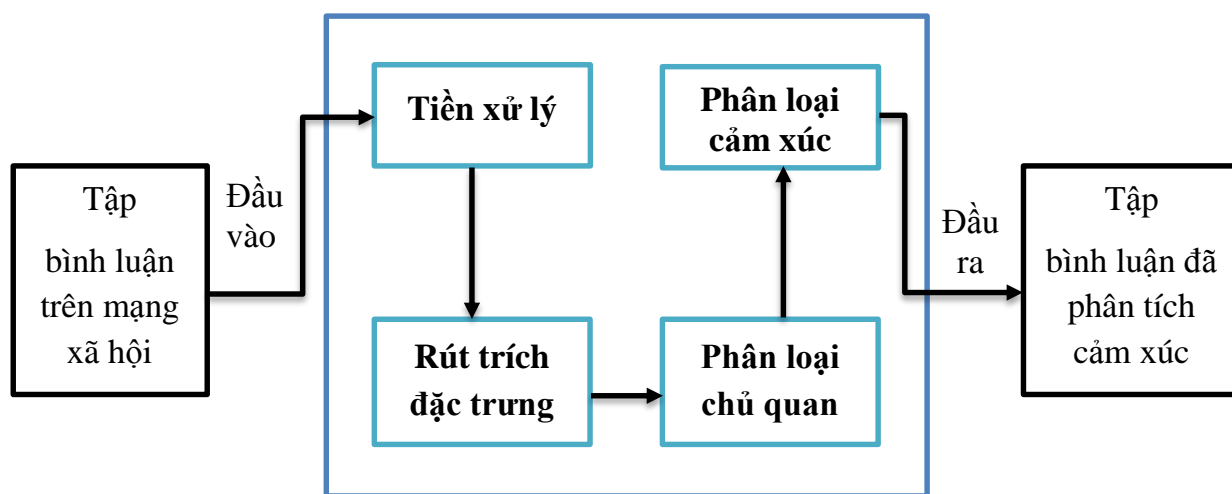
Ví dụ: *“Thời lượng pin và chất lượng cuộc gọi của Iphone rất tốt nhưng khả năng bắt sóng Wifi của nó kém.”*

Ví dụ trên đưa ra ý kiến về ba khía cạnh là thời lượng pin, chất lượng cuộc gọi và khả năng bắt sóng Wifi của thực thể Iphone. Khía cạnh thời lượng pin và chất

lượng cuộc gọi nhận được cảm xúc tích cực còn khía cạnh bất sóng Wifi nhận được cảm xúc tiêu cực. Căn cứ vào các thông tin được thu thập từ mức độ này, một bản tổng hợp ý kiến, quan điểm về các thực thể và khía cạnh của thực thể sẽ được xây dựng phục vụ cho việc biến văn bản phi cấu trúc thành dữ liệu có cấu trúc. Sau này, có thể dùng các dữ liệu này để tiến hành các phân tích định tính định lượng.

Nếu mức văn bản và mức câu đã là những thử thách khó thì mức thực thể và khía cạnh thậm chí còn khó hơn. Đòi hỏi nhiều thời gian điều tra, khảo sát và tổng hợp để xây dựng được tập các thực thể và khía cạnh của chúng. Đồng thời mức thực thể và khía cạnh đưa ra các bài toán đòi hỏi năng lực xử lý ngôn ngữ tự nhiên sâu và chi tiết hơn. Do thời gian và chi phí có hạn của một luận văn cử nhân, chúng tôi quyết định chỉ dừng lại ở phân tích cảm xúc mức câu và sẽ dành mức thực thể và khía cạnh cho những nghiên cứu sau này.

1.2.3. Mô hình tổng quan



Hình 1-1 Mô hình tổng quan hệ thống phân tích cảm xúc dựa vào bình luận trên mạng xã hội

Mô hình tổng quan của hệ thống phân tích cảm xúc gồm ba phần:

- Đầu vào: Tập bình luận tiếng Việt “thô” trên mạng xã hội.
- Hệ thống phân tích cảm xúc: Gồm có bốn hoạt động chính
 - Tiền xử lý.
 - Rút trích đặc trưng.
 - Phân loại chủ quan.

- Phân loại cảm xúc.
- Đầu ra: Tập bình luận tiếng Việt sau khi được hệ thống phân tích cảm xúc đã được phân thành 3 loại: Không có cảm xúc, có cảm xúc tích cực và có cảm xúc tiêu cực.

Đầu tiên dữ liệu đầu vào sẽ là tập các bình luận tiếng Việt “thô” trên mạng xã hội. Đánh giá các bình luận này “thô” bởi vì trước khi có thể sử dụng được chúng, chúng ta cần phải giải quyết nhiều vấn đề như xử lý lỗi tiếng Việt có dấu, xử lý biểu tượng cảm xúc, xử lý “stop words”,... gọi chung là tiền xử lý. Sau khi tiền xử lý xong thu được tập bình luận đã được chuẩn hoá, hệ thống bắt đầu rút trích các đặc trưng của từng câu dựa vào từ điển cảm xúc và các yếu tố ảnh hưởng đến cảm xúc trong câu. Từ các đặc trưng thu được tiến hành phân loại chủ quan và phân loại cảm xúc tập bình luận để cuối cùng xuất ra tập các câu bình luận được phân thành ba loại: không có cảm xúc, có cảm xúc tích cực và có cảm xúc tiêu cực.

1.3. Tổng quan tình hình nghiên cứu trong nước và trên thế giới

1.3.1. Trong nước

Hiện nay, trong nước có rất ít đề tài nghiên cứu về chủ đề trích xuất thông tin từ mạng xã hội (hoặc có nghiên cứu nhưng chưa được công bố rộng rãi). Tuy nhiên, việc nghiên cứu đề tài liên quan đến mạng xã hội ở Việt Nam lại được tập trung vào việc lấy dữ liệu từ mạng xã hội và phân tích tách từ tiếng Việt. Một số đề tài nổi bật:

- Nhóm tác giả Lê Hồng Phương xây dựng công cụ “vnTokenizer” [20]. Theo đó, công cụ này là sự kết hợp giữa từ điển Tiếng Việt và giải thuật ngram.
- Công cụ “vnTagger” [21] để phân loại từ. Xây dựng trên phương pháp gán nhãn từ loại tiếng Việt.
- Sentiment classification using Enhanced Contextual Valence Shifters [3]. Nhóm tác giả Võ Ngọc Phú và Phan Thị Tươi trình bày một phương pháp phân loại cảm xúc tiếng Việt dựa vào giá trị cảm xúc và ngữ cảnh của văn bản. Nhóm đã xây dựng bộ từ điển cảm xúc Tiếng Việt và liệt kê các ngữ cảnh ảnh hưởng đến giá trị cảm xúc của các từ và câu văn trong văn bản. Trong hầu hết các ngữ cảnh, nhóm tác giả đều đưa ra phương pháp giải quyết cụ thể góp phần

nâng cao độ chính xác của quá trình tính toán giá trị cảm xúc trong văn bản tiếng Việt.

- Tóm tắt ý kiến trên cơ sở phân loại cảm xúc [2]. Tác giả Nguyễn Ngọc Duy đã xây dựng mô hình tóm tắt các ý kiến trên cơ sở phân loại cảm xúc từ ý kiến của bạn đọc trên các trang báo mạng và của người dùng trên các trang mạng xã hội tiếng Việt. Với kho ngữ liệu gồm 220 ý kiến từ hai chủ đề là xã hội và kinh doanh, mô hình của tác giả đã đạt những kết quả tích cực.
- Ngày 11-06-2015, Thạc sĩ Võ Ngọc Phú đã tổ chức seminar trình bày về chủ đề "Phân loại cảm xúc văn bản" tại trường Đại học Công nghệ thông tin. Buổi seminar cung cấp cái nhìn khái quát về phân tích cảm xúc trong văn bản tiếng Việt.

1.3.2. Nước ngoài

Bài toán phát hiện, trích xuất và phân tích thông tin trên mạng xã hội được quan tâm bởi rất nhiều nhà khoa học và nghiên cứu sinh trên toàn thế giới. Đề tài liên quan đến việc phân tích thông tin, cảm xúc từ mạng xã hội đã được nghiên cứu nhiều năm ở nhiều nước với nhiều ngôn ngữ khác nhau, trong đó phổ biến nhất vẫn là dữ liệu bằng tiếng Anh. Việc tìm hiểu, tham khảo và đánh giá thành công cũng như hạn chế của những nghiên cứu này trên thế giới cung cấp cái nhìn tổng quan về đề tài.

Dưới đây là một số bài báo liên quan đến đề tài mà chúng tôi đã tìm hiểu và tham khảo:

a) A Sentimental Education: Sentiment Analysis Using Subjective Summarization Based on Minimum Cuts [7]

Phân tích tâm lý, tình cảm là phương pháp tìm cách xác định những quan điểm nằm bên dưới một chuỗi ký tự. Để xác định được tình cảm này, nhóm tác giả đề xuất một phương pháp học máy tiểu thuyết mà áp dụng các kỹ thuật phân loại văn bản để chỉ ra các phần chủ quan của tài liệu. Giải nén những phần có thể được thực hiện bằng cách sử dụng các kỹ thuật hiệu quả cho việc tìm kiếm cắt giảm tối thiểu trong đồ thị.

Phương pháp này tạo thuận lợi lớn cho việc phân tích dữ liệu sử dụng câu trong từng ngữ cảnh cụ thể, xác định.

b) Large-Scale Sentiment Analysis for News and Blogs [8]

Các cơ quan truyền thông: báo, công ty truyền thông, truyền hình v.v... thể hiện ý kiến của họ về những sự vật, hiện tượng của mình thông qua những bài viết. Tác giả trình bày một hệ thống gán điểm cho thấy quan điểm tích cực hay tiêu cực cho từng đối tượng riêng biệt trong ngữ liệu văn bản. Hệ thống được xây dựng bao gồm một giai đoạn xác định tâm lý cộng với việc bày tỏ ý kiến với từng đối tượng có liên quan, và một tập hợp những tâm lý được ghi lại cụ thể qua từng giai đoạn, trong đó điểm số mỗi thực thể liên quan đến những người dùng khác nhau trong cùng một chủ đề được đề cập.

Cuối cùng, nhóm tác giả đánh giá tầm quan trọng của kỹ thuật này lên một bộ ngữ liệu lớn các tin tức và bài viết được công khai trên Internet.

c) Sentiment Analysis: A Combined Approach [9]

Hiện nay, phân tích tâm lý, cảm xúc là một lĩnh vực nghiên cứu quan trọng và có ý nghĩa ứng dụng. Tác giả kết hợp phân loại dựa trên nguyên tắc học có giám sát và máy học thành một phương pháp mới. Phương pháp này được thử nghiệm trên đánh giá phim, đánh giá sản phẩm và ý kiến của người dùng trên “MySpace”. Kết quả cho thấy một phân loại “hybird” có thể nâng cao hiệu quả phân loại về mặt vi mô và vĩ mô trung bình. Ngoài ra, nhóm tác giả còn đề xuất một cách tiếp cận bổ sung bán tự động, trong đó mỗi phân lớp có thể đóng góp vào các phân loại khác nhau để đạt được mức độ hiệu quả tốt nhất.

d) Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone [10]

Hầu hết các nghiên cứu trước đây về việc thể hiện cảm xúc chủ yếu được xác định qua các từ loại: tính từ, động từ và danh từ. Tác giả đề xuất một kỹ thuật phân tích tâm lý AAC (sử dụng kết hợp trạng từ, tính từ) dựa trên việc sử dụng một phân tích ngôn ngữ của phó từ chỉ mức độ. Tác giả định nghĩa một tập hợp các tiên đề chung (dựa trên phân loại phó từ chỉ mức độ thành 5 loại) mà tất cả các kỹ thuật sử

dụng trạng từ phải đáp ứng. Thay vì tính tổng điểm của cả hai trạng từ và tính từ, tác giả đề xuất một phương pháp phân loại ngôn ngữ của trạng từ.

e) Twitter Sentiment Analysis [11]

Những dòng trạng thái của người dùng Twitter được gọi là tweet. Những tweets đôi khi bày tỏ ý kiến về các chủ đề khác nhau. Mục đích của dự án là xây dựng một thuật toán mà có thể phân loại chính xác các thông điệp Twitter là tích cực hay tiêu cực, đối với một thuật ngữ truy vấn mới. Giả thiết của tác giả là dựa vào các kỹ thuật học máy để xây dựng thuật toán có được độ chính xác cao về phân loại tình cảm trong các tin nhắn Twitter.

Nói chung, loại hình này phân tích tình cảm là rất hữu ích cho người sử dụng để nghiên cứu một sản phẩm hoặc dịch vụ, hoặc lấy ý kiến đánh giá của dư luận của công ty họ.

f) Twitter Sentiment Analysis: The Good the Bad and the OMG! [12]

Tác giả thực hiện một điều tra các tiện ích của tính năng ngôn ngữ để phát hiện tình cảm của các thông điệp trên mạng xã hội Twitter. Sau đó, họ đánh giá tính hữu ích của các nguồn tài nguyên từ vựng hiện tại cũng như các tính năng mà nắm bắt thông tin về các ngôn ngữ chính thức và sáng tạo trong sử dụng microblogging. Tác giả có một cách tiếp cận để giám sát các vấn đề, cơ sở hiện nay là những “hashtags” trong các dữ liệu Twitter để xây dựng dữ liệu huấn luyện.

1.4. Bố cục khóa luận

Khóa luận gồm có 05 chương và có bố cục như sau:

Chương 1: Tổng quan về đề tài

Chương 2: Trình bày cơ sở lý thuyết.

Chương 3: Xây dựng hệ thống thử nghiệm.

Chương 4: Trình bày kết quả thử nghiệm.

Chương 5: Kết luận và hướng phát triển cho đề tài.

Chương 2. CƠ SỞ LÝ THUYẾT

Trong chương này chúng tôi sẽ trình bày cơ sở lý thuyết mà chúng tôi đã áp dụng để xây dựng hệ thống thực nghiệm nhằm phân tích cảm xúc dựa vào nguồn dữ liệu trên mạng xã hội. Đồng thời, hướng đi mà chúng tôi lựa chọn khi tiếp cận vấn đề đó là chia bài toán lớn thành 02 bài toán thành phần. Bài toán thứ nhất là phân biệt câu có (hoặc không có) hàm ý cảm xúc. Bài toán thứ hai là từ những câu có cảm xúc đó làm thế nào để nhận biết câu có hàm ý cảm xúc gì (hàm chứa cảm xúc tích cực hay hàm chứa cảm xúc tiêu cực)? Do đó, chúng tôi sẽ trình bày về 04 vấn đề được áp dụng để giải quyết bài toán trên. Bao gồm:

- Bộ từ điển cảm xúc SO-CAL tiếng Anh.
- Phương pháp phân loại chủ quan.
- Phương pháp phân loại cảm xúc.
- Phương pháp phân lớp Support Vector Machine.

2.1. Bộ từ điển cảm xúc SO-CAL tiếng Anh

Bộ từ điển SO-CAL [19] bao gồm 5 bộ từ điển nhỏ là: từ điển danh từ, từ điển động từ, từ điển tính từ, từ điển trạng từ và từ điển từ tăng cường (intensifier). Số lượng từ của các bộ từ điển danh từ, động từ, tính từ và trạng từ lần lượt là 1142 từ, 903 từ, 2252 từ, 745 từ và kèm theo mỗi từ là một số nguyên thể hiện giá trị SO tương ứng trong phạm vi từ -5 cho hết sức tiêu cực đến +5 cho hết sức tích cực và không có từ nào có giá trị SO là 0. Các từ trong bộ từ điển này được lấy từ nhiều nguồn khác nhau và 3 nguồn lớn nhất là:

- Epinions 1: bộ sưu tập gồm 400 văn bản về 8 chủ đề khác nhau: sách, xe hơi, máy vi tính, đồ nấu nướng, khách sạn, phim ảnh, âm nhạc và điện thoại, và được chia đều một nửa tiêu cực và một nửa tích cực [22].
- Một tập hợp con 100 văn bản chứa 2000 bình luận phim trong tập dữ liệu Polarity (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2004, 2005) [23].
- Từ tích cực và tiêu cực từ từ điển General Inquirer (Stone et al. 1966; Stone 1997).

Từ điển từ tăng cường gồm hơn 200 từ được chia làm thành 2 loại là những từ làm tăng mức độ ngữ nghĩa (amplifiers) và những từ làm giảm mức độ ngữ nghĩa (downtoners).

Bảng 2-1 Một số từ tăng cường

Từ tăng cường		Mức độ
Tiếng Anh	Tiếng Việt	
Slightly	hơi	-0.5
somewhat	một chút	-0.3
Pretty	khá	-0.1
Really	thật sự	+0.15
Very	rất	+0.25
extraordinarily	cực kỳ	+0.5
(the) most	nhất	+1

Ví dụ: Từ “sleazy” (nhếch nhác) có giá trị SO là -3 thì “pretty sleazy” (khá nhếch nhác) có giá trị SO là $-3 \times (1 - 0,1) = -2,7$. Từ “excellent” (xuất sắc) có giá trị SO là 5 thì “most excellent” (xuất sắc nhất) có giá trị SO là $5 \times (1 + 1) = 10$.

Từ phủ định được chia làm 2 loại:

- Switch negation (từ phủ định chuyển đổi)
 - Các từ Switch negation như not (không), never (không bao giờ), nobody (không ai),... chỉ đơn giản là đảo ngược cực của một từ hay dễ hiểu hơn là đổi dấu giá trị SO của từ.
 - **Ví dụ:** “Tốt” có giá trị SO là +3 thì “không tốt” có giá trị SO là -3.
- Shift negation (từ phủ định thay đổi)
 - Nếu sử dụng Switch negation thì “xuất sắc” sẽ có giá trị SO là 5, “không xuất sắc” sẽ có giá trị SO là -5. Tương tự “không tốt” sẽ có giá trị SO

là -3. Trên thực tế, thì “không xuất sắc” sẽ có cảm xúc tích cực hơn “không tốt”. Để tránh trường hợp đó, Shift negation sẽ thay đổi giá trị SO của từ phủ định cho phù hợp với thực tế.

- **Ví dụ:** *Cruise là không tốt* (giá trị cảm xúc: $4 - 4 = 0$), *nhưng tôi phải thừa nhận ông không phải xấu tính* (giá trị cảm xúc: $-3 + 4 = 1$).

Đánh giá từ điển SO-CAL so với các từ điển khác

- Bộ thử nghiệm:
 - Epinions 1: Bộ sưu tập gồm 400 văn bản đánh giá về sách, xe hơi, máy tính, nấu ăn, khách sạn, phim, âm nhạc và điện thoại.
 - Epinions 2: Bộ sưu tập mới từ 400 văn bản ở trang web epinions.com tương tự như Epinions 1.
 - Movie: 1,900 văn bản từ kho dữ liệu phân cực (Pang and Lee 2004).
 - Camera: 2,400 văn bản về máy ảnh, máy in, đánh giá xe của Bloom, Garg, and Argamon (2007).
- Các bộ từ điển sử dụng:
 - Từ điển Google-generated PMI-based theo mô tả của Taboada, Anthony, and Voll (2006).
 - Từ điển Maryland (Mohammad, Dorr, và Dunne 2009) là một bộ sưu tập rất lớn khoảng 70000 từ và cụm từ.
 - Từ điển General Inquirer.
 - Từ điển Subjective được Wilson, Wiebe, and Hoffmann công bố năm 2005.
 - Từ điển SentiWordNet (Esuli and Sebastiani 2006; Baccianella, Esuli, and Sebastiani 2010).
 - Từ điển SO-CAL cơ bản (SO-CAL-Basic) và từ điển SO-CAL đầy đủ (SO-CAL-Full).
 - Từ điển SO-CAL cơ bản bao gồm 4 bộ từ điển cơ bản là từ điển danh từ, động từ, tính từ và trạng từ.

- Từ điển SO-CAL đầy đủ bao gồm cả 4 bộ từ điển cơ bản cộng thêm từ điển Intensifier, từ điển từ phủ định và các tính năng đặc biệt của SO-CAL như Irrealis Blocking (ngăn chặn phi thực tế), negative weighting (tăng 50% giá trị SO của từ tiêu cực), repetition weighting (sự xuất hiện lần thứ n của một từ trong văn bản có giá trị SO là (giá trị SO của từ đó) / n).
- So sánh hiệu suất của các bộ từ điển khác nhau với từ điển SO-CAL

Bảng 2-2 Bảng so sánh hiệu suất của các bộ từ điển khác nhau với từ điển SO-CAL

Từ điển	Hiệu suất trên bộ thử nghiệm				
	Epinion 1	Epinion 2	Movie	Camera	Tổng thể
Google-Full	62.00	58.50	66.31	61.25	62.98
Google-Basic	53.25	53.50	67.42	51.40	59.25
Maryland-Full-NoW	58.00	63.75	67.42	59.46	62.65
Maryland-Basic	56.50	56.00	62.26	53.79	58.16
General Inquirer-Full	68.00	70.50	64.21	72.33	68.02
General Inquirer-Basic	62.50	59.00	65.68	63.87	64.23
SentiWordNet-Full	66.50	66.50	61.89	67.00	65.02
SentiWordNet-Basic	59.25	62.50	62.89	59.92	61.47
Subjective-Full	72.75	71.75	65.42	77.21	72.04
Subjective-Basic	64.75	63.50	68.63	64.83	66.51
SO-CAL-Full	80.25	80.00	76.37	80.16	78.74
SO-CAL-Basic	65.50	65.25	68.05	64.70	66.04

2.2. Phương pháp phân loại chủ quan

Phân loại chủ quan là bước đầu tiên cần thiết để phân tích cảm xúc. Trong phần này, công việc cần thực hiện là đánh giá và phân lớp dữ liệu sau khi tiền xử lý thành 02 lớp: lớp chủ quan và lớp khách quan.

2.2.1. Câu có từ hàm chứa cảm xúc

Hiện nay trên thế giới cũng như trong nước, việc phân loại chủ quan chủ yếu dựa vào phương pháp so khớp với bộ từ điển cảm xúc. Do đó, chúng tôi lựa chọn phương pháp so khớp từ với bộ từ điển cảm xúc SO-CAL.

Một câu chủ quan (có cảm xúc) thường có từ hàm chứa cảm xúc.

Ví dụ:

- “*Ngôi nhà màu xanh*” là một câu khách quan vì nó không có từ hàm chứa cảm xúc trong đó.
- “*Ngôi nhà đẹp*” là một câu chủ quan vì nó có từ hàm chứa cảm xúc là từ “*đẹp*”.

Đây là phương pháp cơ bản và đơn giản nhất để phân loại một câu là chủ quan hay khách quan. Theo đó, việc lựa chọn những đặc trưng tốt nhất để đánh giá câu chủ quan là việc chúng tôi cần nghiên cứu để có được kết quả tối ưu.

2.2.2. Các trường hợp ngoại lệ

Phương pháp phân loại câu dựa vào từ hàm chứa cảm xúc là phương pháp chủ đạo để phân loại câu chủ quan. Tuy nhiên, mức độ chính xác chưa cao bởi vì có những trường hợp ngoại lệ là những trường hợp câu có từ hàm chứa cảm xúc nhưng không thể hiện cảm xúc. Cụ thể, đó là câu nghi vấn và câu điều kiện.

Câu nghi vấn:

- Đặc trưng cơ bản của câu nghi vấn là thường có những từ “*gì*”, “*như thế nào*”, “*thế nào*”, “*vì sao*”, “*tại sao*”, “*là sao*”. Những câu này dù có từ hàm chứa cảm xúc nhưng nó vẫn là câu không có cảm xúc.

Ví dụ:

- “*Tại sao bạn lại mặc bộ đồ thiếu tinh tế đến vậy?*” là một câu nghi vấn và không có xúc cảm. Mặc dù trong câu có từ hàm chứa cảm xúc “tinh tế” nhưng thực tế câu này không hề có cảm xúc. Đó chỉ là một nghi vấn mà người nói yêu cầu người nghe trả lời.

Câu điều kiện:

- Đặc trưng của câu điều kiện là thường có những từ: “nếu...thì...”, “giả như...thì ...”,... Ở cả hai trường hợp thì câu đều không chứa cảm xúc mặc dù chúng chứa từ cảm xúc.

Ví dụ:

- “*Nếu ngày mai trời mưa thì tôi sẽ rất buồn.*”. Trong câu có từ “rất buồn” có giá trị SO là $(-2) \times (1+0.2) = (-2.4)$ nhưng câu trên chưa chắc diễn ra trong thực tế mà chỉ là suy đoán của người nói. Có thể ngày mai trời mưa nhưng người nói chưa chắc chắn buồn. Nên câu sẽ không có cảm xúc.
- “*Giả như con học giỏi thì mẹ sẽ cho con đi chơi.*”. Trong câu có từ “giỏi” có giá trị SO là $(+3)$ nhưng sự việc trên đã không diễn ra. Vì vậy câu trên sẽ không có cảm xúc.

Ngoài những trường hợp bên trên, chúng tôi nhận thấy một câu có chứa cảm xúc nếu đó là một câu dài. Thông thường, những câu ngắn chỉ là những danh từ (người, vật, địa điểm,...), động từ hoặc trạng từ và các câu này thường không hàm chứa cảm xúc. Khi người nói đã có ý thể hiện một câu dài thì hầu hết sẽ đặt yếu tố cảm xúc trong đó. Tuy nhiên, việc đánh giá một câu như thế nào là đủ dài và mang yếu tố cảm xúc thì cần thời gian thực nghiệm hơn và có các nghiên cứu riêng về vấn đề này. Trong phạm vi khóa luận, chúng tôi lựa chọn giá trị 5 đơn vị từ để làm mốc cho một câu dài và ngắn.

2.3. Phương pháp phân loại cảm xúc

Sau khi xác định được câu có cảm xúc, chúng tôi tiếp tục dựa vào bộ từ điển cảm xúc SO-CAL tiếng Việt và các đặc trưng được rút trích dựa vào những đặc điểm

câu văn của tiếng Việt để tính toán giá trị cảm xúc của câu. Dựa vào giá trị này để phân loại câu có cảm xúc thành câu có cảm xúc tích cực và câu có cảm xúc tiêu cực.

2.3.1. Giá trị cảm xúc của câu phụ thuộc vào từ hàm chứa cảm xúc

Từ hàm chứa cảm xúc (sentiment word) là thành phần có ảnh hưởng lớn nhất đến giá trị cảm xúc của câu. Hiểu đơn giản, từ hàm chứa cảm xúc là từ chứa cảm xúc và thường được sử dụng để thể hiện cảm xúc tiêu cực hoặc tích cực. Chẳng hạn các từ “tốt”, “tuyệt vời”, “đẹp” là những từ chứa cảm xúc tích cực và “xấu xí”, “kinh khủng”, “tệ hại” là những từ chứa cảm xúc tiêu cực. Ngoài những từ riêng lẻ, còn có cụm từ chứa cảm xúc như “không thể tin được”, “như một giấc mơ”,... Một danh sách các từ và cụm từ như vậy được gọi là từ điển cảm xúc.

Cách đơn giản nhất để tính giá trị cảm xúc của một câu là tính tổng giá trị cảm xúc của các từ hàm chứa cảm xúc trong câu đó.

Ví dụ:

- “*Anh ấy thông minh và đẹp trai*”. Từ “thông minh” có giá trị SO là (+4) và “đẹp trai” có giá trị SO là (+4) nên tổng giá trị SO của câu là (+8).
- “*Chiếc áo này hợp thời trang*”. Câu trên chỉ có một cụm từ mang cảm xúc là “hợp thời trang” nên tổng giá trị SO của câu cũng bằng giá trị SO của từ này là (+2).

Mặc dù từ điển cảm xúc là thành phần quan trọng trong quá trình tính toán giá trị cảm xúc của câu nhưng chỉ sử dụng nó thôi là chưa đủ. Cảm xúc con người rất phức tạp. Có nhiều trường hợp mà chỉ sử dụng từ điển cảm xúc không thể đánh giá chính xác giá trị cảm xúc trong câu. Một số trường hợp cụ thể:

- Từ mang giá trị cảm xúc chịu ảnh hưởng của từ tăng cường. Ví dụ như, “đẹp”, “hơi đẹp”, “rất đẹp” và “đẹp nhất” nếu chỉ dựa vào từ điển cảm xúc thì những từ, cụm từ trên sẽ có giá trị SO như nhau. Nhưng trên thực tế lại không như vậy. Tất cả chúng đều mang cảm xúc tích cực nhưng được xếp theo giá trị cảm xúc tăng dần lần lượt là “hơi đẹp”, “đẹp”, “rất đẹp”, “đẹp nhất”.

- Dễ nhầm lẫn giữa tích cực và tiêu cực. Một số từ có khả năng làm đổi cực của từ hay cụm từ cảm xúc như “không”, “không được”, “không phải”, “không bao giờ”,... Ví dụ: từ “tốt” mang cảm xúc tích cực thì “không tốt” mang cảm xúc tiêu cực.

Để giải quyết những vấn đề nêu trên cần đi sâu phân tích tiếp tục các đặc điểm khác của câu. Mỗi đặc điểm sẽ dần dần giải quyết từng vấn đề cụ thể.

2.3.2. Giá trị cảm xúc của câu phụ thuộc vào từ tăng cường

Quirk et al. (1985) đã chia từ tăng cường (intensifier) thành hai loại là làm tăng mức độ ngữ nghĩa (amplifiers) và làm giảm mức độ ngữ nghĩa (downtoners). Năm 2006, một số nhà nghiên cứu xử lý ngôn ngữ tự nhiên (Kennedy và Inkpen; Polanyi và Zaenen) đã sử dụng từ tăng cường để đơn giản sự tăng và giảm giá trị cảm xúc. Trong SO-CAL cũng bổ sung từ điển từ tăng cường. Những từ chịu ảnh hưởng bởi các từ tăng cường sẽ có giá trị cảm xúc thay đổi tùy thuộc vào giá trị tăng hay giảm mức độ ngữ nghĩa của từ tăng cường đó.

Ví dụ:

- Từ “*mệt mỏi*” mang giá trị SO (-3). Nhưng nếu phía trước nó có từ tăng cường “*hơi*” (-0.5) thì giá trị SO của “*hơi mệt mỏi*” là: $(-3) * (1 - 0.5) = (-1.5)$.
- Từ “*đẹp*” mang giá trị SO là (+4) thì “*rất đẹp*” có giá trị SO là:

$$(+4) * (1 + 0.2) = (+4.8)$$

- Từ “*giỏi*” mang giá trị SO là (+3) thì “*giỏi nhất*” có giá trị SO là:

$$(+3) * (1 + 1) = (+6)$$

2.3.3. Giá trị cảm xúc của câu phụ thuộc vào từ phủ định

Tương tự như việc tăng cường giá trị cảm xúc khi từ hàm chứa cảm xúc chịu ảnh hưởng của từ nằm trong từ điển từ tăng cường thì việc từ cảm xúc chịu ảnh hưởng của những từ phủ định cũng làm thay đổi giá trị cảm xúc của từ hàm chứa cảm xúc đó. Lúc nói hoặc viết, chúng ta thường dùng các từ phủ định bao gồm: “không”, “không được”, “không phải”,... để thể hiện một mức độ cảm xúc đối nghịch so với từ hàm chứa cảm xúc theo sau từ phủ định đó.

Do đó, đối với các từ cảm xúc mà đằng trước có từ phủ định thì giá trị cảm xúc từ đó sẽ được đảo ngược cực hay dễ hiểu hơn là đổi dấu giá trị cảm xúc của từ.

Ví dụ:

- Từ “*tốt*” có giá trị SO là (+3) thì “*không tốt*” có giá trị SO là (-3).
- Từ “*bịa đặt*” có giá trị SO là (-2) thì “*không bịa đặt*” có giá trị SO là (+2).

2.3.4. Giá trị cảm xúc của câu phụ thuộc vào từ khiếm khuyết

Những từ khiếm khuyết bao gồm: “nên”, “phải” và “có thể”. Những câu có chứa từ khiếm khuyết thường thể hiện mức độ cảm xúc giảm nhẹ hơn so với những câu tương tự nhưng không chứa từ khiếm khuyết.

Rõ ràng ta có thể dễ dàng nhận thấy câu: “*Bạn có thể làm tốt*” thì đối tượng được nói đến ở đây thực sự chưa làm tốt nhất khả năng của mình, và ý nghĩa cảm xúc sẽ giảm hơn so với câu: “*Bạn làm tốt*”. Do đó, việc lựa chọn một mức độ giảm nhẹ cảm xúc trong câu có từ khiếm khuyết là thực tế cần quan tâm, tuy nhiên giá trị giảm nhẹ đó là bao nhiêu là thích hợp thì cần thời gian để khảo sát và nghiên cứu thêm. Trong đề tài này, giá trị giảm nhẹ mà chúng tôi lựa chọn là 50%. Theo đó, những câu có chứa từ khiếm khuyết thì giá trị cảm xúc của câu giảm 50% so với giá trị cảm xúc của tất cả các từ mang ý nghĩa cảm xúc trong câu.

Dưới đây là một số ví dụ cụ thể về việc tính toán giá trị cảm xúc trong câu có từ khiếm khuyết:

- Câu “*Bạn có thể làm tốt hơn.*”. Cụm từ “tốt hơn” có giá trị SO là (+2) nhưng trong câu có từ khiếm khuyết “có thể” nên giá trị SO của “tốt hơn” giảm xuống còn (+1).
- Câu “*Chúng ta phải thật mạnh mẽ.*”. Cụm từ “thật mạnh mẽ” có giá trị SO là $(+2) * (1 + 0.3) = (+2.6)$ nhưng trong câu có từ khiếm khuyết “phải” nên giá trị SO của “thật mạnh mẽ” sẽ còn (+1.3).

2.3.5. Giá trị cảm xúc của câu có xu hướng tích cực

Phân loại cảm xúc dựa vào từ điển cảm xúc thường cho thấy một xu hướng tích cực (Kennedy and Inkpen [14], 2006). Trên thực tế thì con người có xu hướng

sử dụng từ ngữ tích cực nhiều hơn. Để cân bằng giữa tích cực và tiêu cực có rất nhiều cách. Trong đó, việc tăng giá trị cảm xúc của từ mang hàm ý tiêu cực được cho là có hiệu quả hơn cả. Chúng tôi đã thử nghiệm nhiều mức độ gia tăng giá trị cảm xúc của từ mang hàm ý tiêu cực và kết quả trả về khi tăng 50% giá trị cảm xúc của từ tiêu cực là tốt nhất.

Ví dụ: Câu “*Hôm nay giá vàng tăng và giá đô la giảm*”. Từ “giảm” có giá trị SO là (-2) sẽ được tăng 50% giá trị thành $(-2) \cdot (1+0.5) = (-3)$.

Vì vậy, trong phạm vi đề tài chúng tôi lựa chọn phương pháp tăng 50% giá trị cảm xúc của từ tiêu cực để xây dựng trong chương trình thử nghiệm.

2.4. Phương pháp phân lớp Support Vector Machine (SVM)

Support Vector Machines (SVM) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học máy có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. Nó là một công cụ mạnh mẽ cho các bài toán phân lớp phi tuyến tính được Cortes và Vapnik giới thiệu vào năm 1995 để giải quyết vấn đề nhận dạng mẫu hai lớp sử dụng nguyên lý cực tiểu hoá rủi ro cấu trúc (Structural Risk Minimization – SRM)

Các bước chính của phương pháp SVM:

- Tiền xử lý dữ liệu: thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán.
- Chọn hàm hạt nhân: lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể.
- Thực hiện kiểm tra để xác định các tham số cho ứng dụng. Điều này cũng quyết định đến tính chính xác của phương pháp.
- Sử dụng các tham số cho việc huấn luyện các tập mẫu: trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp bằng cách ánh xạ chúng vào không gian đặc trưng bằng các hàm hạt nhân.
- Kiểm thử dữ liệu test.

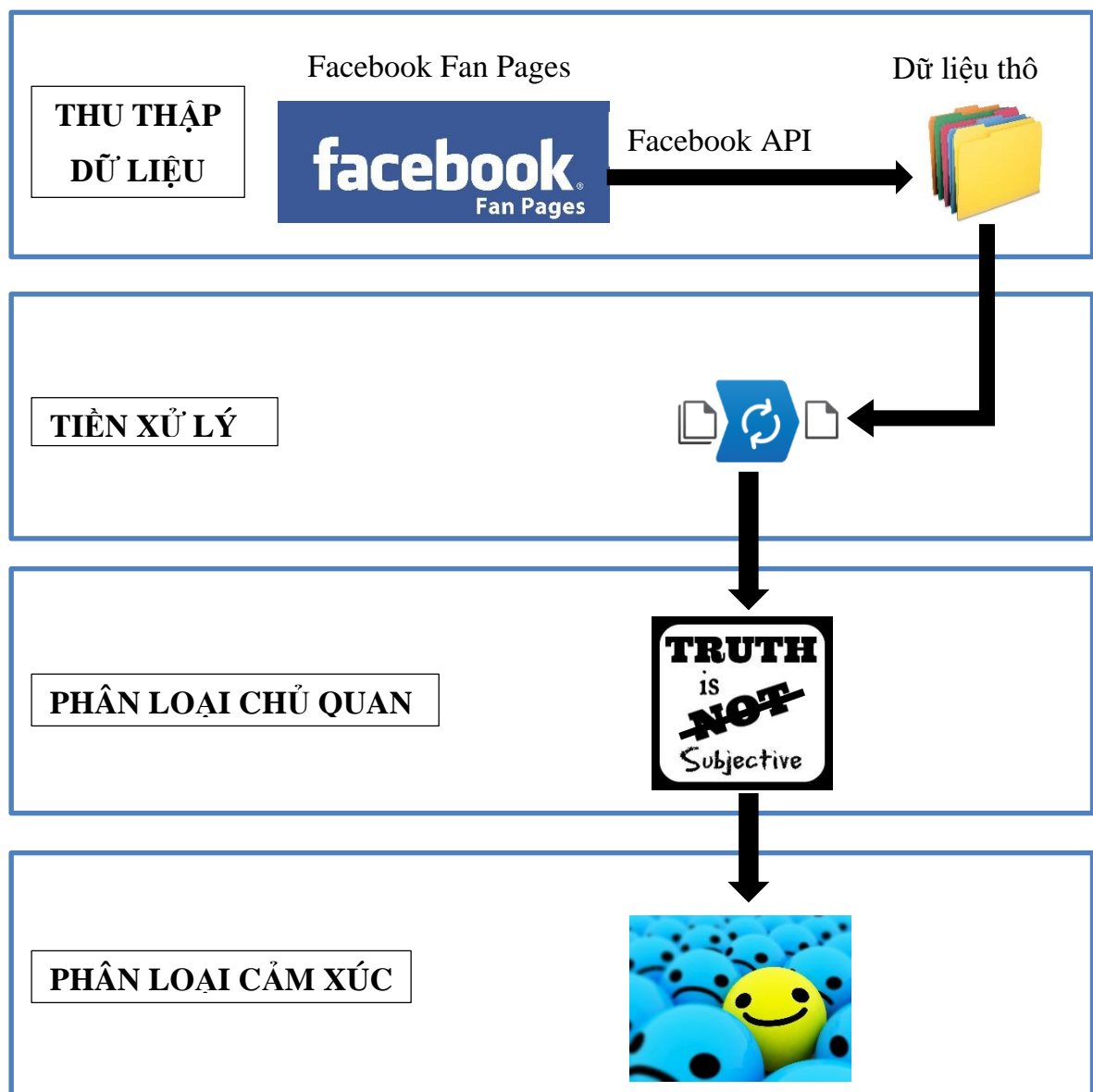
SVM có rất nhiều ứng dụng thiết thực giúp giải quyết các vấn đề trong thực tế như:

- Chuẩn đoán virus máy tính.
- Lọc thư hoặc tin rác.
- Nhận diện khuôn mặt, giọng nói, chữ viết tay, biển số xe.
- Phân loại gien.
- Phân loại văn bản.

Chương 3. XÂY DỰNG HỆ THỐNG THỬ NGHIỆM

3.1. Giới thiệu

Dựa vào những nghiên cứu lý thuyết và thực tế người dùng mạng xã hội thực tại ở Việt Nam. Chúng tôi xây dựng một hệ thống thực nghiệm phân tích, đánh giá cảm xúc dựa vào bình luận trên mạng xã hội Facebook. Cụ thể, dưới đây là mô hình hệ thống thực nghiệm mà chúng tôi đã xây dựng:



Hình 3-1 Mô hình hệ thống thực nghiệm

Theo mô hình trên, công việc cụ thể sẽ được thực hiện trong hệ thống thực nghiệm như sau:

- Thu thập dữ liệu: thu thập dữ liệu là những câu bình luận tiếng Việt trên mạng xã hội Facebook.
- Tiền xử lý và rút trích đặc trưng:
 - Từ dữ liệu tải về, tiền xử lý: loại stop word, emotion icon, v.v...
 - Gán nhãn từ loại.
 - Rút trích đặc trưng cho câu theo danh sách đặc trưng đã chọn.
- Phân loại chủ quan (subjectivity classification): dựa vào đặc trưng đã được phân tích, đánh giá bình luận là chủ quan (có cảm xúc) hay khách quan (không có cảm xúc).
- Phân loại cảm xúc (sentiment classification): sau khi đánh giá bình luận là có cảm xúc, dựa vào đặc trưng được phân tích bên trên, đánh giá bình luận đó là hàm chứa cảm xúc tích cực hay hàm chứa cảm xúc tiêu cực.

3.2. Bộ từ điển cảm xúc SO-CAL tiếng Việt

Để dịch bộ từ điển SO-CAL tiếng Anh chúng tôi đã sử dụng kết hợp hai bộ từ điển Viettien Dictionary [24] và Google Translate.

- Viettien Dictionary được Nguyễn Việt Khoa – Viện ngoại ngữ, Đại học Bách khoa Hà Nội công bố bản đầu tiên v1.0 vào tháng 8/2010 và bản cập nhật mới nhất tính đến tháng 3/2015 là phiên bản v4.0b công bố vào tháng 7/2014 trên nền tảng Mac OS cũng chính là phiên bản mà chúng tôi sử dụng. Tính đến thời điểm 7/2014, bộ từ điển Anh-Việt của Viettien đã có hơn 390,000 từ. Cơ sở dữ liệu của Viettien được bổ sung, biên tập và chỉnh sửa từ nhiều nguồn như:
 - Dự án Từ điển tiếng Việt miễn phí của Hồ Ngọc Đức.
 - Dự án Từ điển tiếng Việt mở của Trần Bình An.
 - Phần mềm từ điển Super Power Dict của Bùi Đức Tiến.
 - Dữ liệu từ điển dành cho phần mềm Babylon của Đào Công Tiến.
 - Một số dữ liệu từ điển do các thành viên Diễn đàn Tinh tế cung cấp.
- Google Translate: Dịch vụ này tính đến thời điểm tháng 2 năm 2010 đã hỗ trợ 52 ngôn ngữ trong đó có tiếng Việt. Chất lượng dịch ban đầu của Google Translate không được tốt. Nhưng do được trang bị tính năng tương tác giúp

mọi người có thể thay đổi nghĩa của từ cho phù hợp nhất nên chất lượng đã ngày càng cải thiện. Tốc độ dịch của Google là rất tốt so với các dịch vụ trực tuyến tương tự khác dành cho người Việt và nhất là ở khả năng dịch văn bản dài.

Chúng tôi dựa vào hai từ điển Anh-Việt trên để dịch bộ từ điển SO-CAL tiếng Anh sang tiếng Việt. Quá trình dịch từ điển được thực hiện tuần tự từ đầu đến cuối mỗi từ điển trong bộ từ điển SO-CAL. Những trường hợp xảy ra trong quá trình dịch:

- Một từ tiếng Anh chỉ có một nghĩa tiếng Việt: Chúng tôi sẽ thêm nghĩa tiếng Việt và giá trị cảm xúc của từ này vào từ điển SO-CAL tiếng Việt.

Ví dụ: Cụm từ “mega-star” (+3) được dịch sang tiếng Việt thành “siêu sao” (+3). Cụm từ “queen-sized” (+3) được dịch sang tiếng Việt thành “cỡ lớn” (+3).

- Một từ tiếng Anh có nhiều nghĩa tiếng Việt: Chúng tôi sẽ chọn nghĩa tiếng Việt thường được sử dụng và ngắn gọn dễ hiểu nhất để thêm vào từ điển SO-CAL tiếng Việt. Nếu trong các nghĩa còn lại có nghĩa ngắn gọn và đồng nghĩa với nghĩa được chọn trước đó thì chúng tôi cũng thêm nghĩa đó vào từ điển SO-CAL tiếng Việt. Các nghĩa được chọn thường có độ dài ngắn từ một đến ba từ và sau khi được thêm vào bộ từ điển cảm xúc SO-CAL tiếng Việt thì chúng sẽ giữ nguyên giá trị cảm xúc của từ tiếng Anh được dịch trong bộ từ điển SO-CAL tiếng Anh.

Ví dụ: Từ “exceptional” (+5) được dịch sang tiếng Việt thành “xuất chúng” (+5) và “vượt trội” (+5). Từ “glorious” được dịch sang tiếng Việt thành “vẻ vang” (+5) và “vinh quanh” (+5).

- Một từ hay cụm từ tiếng Anh không có trong từ điển Anh-Việt hoặc được ghép từ nhiều từ dẫn tới có nghĩa tiếng Việt quá dài: Khi có một từ tiếng Anh nào không có trong cả hai bộ từ điển Anh-Việt ở trên thì chúng tôi sẽ bỏ qua từ tiếng Anh đó. Còn khi gặp một từ hay cụm từ tiếng Anh được ghép bởi nhiều từ dẫn tới có nghĩa tiếng Việt quá dài, chúng tôi sẽ cố gắng rút ngắn nghĩa của chúng xuống ngắn nhất có thể. Nếu không được chúng tôi sẽ bỏ từ hay cụm từ đó đi.

Ví dụ: Từ “ritz-carlton” không có trong cả hai từ điển Anh-Việt ở trên nên chúng tôi bỏ qua cụm từ này. Từ “all-too-rare” được dịch thành “tất cả quá hiếm” quá dài nên chúng tôi bỏ qua cụm từ này. Từ “well-fitting” (+4) được dịch sang tiếng Việt thành “vừa vặn” (+4).

- Từ hay cụm từ tiếng Việt được thêm đã có trong bộ từ điển SO-CAL tiếng Việt: Khi có từ hay cụm từ tiếng Việt được thêm đã có trong bộ từ điển SO-CAL tiếng Việt thì chúng tôi sẽ bỏ từ đó.

Ví dụ: Từ “perfect” có nghĩa là “hoàn hảo” nhưng trước nó có từ “impeccable” được dịch là “hoàn hảo” trước rồi. Nên cụm từ “hoàn hảo” của từ “perfect” không được thêm vào từ điển SO-CAL tiếng Việt nữa.

- Ngoài ra để phù hợp với ngữ pháp tiếng Việt và cách viết ngắn gọn của các bình luận trên mạng xã hội, chúng tôi bổ sung thêm một số từ và cụm từ có số từ ít hơn nhưng vẫn đồng nghĩa với các từ hay cụm từ trong từ điển SO-CAL tiếng Việt.

Ví dụ: Chúng tôi thấy từ “may” có số từ ít hơn nhưng vẫn đồng nghĩa với từ “may mắn” (+2) nên chúng tôi thêm từ “may” (+2) vào từ điển SO-CAL tiếng Việt.

Sau khi dịch xong bộ từ điển SO-CAL tiếng Anh sang tiếng Việt, chúng tôi đã thu được bộ từ điển SO-CAL tiếng Việt bao gồm 5 bộ từ điển nhỏ: Từ điển danh từ (1546 từ), từ điển động từ (1108 từ), từ điển tính từ (2357 từ), từ điển trạng từ (749 từ) và từ điển từ tăng cường (intensifier) (185 từ).

Dưới đây là một số từ trong các từ điển trong bộ từ điển SO-CAL tiếng Việt mà chúng tôi xây dựng.

- Một số từ trong bộ từ điển danh từ.

Bảng 3-1 Một số từ trong bộ từ điển danh từ

Danh từ	Giá trị cảm xúc
hoàn hảo	5
lộng lẫy	4

chiến thắng	3
phước lành	2
độc lập	1
tội phạm	-1
điểm yếu	-2
tai ương	-3
thảm họa	-4
kỳ quái	-5

- Một số từ trong bộ từ điển động từ.

Bảng 3-2 Một số từ trong bộ từ điển động từ

Động từ	Giá trị cảm xúc
tôn kính	4
hoan hỉ	4
thành công	3
sáng tạo	2
tăng	1
vùi dập	-1
xấu hổ	-2
nguyên rủa	-3
ghét	-4
ghê tởm	-5

- Một số từ trong bộ từ điển tính từ.

Bảng 3-3 Một số từ trong bộ từ điển tính từ

Tính từ	Giá trị cảm xúc
tuyệt vời	5
cao cấp	4
bổ ích	3
chặt chẽ	2
hợp lý	1
cũ	-1
đần độn	-2
bẩn	-3
tai hại	-4
thảm khốc	-5

- Một số từ trong bộ từ điển trạng từ.

Bảng 3-4 Một số từ trong bộ từ điển trạng từ

Trạng từ	Giá trị cảm xúc
thú vị	5
huy hoàng	4
giỏi	3
tươi	2
sạch	1
kỳ quặc	-1

thô	-2
kém cỏi	-3
tàn bạo	-4
khiếp	-5

- Một số từ trong bộ từ điển từ tăng cường.

Bảng 3-5 Một số từ trong bộ từ điển từ tăng cường

Từ tăng cường	Giá trị cảm xúc
ít	-1.5
chút ít	-0.9
hơi	-0.5
khá	-0.2
chắc	0.2
siêu	0.4
hoàn toàn	0.5
nhất	1

3.3. Thu thập dữ liệu

Mạng xã hội mà chúng tôi tiếp cận để thu thập dữ liệu là Facebook. Hiện nay, Facebook là mạng xã hội phổ biến nhất ở Việt Nam. Do đó nguồn dữ liệu ở đây vô cùng phong phú. Để giúp các nhà phát triển dễ dàng khai thác và xây dựng các chương trình liên quan đến nguồn dữ liệu này, Facebook cung cấp một công cụ hỗ trợ có tên là Graph API.



Hình 3-2 Mô hình Graph API

Graph API là cách cơ bản để lấy ra và đưa dữ liệu vào social graph của Facebook. Đó là một API HTTP-based cấp thấp sử dụng để truy vấn dữ liệu, cập nhật trạng thái, tải lên các bức ảnh và nhiều hành động liên quan khác. Graph API có nhiều phiên bản. Chúng tôi sử dụng phiên bản Graph API v2.2 – là phiên bản mới nhất tính từ tháng 3/2015 trở về trước. Graph API được đặt tên dựa theo ý tưởng “social graph”. Thông tin trong Facebook gồm 3 phần:

- Các node – đầu mút (những thứ cơ bản như là người dùng, hình ảnh, trang, bình luận).
- Các edge – cạnh (những mối liên kết giữa các thứ cơ bản ở trên ví dụ như các hình ảnh của trang, hoặc một hình ảnh của các bình luận)
- Các field (thông tin về các node như ngày sinh của người dùng, tên của một trang).

Graph API dựa trên HTTP do đó nó hoạt động với các ngôn ngữ có thư viện HTTP như là cURL, url lib. Để truy cập vào API này, chúng ta sẽ tạo ra các HTTP GET request để truy cập tới các đầu nút hoặc các cạnh của đầu nút. Ngoài trừ việc tải lên video sử dụng `graph.video.facebook.com` còn lại các request đều sử dụng `graph.facebook.com`

Trước khi sử dụng Graph API, ta cần lấy quyền truy cập (get access token). Đơn giản nhất là truy cập vào Graph API Explorer trên trang <https://developers.facebook.com> để lấy quyền truy cập. Cách này giúp nhanh chóng lấy được quyền truy cập nhưng có một nhược điểm là chỉ tồn tại trong thời gian nhất định nên chúng tôi đã chọn cách phức tạp hơn là tạo một ứng dụng trên trang <https://developers.facebook.com>. Sau đó xin quyền truy cập của ứng dụng này. Tuy cách này tốn thời gian và công sức hơn nhưng đổi lại sẽ lấy được một quyền truy cập (access token) tĩnh.

Sau khi lấy được quyền truy cập, ta có thể đọc được tất cả các node và edge trong Graph API bằng các câu lệnh HTTP request với endpoint thích hợp.

Ví dụ dưới đây là quá trình chúng tôi lấy dữ liệu từ trang `VnExpress.net`. Chúng tôi muốn lấy tất cả tin tức kèm ngày tháng, người like, các bình luận,... của trang `VnExpress.net`

- Đầu tiên chúng tôi sử dụng graph API:
https://graph.facebook.com/congdongvnexpress/posts?access_token=792661064152453|sO-FOihVUkXtL9eUVPlom3XbMQ
- Chương trình tải dữ liệu sẽ sử dụng đường dẫn này để truy vấn dữ liệu và ghi vào tệp.

- Cấu trúc dữ liệu khi truy vấn sẽ trả về là định dạng json lưu tất cả thông tin.

Hình 3-3 Ví dụ về dữ liệu thô chưa xử lý

Công việc kế tiếp là sử dụng một ngôn ngữ lập trình để lấy những dữ liệu cần thiết từ file json này.

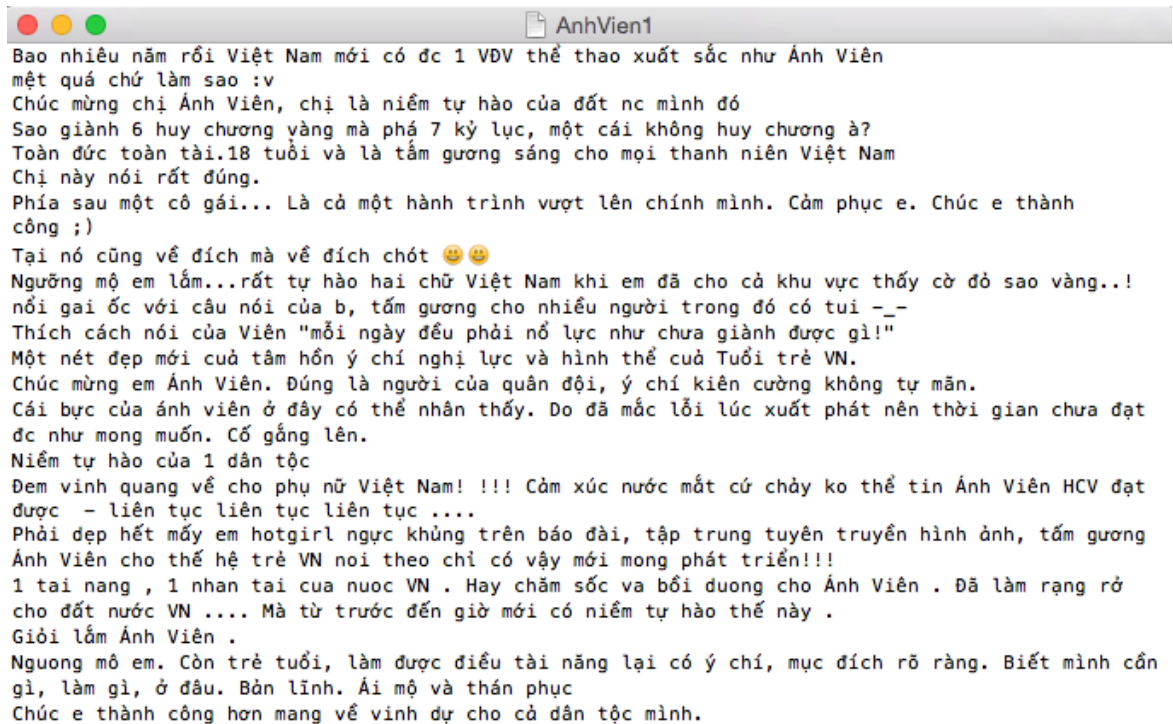
Dữ liệu thử nghiệm chúng tôi lựa chọn là những bình luận tiếng Việt từ các trang (fanpage) trên mạng xã hội Facebook như VnExpress.net, CGV Cinemas Vietnam, Mann up,...

Các bài viết sẽ chứa rất nhiều thông tin: nội dung bài viết, số lượt thích, tác giả, thời gian, các bình luận, tên người viết bình luận,... Nhưng hệ thống chỉ cần lựa chọn mã bài viết (ID post) để từ đó lấy được nội dung các bình luận của những bài viết đó.



Hình 3-4 Những bình luận của trang VnExpress.net trên mạng xã hội Facebook

Từ những bài viết như thế này, chúng tôi xử lý dữ liệu lấy về để trích xuất những bình luận và lấy đó là dữ liệu cơ bản nhất cho hệ thống thực nghiệm.



Hình 3-5 Nội dung các bình luận được lấy về thông qua thư viện Facebook Graph API

3.4. Đặc trưng dữ liệu từ mạng xã hội

Mạng xã hội là nguồn dữ liệu vô cùng lớn và chứa hầu hết thông tin trong thế giới hiện đại. Ngày nay chúng ta có thể dễ dàng nắm bắt được mọi thông tin mới nhất mà không nhất thiết phải sử dụng đến truyền hình hoặc báo vì mọi sự kiện đều dễ dàng được chia sẻ trên mạng xã hội. Mức độ tăng trưởng của mạng xã hội trên thế giới nói chung và Việt Nam nói riêng là hết sức nhanh chóng. Và việc bày tỏ cảm xúc, ý kiến cá nhân và sự đánh giá của người dùng mạng xã hội đối với các sự kiện, sản phẩm được chia trên mạng xã hội là nhu cầu thiết yếu và cơ bản của họ. Những lời bình luận đó thường có một số đặc trưng sau:

- Câu bình luận trên mạng xã hội thường ngắn, súc tích. Đây là đặc trưng cơ bản, dễ nhận thấy của những bình luận trên mạng xã hội. Người dùng mạng xã hội thường bày tỏ một cách thẳng thắn và trực tiếp vấn đề được bàn luận. Do đó, câu chữ thường ngắn gọn, dễ hiểu và đơn nghĩa.
- Mặc dù đa số những lời bình luận là những câu văn ngắn gọn nhưng người dùng mạng xã hội vẫn mang nét văn hóa ngôn ngữ Việt Nam từ trước đến nay vào những lời bình luận này. Đó là lối nói chuyện tinh tế, lòng vòng và đa

nghĩa khi bày tỏ ý kiến của mình. Do đó, bài toán cũng đặt ra thêm vấn đề giải quyết những câu bình luận mang nhiều nghĩa khác nhau.

- Với việc lượng người dùng chủ yếu là người trẻ và việc ngôn ngữ nước ngoài ngày càng được sử dụng rộng rãi thì một đặc trưng rất dễ nhận thấy ở người dùng mạng xã hội Việt Nam là việc lồng ghép ngôn ngữ nước ngoài vào những bình luận. Ngoài ra, một đặc thù nữa đối với người dùng ở tuổi vị thành niên đó là việc kết hợp những câu chữ viết tắt, và biểu tượng cảm xúc được sử dụng hết sức phổ biến. Từ khi hệ thống điện thoại di động phát triển chúng tôi người dùng này đã hình thành nên một hệ thống ký tự viết tắt đặc biệt để tiết kiệm thời gian viết và thể hiện cảm xúc cũng như cá tính của bản thân. Bộ ký tự này thường được gọi là “teen code”.
- Tính vùng miền, địa phương của bình luận trên mạng xã hội. Người dùng mạng xã hội dễ dàng nhận biết chủ nhân của một bình luận trên mạng xã hội đến từ vùng miền, địa phương nào vì tính vùng miền, địa phương được thể hiện trong bình luận đó. Đặc biệt là đối với người dùng mạng xã hội ở khu vực miền Trung là hết sức rõ nét vì từ ngữ địa phương ở khu vực này là hết sức đa dạng và được sử dụng rất rộng rãi, phổ biến.

Từ những đặc trưng trên và tham khảo những nghiên cứu có liên quan đến đề tài trong những năm gần đây ở trong và ngoài nước. Chúng tôi đã quyết định lựa chọn phương pháp phân tích cảm xúc dựa vào bộ từ điển SO-CAL, đồng thời xây dựng dựa trên những đặc trưng của ngôn ngữ tiếng Việt và văn hóa ngôn ngữ ở Việt Nam cũng như đặc trưng cơ bản của người dùng mạng xã hội ở Việt Nam.

3.5. Tiền xử lý dữ liệu

Sau khi lấy được dữ liệu là những bình luận trên mạng xã hội, chúng tôi tiến hành tiền xử lý dữ liệu bởi vì dữ liệu này chưa phải là dữ liệu chuẩn tiếng Việt.

Trong quá trình thực hiện đề tài, chúng tôi nhận thấy có 02 vấn đề dẫn tới lỗi trong chương trình, đó là việc gán nhãn trong câu văn có biểu tượng cảm xúc và một lỗi khác liên quan đến lỗi tiếng Việt có dấu.

"yêu Tô' quô'c, yêu đồ'ng ba'o.
Khiêm tô'n, thâ.t tha', du'ng ca'm"! Nhe'!

Hình 3-6 Ví dụ dữ liệu dạng mã UNICODETH

Đây là dữ liệu dạng mã UNICODETH. Do đó, trước tiên chúng tôi tiến hành chuyển mã của tất cả dữ liệu về dạng mã UNICODE.

"yêu Tổ quốc, yêu đồng bào
Khiêm tốn, thật thà, dũng cảm"

Hình 3-7 Ví dụ dữ liệu sau khi chuyển mã

Đối với câu văn có biểu tượng cảm xúc, chúng tôi tiến hành loại bỏ các biểu tượng cảm xúc này ra khỏi dữ liệu. Theo đó, danh sách biểu tượng cảm xúc chúng tôi xây dựng bao gồm bộ biểu tượng cảm xúc chuẩn của mạng xã hội Facebook và một số biểu tượng cảm xúc mà người dùng mạng xã hội ở Việt Nam thường được sử dụng.

Sau khi loại bỏ những vấn đề dẫn tới lỗi, hệ thống sẽ tự động thực hiện các chức năng sau:

- Cắt dữ liệu thành câu.
- Gán nhãn từ loại.

Chúng tôi sử dụng thư viện mã nguồn mở vnTagger [19] để gán nhãn từ loại tiếng Việt cho dữ liệu.

Ví dụ: Đối với một bình luận như sau: *"Bạn thật tuyệt. Tôi thích bạn. Một con người tốt bụng :D."*.

- Tiền xử lý, cắt dữ liệu thành câu:

Bạn thật tuyệt

Tôi thích bạn

Một con người tốt bụng

- Gán nhãn từ loại:

```
<doc>
  <s>
    <w pos="N">Bạn</w>
    <w pos="A">thật</w>
    <w pos="A">tuyệt</w>
  </s>
  <s>
    <w pos="P">Tôi</w>
    <w pos="V">thích</w>
    <w pos="N">bạn</w>
  </s>
  <s>
    <w pos="M">Một</w>
    <w pos="N">con người</w>
    <w pos="A">tốt bụng</w>
  </s>
</doc>
```

Kết quả của quá trình tiền xử lý là một tập tin có cấu trúc html mô tả sự gán nhãn từ loại. Chúng tôi gọi tập tin này là tập tin tagger, tức là tập tin đã được gán nhãn. Từ tập tin này, chương trình tiếp tục quá trình rút trích các đặc trưng của từng câu văn để phân tích trong các bước tiếp theo.

3.6. Bộ dữ liệu huấn luyện

3.6.1. Gán nhãn câu bằng tay

Để xây dựng bộ dữ liệu huấn luyện. Chúng tôi tiến hành gán nhãn cho câu. Đây là quá trình mang tính chủ quan của từng cá nhân. Cảm xúc của con người rất phức tạp. Phân tích cảm xúc mức văn bản chứa nhiều quan điểm về nhiều đối tượng

thì quá thô đối với hầu hết các ứng dụng. Phân tích cảm xúc mức câu đơn giản hơn, thường chỉ chứa một quan điểm duy nhất. Ở mức này, có hai vấn đề cần giải quyết:

- ***Vấn đề đầu tiên là phân loại xem câu có chứa cảm xúc hay không chứa cảm xúc (thường được gọi là phân loại chủ quan).***

Khó khăn ở vấn đề này: Một câu khi được viết hay nói thường có một mục đích nói nhất định: trần thuật (dùng để miêu tả, kể hay giới thiệu về một sự vật, sự việc), nghi vấn (dùng để hỏi), cầu khiến (dùng để đề nghị, yêu cầu), cảm thán (dùng để bộc lộ cảm xúc),... Hoàn thành phân loại mục đích nói sẽ giúp phân loại chủ quan dễ dàng và chính xác hơn.

Quá trình gán nhãn câu chủ quan và câu khách quan: Dựa theo định nghĩa câu chủ quan, câu khách quan ở mục phân tích cảm xúc mức câu của chương tổng quan đề tài và phương pháp phân loại chủ quan của chương cơ sở lý thuyết, chúng tôi thực hiện gán nhãn câu chủ quan và câu khách quan bằng tay. Câu không chứa từ hàm chứa cảm xúc và chỉ miêu tả một số thông tin thực tế là câu khách quan. Câu hỏi hay câu điều kiện dù chứa từ hàm chứa cảm xúc vẫn là câu khách quan. Câu chứa từ hàm chứa cảm xúc và không phải câu hỏi hay câu cầu khiến là câu chủ quan. Câu chủ quan thường mang tới góc nhìn hay ý kiến cá nhân của người viết.

Ví dụ:

(1) *Iphone có thiết kế tốt.*

(2) *Thiết kế của Iphone có tốt không?*

(3) *Nếu Iphone có thiết kế xấu thì mọi người sẽ không mua nó.*

Ở ba ví dụ trên, có thể dễ dàng nhận thấy câu (1) mang cảm xúc (tích cực) về thiết kế của Iphone. Câu (2) và (3) không mang cảm xúc. Câu (2) đặt ra câu hỏi nghi ngờ về chất lượng của Iphone còn câu (3) đặt ra một giả định chưa chắc có thực. Nên câu (1) là câu chủ quan và câu (2), (3) là câu khách quan.

- ***Vấn đề thứ hai là phân loại những câu chứa cảm xúc là tích cực hay tiêu cực (thường được gọi là phân loại cảm xúc).***

Nếu câu chỉ có những từ thông thường không mang cảm xúc và những từ mang cảm xúc tích cực hay tiêu cực thôi thì vấn đề này sẽ được giải quyết dễ dàng. Nhưng trong thực tế, câu còn có thêm những từ phủ định (negation), từ làm tăng mức độ ngữ nghĩa (amplifiers), từ làm giảm mức độ ngữ nghĩa (downtoners), động từ khiếm khuyết,... Làm sao đánh giá ảnh hưởng của các từ này đến cảm xúc trong câu đồng thời kết hợp chúng với các từ cảm xúc để đưa ra kết luận chính xác nhất là câu mang cảm xúc tiêu cực hay tích cực là khó khăn gặp phải ở vấn đề này.

Quá trình gán nhãn câu chứa cảm xúc tích cực và câu chứa cảm xúc tiêu cực: Các câu trong những bình luận trên mạng xã hội thường ngắn gọn và chứa một quan điểm duy nhất. Nếu câu chỉ chứa một quan điểm, ý kiến duy nhất thì chúng tôi xác định xem quan điểm hay ý kiến đó là tích cực hay tiêu cực. Sau đó, gán nhãn câu chứa quan điểm tích cực là câu tích cực và câu chứa quan điểm tiêu cực là câu tiêu cực. Ngoài ra còn có một số câu chứa nhiều hơn một quan điểm hay ý kiến. Chúng tôi sẽ dựa vào phương pháp phân loại cảm xúc ở chương cơ sở lý thuyết sau đó xét đến tất cả các yếu tố ảnh hưởng đến mức độ cảm xúc tích cực và tiêu cực. Cuối cùng, tổng hợp lại để đánh giá trong câu quan điểm tích cực hay quan điểm tiêu cực có mức độ cảm xúc lớn hơn. Từ đó gán nhãn câu dựa vào quan điểm có mức độ cảm xúc lớn hơn.

Ví dụ:

(4) Đây là một bộ phim hay.

(5) Đây là một bộ phim không hay

(6) Trong hoàn cảnh khó khăn, anh ấy vẫn cố gắng vượt qua và gặt hái nhiều thành công.

Câu (4) mang cảm xúc tích cực. Chỉ cần thêm một từ phủ định “không” vào trước từ “hay” ở câu (4), câu (5) đã mang cảm xúc tiêu cực. Câu (6) quan điểm tích cực “cố gắng vượt qua” và “gặt hái nhiều thành công” có mức độ cảm xúc lớn hơn quan điểm tiêu cực “hoàn cảnh khó khăn” nên câu (6) là câu mang cảm xúc tích cực.

3.6.2. Mô tả bộ dữ liệu huấn luyện

Chúng tôi quyết định lựa chọn 3 bộ dữ liệu ở 3 chủ đề: giáo dục, phim ảnh và thể thao [Phụ lục III]. Mỗi bộ dữ liệu bao gồm từ hơn 250 cho 350 bình luận về các chủ đề trên. Sau đó, từ 03 bộ dữ liệu trên chúng tôi gộp tất cả các chủ đề lại để xây dựng một bộ dữ liệu huấn luyện lớn hơn gồm 885 câu. Đây là bộ dữ liệu tổng hợp.

Đầu tiên chúng tôi thực hiện phân loại chủ quan bằng tay với các bộ dữ liệu trên. Kết quả phân loại chủ quan bằng tay được thể hiện trong bảng sau:

Bảng 3-6 Kết quả phân loại chủ quan bằng tay

STT	Chủ đề	Dữ liệu huấn luyện	
		Câu chủ quan	Câu khách quan
1	Giáo dục	173	99
2	Phim ảnh	194	95
3	Thể thao	248	76
4	Tổng hợp	615	270

Sau khi phân loại chủ quan, chúng tôi lựa chọn những câu chủ quan (có cảm xúc) để tiếp tục phân loại cảm xúc. Kết quả phân loại được trình bày theo bảng sau:

Bảng 3-7 Kết quả phân loại cảm xúc bằng tay

STT	Chủ đề	Dữ liệu huấn luyện	
		Câu tích cực	Câu tiêu cực
1	Giáo dục	133	40
2	Phim ảnh	115	79
3	Thể thao	201	47
4	Tổng hợp	449	166

3.7. Phương pháp phân loại chủ quan

Từ tập tin tagger và từ điển SO-CAL tiếng Việt, chúng tôi tiến hành rút trích các đặc trưng dựa vào những cơ sở lý thuyết đã được trình bày ở mục 2.3. Theo đó, để đánh giá một câu có hay không có cảm xúc chúng tôi lựa chọn những đặc trưng sau:

- **Đặc trưng số 1:** số lượng từ trong câu. Số lượng từ trong câu cũng thể hiện cảm xúc mà người nói, người viết muốn biểu lộ với người nghe, người đọc. Nếu số lượng từ lớn thông thường đó sẽ là một câu có cảm xúc vì người nói, người viết đã đầu tư một công sức đáng kể và rõ ràng là họ quan tâm đến chủ đề đang được nhắc đến. Ngược lại, nếu số lượng từ quá ít thì có thể đó là một danh từ chỉ người, chỉ vật, v.v...
- **Đặc trưng số 2, 3, 4 và 5:** tổng giá trị cảm xúc của các từ loại: tính từ, trạng từ, danh từ và động từ trong câu. Giá trị cảm xúc trong câu phụ thuộc vào loại từ và giá trị cảm xúc của loại từ đó được so khớp với bộ từ điển SO-CAL tiếng Việt. Chúng tôi nhận thấy, giá trị cảm xúc trong câu chủ yếu phụ thuộc vào các loại từ sau: trạng từ, tính từ, danh từ và động từ. Theo đó, ứng với tổng giá trị cảm xúc của mỗi loại từ chúng tôi chọn thành một đặc trưng.
 - Tổng giá trị cảm xúc của trạng từ trong câu. Sau khi được gán nhãn, những thẻ trạng từ được duyệt và so khớp với từ điển trạng từ trong bộ từ điển SO-CAL tiếng Việt. Nếu giống nhau thì giá trị này được cộng dồn vào tổng giá trị cảm xúc trạng từ. Nếu trong câu không có trạng từ hoặc không khớp với từ điển, giá trị này mặc định bằng 0.
 - Hoàn toàn tương tự đối với tính từ, danh từ và động từ. Những thẻ loại từ này trùng khớp với từ điển tương ứng trong bộ từ điển SO-CAL tiếng Việt. Nếu không có giá trị nào trùng khớp hoặc câu không chứa những loại từ này, giá trị mặc định sẽ là 0.
- **Đặc trưng số 6:** tổng giá trị cảm xúc của câu. Đặc trưng này thể hiện tổng giá trị cảm xúc của câu. Giá trị của đặc trưng này về cơ bản là tổng của 04 đặc trưng phía trên mà chúng tôi xây dựng. Mặc dù chúng có liên quan với nhau và tưởng chừng giá trị này dư thừa, nhưng thực tế việc tính tổng này là hết sức

cần thiết vì nếu tổng những giá trị phía trên bằng 0 thì việc đánh giá chủ quan còn chưa chắc chắn là đúng đắn. Ngoài ra, giá trị cảm xúc trong một câu không chỉ phụ thuộc vào từ hàm chứa cảm xúc, một câu chủ quan còn phụ thuộc vào loại câu của nó nữa. Nếu là một câu nghi vấn hoặc một câu cầu khiến thì câu đó hoàn toàn không có giá trị cảm xúc. Do đó, tổng giá trị cảm xúc của câu còn có thể bằng 0 nếu như câu đó thuộc một trong hai loại câu bên trên.

Khái quát phương pháp phân loại chủ quan:

Input: tập tin tagger và bộ từ điển SO-CAL tiếng Việt.

Output: tập tin có cấu trúc vector, với mỗi dòng là 01 vector đặc trưng.

Các thao tác áp dụng:

Với mỗi câu trong bộ dữ liệu, rút trích các giá trị

- 1) *Tổng số từ.*
- 2) *Tổng giá trị cảm xúc của các tính từ.*
- 3) *Tổng giá trị cảm xúc của các trạng từ.*
- 4) *Tổng giá trị cảm xúc của các danh từ.*
- 5) *Tổng giá trị cảm xúc của các động từ.*
- 6) *Giá trị cảm xúc của cả câu:*

Nếu câu thuộc câu nghi vấn hoặc câu điều kiện thì trả về 0

Ngược lại, trả về tổng của các đặc trưng số 2, 3, 4 và 5.

Trả về vector đặc trưng

Từ tập tin kết quả của quá trình rút trích đặc trưng bên trên. Chúng tôi sử dụng phương pháp phân lớp SVM được trình bày ở mục 2.5 với bộ dữ liệu huấn luyện được trình bày ở mục 3.6 để tiến hành phân lớp. Chương trình sẽ tiến hành phân lớp cho từng vector bằng phương pháp học máy SVM. Kết quả trả về của quá trình này là kết quả phân lớp cho câu văn vào 02 lớp: chủ quan (subjectivity) và khách quan (objective).

Ở đây, chúng tôi chỉ sử dụng những thông số cơ bản của phương pháp SVM để phân lớp.

Những ví dụ dưới đây trình bày chi tiết về quá trình rút trích đặc trưng của chương trình đối với một câu văn cụ thể.

Ví dụ:

Câu “*Cô ấy vừa đẹp mà vừa học giỏi nữa.*” sẽ được rút trích đặc trưng và trả về các giá trị như sau: “1:9.0 2:7.0 3:0.0 4:0.0 5:0.0 6:7.0”. Các giá trị này có ý nghĩa như sau:

- Đặc trưng số 1 là số từ trong câu. Ở đây giá trị là 9.0.
- Đặc trưng số 2, 3, 4 và 5 lần lượt là tổng giá trị cảm xúc của các loại tính từ, trạng từ, danh từ và động từ trong câu. Tổng giá trị cảm xúc của các tính từ trong câu là 7.0, bao gồm: “đẹp” mang giá trị (+4) và “giỏi” (+3). Tổng giá trị cảm xúc các loại từ: trạng từ, danh từ và động từ trong câu này bằng 0 vì câu không có trạng từ, danh từ và động từ.
- Đặc trưng số 6 là tổng giá trị cảm xúc của tất cả các loại từ ở các đặc trưng 2, 3, 4 và 5. Giá trị này là 7.0 bao gồm: tính từ (+7.0), trạng từ (0), danh từ (0) và động từ (0).

Từ những đặc trưng trên, câu “*Cô ấy vừa đẹp mà vừa học giỏi nữa.*” là một câu chủ quan có hàm chứa cảm xúc.

Ví dụ:

Câu “*Nếu học tốt hơn thì tôi sẽ đăng ký kỳ thi tới.*” sau khi được rút trích đặc trưng sẽ có kết quả như sau: “1:10.0 2:3.0 3:0.0 4:0.0 5:0.0 6:0.0”. Mặc dù câu trên có giá trị cảm xúc của tính từ là (+3) nhưng tổng giá trị cảm xúc lại là (0) vì đây là một câu điều kiện. Do đó, đây là một câu khách quan không hàm chứa cảm xúc.

3.8. Phương pháp phân loại cảm xúc

Sau khi phân loại những câu chủ quan có hàm chứa cảm xúc. Chương trình sẽ tiếp tục việc phân loại cảm xúc cho những câu này.

Từ những câu chủ quan có hàm chứa cảm xúc, chúng tôi tiến hành gán nhãn từ loại một lần nữa cho những câu này để thành một tập tin tagger mới. Sau đó từ tập tin tagger mới bên trên và bộ từ điển SO-CAL tiếng Việt để phân loại cảm xúc. Việc phân loại cảm xúc của một câu thực tế là việc lựa chọn bộ đặc trưng tốt để đạt được độ chính xác cao. Bộ đặc trưng sau đây chúng tôi lựa chọn được kế thừa từ phương pháp phân tích cảm xúc ở tiếng Anh được trình bày ở mục 2.4 đồng thời có sự phát triển và chỉnh sửa cho phù hợp với đặc trưng ngôn ngữ tiếng Việt.

Giá trị cảm xúc của câu phụ thuộc vào từ hàm chứa cảm xúc:

- Đầu tiên, những đặc trưng cơ bản nhất là sự kế thừa từ phương pháp phân tích chủ quan. Bao gồm:
 - Giá trị cảm xúc của các loại từ trong câu: tính từ, trạng từ, danh từ và động từ.
 - Tổng giá trị cảm xúc của tất cả các loại từ bên trên.

Giá trị cảm xúc của câu phụ thuộc vào từ tăng cường:

- Đặc trưng tiếp theo là giá trị cảm xúc trong câu chịu ảnh hưởng của từ tăng cường. Hệ thống sẽ duyệt tìm những từ trong câu trùng khớp với từ điển từ tăng cường. Sau đó, những từ liền kề trước và liền kề sau của từ tăng cường đó được duyệt theo những bộ từ điển: từ điển tính từ, từ điển trạng từ, từ điển danh từ và từ điển động từ. Nếu những từ này trùng khớp với từ thuộc bộ từ điển bên trên thì giá trị cảm xúc của nó được tính theo công thức:

$$\text{Giá trị cảm xúc} = \text{giá trị từ tăng cường} * \text{giá trị cảm xúc của từ}$$

- Tổng những giá trị này sẽ là giá trị cảm xúc mới của câu sau khi xét từ tăng cường. Trong trường hợp không có từ tăng cường trong câu, giá trị này chính là giá trị của tổng giá trị cảm xúc của tất cả các loại từ trong câu.

Giá trị cảm xúc của câu phụ thuộc vào từ phủ định:

- Tương tự như đặc trưng về từ tăng cường trong câu. Hệ thống cũng sẽ duyệt tìm những từ nằm trong danh sách từ phủ định (bao gồm: “không”, “không có”, “không phải”, “không được”, “chẳng”, “chẳng có” và “chẳng phải”) sau

đó xét các từ liền kề sau của những từ phủ định này xem chúng có xuất hiện trong từ điển cảm xúc không? Nếu có, giá trị cảm xúc của nó được thay đổi như sau:

$$\text{Giá trị cảm xúc} = (-1) * \text{giá trị cảm xúc của từ}$$

- Trong trường hợp câu không có từ phủ định thì giá trị này chính là giá trị của tổng giá trị cảm xúc của các loại từ trong câu.

Giá trị cảm xúc của câu phụ thuộc vào từ khiếm khuyết:

- Trong trường hợp này, hệ thống chỉ duyệt xem trong câu có chứa từ khiếm khuyết hay không. Nếu có thì giá trị cảm xúc trong câu được tính theo công thức:

$$\text{Giá trị cảm xúc của câu} = (0.5) * \text{tổng giá trị cảm xúc các loại từ trong câu}$$

- Đây là đặc trưng được chúng tôi tự tìm hiểu và áp dụng cho đặc trưng ngôn ngữ tiếng Việt và người dùng mạng xã hội Facebook tại Việt Nam.

Giá trị cảm xúc của câu có xu hướng tích cực:

- Trong thực tế và văn hóa Việt Nam. Việc sử dụng từ ngữ nói giảm, nói tránh để thể hiện cảm xúc là hết sức phổ biến. Người dùng thường tránh nói ra những từ ngữ tiêu cực, do đó dẫn đến việc những từ ngữ tiêu cực thường ít gặp hơn so với từ tích cực. Đặc trưng này được xây dựng từ lý do trên. Theo đó, những từ ngữ hàm chứa cảm xúc tiêu cực (mang giá trị cảm xúc âm) sẽ được tính theo công thức:

$$\text{Giá trị cảm xúc} = (1 + 0.5) * \text{giá trị cảm xúc của từ}$$

- Trong trường hợp câu không có từ hàm chứa cảm xúc tiêu cực thì giá trị này chính là giá trị của tổng giá trị cảm xúc của các loại từ trong câu.

Ngoài những đặc trưng cơ bản bên trên, chúng tôi còn xây dựng thêm một đặc trưng khác dựa vào đặc điểm sử dụng ngôn ngữ của người dùng mạng xã hội tại Việt Nam. Đó là đặc trưng về câu có từ liên kết mang ý nghĩa trái ngược (bao gồm: “nhưng”, “nhưng mà”, “mà” và “cơ mà”).

Đối với những câu có chứa những từ liên kết mang ý nghĩa trái ngược được nêu bên trên thì giá trị cảm xúc của câu không phải là giá trị của tổng giá trị cảm xúc các loại từ trong câu mà chỉ là giá trị cảm xúc của về phía sau từ liên kết đó. Do đó, chúng tôi đánh giá giá trị cảm xúc của loại câu này bằng cách bỏ đi phần giá trị cảm xúc của về phía trước từ liên kết.

Khái quát phương pháp phân loại chủ quan:

Input: tập tin tagger và bộ từ điển SO-CAL tiếng Việt.

Output: tập tin có cấu trúc vector, với mỗi dòng là 01 vector đặc trưng.

Các thao tác áp dụng:

Với mỗi câu trong bộ dữ liệu, rút trích các giá trị

- 1) *Tổng giá trị cảm xúc của các tính từ.*
- 2) *Tổng giá trị cảm xúc của các trạng từ.*
- 3) *Tổng giá trị cảm xúc của các danh từ.*
- 4) *Tổng giá trị cảm xúc của các động từ.*
- 5) *Giá trị cảm xúc của cả câu: tổng của các đặc trưng số 2, 3, 4 và 5.*
- 6) *Giá trị cảm xúc phụ thuộc vào từ tăng cường.*
- 7) *Giá trị cảm xúc phụ thuộc vào từ liên kết mang nghĩa trái ngược.*
- 8) *Giá trị cảm xúc phụ thuộc vào từ khiếm khuyết.*
- 9) *Giá trị cảm xúc của câu có xu hướng tích cực.*
- 10) *Giá trị cảm xúc phụ thuộc vào từ phủ định thay đổi.*

Trả về vector đặc trưng

Sau đó, tương tự như phương pháp phân loại chủ quan. Hệ thống dựa vào đặc trưng được rút trích sẽ sử dụng phương pháp học máy với bộ dữ liệu huấn luyện bên trên để phân lớp cho từng câu: lớp tích cực (positive) và lớp tiêu cực (negative). Kết

quả cuối cùng nhận được đó là dữ liệu được phân loại thành 02 loại: tích cực và tiêu cực.

Dưới đây là ví dụ tổng quát về quá trình phân loại cảm xúc cho một câu bình luận. Để có thể phân loại cảm xúc, trước đó phải phân loại chủ quan xem câu văn có hàm chứa cảm xúc hay không. Do đó, ở ví dụ này chúng tôi trình bày cả 02 phần phân loại chủ quan và phân loại cảm xúc để có cái nhìn tổng quan nhất về toàn bộ quá trình thực thi của chương trình.

Ví dụ:

Phân tích cảm xúc đối với bình luận: “*Chúc mừng em một nhân tài trong tương lai. Hãy cố gắng học tốt nhất, để trở thành nhân tài cho đất nước Việt Nam nhé.*”. Sau khi tiền xử lý và gán nhãn dữ liệu trả về như sau:

```
<doc>
  <s>
    <w pos="V">Chúc mừng</w>
    <w pos="N">em</w>
    <w pos="M">một</w>
    <w pos="N">nhân tài</w>
    <w pos="E">trong</w>
    <w pos="N">tương lai</w>
  </s>
  <s>
    <w pos="R">Hãy</w>
    <w pos="V">cố gắng</w>
    <w pos="V">học</w>
    <w pos="A">tốt</w>
    <w pos="R">nhất</w>
    <w pos="," ">,</w>
    <w pos="E">để</w>
    <w pos="V">trở thành</w>
    <w pos="N">nhân tài</w>
  </s>
</doc>
```

```
<w pos="E">cho</w>
<w pos="N">đất nước</w>
<w pos="Np">Việt Nam</w>
<w pos="I">nhé</w>

</s>

</doc>
```

Sau khi tiền xử lý, hệ thống tiến hành rút trích đặc trưng. Dữ liệu trả về là đặc trưng của từng câu trong bình luận trên như sau.

Đối với câu: “*Chúc mừng em một nhân tài trong tương lai.*”.

- Kết quả rút đặc trưng đối với quá trình phân tích chủ quan là:

1:6.0 2:0.0 3:0.0 4:3.0 5:1.0 6:4.0

Trong đó:

- Đặc trưng số 1 (đặc trưng về số từ trong câu) có giá trị là 6.0 vì câu có 6 từ.
- Các đặc trưng số 2, 3, 4 và 5 lần lượt là tổng giá trị cảm xúc của các loại từ trong câu theo thứ tự sau:
 - Đặc trưng số 2 và 3 đều có giá trị là 0.0 vì trong câu không có tính từ (thẻ A) và trạng từ (thẻ R).
 - Đặc trưng số 4 có giá trị là 3.0. Danh từ (thẻ N) “nhân tài” trong câu có giá trị cảm xúc là 3.0.
 - Đặc trưng số 5 có giá trị là 1.0. Động từ (thẻ V) “chúc mừng” có giá trị cảm xúc là 1.0.
- Đặc trưng số 6 (đặc trưng về tổng giá trị cảm xúc của cả câu) có giá trị là 4.0 ($0.0 + 0.0 + 3.0 + 1.0$). Ta thấy, đây là một câu bình thường và không thuộc vào những trường hợp ngoại lệ. Do đó tổng giá trị cảm xúc trong câu bằng tổng giá trị cảm xúc của các loại từ trong câu. Tức là trong trường hợp này giá trị của đặc trưng số 6 bằng tổng giá trị của các đặc trưng số 2, 3, 4 và 5 cộng lại.

- Kết quả phân loại chủ quan trả về đây là một câu chủ quan có hàm chứa cảm xúc.
- Sau khi phân loại câu trên là một câu chủ quan có hàm chứa cảm xúc, chương trình tiếp tục rút trích đặc trưng đối với quá trình phân loại cảm xúc là:

1:0.0 2:0.0 3:3.0 4:1.0 5:4.0 6:4.0 7:4.0 8:4.0 9:4.0 10:4.0

Trong đó:

- Các đặc trưng số 1, 2, 3, 4 và 5 được kế thừa từ các đặc trưng số 2, 3, 4, 5 và 6 ở phần phân tích chủ quan.
- Sau khi phân tích, câu văn bên trên không có các yếu tố đặc biệt như: từ tăng cường, từ liên kết mang nghĩa trái ngược, từ khiếm khuyết, từ tiêu cực và từ phủ định thay đổi. Do đó các đặc trưng số 6, 7, 8, 9 và 10 đều có giá trị là 4.0 và bằng giá trị đặc trưng số 5. Tức là, giá trị cảm xúc của câu văn này chỉ phụ thuộc vào từ hàm chứa cảm xúc chứ không phụ thuộc vào các yếu tố khác.
- Kết quả phân loại cảm xúc trả về cho câu này là một câu tích cực vì các đặc trưng đều mang giá trị dương.

Tương tự, đối với câu “*Hãy cố gắng học tốt nhất, để trở thành nhân tài cho đất nước Việt Nam nhé.*”, kết quả lần lượt là:

Kết quả phân tích chủ quan:

1:12.0 2:3.0 3:0.0 4:3.0 5:2.0 6:8.0

Trong đó:

- Đặc trưng số 1 (đặc trưng về số từ trong câu) có giá trị là 12.0 vì câu có 12 từ.
- Các đặc trưng số 2, 3, 4 và 5 lần lượt là tổng giá trị cảm xúc của các loại từ trong câu theo thứ tự sau:
 - Đặc trưng số 2 có giá trị là 3.0. Tính từ (thẻ A) “tốt” có giá trị cảm xúc là 3.0.

- Đặc trưng số 3 có giá trị là 0.0 vì trong câu có trạng từ (thẻ R) “nhất” nhưng không khớp với từ điền trạng từ.
- Đặc trưng số 4 có giá trị là 3.0. Danh từ (thẻ N) “nhân tài” trong câu có giá trị cảm xúc là 3.0.
- Đặc trưng số 5 có giá trị là 2.0. Động từ (thẻ V) “cố gắng” có giá trị cảm xúc là 2.0.
- Đặc trưng số 6 (đặc trưng về tổng giá trị cảm xúc của cả câu) có giá trị là 8.0 ($3.0 + 0.0 + 3.0 + 2.0$). Ta thấy, đây là một câu bình thường và không thuộc vào những trường hợp ngoại lệ. Do đó tổng giá trị cảm xúc trong câu bằng tổng giá trị cảm xúc của các loại từ trong câu. Tức là trong trường hợp này giá trị của đặc trưng số 6 bằng tổng giá trị của các đặc trưng số 2, 3, 4 và 5 cộng lại.
- Kết quả phân loại chủ quan trả về đây là một câu chủ quan có hàm chứa cảm xúc.
- Sau khi phân loại câu trên là một câu chủ quan có hàm chứa cảm xúc, chương trình tiếp tục rút trích đặc trưng đối với quá trình phân loại cảm xúc là:

1:3.0 2:0.0 3:3.0 4:2.0 5:8.0 6:11.0 7:8.0 8:8.0 9:8.0 10:8.0

Trong đó:

- Các đặc trưng số 1, 2, 3, 4 và 5 được kế thừa từ các đặc trưng số 2, 3, 4, 5 và 6 ở phần phân tích chủ quan.
 - Đặc trưng số 6 có giá trị là 10.0. Trong câu văn trên có từ tăng cường “nhất” mang giá trị 1.0. Từ chịu ảnh hưởng của từ tăng cường là tính từ “tốt” có giá trị 3.0. Do đó, giá trị cảm xúc của câu khi phụ thuộc vào từ tăng cường được tính:
- $$(3.0 * (1.0 + 1.0)) + 0.0 + 3.0 + 2.0 = 11.0$$
- Câu trên chỉ có duy nhất một yếu tố từ tăng cường là yếu tố đặc biệt. Do đó các đặc trưng 7, 8, 9 và 10 đều có giá trị là 8.0 bằng với đặc trưng số 5.

- Kết quả phân loại cảm xúc trả về cho câu này là một câu tích cực vì các đặc trưng đều mang giá trị dương.
- ❖ Việc phân loại chủ quan và phân loại cảm xúc cho một câu văn tiếng Việt thực tế là việc lựa chọn bộ đặc trưng tốt để đạt được kết quả cao. Chúng tôi lựa chọn 02 bộ đặc trưng bên trên sau khi học hỏi được từ những kết quả nghiên cứu đã được công bố đồng thời có phát triển và xây dựng những đặc trưng mới phù hợp với đặc điểm ngôn ngữ của người dùng mạng xã hội nói riêng và thói quen sử dụng ngôn ngữ của người Việt Nam nói chung. Trong giới hạn đề tài này, chúng tôi chưa tiến hành các phương pháp cải tiến công cụ phân lớp SVM để tìm kiếm kết quả tốt hơn.

3.9. Giao diện hệ thống thực nghiệm

Dựa vào những chức năng bên trên, chúng tôi xây dựng hệ thống gồm 03 chức năng sau:

- Đánh giá độ chính xác của dữ liệu thử nghiệm.
- Thu thập dữ liệu.
- Phân loại cảm xúc, đánh giá mức độ quan tâm của người dùng đến chủ đề.

Theo đó, giao diện các chức năng chính của chương trình như sau:

- Giao diện đánh giá độ chính xác của dữ liệu thử nghiệm.

The screenshot displays the 'Main Processing' application window. It features a 'Data Train' table with columns for 'No.', 'Sentences', 'Subjectivity', and 'Sentiment'. The table contains 20 rows of data. Below the table, there is a text input field with the sentence: 'Nước mắt của niềm hạnh phúc, xem Ánh Viên thi mà mình cũng khóc, khóc vì tự hào cho dân tộc Việt Nam'. At the bottom, there are checkboxes for 'Subjectivity' (checked), 'Passive' (unchecked), 'Positive' (checked), and 'Negative' (unchecked). On the right side, there is a 'Dictionary' button, a 'Datasets' section with 'Default' (checked), 'Education' (unchecked), 'Movie' (unchecked), and 'Sport' (unchecked), a 'Run SVM' button, and an 'Accuracy' section showing 'Subjectivity' at 89.8% and 'Sentiment' at 89.5%. At the bottom right, there is a 'Sentiment Analysis' button.

No.	Sentences	Subjectivity	Sentiment
571	Chúc em gặt hái thật nhiều Huy chương	1	1
572	câu này ra đề văn quốc gia được ý nhỉ	0	---
573	Truyền nhân của Yết Kiêu rồi đó	0	---
574	vậy là tốt nhưng đừng quá áp lực nhé	1	1
575	mặt em ấy có gì đấy khiến cho mình cảm thấy rất rất dễ c...	1	1
576	phải có nhiều thật nhiều những bài báo như thế này chứ ...	1	0
577	Việt Nam tự hào về em, Ánh Viên	1	1
578	chị giỏi quá những nỗ lực của chị đã được đền đáp xứng ...	1	1
579	Mới 18 tuổi được đeo hàm đại uý và chuẩn bị phong thiế...	1	1
580	Đã từng là 1 vận động viên, đã từng có suy nghĩ như cô g...	1	0
581	Mỗi lần đọc bài viết về Ánh Viên là máu sôi lên	0	---
582	Tinh thần thể thao thì ai cũng muốn vậy, nhưng đạt được ...	1	1
583	Nửa tỷ nghe ghê, nhưng thật ra chỉ có 500tr	0	---
584	Với 1 người hy sinh máu và nước mắt như Ánh Viên thì th...	1	0
585	Em xứng đáng được nhiều hơn nữa, hơn nhiều nữa	1	1
586	Nước mắt của niềm hạnh phúc, xem Ánh Viên thi mà mìn...	1	1
587	Con gái của tôi Cũng đang trong môi trường rèn luyện nh...	0	---
588	chúc Ánh Viên luôn vượt qua những thành tích mình đạt đ...	1	1
589	Xin mọi người hãy ủng hộ cho con tôi vì Ánh Viên cũng là...	1	1
590	Cần khen thưởng ở mức cao nhất, kể cả tuyên dương an...	1	1
591	Ánh Viên thật tuyệt vời đúng là cô gái vàng	1	1
592	Ngành thể thao và nhà nước Việt Nam nên giành phần th...	1	1
593	Dự là sẽ sớm xuất hiện đề văn nghị luận xã hội câu nói c...	0	---
594	không hài lòng với bản thân để có thể đạt được đỉnh cao ...	1	1
595	Cho em hỏi tại sao phá kỷ lục t7 mà chỉ có 6 HCV a?	0	---

Nước mắt của niềm hạnh phúc, xem Ánh Viên thi mà mình cũng khóc, khóc vì tự hào cho dân tộc Việt Nam

☒ Subjectivity ☐ Passive ☒ Positive ☐ Negative

Dictionary

Datasets

☒ Default ☐ Education ☐ Movie ☐ Sport

Run SVM

Accuracy

Subjectivity
89.8%

Sentiment
89.5%

Sentiment Analysis

Hình 3-8 Giao diện đánh giá độ chính xác của dữ liệu thử nghiệm

- Giao diện phân loại cảm xúc, đánh giá mức độ quan tâm của người dùng.

Sentiment Analysis

Type: Sport

File Clean Analyze

Data

Test Data Cleaned Result

No.	Sentences	Subjectivity	Sentiment
1	ông ad đưa tin về Lý Hoàng Nam đi	0.0	---
2	Chưa cần suy xét xấu-tốt,có tài là rất đắ...	1.0	1.0
3	Ai cũng đắg để học tập đừg mong chờ n...	1.0	1.0
4	Con cái là nền móng của gia đình và x...	1.0	0.0
5	không thành công cũng thành danh, cả...	1.0	1.0
6	Một vị Thủ tướng một nước mà sống tr...	1.0	1.0
7	Một bài học thật hay :hương pháp dạy c...	1.0	1.0
8	Ở Singapore sự tiết kiệm dc tất cả mọi ...	1.0	1.0
9	như gd e cũng vậy khi ăn cơm đến hạt ...	1.0	0.0
10	và khi ra khỏi nhà lúc nào cũng tắt các t...	1.0	1.0
11	Ví như thế mình mới tiết kiệm dc nhiều ...	1.0	1.0
12	Mình cũng mần được vài điều, nhất là "	0.0	---
13	Trong cuộc sống, luôn tồn tại luật nhân...	1.0	1.0
14	Những gì con làm hôm nay là cần nguy...	1.0	1.0
15	CỤ LÝ QUANG ĐIỀU THÌ TUYỆ VỚI VẢ ...	0.0	---
16	"yêu Tổ quốc,yêu đồng bào	1.0	1.0
17	Khiêm tốn,thật thà,dũng cảm"	1.0	1.0
18	Nhé	0.0	---
19	còn thiếu: gió chiều nào thì ngã theo ch...	1.0	0.0
20	COCC thì hách dịch hơi bị ghê	0.0	---
21	Ong cung noi voi cac lanh dao Viet Na...	0.0	---
22	Tiết kiệm, vì sinhgapo ko có tiền chùa	1.0	1.0
23	hay nhất là đừng dạy dút về quá khứ	1.0	1.0

Một vị Thủ tướng một nước mà sống trong sạch thể

☒ Subjectivity ☐ Passive ☒ Positive ☐ Negative

Results

Subjectivity

Sentence: 97

Subjective: 38 ~ 39.2%

Passive: 59 ~ 60.8%

Attention: Low

Sentiment

Sentence: 38

Positive: 30 ~ 78.9%

Negative: 8 ~ 21.1%

Attention: Good

Hình 3-9 Giao diện phân loại cảm xúc, đánh giá mức độ quan tâm của người dùng

- Giao diện thu thập dữ liệu.

Get Data

Facebook Page Name:

Data

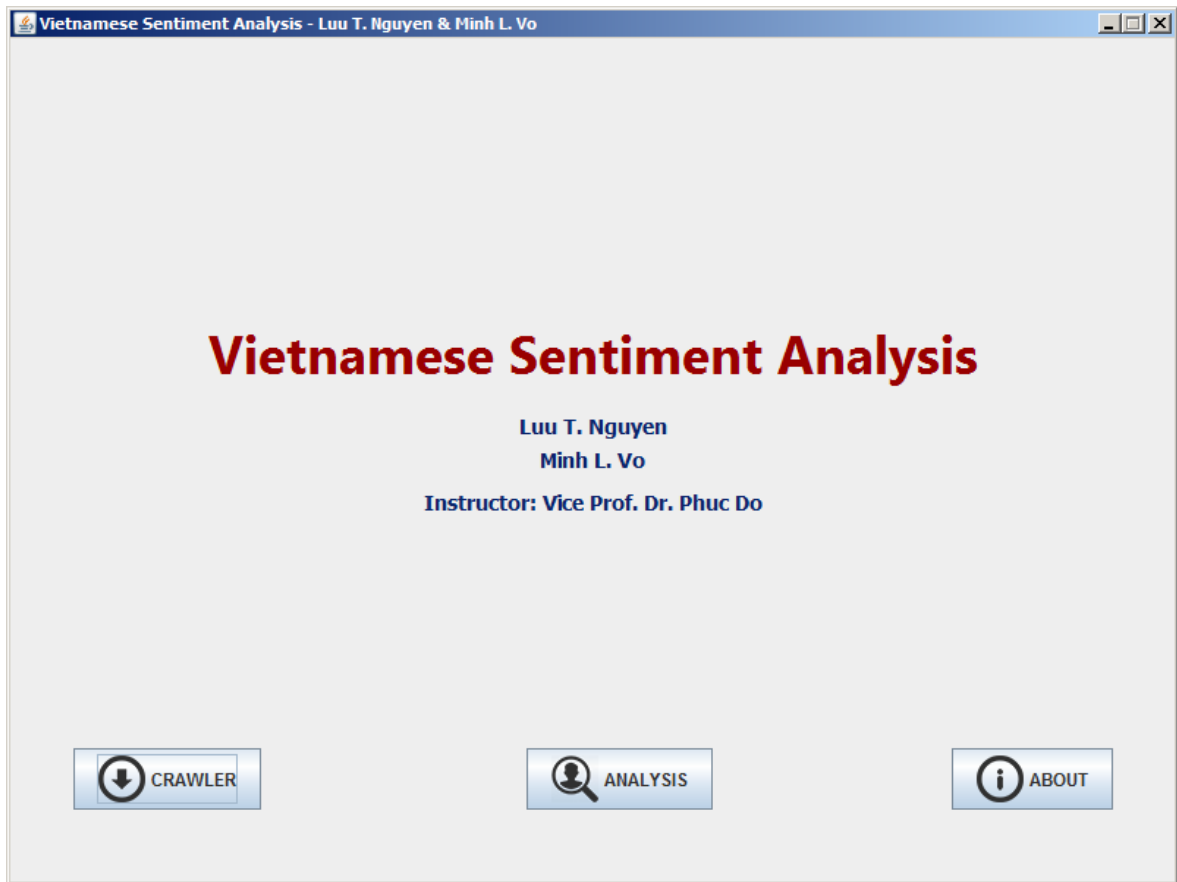
Posts ID **Posts**

1. ít nói, hiền lành lại không nhậu nhẹt nhưng hàng xóm cho biết, những ngày gần đây Tiến chủ động rủ mọi người uống rượu đến say khướt. C
2. Dương từng có tình cảm với Linh - con gái ông Mỹ, đại gia ngành gỗ ở Bình Phước. Biết gia đình người yêu cũ giàu có, hẳn cùng Tiến lên kế h
3. 21h30, ôtô của Công an Bình Phước đã đến khu trọ gần cầu Nhum, xã Nhị Bình, huyện Hóc Môn, TPHCM, chờ theo nghi can gây ra vụ thảm s
4. Bị cho là có liên quan đến vụ sát hại 6 người trong gia đình đại gia ngành gỗ tại Bình Phước, 2 người vừa bị nhà chức trách bắt giữ, chiều 10/
5. Cùng đối tác Sumit Nagai, tay vợt số một Việt Nam đánh bại bộ đôi người Nhật Bản Yusuke Takahashi - Jumpei Yamasaki sau hai set ở tứ k
6. Khoác hờ blazer trên vai, chọn áo cardigan có điểm nhấn, tạo eo bằng áo khoác ngắn... là một số bí quyết giúp bạn ghi điểm trong mắt người
7. Một ngư dân Mỹ bắt được con tôm hùm hiếm, thân hai màu cam và xám. Các nhà khoa học khẳng định, xác suất nhìn thấy loại tôm hiếm như
8. Sau khi đâm nhún viên xe buýt vì lỡ điểm xuống, biết bị nhận diện qua camera, Đức nhuộm tóc, thay đổi diện mạo, đi từ Vinh vào Đắk Lắk che
9. Nữ MC tự nhận mình thiếu kinh nghiệm xử lý khủng hoảng hôn nhân. Hiện cô và gia đình chồng đã ngồi lại bàn bạc, đưa ra cá
10. Nghề diễn xuất với thu nhập bấp bênh không khiến Lê Quang bỏ nghề. Anh đi hát hội chợ, nhà hàng, quán bar... để mưu sinh và nuôi dưỡng
11. Tàu vũ trụ Solar Dynamics Observatory của Cơ quan Hàng không Vũ trụ Mỹ (NASA) ghi lại những hình ảnh tuyệt đẹp của một vụ phun trào ló
12. Chỉ cách thiên đường nghỉ dưỡng Nha Trang 60 km, đến đảo Bình Ba, dân phượt sẽ được tham gia vào nhiều hoạt động giải nhiệt mùa hè
13. Chuyến du lịch bằng tàu hỏa khởi hành vào tháng 10 mang tới cho du khách cơ hội khám phá những thành phố "chưa từng được trông thấy
14. Một cơ trưởng hãng United Airlines vứt đũa vào một thùng rác, và sau đó xả chúng đi trong toilet trên máy bay ông đang điều khiển đến Đức.
15. Sáng nay, hàng chục cảnh sát mặc thường phục đã lục tung từng góc ngách, lùm cây, đường ống... xung quanh khu biệt thự 6 người bị thả
16. Lo lắng tiền gửi có thể được dùng để giải cứu ngân hàng, người dân nước này đang tích cực trả nợ, trả thuê và mua mọi đồ dùng có giá trị đ
17. "Sáu vết dao cắt đau vào lòng nhân loại Một nhất lượng tri theo máu đỏ đầu rơi...".
18. Hơn 10 năm làm nghề giao đá, anh Trần Văn Ái (ngụ quận Tân Phú, TP HCM) vẫn dành thời gian cho âm nhạc sau những giờ lao động mệt
19. Kỳ thi THPT quốc gia năm nay sẽ xuất hiện tình trạng thí sinh rút tốt nghiệp THPT nhưng lại "đầu" đại học do một số tổ hợp môn xét tuyển có
20. Với nhiều dấu vết trên tường rào cao hơn 2 m, cảnh sát nhận định hung thủ đột nhập vào căn biệt thự thăm sát 6 người và tẩu thoát bằng ch
21. Dinh thự Morris-Jumel được mệnh danh là một trong những bí mật được giữ kín nhất của New York, nơi ám ảnh kinh hoàng mà ai cũng ph

Hình 3-10 Giao diện thu thập dữ liệu

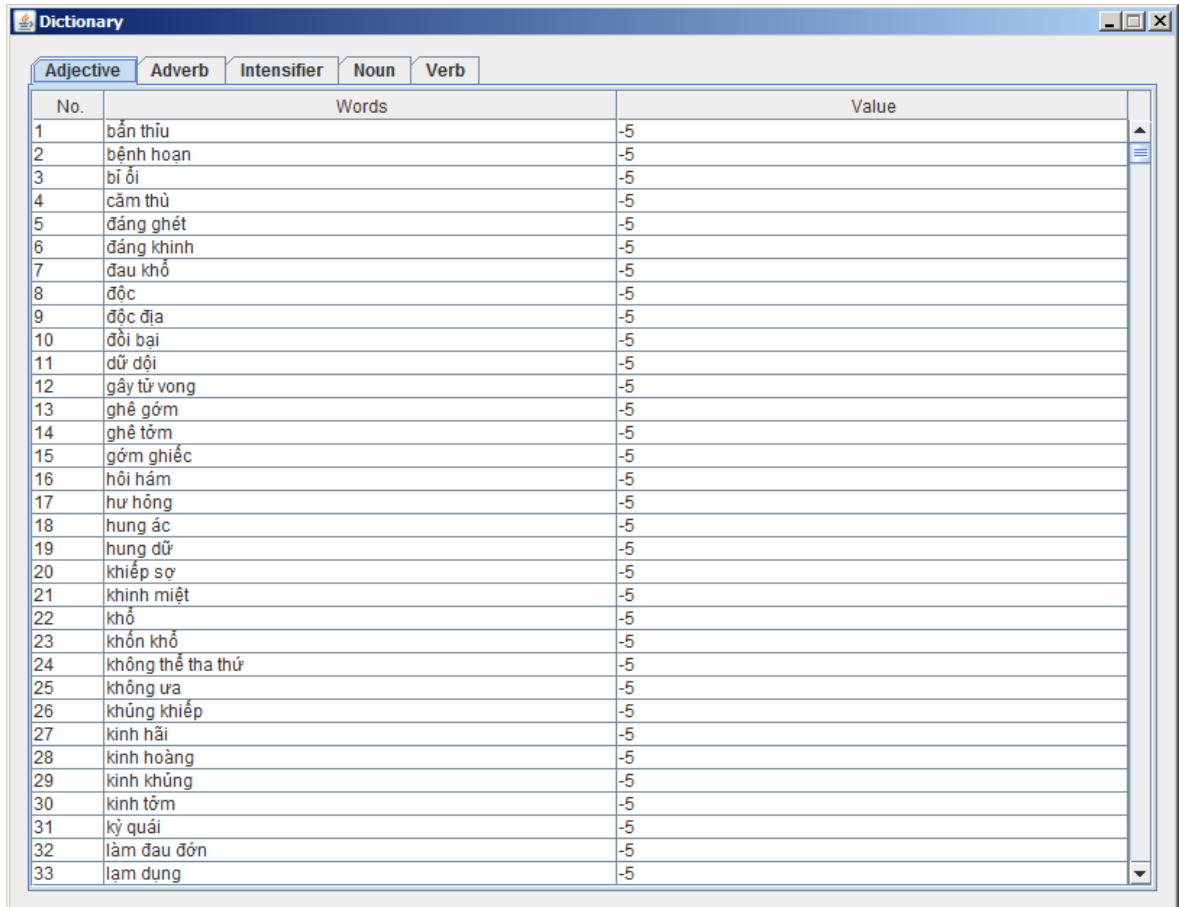
Ngoài những giao diện cho những chức năng chính trên, chương trình còn một số các giao diện hiển thị thông tin khác, bao gồm:

- Giao diện màn hình bắt đầu.



Hình 3-11 Giao diện màn hình bắt đầu

- Giao diện hiển thị từ điển.



The screenshot shows a window titled "Dictionary" with a menu bar containing "Adjective", "Adverb", "Intensifier", "Noun", and "Verb". Below the menu bar is a table with three columns: "No.", "Words", and "Value". The table contains 33 rows of data, each with a number, a Vietnamese word, and the value "-5".

No.	Words	Value
1	bản thù	-5
2	bệnh hoạn	-5
3	bĩ ối	-5
4	cảm thù	-5
5	đáng ghét	-5
6	đáng khinh	-5
7	đau khổ	-5
8	độc	-5
9	độc địa	-5
10	đổi bại	-5
11	dữ dội	-5
12	gây tử vong	-5
13	ghê gớm	-5
14	ghê tởm	-5
15	gớm ghiếc	-5
16	hôi hám	-5
17	hư hỏng	-5
18	hung ác	-5
19	hung dữ	-5
20	khíếp sợ	-5
21	khinh miệt	-5
22	khổ	-5
23	khốn khổ	-5
24	không thể tha thứ	-5
25	không ưa	-5
26	khủng khiếp	-5
27	kinh hãi	-5
28	kinh hoàng	-5
29	kinh khủng	-5
30	kinh tởm	-5
31	kỳ quái	-5
32	lâm đau đớn	-5
33	lạm dụng	-5

Hình 3-12 Giao diện hiển thị từ điển

- Giao diện thông tin tác giả.



Hình 3-13 Giao diện thông tin tác giả

Chương 4. KẾT QUẢ THỬ NGHIỆM

4.1. Bộ dữ liệu thử nghiệm

Chúng tôi thu thập 3 bộ dữ liệu từ 3 chủ đề: giáo dục, phim ảnh và thể thao. Mỗi bộ dữ liệu bao gồm hơn 100 câu bình luận về các chủ đề:

- Giáo dục: gồm 135 câu bình luận thu thập ở tin “Phạm Minh Hiếu (19 tuổi) vừa được Stanford University (Mỹ), đứng thứ tư trong danh sách đại học hàng đầu thế giới, đồng ý cấp học bổng. Đại học Chicago, Columbia (top 15 thế giới) cũng mời Hiếu sang học.” trên trang VnExpress.net của mạng xã hội facebook.
- Phim ảnh: gồm 146 câu bình luận thu thập chủ yếu từ tin “Cha là người hùng đầu tiên của con trai và là tình yêu đầu tiên của con gái” về phim “La Vita e Bella”, tin “Thế giới quan của mỗi con người đều hạn hẹp, ai cũng đầy trong tâm tưởng những định kiến về một phần còn lại của thế giới.” về phim “Intouchables” trên trang “Mann up” và một số tin trên trang “CGV Cinemas Vietnam” của mạng xã hội facebook.
- Thể thao: gồm 162 câu bình luận thu thập từ tin “Nếu tôi hài lòng với những gì đã đạt được, tôi là kẻ thất bại ngay từ bây giờ, chứ không phải chờ tới ngày mai.” về vận động viên Ánh Viên và tin “Điều mà nhà vô địch điền kinh SEA Games năm nào cần bây giờ là chữa khỏi cái lưng, đi lại được và tiếp tục công tác huấn luyện, để chăm sóc cho cậu út vẫn còn bệnh tật và trả nợ cho ngôi nhà.” về vận động viên Vũ Bích Hương trên trang VnExpress.net của mạng xã hội facebook.

Tương tự như bộ dữ liệu huấn luyện, chúng tôi tổng hợp các chủ đề trên để xây dựng bộ dữ liệu lớn hơn bao gồm 443 câu.

Sau đó, chúng tôi tiến hành phân loại chủ quan và phân loại cảm xúc bằng tay. Kết quả được trình bày theo bảng sau:

Bảng 4-1 Kết quả phân loại bằng tay bộ dữ liệu thử nghiệm

STT	Chủ đề	Dữ liệu thử nghiệm			
		Câu chủ quan	Câu khách quan	Câu tích cực	Câu tiêu cực
1	Giáo dục	87	48	76	11
2	Phim ảnh	106	40	80	26
3	Thể thao	121	41	113	8
4	Tổng hợp	314	129	269	45

4.2. Kết quả đánh giá phương pháp phân loại chủ quan

Từ bộ dữ liệu thử nghiệm phân loại bằng tay, phương pháp phân lớp SVM và bộ dữ liệu huấn luyện ở mục 3.6 chúng tôi tiến hành kiểm tra độ chính xác của phương pháp phân loại chủ quan. Kết quả đánh giá mức độ chính xác theo bảng sau:

Bảng 4-2 Kết quả đánh giá độ chính xác phương pháp phân loại chủ quan

STT	Chủ đề	Kết quả thử nghiệm (độ chính xác: %)
1	Giáo dục	92.6%
2	Phim ảnh	89.7%
3	Thể thao	89.5%
4	Tổng hợp	89.8%

4.3. Kết quả đánh giá phương pháp phân loại cảm xúc

Chúng tôi tiếp tục tiến hành đánh giá độ chính xác của phương pháp phân loại cảm xúc. Kết quả được trình bày trong bảng sau:

Bảng 4-3 Kết quả đánh giá độ chính xác phương pháp phân loại cảm xúc

STT	Chủ đề	Kết quả thử nghiệm (độ chính xác: %)
1	Giáo dục	90.8%
2	Phim ảnh	79.2%
3	Thể thao	95.0%
4	Tổng hợp	89.5%

Kết quả này nằm ngoài mong đợi của chúng tôi vì với bộ từ điển chỉ được dịch trong thời gian ngắn thì kết quả này rất tốt. Qua đó chúng tôi phương pháp chúng tôi lựa chọn là có hiệu quả.

Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được

Xuyên suốt quá trình thực hiện, chúng tôi được tiếp cận với nhiều nghiên cứu liên quan cả trong và ngoài nước. Điều đó giúp cho chúng tôi hiểu được rộng hơn, sâu hơn và kỹ càng hơn về đề tài.

Chúng tôi đã cố gắng xây dựng một phương pháp khả thi để phân tích cảm xúc trên ngôn ngữ tiếng Việt dựa vào những đặc trưng sử dụng ngôn từ của người dùng mạng xã hội ở Việt Nam. Tuy nhiên, với giới hạn của một khóa luận tốt nghiệp chúng tôi mới chỉ xây dựng được một mô hình phân loại cảm xúc có dựa vào phương thức học máy một cách rập khuôn mà chưa xét đến các vấn đề về xử lý ngôn ngữ tự nhiên. Đồng thời, bộ dữ liệu huấn luyện và bộ dữ liệu thử nghiệm mà chúng tôi xây dựng vẫn còn ít nên chắc chắn sẽ có những thiếu sót các trường hợp. Ngoài ra, việc dịch bộ từ điển cảm xúc từ bộ từ điển tiếng Anh thì độ chính xác sẽ không cao nhưng với thời gian có hạn nên chúng tôi chỉ có thể thực hiện ở mức độ này.

Những vấn đề trên là những hạn chế dễ thấy nhất của đề tài. Nếu có cơ hội tiếp tục phát triển đề tài này hy vọng những hạn chế bên trên sẽ dần được khắc phục để có thể xây dựng một hệ thống tốt hơn, hoàn thiện hơn.

Mặc dù còn gặp nhiều khó khăn nhưng với sự hướng dẫn tận tình của PGS. TS. Đỗ Phúc, sự trợ giúp nhiệt tình của Thạc sĩ Trịnh Quốc Sơn và những chia sẻ chân thành của Thạc sĩ Nguyễn Ngọc Duy, Thạc sĩ Võ Ngọc Phú, chúng tôi đã đạt được kết quả hết sức khả quan trên cả mong đợi ban đầu. Với kết quả này, chúng tôi hy vọng có thể phát triển đề tài lên một mức cao hơn và áp dụng vào thực tiễn cuộc sống cũng như đóng góp vào các nghiên cứu khoa học khác có liên quan.

5.2. Hướng phát triển

Việc có thể phát triển một phương pháp phân tích cảm xúc tiếng Việt, đặc biệt là đối với dữ liệu từ mạng xã hội vốn đã không phải là dạng dữ liệu chuẩn tiếng Việt thì còn cần nhiều cải tiến và nhiều nghiên cứu khác: xây dựng bộ từ điển cảm xúc đủ lớn và độ chính xác cao, xây dựng bộ dữ liệu huấn luyện và bộ dữ liệu thử nghiệm

đạt chuẩn về độ lớn và độ chính xác, áp dụng các phương pháp xử lý ngôn ngữ tự nhiên, phương pháp chuẩn hóa dữ liệu từ mạng xã hội, giải quyết bài toán big data khi chương trình thực thi trên bộ dữ liệu lớn, v.v... nhằm đạt được độ chính xác tốt hơn và hiệu năng hệ thống tốt hơn đối với khối lượng dữ liệu lớn hơn.

Đồng thời, khi được cải tiến và nâng cấp hệ thống thực nghiệm, chúng tôi hy vọng đề tài có thể được áp dụng trong thực tiễn cuộc sống và đóng góp cho các nghiên cứu khác có liên quan. Thiết thực nhất là việc đánh giá ý kiến khách hàng trong lĩnh vực kinh tế.

TÀI LIỆU THAM KHẢO

TÀI LIỆU TIẾNG VIỆT

- [1] Thái Sơn, “Luận văn thạc sĩ khoa học: Kỹ thuật Support Vector Machines và ứng dụng.”, ngành toán tin ứng dụng, Đại học Bách khoa Hà Nội, 2006.
- [2] Nguyễn Ngọc Duy, “Luận văn thạc sĩ khoa học: Tóm tắt ý kiến trên cơ sở phân loại cảm xúc”, ngành Khoa học máy tính, Đại học Bách khoa Hồ Chí Minh, 2014.

TÀI LIỆU TIẾNG ANH

- [3] Vo Ngoc Phu and Phan Thi Tuoi, “Sentiment classification using Enhanced Contextual Valence Shifters”, *Proceedings of International Conference on Asian Language Processing*, Malaysia, 2014.
- [4] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede, “Lexicon-Based Methods for Sentiment Analysis”, *Association for Computational Linguistics*, 2011.
- [5] Bing Liu, “Sentiment Analysis and Opinion Mining”, *Morgan & Claypool Publishers*, May 2012.
- [6] Wiebe, Janyce, Rebecca F. Bruce, and Thomas P. O'Hara, “Development and use of a gold-standard data set for subjective classifications”, *Proceedings of the Association for Computational Linguistics (ACL-1999)*, 1999.
- [7] Bo Pang and Lillian Lee, “A Sentimental Education: Sentiment Analysis Using Subjective Summarization Based on Minimum Cuts”, *Proceedings of ACL*, pp. 271-278, 2004.
- [8] Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena, “Large-Scale Sentiment Analysis for News and Blogs”, ICWSM '2007 Boulder, Colorado, USA.
- [9] Rudy Prabowo and Mike Thelwall, “Sentiment Analysis: A Combined Approach”, *Journal of Informetrics Volume 3*, Issue 2, Pages 143–157, April 2009.
- [10] Farah Benamara, Carmine Cesarano and Diego Reforgiato, “Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone”, *ICWSM '2006 Boulder*, CO USA, 2006.

- [11] A Go, L Huang, R Bhayani – Entropy, “Twitter Sentiment Analysis”, *CS224N - Final Project Report*, June 6, 2009.
- [12] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”, *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- [13] Thanh Ho, Duy Doan and Phuc Do, “Discovering Hot Topics On Social Network Based On Improving The Aging Theory”, *ACSIJ Advances in Computer Science: an International Journal*, Vol. 3, Issue 3, No.9 , May 2014.
- [14] Kennedy, Alistair and Diana Inkpen, “Sentiment classification of movie and product reviews using contextual valence shifters”, *Computational Intelligence*, 2006.
- [15] Mrutyunjaya Panda, Satchidananda Dehuri and Gi-Nam Wang, “Social Networking Mining, Visualization, and Security”, *Springer International Publishing*, Switzerland, 2014.

TÀI LIỆU TRÊN MẠNG INTERNET

- [16] Digital, Social and Mobile in 2015, <http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/>, truy cập ngày 01/04/2015.
- [17] Twitter statistics, <http://www.statisticbrain.com/twitter-statistics/> , truy cập ngày 16/01/2015.
- [18] Facebook statistics, <http://www.statisticbrain.com/facebook-statistics/> , truy cập ngày 16/01/2015.
- [19] Dictionaries for the Semantic Orientation CALculator, <https://github.com/DrOttenssooser/BiblicalNLPworks/tree/master/SkyDrive/NLP/CommonWorks/Data/Opion-Lexicon-English/SO-CAL>, truy cập ngày 01/05/2015
- [20] vnTokenizer, <http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer>, truy cập ngày 15/05/2015.
- [21] vnTagger, <http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger> , truy cập ngày 15/05/2015.

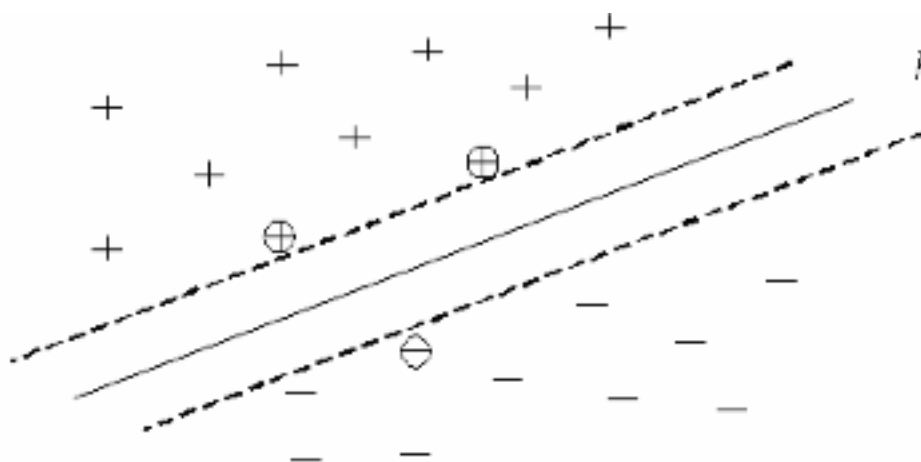
- [22] Epinions 1, https://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html, truy cập ngày 18/05/2015.
- [23] Polarity Dataset, www.cs.cornell.edu/People/pabo/movie-review-data/, truy cập ngày 03/06/2015.
- [24] VIETTIEN Dictionary for Mac, http://nguyenvietkhoa.edu.vn/?page_id=346, truy cập ngày 01/05/2015.

PHỤ LỤC

I. Phương pháp SVM

1. Ý tưởng

Ý tưởng chính của thuật toán này là cho trước một tập huấn luyện được biểu diễn trong không gian vector trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một mặt phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp $+$ và lớp $-$. Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm ra được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt.



Hình I-1 Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp “+” và “-” với khoảng cách biên lớn nhất

Các điểm gần h nhất là các Support Vector. Xem dữ liệu đầu vào như hai tập vector n chiều, một SVM sẽ xây dựng một mặt phẳng riêng biệt trong không gian đó sao cho nó tối đa hóa biên lề giữa hai tập dữ liệu. Để tính lề, hai siêu phẳng song song được xây dựng, mỗi cái nằm ở 1 phía của siêu phẳng phân biệt và chúng được đẩy về phía hai tập dữ liệu.

Sau quá trình huấn luyện nếu hiệu suất tổng quát hoá của bộ phân lớp cao thì thuật toán huấn luyện được đánh giá là tốt. Hiệu suất tổng quát hoá phụ thuộc vào hai

tham số là sai số huấn luyện hay và năng lực của máy học. Trong đó sai số huấn luyện là tỷ lệ lỗi phân lớp trên tập dữ liệu huấn luyện. Còn năng lực của máy học được xác định bằng kích thước Vapnik-Chervonenkis (kích thước VC). Kích thước VC là một khái niệm quan trọng đối với một họ hàm phân tách (hay là tập phân lớp). Đại lượng này được xác định bằng số điểm cực đại mà họ hàm có thể phân tách hoàn toàn trong không gian đối tượng. Một tập phân lớp tốt là tập phân lớp có năng lực thấp nhất (có nghĩa là đơn giản nhất) và đảm bảo sai số huấn luyện nhỏ.

2. Cơ sở lý thuyết

Xét bài toán phân lớp đơn giản nhất – phân lớp hai lớp với tập dữ liệu mẫu:

$$\{(x_i, y_i) | i = 1, 2, \dots, n; x_i \in R^m\}$$

Trong đó mẫu là các vector đối tượng được phân lớp thành các mẫu dương và mẫu âm như trong hình 3.1:

- 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 $y_i = -1$

Thực chất phương pháp này là một bài toán tối ưu, mục tiêu là tìm ra một không gian H và siêu mặt phẳng quyết định h trên H sao cho sai số phân lớp là thấp nhất.

Trong trường hợp này, tập phân lớp SVM là mặt siêu phẳng phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại, trong đó độ chênh lệch – còn gọi là Lề (margin) xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất (hình 1). Mặt siêu phẳng này được gọi là mặt siêu phẳng lề tối ưu.

Các mặt siêu phẳng trong không gian đối tượng có phương trình là:

$$C + w_1x_1 + w_2x_2 + \dots + w_nx_n = 0$$

Tương đương với công thức

$$C + \sum_{i=1}^n w_i x_i = 0$$

Với:

$w = w_1 + w_2 + \dots + w_n$ là bộ hệ số siêu phẳng hay là vector trọng số,

C là độ dịch, khi thay đổi w và C thì hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi.

Tập phân lớp SVM được định nghĩa như sau:

$$f(x) = \sin\left(C + \sum_{i=1}^n w_i x_i\right)$$

Trong đó,

$\sin(z) = 1$ nếu $z \geq 0$,

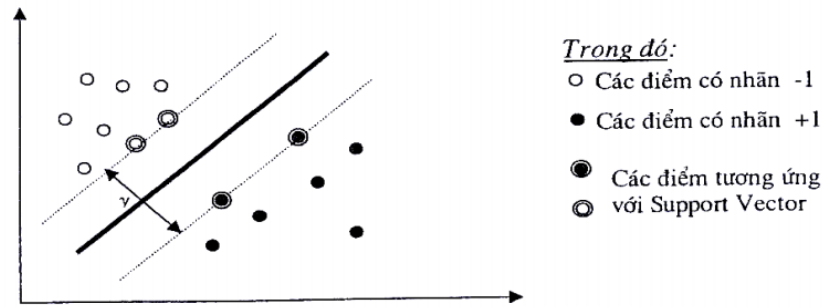
$\sin(z) = -1$ nếu $z < 0$.

Nếu $f(x) = 1$ thì x thuộc về lớp dương (lĩnh vực được quan tâm), và ngược lại, nếu $f(x) = -1$ thì x thuộc về lớp âm (các lĩnh vực khác).

a) Bài toán phân hai lớp SVM

Bài toán đặt ra là xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới x_i thì cần phải xác định x_i được phân vào lớp +1 hay lớp -1.

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tích hai lớp này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được. Việc phân tách này được gọi là phân tách tuyến tính.



Hình I-2 Minh hoạ bài toán phân hai lớp với phương pháp SVM

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

b) Bài toán phân nhiều lớp với SVM

Việc phân nhiều lớp với SVM thì cũng giống như quá trình chia không gian thành 2 phần và quá trình này được lặp lại nhiều lần. Khi đó hàm quyết định phân lớp dữ liệu vào lớp thứ i của tập n -lớp sẽ là:

$$f_i(x) = w_i^T x + b_i$$

Những phần tử x là support vector sẽ thỏa điều kiện:

$$\begin{cases} f_i(x) = 1 \\ f_i(x) = -1 \end{cases}$$

Trong đó

$$f_i(x) = 1 \text{ nếu thuộc lớp } i,$$

$$f_i(x) = -1 \text{ nếu thuộc các lớp còn lại.}$$

Như vậy, bài toán phân nhiều lớp sử dụng phương pháp SVM hoàn toàn có thể thực hiện giống như bài toán hai lớp. Bằng cách sử dụng chiến lược "một đối một" (one - against - one). Giả sử bài toán cần phân loại có k lớp ($k > 2$), chiến lược "một đối một" sẽ tiến hành $\frac{k(k-1)}{2}$ lần phân lớp nhị phân sử dụng phương pháp SVM.

Mỗi lớp sẽ tiến hành phân tách với $k - 1$ lớp còn lại để xác định $k - 1$ hàm phân tách dựa vào bài toán phân hai lớp bằng phương pháp SVM.

II. Bộ từ điển cảm xúc SO-CAL tiếng Việt

Một số từ trong các bộ từ điển từ loại

<i>Danh từ</i>		<i>Động từ</i>		<i>Tính từ</i>		<i>Trạng từ</i>	
kiệt tác	5	cảm phục	4	vượt trội	5	lộng lẫy	5
cực điểm	4	hoan hỉ	4	xuất sắc	5	xuất sắc	5
kỳ quan	4	thành đạt	3	nổi bật	5	sáng chói	5
thiên tài	4	hân hoan	3	ưu việt	4	thần kỳ	4
công lao	3	đăng quang	3	thông minh	4	tuyệt hảo	4
cống hiến	3	thắng	2	sôi nổi	4	tích cực	3
huy chương	3	quyết tâm	2	thuận lợi	3	nhộn nhịp	3
chiến công	2	hợp tác	2	tế nhị	3	nhất quán	2
học bổng	2	cảm ơn	2	nổi tiếng	3	phù hợp	2
phúc lợi	2	tốt nghiệp	1	trẻ	2	lôi cuốn	2
niềm tin	2	miễn phí	1	tin xảo	2	khiêm tốn	1
ấn tượng	1	giúp đỡ	1	trong sáng	1	mãnh liệt	1
độc lập	1	chán	-1	hợp lý	1	nhiệt tình	1
vận may	1	hủy	-1	thích hợp	1	tồi tệ	-1
cô lập	-1	đói	-2	yêu ót	-1	nặng nề	-1
kẻ cướp	-1	buồn	-2	mờ mịt	-1	khó chịu	-2
bạo lực	-2	ám sát	-2	tồi	-2	ngu ngốc	-2
ác mộng	-3	hối tiếc	-3	thô bỉ	-3	thô thiển	-3
rác rưởi	-3	đánh	-3	ngu	-3	quanh co	-3

Một số từ trong từ điển từ tăng cường

<i>Từ tăng cường</i>	<i>Giá trị</i>
ít nhất	-3
vài	-2
thấp	-2
vài	-2
hầu như	-1.5
ít hơn	-1.5
chỉ	-0.5
một chút	-0.5
một phần	-0.3
tương đối	-0.3
vừa phải	-0.3
chắc chắn	0.2
ngay	0.1
chính	0.2
đáng kinh ngạc	0.3
tuyệt	0.3
khó tin	0.4
quá chừng	0.4
rất đối	0.4
vô cùng	0.4
không lồ	0.5
nhiều hơn	0.5
phi thường	0.5
tuyệt đối	0.5
hết sức	0.5

III. Bộ dữ liệu thử nghiệm

1. Bộ dữ liệu về chủ đề giáo dục

Câu	Nội dung các ý kiến
Khách quan	Con nhà người ta
	Trước năm 75, ở miền Nam đa phần sinh viên du học đều trở về quê hương làm việc
	Còn ngày nay, sinh viên ra đi du học là không muốn quay về đất nước nữa
	Đừng về Việt Nam nha em
	con nhà người ta là đây
	nếu mà ở Việt Nam thì chắc là mơ đi cung
Chủ quan	đi mà ở bên đó luôn cho sướng, đừng về
	về là bị vui dập dấy
	Học xong cố gắng ở bên đó công tác
	Nhân tài mà về Việt Nam bọn nó hành cho ra bả
	chúc em sau này sẽ là 1 công dân Mỹ thành đạt
	Nước ngoài người ta hậu đãi thế đấy Việt Nam còn ảo tưởng đến bao giờ
	Thật buồn cho Việt Nam
	Thật là giỏi, khâm phục bạn
	Hy vọng đất nước ta không phải "chảy máu chất xám"
	cố lên
	phát triển sự nghiệp đổi mới cho nước người ta
	nước ngoài họ luôn biết cách trọng dụng nhân tài

Câu	Nội dung các ý kiến
Tiêu cực	Thật buồn cho Việt Nam
	Giỏi vậy chứ thi công chức Việt Nam rớt chắc 100%, em đi đi nha
	Quản lý viết văn lủng củng quá
	Đọc bình luận mà buồn
	xấu trai
Tích cực	Hãy học cái hay cái tốt những điều mới mẻ của người ta để phục vụ cho quê hương mình thế mới là một người thông minh yêu nước em nhé
	Chúc em bình an và thành tài
	Ngưỡng mộ quá
	Tự hào 2 tiếng Việt Nam
	Hãy biết ơn đất nước Việt Nam và hãy học thật giỏi để mang vinh quan về cho đất nước, cho Đảng em nhé
	Ngưỡng mộ em, cố gắng thành người có ích nhé em
	Chúc em thành công
	Thật tự hào nhân tài đất Việt
	Chúc em bình an và thành tài
	Chúc em học thành tài mai sau dựng xây đất nước thế mới gọi là một người có tài có tâm em nhé

2. Bộ dữ liệu về chủ đề phim ảnh

Câu	Nội dung các ý kiến
Khách quan	Nguyễn Anh Tú hồi đấy tao ra rạp để xem phim này luôn đấy.
	Ý kiến riêng tôi
	Có bạn nào nghiện bản nhạc phim này giống mình không?
	ai có đường dẫn xem phim không?
	phim này con xem cũng khoảng 4 lần.
	Thì ra " Anh Da Đen " từ đây mà ra.
Chủ quan	bộ phim hay nhất từng xem.
	phim này hay, mình xem không dưới 5 lần và nhớ đến từng chi tiết, cả phim about time nữa.
	Đây là một trong những bộ phim tôi thích.
	Đúng.
	Kéo nhau ra khỏi nguyên tắc của nhau nên tình cảm của chúng tôi mới nảy sinh.
	Nói thiệt là em yêu giọng văn của mấy cái anh quản lý trang này quá à.
	Vừa lãng mạn vừa có chiều sâu đôi khi lại rất đáng yêu.
	Với Driss anh không xem Phillip là 1 người tàn tật mà đối xử với ông như 1 người bình thường.
	Công nhận là hay thật
	Tôi thấy được một bài học ở phim là phải luôn giữ cho mình khoẻ mạnh dù có không giàu nhưng khoẻ là được
	phim này ai chưa xem thì nên xem rất ý nghĩa
	Phim này xem lại mấy lần vẫn hay như lần đầu.

Câu	Nội dung các ý kiến
Tiêu cực	Ban đầu tôi thấy buồn cười và khó hiểu với tình cảm của mình.
	Nhưng nói thật tạo hình yêu quái quá xấu.
	Nội dung phim ổn nhưng nhìn mấy con yêu quái chán quá.
	riết rồi làm phim thua xa phim hồi xưa xem nữa chứ.
	Đồ hoạ phim tệ vậy.
	Phim tệ quá nghỉ đi
	nhân vật đọc lời thoại chậm thật không có mạch lạc.
Tích cực	bộ phim hay nhất từng xem.
	phim này hay, mình xem không dưới 5 lần và nhớ đến từng chi tiết, cả phim about time nữa.
	Đây là một trong những bộ phim tôi thích.
	Nói thiệt là em yêu giọng văn của mấy cái anh quản lý trang này quá à.
	Vừa lãng mạn vừa có chiều sâu đôi khi lại rất đáng yêu.
	Với Driss anh k xem Phillip là 1 người tàn tật mà đối xử với ông như 1 người bình thường.
	Điều đó đã gắn kết họ với nhau bằng 1 tình bạn giữa những kẻ đang tìm cho bản thân mình 1 lẽ sống !
	Thích nhất đoạn này.
	Thưa các bạn , đây là một trong những bộ phim hay nhất tôi từng xem
	Tôi thấy được 1 bài học ở phim là phải luôn giữ cho mình khoẻ mạnh dù có không giàu nhưng khoẻ là được

3. Bộ dữ liệu về chủ đề thể thao

Câu	Nội dung các ý kiến
Khách quan	ánh viên + công phượng = công viên
	Người con của đất Phong Điền Cần thơ
	câu này ra đề văn quốc gia được ý nhi
	Truyền nhân của Yết Kiêu rồi đó
	Con gái của tôi cũng đang trong môi trường rèn luyện như Ánh Viên
	Trời ơi
Chủ quan	đây là tính cách của 1 nhà vô địch, Ánh Viên sẽ còn tiến xa.
	xin chúc mừng Ánh Viên, chúc mừng Việt Nam
	Thay vì suốt ngày tập trung vào mấy hot girl hot boy ca sĩ diễn viên suốt ngày trưng "hàng" khoe tài sản, giới trẻ Việt Nam nên thần tượng bản lĩnh và ý chí của cô gái trẻ này
	Viên đẹp từ trong chính tâm hồn của em, không cần phải trang điểm điểm tô
	Yêu và tự hào về em lắm
	Xem các vận động viên thi đấu sừng sật
	Họ rèn luyện, hi sinh nhiều thứ chỉ để tỏa sáng trong khoảnh khắc
	Chúc mừng Đại Uý Ánh Viên
	Bơi nhanh chẳng kém Thủy Thần Yết Kiêu
	Triệu Fan cả nước mến yêu
	Chúc em gặt hái thật nhiều huy chương
	Việt Nam tự hào về em, Ánh Viên

Câu	Nội dung các ý kiến
Tiêu cực	Tội chi quá.
	Sau ánh hào quang thể thao là bệnh tật, đau bệnh.
	Còn khi đã hết thời rồi thì bị đối xử quá tệ bạc.
	Nghĩ mà buồn cho những số phận như thế này.
	Việt Nam là vậy, vắt chanh bỏ vỏ.
	Cái nghiệp bạc bẽo nhất là nghiệp thể thao.
	Thương xót.
Tích cực	đây là tính cách của 1 nhà vô địch, Ánh Viên sẽ còn tiến xa.
	xin chúc mừng Ánh Viên, chúc mừng Việt Nam
	Thay vì suốt ngày tập trung vào mấy hot girl hot boy ca sĩ diễn viên suốt ngày trưng "hàng" khoe tài sản, giới trẻ Việt Nam nên thần tượng bản lĩnh và ý chí của cô gái trẻ này
	Viên đẹp từ trong chính tâm hồn của em, không cần phải makeup điểm tô
	Yêu và tự hào về em lắm
	Xem các vận động viên thi đấu sướng thật
	Họ rèn luyện, hi sinh nhiều thứ chỉ để tỏa sáng trong khoảnh khắc
	Chúc mừng Đại Uý Ánh Viên
	Đây mới là tấm gương để nỗ lực
	Mang lại vinh quang cho tổ quốc, cho gia đình và Ánh Viên là niềm tự hào của dân tộc
	Thật sự rất yêu quý và khâm phục chị!

IV. Thử nghiệm phân tích dữ liệu

Phân tích bình luận “*Đề dài, chỉ sợ viết không kịp chứ mình cảm thấy đề không khó lắm. Các sĩ tử làm bài như thế nào rồi nhỉ?*”

Tiền xử lý, cắt câu:

Đề dài, chỉ sợ viết không kịp chứ mình cảm thấy đề không khó lắm

Các sĩ tử làm bài như thế nào rồi nhỉ

Gán nhãn:

```
<doc>
  <s>
    <w pos="Np">Đề</w>
    <w pos="A">dài</w>
    <w pos="," ">,</w>
    <w pos="R">chỉ</w>
    <w pos="V">sợ</w>
    <w pos="V">viết</w>
    <w pos="R">không</w>
    <w pos="A">kịp</w>
    <w pos="C">chứ</w>
    <w pos="P">mình</w>
    <w pos="V">cảm thấy</w>
    <w pos="N">đề</w>
    <w pos="R">không</w>
    <w pos="A">khó</w>
    <w pos="R">lắm</w>
  </s>
  <s>
    <w pos="L">Các</w>
    <w pos="N">sĩ tử</w>
    <w pos="V">làm</w>
```

```

        <w pos="N">bài</w>
        <w pos="X">như thế nào</w>
        <w pos="T">rồi</w>
        <w pos="T">nhỉ</w>

    </s>
</doc>

```

Rút đặc trưng:

- Chủ quan:

- Câu: “Đề dài, chỉ sợ viết không kịp chứ mình cảm thấy đề không khó lắm”

1:14.0 2:-2.0 3:-2.0 4:0.0 5:-1.0 6:-5.0

Trong đó:

- Đặc trưng số 1 là tổng số từ trong câu có giá trị là 14.0 vì câu có 14 từ.
- Đặc trưng số 2 là tổng giá trị cảm xúc của các tính từ có giá trị là -2.0 do trong câu có một tính từ chứa cảm xúc là “khó” (-2).
- Đặc trưng số 3 là tổng giá trị cảm xúc của các trạng từ có giá trị là -2.0 do trong câu có một trạng từ chứa cảm xúc là “chỉ” (-2).
- Đặc trưng số 4 là tổng giá trị cảm xúc của các danh từ có giá trị là 0.0 vì trong câu không có danh từ nào chứa cảm xúc.
- Đặc trưng số 5 là tổng giá trị cảm xúc của các động từ có giá trị là -1.0 vì trong câu có một động từ chứa cảm xúc là “sợ” (-1).
- Đặc trưng số 6 là tổng giá trị cảm xúc trong câu có giá trị là (-5.0). Ta thấy, đây là một câu bình thường và không thuộc vào những trường hợp ngoại lệ. Do đó tổng giá trị cảm xúc trong câu bằng tổng giá trị cảm xúc của các loại từ trong câu hay nói cách khác bằng tổng giá trị của các đặc trưng số 3, 4, 5 và 6 cộng lại: $(-2) + (-2) + (0) + (-1) = (-5)$.
- Câu: “Các sĩ tử làm bài như thế nào rồi nhỉ”

1:7.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0

Trong đó:

- Đặc trưng số 1 có giá trị là 7.0 vì câu có 7 từ.
- Đặc trưng số 2, 3, 4, 5 đều có giá trị là 0.0 vì trong câu không có tính từ, trạng từ, danh từ hay động từ nào chứa cảm xúc.
- Đặc trưng số 6 có giá trị là 0.0 do đây là một câu nghi vấn (vì có cụm từ “như thế nào”) nên tổng giá trị cảm xúc của cả câu sẽ là 0.

- Phân tích cảm xúc:

- Câu: “*Đề dài, chỉ sợ viết không kịp chứ mình cảm thấy đề không khó lắm*”

1:-2.0 2:-2.0 3:0.0 4:-1.0 5:-5.0 6:-4.5 7:-1.0 8:-5.0 9:-7.5 10:-5.0

Trong đó:

- Các đặc trưng số 1, 2, 3 và 4 được kế thừa từ các đặc trưng số 3, 4, 5 và 6 ở phần phân tích chủ quan.
- Đặc trưng số 5 có giá trị bằng tổng giá trị của các đặc trưng 1, 2, 3 và 4 cộng lại là (-5.0)
- Đặc trưng số 6 là giá trị cảm xúc của câu sau khi hệ thống xét trường hợp có từ tăng cường. Đặc trưng số 6 có giá trị là -4.5 vì trong câu có từ “chỉ” mang giá trị cảm xúc là (-0.5) trong từ điển từ tăng cường và từ “sợ” mang giá trị cảm xúc là (-1). Do đó, giá trị cảm xúc trong cả câu được tính như sau: $(-2) + (-0.5)*(-1) + (-1) + (-2) = (-4.5)$.
- Đặc trưng số 7 là giá trị cảm xúc của câu khi chúng tôi xét đến trường hợp giá trị cảm xúc trong câu thay đổi nếu câu có 2 vế và liên kết với nhau bằng từ liên kết mang nghĩa phủ định. Trong câu trên có từ liên kết mang nghĩa phủ định là “không” và tính từ “khó” mang giá trị cảm xúc là (-2) ở sau nên giá trị cảm xúc của câu được tính như sau: $(-1) + (-2) + (-1)*(-2) = (-1)$. Vì vậy đặc trưng số 7 có giá trị là (-1.0).
- Đặc trưng số 8 là giá trị cảm xúc của câu khi chúng tôi xét trường hợp giá trị cảm xúc của câu thay đổi khi chịu ảnh hưởng của từ khiếm

khuyết. Câu này không có từ khiếm khuyết cho nên giá trị cảm xúc của nó không thay đổi và bằng với giá trị của đặc trưng số 5.

- Đặc trưng số 9 là giá trị cảm xúc của câu khi chúng tôi xét trường hợp giá trị cảm xúc trong câu thay đổi khi tăng 50% giá trị cảm xúc đối với từ tiêu cực. Trong trường hợp này, câu có ba từ tiêu cực là từ “chỉ”, từ “sợ” và từ “khó”. Ba từ này có giá trị cảm xúc lần lượt là (-2), (-1) và (-2) nên giá trị cảm xúc của câu được tính như sau: $(-2) \cdot 1.5 + (-1) \cdot 1.5 + (-2) \cdot 1.5 = (-7.5)$.
- Đặc trưng số 10 là giá trị cảm xúc của câu thay đổi khi xét đến trường hợp từ phủ định thay đổi. Vì không có từ phủ định thay đổi nên giá trị cảm xúc trong câu vẫn giữ nguyên và bằng với giá trị của đặc trưng số 5.