

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

PHẠM NGUYỄN HÙNG
TRƯỜNG HOÀNG DIỄM HUYỀN

KHÓA LUẬN TỐT NGHIỆP
ỨNG DỤNG KHAI PHÁ QUAN ĐIỂM TRONG VIỆC
PHÂN LOẠI ĐÁNH GIÁ CỦA NGƯỜI DÙNG VỚI SẢN
PHẨM ĐIỆN TỬ

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2016

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

PHẠM NGUYỄN HÙNG – 12520167
TRƯƠNG HOÀNG ĐIỂM HUYỀN – 12520855

KHÓA LUẬN TỐT NGHIỆP
ỨNG DỤNG KHAI PHÁ QUAN ĐIỂM TRONG VIỆC
PHÂN LOẠI ĐÁNH GIÁ CỦA NGƯỜI DÙNG VỚI SẢN
PHẨM ĐIỆN TỬ

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

GIẢNG VIÊN HƯỚNG DẪN
PGS. TS. ĐỖ VĂN NHƠN

TP. HỒ CHÍ MINH, 2016

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. Nguyễn Hoàng Tú Anh – Chủ tịch.
2. ThS. Huỳnh Thị Thanh Thương – Thư ký.
3. TS. Ngô Thanh Hùng – Ủy viên.

MỤC LỤC

LỜI CẢM ƠN	1
MỞ ĐẦU	2
Chương 1. TỔNG QUAN VỀ ĐỀ TÀI.....	4
1.1. Đặt vấn đề.....	4
1.1.1. Thuật ngữ khai phá quan điểm	4
1.1.2. Nhu cầu của việc khai phá quan điểm trong đời sống.....	5
1.1.3. Ứng dụng của khai phá quan điểm lên các lĩnh vực khác nhau	5
1.1.4. Các khó khăn, thách thức trong lĩnh vực khai phá quan điểm	6
1.1.5. Những nghiên cứu, sản phẩm của khai phá quan điểm.....	7
1.2. Mục tiêu và giới hạn của đề tài.....	10
1.3. Ý nghĩa của đề tài	11
1.4. Nội dung thực hiện đề tài	11
Chương 2. CƠ SỞ LÝ THUYẾT	13
2.1. Các vấn đề trong khai phá quan điểm	13
2.1.1. Các mức độ trong việc khai phá quan điểm	13
2.1.2. Các loại câu quan điểm chính.....	14
2.1.3. Các thành phần cấu tạo của một câu quan điểm	14
2.2. Những kĩ thuật, phương pháp khai phá quan điểm đã được phát triển để phân loại quan điểm	16
2.2.1. Phân loại quan điểm dựa vào phương pháp phân lớp quan điểm	16
2.2.2. Phân loại quan điểm dựa vào cụm từ thể hiện quan điểm.....	17
2.2.3. Phương pháp tính điểm dựa vào hàm số	18
2.3. Kho ngữ liệu khai phá quan điểm.....	19

2.3.1.	Từ điển SentiWordNet	19
2.3.2.	Từ điển Negdic	21
2.3.3.	Từ điển thể hiện mức độ sắc thái.....	21
2.4.	Ontology	22
2.4.1.	Khái niệm Ontology	22
2.4.2.	Tính chất của Ontology	22
2.4.3.	Vai trò của Ontology	23
2.4.4.	Các thành phần chính của Ontology	24
2.4.5.	Phân loại Ontology	25
2.4.6.	Ngôn ngữ Ontology	26
2.5.	Đồ thị khái niệm	27
2.5.1.	Định nghĩa đồ thị khái niệm	27
2.5.2.	Loại, cá thể và tên.....	28
2.5.3.	Mô hình Đồ thị khái niệm cơ bản.....	29
2.5.4.	Đồ thị G	32
2.5.5.	Định mệnh đề trên Đồ thị khái niệm	33
2.5.6.	Các phép toán cơ bản trên Đồ thị khái niệm	34
Chương 3.	MÔ HÌNH VÀ GIẢI PHÁP	39
3.1.	Phát biểu bài toán	39
3.2.	Mô hình Ontology của ứng dụng.....	40
3.2.1.	Mô hình Ontology cho lĩnh vực Sản Phẩm Điện Tử.....	40
3.2.2.	Quy trình xây dựng Ontology	46
3.3.	Mô hình đồ thị khái niệm biểu diễn câu và vế câu trong một đánh giá của người dùng	56

3.3.1.	Mô hình đồ thị khái niệm biểu diễn cho mức vế câu	57
3.3.2.	Mô hình đồ thị khái niệm biểu diễn cho mức câu	61
3.4.	Mô hình đề xuất	62
3.4.1.	Pha kiểm tra và thu thập dữ liệu liên quan	64
3.4.2.	Pha phân tích và trích xuất quan điểm	67
3.4.3.	Pha tổng hợp kết quả	74
Chương 4.	CÀI ĐẶT VÀ THỬ NGHIỆM	78
4.1.	Ngôn ngữ sử dụng và công cụ hỗ trợ	78
4.1.1.	Ngôn ngữ sử dụng	78
4.1.2.	Công cụ hỗ trợ	79
4.2.	Tổ chức lưu trữ Ontology và dữ liệu mẫu trên máy tính	81
4.2.1.	Bảng Product	81
4.2.2.	Bảng Feature	81
4.2.3.	Bảng SentiWord	81
4.2.4.	Bảng DegreeWord	82
4.2.5.	Bảng DeniedWord	82
4.2.6.	Bảng ComparisionWord	82
4.2.7.	Bảng ReferWord	82
4.2.8.	Bảng Relation	83
4.2.9.	Bảng Rule	83
4.2.10.	Bảng Comment	83
4.3.	Giới thiệu giao diện chương trình ứng dụng	84
4.3.1.	Frame hiển thị kết quả phân tích và rút trích quan điểm của một bình luận	84

4.3.2.	Frame hiển thị nội dung của pha tổng hợp kết quả	86
4.4.	Thử nghiệm.....	89
4.4.1.	Pha phân tích một bình luận	89
4.4.2.	Phương pháp chọn mẫu thử.....	92
4.4.3.	Đánh giá kết quả thử nghiệm	92
Chương 5.	TỔNG KẾT	95
5.1.	Kết quả đạt được.....	95
5.2.	Hạn chế.....	96
5.3.	Hướng phát triển.....	96
TÀI LIỆU THAM KHẢO.....		97
PHỤ LỤC.....		99

DANH MỤC HÌNH VẼ

Hình 1.1: Giao diện trang Sentiment140 sau khi nhập từ khóa “Ipad”	8
Hình 1.2: Giao diện trang Tweet Sentiment Visualization sau khi nhập từ khóa “Ipad”	9
Hình 2.1: Bảng các nhãn từ loại của Penn Treebank	18
Hình 2.2: Một phần của file SentiWordNet	20
Hình 2.3: Các từ phủ định trong tiếng Việt và trọng số tương ứng	21
Hình 2.4: Ví dụ minh họa đồ thị biểu diễn quan hệ một ngôi.....	28
Hình 2.5: Ví dụ minh họa đồ thị biểu diễn quan hệ hai ngôi	28
Hình 2.6: Ví dụ minh họa đồ thị biểu diễn quan hệ ba ngôi	28
Hình 2.7: Ví dụ minh họa dùng biến để biểu diễn đồ thị.....	29
Hình 2.8: Sơ đồ phân cấp “Tứ giác”	31
Hình 2.9: Ví dụ minh họa đỉnh mệnh đề.....	34
Hình 2.10: Đồ thị A trong phép Copy.....	34
Hình 2.11: Đồ thị B trong phép Copy	34
Hình 2.12: Đồ thị X trong phép Join.....	35
Hình 2.13: Đồ thị Y trong phép Join.....	35
Hình 2.14: Đồ thị Z trong phép Join	36
Hình 2.15: Đồ thị X trong phép Simplify	37
Hình 2.16: Đồ thị Y trong phép Simplify	37
Hình 2.17: Đỉnh phủ định neg.....	37
Hình 2.18: Ví dụ về phép Not	37
Hình 2.19: Đỉnh quan hệ Or.....	38
Hình 2.20: Ví dụ về phép Or.....	38

Hình 3.1: Ví dụ minh họa tập hợp Rules của Ontology.....	46
Hình 3.2: Kết quả xây dựng tập Feature	49
Hình 3.3: Một mẫu của tập DegreeWord	53
Hình 3.4: Các luật cơ bản trong tập Rules	56
Hình 3.6: Ví dụ minh họa đồ thị về câu nhóm 1	58
Hình 3.7: Ví dụ minh họa đồ thị về câu nhóm 2.....	59
Hình 3.8: Ví dụ minh họa đồ thị về câu nhóm 3	59
Hình 3.9: Ví dụ minh họa đồ thị về câu nhóm 4.....	60
Hình 3.10: Ví dụ minh họa đồ thị về câu nhóm 5	60
Hình 3.11: Ví dụ minh họa đồ thị mức câu.....	61
Hình 3.12: Mô hình đề xuất	63
Hình 3.13: Pha kiểm tra và thu thập tài liệu liên quan.....	64
Hình 3.14: Kết quả tìm kiếm cho “Bphone” trên trang VnExpress.Net	65
Hình 3.15: Chuỗi kết quả rút trích bình luận	66
Hình 3.16: Mô hình pha phân tích và rút trích quan điểm	67
Hình 4.1: Giao diện Eclipse	79
Hình 4.2: Giao diện SQLite Expert Personal (version 3.5.78.2498)	80
Hình 4.3: Bảng Product.....	81
Hình 4.4: Bảng Feature	81
Hình 4.5: Bảng SentiWord.....	82
Hình 4.6: Bảng DegreeWord.....	82
Hình 4.7: Bảng DeniedWord.....	82
Hình 4.8: Bảng ComparisionWord	82
Hình 4.9: Bảng ReferWord	83

Hình 4.10: Bảng Relation.....	83
Hình 4.11: Bảng Rule.....	83
Hình 4.12: Bảng Comment	84
Hình 4.13: Giao diện Frame hiển thị kết quả phân tích một bình luận.....	84
Hình 4.14: JTextPane tpComment	85
Hiển thị nội dung nguyên thủy của câu bình luận.....	85
Hình 4.15: JList lSentence	85
Hình 4.16: JTextPane tpExplain	86
Hình 4.17: JPane pGraph	86
Hình 4.18: Giao diện Frame hiển thị nội dung của pha tổng hợp kết quả	87
Hình 4.19: JComboBox cbbProductName	87
Hình 4.20: JTextField tfProductName	88
Hình 4.21: JButton btnStart.....	88
Hình 4.22: JList lCmts	88
Hình 4.23: JTextPane tpStatement.....	89
Hình 4.24: Kết quả phân tách câu của một bình luận	90
Hình 4.25: Kết quả phân tích câu đầu tiên trong bình luận	92
Hình 4.26: Đồ thị của câu đầu tiên trong bình luận	92

DANH MỤC BẢNG

Bảng 2.1: Một ví dụ đầy đủ của 1 synnet bao gồm 6 cột.....	21
Bảng 2.2: Phân loại Ontology.....	26
Bảng 3.1: Tập hợp Concepts của Ontology lĩnh vực Sản Phẩm Điện Tử.....	43
Bảng 3.2: Tập hợp Relations của Ontology lĩnh vực Sản Phẩm Điện Tử.....	44
Bảng 3.3: Kết quả tập mẫu bình luận.....	47
Bảng 3.4: Ví dụ minh họa bước phân tách về câu.....	73
Bảng 4.1: Các thành phần của Frame hiển thị kết quả phân tích và rút trích quan điểm của một bình luận.....	85
Bảng 4.2: Các thành phần của Frame hiển thị nội dung của pha tổng hợp kết quả.....	87
Bảng 4.3: Kết quả đánh giá pha phân tích và rút trích “cụm quan điểm”.....	93
Bảng 4.4: Kết quả đánh giá sau khi tính trọng số và phân loại chiều hướng quan điểm	94

LỜI CẢM ƠN

Chúng em xin được gửi đến PGS. TS. Đỗ Văn Nhơn lời cảm ơn chân thành nhất. Nhờ sự dìu dắt, hướng dẫn tận tình và định hướng nghiên cứu của thầy trong suốt thời gian qua cũng như tài liệu khoa học quý báu mà thầy cung cấp đã giúp chúng em thuận lợi hoàn thành khóa luận này.

Chúng em cũng xin được bày tỏ lòng biết ơn chân thành tới các thầy cô trường Đại học Công Nghệ Thông Tin nói chung và các thầy cô, anh chị khoa Khoa Học Máy Tính nói riêng đã truyền đạt cho chúng em kiến thức và tài liệu nghiên cứu vô cùng quý báu để hình thành nên cơ sở vững chắc giúp chúng em thực hiện khóa luận này.

Chúng em cũng xin cảm ơn gia đình, bạn bè và tập thể lớp KHTN2012, những người đã ở bên khích lệ và động viên chúng em rất nhiều.

Cuối cùng, mặc dù đã cố gắng hoàn thiện chuyên đề với tất cả sự nỗ lực của bản thân nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Chúng em kính mong nhận được sự thông cảm và chỉ bảo của quý Thầy Cô và các bạn.

Chúng em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 6 năm 2016

Nhóm sinh viên thực hiện

Phạm Nguyên Hưng

Trương Hoàng Diễm Huyền

MỞ ĐẦU

Hiện nay, với sự bùng nổ công nghệ, các thiết bị cầm tay hay các máy tính cá nhân không còn quá xa lạ với hầu hết mọi người. Bên cạnh đó là sự phát triển của internet đã ra đời hàng loạt các mạng xã hội, diễn đàn, blog. Mọi người có thể dễ dàng thể hiện quan điểm cá nhân của mình bằng cách bình luận, đánh giá trên các mạng xã hội, diễn đàn. Có rất nhiều các chủ đề được thảo luận, từ các dịch vụ, sản phẩm cho đến các vấn đề về giải trí, văn hóa, chính trị.

Sự ra đời của nhiều trang web về các thiết bị công nghệ đã giúp cho người dùng có thể tìm hiểu, đánh giá, so sánh các thiết bị công nghệ. Trang báo điện tử của Việt Nam VnExpress.Net đã có riêng một mục “số hóa” dành cho các thiết bị công nghệ hiện đại, gắn liền với đời sống con người như: điện thoại, máy tính bảng, laptop... Ở đó có rất nhiều bài viết về các tính năng, giá cả, thông tin bên lề của các thiết bị đó. Bên cạnh mỗi bài viết còn có các chức năng nhận xét đánh giá để độc giả có thể để lại bình luận để thể hiện quan điểm của mình. Mỗi ngày xuất hiện rất nhiều bài viết về các loại thiết bị, hãng sản xuất khác nhau, mỗi bài viết có thể lên đến hàng chục, hàng trăm các bình luận đánh giá. Vậy với một lượng thông tin phong phú, dồi dào như vậy, làm sao chúng ta có thể thống kê, tổng hợp và kiểm soát được các quan điểm chứa trong những bình luận đó. Đây cũng vừa là một thách thức vừa là một khía cạnh mới mẻ trong lĩnh vực xử lý ngôn ngữ tự nhiên (nlp) nhằm đáp ứng nhu cầu khám phá thông tin ngày càng cao của con người trong thời đại kỹ thuật số hiện nay. Khái niệm khai phá quan điểm lần đầu tiên xuất hiện năm 2003 do 2 nhà nghiên cứu Nasukawa và Yi. Đó là việc nghiên cứu về việc đưa ra ý kiến của mọi người về một vấn đề, chúng ta cần xét xem quan điểm của ý kiến đó, có thể là: thích, không thích, tích cực, tiêu cực, trung lập hoặc thậm chí chỉ là câu bình luận vu vơ, không thể hiện rõ ràng quan điểm.

Mục đích của khóa luận là tìm hiểu, nghiên cứu về khai phá quan điểm của độc giả trên trang tin tức vnexpress.vn với miễn tri thức về các tin tức, bình luận về các loại điện thoại thông minh phổ biến hiện nay. Cài đặt thực nghiệm và thống kê độ chính xác trên miền dữ liệu đó.

Nội dung đồ án bao gồm:

- Tìm hiểu về khai phá quan điểm, các bài toán khai phá quan điểm trong tiếng Việt, tình hình nghiên cứu quan điểm trong và ngoài nước hiện nay.
- Xây dựng bộ từ điển thống kê và các quy tắc để tìm kiếm quan điểm dựa trên miền tri thức về các ý kiến, bình luận của độc giả trên trang VnExpress.Net về các loại điện thoại thông minh hiện nay.
- Xây dựng ứng dụng trực quan để đánh giá và kiểm chứng về tính chính xác, ưu điểm và hạn chế của việc phân tích trên.

Trong bài báo cáo ngoài phần Mở đầu và Tổng kết, bố cục của bài báo cáo sẽ bao gồm 4 chương chính sau:

Chương 1. Tổng quan đề tài: giới thiệu tổng quan về đề tài bao gồm việc ra đời của thuật ngữ khai phá quan điểm; nhu cầu của việc khai phá quan điểm trong các lĩnh vực trong cuộc sống; các khó khăn, thách thức trong việc khai phá quan điểm; các đề tài đã được nghiên cứu trong khai phá quan điểm và một số ứng dụng đã được tạo ra; trình bày tóm lược về mục tiêu của đề tài, ý nghĩa của đề tài và nội dung các bước thực hiện, tiếp cận đề tài.

Chương 2. Cơ sở lý thuyết: trình bày về cơ sở lý thuyết của đề tài bao gồm các vấn đề trong khai phá quan điểm; những kỹ thuật, phương pháp khai phá quan điểm đã được phát triển để phân loại quan điểm; kho ngữ liệu khai phá quan điểm, lý thuyết về Ontology và đồ thị khái niệm.

Chương 3. Mô hình và giải pháp: trình bày những nội dung bao gồm phát biểu bài toán, mô hình Ontology dành cho lĩnh vực Sản Phẩm Điện Tử, mô hình đồ thị khái niệm biểu diễn về câu và câu trong bình luận và đề xuất mô hình giải quyết yêu cầu đặt ra cho đề tài.

Chương 4. Cài đặt và thử nghiệm: trình bày những nội dung về công cụ hỗ trợ, ngôn ngữ sử dụng, giới thiệu giao diện ứng dụng, thử nghiệm và đánh giá.

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Đặt vấn đề

1.1.1. Thuật ngữ khai phá quan điểm

Khai phá quan điểm (sentiment analysis) hay còn được gọi là phân tích quan điểm (opinion mining) là một lĩnh vực nghiên cứu về ý kiến, quan điểm, đánh giá, thái độ và cảm xúc về những lĩnh vực cụ thể nào đó như là: sản phẩm, dịch vụ, vấn đề, sự kiện, ... Thuật ngữ khai phá quan điểm được xuất hiện vào năm 2003, thuật ngữ này có thể giải thích theo nhiều cách khác nhau, liên quan chặt chẽ với tìm kiếm và trích xuất thông tin. Đến năm 2006, Bin Liu, tác giả quyển sách nổi tiếng, 2012 [5] về lĩnh vực khai phá quan điểm đã định nghĩa lại một cách rõ ràng hơn về thuật ngữ khai phá quan điểm. Ông cho rằng quan điểm có thể là bất cứ điều gì về ý kiến của con người như: sản phẩm, dịch vụ, quan điểm chính trị, bầu cử, cá nhân, tổ chức, ... Ông gọi các thực thể được đánh giá là các đối tượng (object). Mỗi đối tượng này gồm có các thành phần (components) và tập các thuộc tính (attributes). Ví dụ: một chiếc điện thoại cũng có nhiều thành phần khác nhau, một cá nhân cũng bao gồm những khía cạnh như tính cách, chiều cao, cân nặng, ...

Làm về khai phá quan điểm là cố gắng làm cho máy tính có thể hiểu được một cách tự động các cảm xúc, quan điểm, ý kiến của con người được viết từ các văn bản bằng ngôn ngữ tự nhiên. Khai phá quan điểm dựa trên sự tính toán, truy vấn thông tin, khai phá tiên đoán, số hóa các văn bản dựa vào bộ dữ liệu được thống kê từ nhiều phương pháp khác nhau. Vì vậy việc khai phá quan điểm chịu ảnh hưởng rất nhiều về mặt ngôn ngữ, chữ viết, quan niệm của mỗi quốc gia. Mỗi văn bản là tập hợp các đoạn văn ngắn thể hiện một chủ đề nào đó và được viết bằng bất cứ ngôn ngữ nào, đây cũng là một khó khăn trong việc thống nhất về phương pháp, thống kê bộ từ điển quan điểm do mỗi ngôn ngữ có quá nhiều quy tắc và cấu trúc riêng.

1.1.2. Nhu cầu của việc khai phá quan điểm trong đời sống

Việc áp dụng nghiên cứu quan điểm bắt đầu từ năm 2003, tuy nhiên từ sớm đã có những nhà nghiên cứu về các quan điểm, ý kiến của con người [4, 7]. Người ta đã sớm nhận ra lợi ích to lớn của việc khai phá quan điểm và nguồn lợi khổng lồ từ kho dữ liệu to lớn trên internet. Nếu bạn là một người tiêu dùng, khi bạn đắn đo có muốn mua một món hàng hay không, bạn sẽ tìm kiếm những lời khuyên, ý kiến từ gia đình, bạn bè và quan trọng hơn hết là ý kiến đánh giá của những người đã từng mua món hàng đó trước bạn. Còn nếu bạn là một nhà buôn bán, một doanh nghiệp, khi bạn tung ra thị trường một sản phẩm nào đó, việc đầu tiên là phải tiến hành khảo sát, lấy ý kiến của người dùng, và khi đã tung sản phẩm đó ra thị trường, lúc này bạn cũng sẽ phải lấy ý kiến, quan điểm của người dùng về các khía cạnh khác nhau của sản phẩm để có những điều chỉnh thích hợp nhất. Nhu cầu thông tin về quan điểm của một cá nhân hay của một tập thể là vô cùng cao và xuất hiện trong rất nhiều khía cạnh khác nhau trong thực tế. Các quan điểm có ảnh hưởng rất lớn đến quyết định của chúng ta trong cuộc sống.

1.1.3. Ứng dụng của khai phá quan điểm lên các lĩnh vực khác nhau

Thực tế đã chứng minh, các ý kiến quan điểm có ảnh hưởng rất lớn đến chiến lược đầu tư, kinh doanh trên thị trường. Nó giúp định hình lại chiến lược kinh doanh sao cho phù hợp nhất với mỗi khu vực, địa điểm khác nhau trong những mốc thời gian khác nhau. Các ý kiến quan điểm đánh giá của người dùng có tác động mạnh mẽ đến hình dáng, chức năng sau này của những sản phẩm, các công ty, doanh nghiệp coi ý kiến người dùng là chìa khóa của sự thành công trong kinh doanh, vì sở thích, tính cách của họ chính là mấu chốt để cạnh tranh giữa những doanh nghiệp khác nhau. Khai phá quan điểm cũng trở nên cần thiết trong giao dục, nhất là ở các trường cao đẳng, đại học. Việc tự động rút trích thông tin, quan điểm trên các diễn đàn, website của các trường đại học về ý kiến phản hồi của sinh viên với nhà trường trở nên vô cùng cần thiết. Ngoài ra còn là việc lấy ý

kiến đánh giá của từng giảng viên về chính sách của ban giám hiệu nhà trường. Sức ảnh hưởng của khai phá quan điểm còn thể hiện rõ ở các lĩnh vực khác, trải dài từ kinh doanh, thương mại cho đến các dịch vụ chăm sóc sức khỏe, các cuộc tranh cử lấy ý kiến của cử tri, các cuộc khảo sát... Nguồn dữ liệu của khai phá quan điểm không chỉ gói gọn trong các trang web, diễn đàn của tổ chức, tập thể mà nó còn có thể ở nhiều nguồn khác nhau như: email, phản hồi, dữ liệu trong nội bộ các tổ chức.

1.1.4. Các khó khăn, thách thức trong lĩnh vực khai phá quan điểm

Đầu tiên, như đã nói, khó khăn của lĩnh vực khai phá quan điểm nằm ở vấn đề ngôn ngữ. Các quốc gia khác nhau có hệ thống ngôn ngữ khác nhau. Hệ thống đó bao gồm các vấn đề về chữ viết, ngữ pháp của từng loại ngôn ngữ. Do đó việc xây dựng, tìm hiểu một phương pháp chung, một kho từ điển quan điểm chung là vô cùng khó khăn. Có thể cách làm này phù hợp với hệ thống ngôn ngữ này tuy nhiên lại không phù hợp với hệ thống ngôn ngữ khác.

Vấn đề tiếp theo nằm ở chính bản thân mỗi từ, mỗi chữ của một ngôn ngữ cụ thể. Ví dụ trong tiếng Việt ta có hai câu quan điểm về laptop “giá cả của các sản phẩm của Apple thường khá cao”, “cấu hình của Dell Vostro 3450 khá là cao”. Cùng là một cụm từ quan điểm “khá cao” tuy nhiên ở câu đầu tiên là lời phàn nàn về mức giá, và có thể xem như đây là một ý kiến tiêu cực, còn ở câu thứ hai là lời khen của người dùng cho dòng laptop Dell Vostro 3450, và đây là một điểm cộng cho dòng laptop này đối với người dùng đó. Ngoài ra việc khai phá quan điểm còn phải đối mặt với những thách thức khác như vấn đề về từ lóng, từ địa phương trong bản thân mỗi hệ thống ngôn ngữ cụ thể.

Tiếp đó là sự phức tạp về các kiểu dùng từ khác nhau, các ý kiến trên những trang mạng thường là ý kiến của rất nhiều người và đến từ rất nhiều tầng lớp khác nhau, cho nên việc các ý kiến đó khi viết, họ thường là không tuân theo một chuẩn mực, quy tắc nhất định. Ngoài ra còn phải đối mặt với các lỗi chính tả, các biểu tượng cảm xúc (emotion) để thể hiện quan điểm. Ngoài ra các trang mạng xã

hội, mỗi một bình luận còn thường chứa các hashtag, đường link, điều này làm cho các ứng dụng phân tích quan điểm phải phân loại ra với các bình luận thông thường.

Ngoài ra các quan điểm không phải trực tiếp nói ra sản phẩm mà người phân tích quan tâm. Ví dụ trong một bài viết về dòng điện thoại Samsung có những nhận xét như: “khá tốt so với các máy chạy android hiện nay” hoặc là “so với Iphone còn thua nhiều lắm”. Những nhận xét so sánh với những sản phẩm khác khiến cho các thông tin bị nhiễu và rất dễ dẫn đến sai lệch trong việc đánh giá quan điểm.

1.1.5. Những nghiên cứu, sản phẩm của khai phá quan điểm

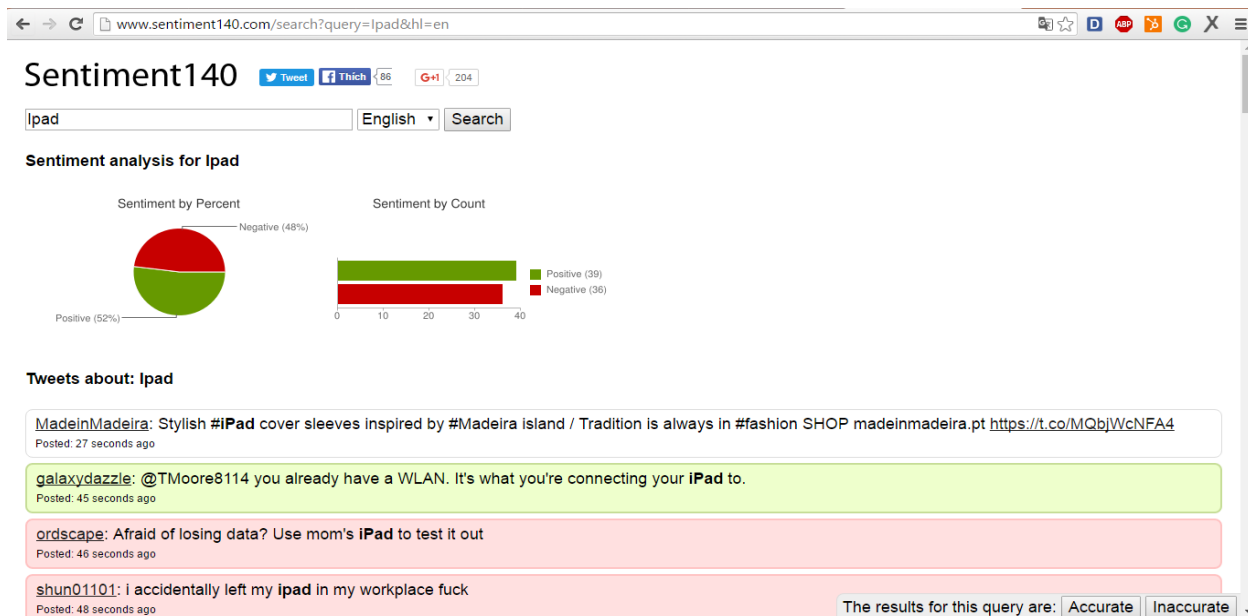
Hiện nay trên thế giới đã có nhiều các nghiên cứu liên quan đến khai phá quan điểm. Có cả những nghiên cứu về mặt lý thuyết dưới dạng các bài báo cáo khoa học cho đến những ứng dụng cụ thể đã được công khai trên những trang mạng và đưa vào sử dụng.

Về mặt sản phẩm ứng dụng hiện nay phải kể đến 2 trang web *Sentiment140* và *Tweet Sentiment Visualization*. Với việc hướng đến kho dữ liệu là các bình luận trên mạng xã hội Twitter, người dùng có thể nhập đầu vào là một thực thể dưới dạng từ khóa mà họ quan tâm, ví dụ: “Iphone”, “Ronaldo”, “Obama” ... hệ thống sẽ tổng hợp, tìm kiếm sau đó đưa ra phân tích và thống kê:

- Sentiment140 [3] trước kia được biết đến với cái tên “Twitter Sentiment” được tạo bởi 3 sinh viên đã tốt nghiệp đại học Stanford: Alec Go, Richa Bhayani và Lei Huang. Sentiment140 cho phép bạn khám phá quan điểm của nhãn hàng, sản phẩm hoặc là một chủ đề nào đó trên Twitter. Sentiment140 sử dụng việc phân lớp dựa trên các thuật toán máy học, đây là điều khác biệt so với việc sử dụng hướng tiếp cận dựa vào từ khóa truyền thống, điều này đã giúp đạt được độ chính xác cao hơn tuy nhiên gây ra việc xử lý khá chậm, ta có thể thấy thời gian chờ cho việc xử lý từ khóa khi ta nhập vào đến lúc xuất ra kết quả khá lâu, đây

cũng chính là nhược điểm chung của các hệ thống phân tích quan điểm dựa trên các thuật toán máy học.

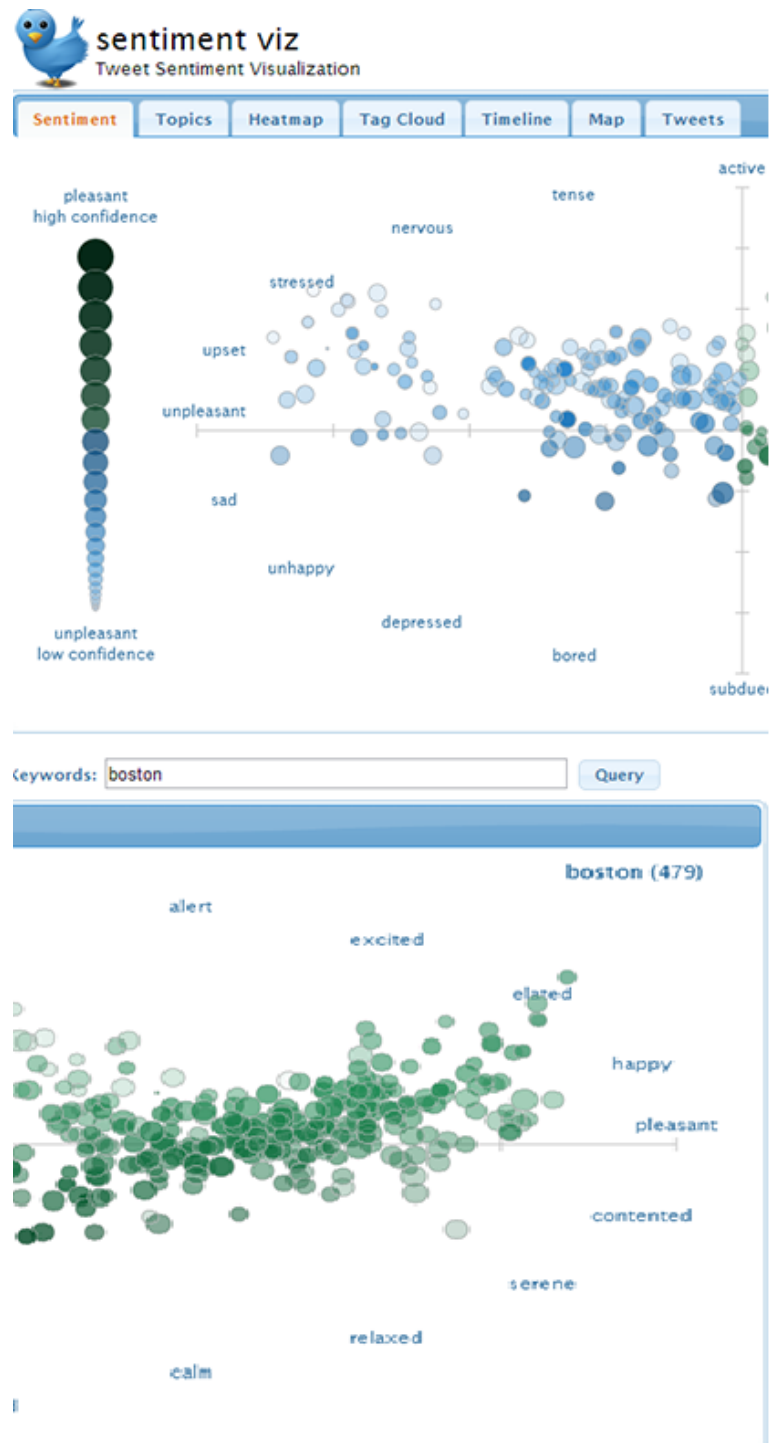
Ví dụ dưới đây là việc đánh giá của Sentiment140 dựa vào từ khóa nhập vào là “Ipad”:



Hình 1.1: Giao diện trang Sentiment140 sau khi nhập từ khóa “Ipad”

- Tweet Sentiment Visualization [8] là sản phẩm nghiên cứu của 2 tác giả Healey và Ramaswamy. Tweet Sentiment Visualization là trang web tổng hợp và theo dõi quan điểm của người dùng trên mạng xã hội Twitter. Không những vậy, trang web còn phân loại quan điểm theo nhiều mức, phân tích theo chủ đề và đưa ra được địa điểm và thời gian của những người đưa ra quan điểm.

Ví dụ dưới đây là việc đánh giá của Tweet Sentiment Visualization dựa vào từ khóa nhập vào là “Ipad”:



Hình 1.2: Giao diện trang Tweet Sentiment Visualization sau khi nhập từ khóa “Ipad”

Ngoài ra có một số bài báo khoa học, nghiên cứu về khai phá dữ liệu về nhiều lĩnh vực khác nhau. Trong lĩnh vực kinh doanh thương mại có công trình nghiên

cứ [9] của tác giả ...đã nghiên cứu về mô hình phân tích cảm xúc đã được đề xuất để dự đoán bán hàng. Ngoài ra còn một số công trình nghiên cứu liên quan đến giáo dục như: nghiên cứu về cảm xúc của miền tri thức là các phản hồi của sinh viên về việc giảng dạy [12] của tác giả, phân tích cảm xúc về đánh giá của sinh viên cho hệ thống bài giảng và phong cách giảng dạy [10] của tác giả.

Đối với tiếng Việt cũng đã có một số công trình nghiên cứu về khai phá quan điểm như công trình nghiên cứu về phân tích cảm xúc người dùng máy tính bàn và laptop [6] của 2 tác giả Kiều Thanh Bình và Phạm Bảo Sơn. Nghiên cứu về việc đánh giá, so sánh của người dùng cho các sản phẩm điện tử [11] của tác giả Nguyễn Thị Duyên.

1.2. Mục tiêu và giới hạn của đề tài

Sau khi cân nhắc về mặt thời gian và nhân lực, nhóm chúng em đã tự đặt ra mục tiêu trong đề tài lần này:

- Tìm hiểu cơ sở lý thuyết về các vấn đề cần giải quyết trong việc ứng dụng khai phá quan điểm vào phân loại đánh giá, bình luận của người dùng bằng ngôn ngữ tiếng Việt trên miền tri thức xét đến là lĩnh vực **Sản Phẩm Điện Tử**, cụ thể là về điện thoại di động
- Ứng dụng các cơ sở lý thuyết để có thể đưa ra được mô hình tổng thể nhằm giải quyết các bài toán liên quan đến việc rút trích quan điểm và xác định chiều hướng quan điểm cho các bình luận của người dùng cho sản phẩm điện thoại di động. Bên cạnh đó, việc rút trích quan điểm còn được phân loại theo tính năng của sản phẩm (pin, màn hình, giá cả, thiết kế, ứng dụng, camera, cấu hình)
- Nghiên cứu để xây dựng mô hình cơ sở tri thức dành riêng cho lĩnh vực quan tâm để áp dụng vào quá trình phân tích và xử lý câu.
- Xây dựng được ứng dụng cụ thể để hiện thực hóa mô hình đề xuất, từ đó kiểm tra lại tính chính xác và tính khả thi của mô hình.

1.3. Ý nghĩa của đề tài

Ý nghĩa thực tiễn của đề tài xuất phát từ nhu cầu sử dụng điện thoại thông minh hiện nay của con người. Hầu như việc mỗi một người hiện nay đề có thể có cho mình từ 1 tới 2 chiếc điện thoại thông minh là rất phổ biến. Cũng từ đó các hãng điện thoại có một cuộc chạy đua công nghệ, họ sẽ rất chú trọng đến những nhận xét đánh giá của người tiêu dùng về những sản phẩm điện thoại làm ra của họ. Việc các doanh nghiệp khảo sát và lấy ý kiến người dùng trước và sau khi tung ra sản phẩm không còn là điều mới mẻ trong thời đại công nghệ hiện nay. Và chính bản thân người tiêu dùng cũng muốn được nghe, được tổng hợp về các nhận xét của những người tiêu dùng khác để có thể lựa chọn cho mình một sản phẩm ưng ý nhất. Có người thì thích một chiếc điện thoại chụp hình đẹp, có người thì thích một chiếc điện thoại pin dùng được lâu... cho nên các khía cạnh cơ bản của một chiếc điện thoại luôn là yếu tố hàng đầu quyết định sự thành công và hữu ích của chiếc điện thoại đó.

1.4. Nội dung thực hiện đề tài

Trước tiên là tìm hiểu về cơ sở lý thuyết và các ứng dụng của khai phá quan điểm, tiếp cận được với một số vấn đề đang tồn tại cũng như các phương pháp cơ bản trong lĩnh vực khai phá quan điểm.

Tiếp theo nhóm chúng em muốn xây dựng được một ứng dụng trực quan để kiểm tra lại những cơ sở lý thuyết đã tìm hiểu được:

- Mẫu thử là những bình luận về các sản phẩm điện thoại di động trên trang báo điện tử VnExpress.Net. Các bình luận sẽ được phân loại một cách thủ công trước để tiện cho việc so sánh với kết quả mà ứng dụng đưa ra.
- Xây dựng được một hệ thống mà trong đó có các bước xử lý để tách từ, gán nhãn tự loại. Việc gán nhãn từ loại ở đây không đơn thuần là gán nhãn các loại thông thường (danh từ, động từ, tính từ, ...) mà sẽ thêm một số nhãn như: `feature_word` (từ chỉ khía cạnh người viết muốn nhận xét, ví dụ: “màn hình”, “pin”, ...) `senti_word` (từ chỉ quan điểm của người viết, ví dụ: “đẹp”,

“xấu”, “mượt mà”, ...), degree_word (từ chỉ mức độ trong một cụm quan điểm, ví dụ: “quá”, “rất”, “khá là”, ...), compare_word (từ dùng trong câu so sánh, ví dụ: “hơn”, “kém”, “thua”, “giống như”, ...), denied_word (từ phủ định), refer_word (từ liên kết, ví dụ: “về”, “có”, ...),...

- Xác định loại câu xem có chứa quan điểm không, nếu có chứa thì là quan điểm ẩn hay quan điểm hiện, xem câu đó có phải đang nói về chính sản phẩm cần phân tích không hay đang nói về sản phẩm khác, hoặc đang so sánh sản phẩm cần phân tích với sản phẩm khác.
- Phân tích các cụm từ quan điểm tìm được và tính toán trọng số cho cả cụm. Việc tính toán sẽ dựa vào nhiều yếu tố như: trọng số của từ quan điểm, các từ đi kèm trong cụm từ quan điểm, ... Sau khi đã có trọng số cho cụm từ quan điểm sẽ phân loại câu bình luận dựa theo 2 yếu tố: khía cạnh người dùng đề cập và mức độ hài lòng về khía cạnh đó. Từ đó thống kê cho cả đoạn tài liệu chứa nhiều câu bình luận.
- Xây dựng mô hình Ontology dành riêng cho lĩnh vực sản phẩm điện tử (cụ thể là điện thoại di động), dựa trên mô hình cơ bản của Ontology gồm các thành phần: Concepts, Relations và Rules.

Chương 2. CƠ SỞ LÝ THUYẾT

2.1. Các vấn đề trong khai phá quan điểm

2.1.1. Các mức độ trong việc khai phá quan điểm

Trong lĩnh vực khai phá quan điểm ba mức độ trong việc khai phá quan điểm:

- **Mức tài liệu:** nhiệm vụ của mức độ này là phân loại toàn bộ tài liệu chứa quan điểm xem đó là tích cực hay tiêu cực. Cho ví dụ, khi một đánh giá về sản phẩm được đưa ra, hệ thống sẽ xem xét liệu đánh giá đó có biểu lộ quan điểm tích cực hay tiêu cực về sản phẩm đó không. Nhiệm vụ này thường được biết đến như là việc phân loại ở mức tài liệu. Và ở mức độ này coi mỗi tài liệu biểu lộ một thực thể duy nhất. Vì vậy không thể áp dụng tài liệu này với nhiều loại thực thể khác nhau.
- **Mức câu:** nhiệm vụ của mức độ này là phân loại câu để xác định xem đây là câu thể hiện ý kiến tích cực hay tiêu cực hoặc cũng có thể là một ý kiến trung lập. Ý kiến trung lập thường có nghĩa là không có ý kiến, nhận định nào về sản phẩm đó.
- **Mức thực thể, khía cạnh:** cả hai mức độ tài liệu và mức độ câu không khám phá ra được chính xác những gì con người thích và không thích. Mức độ khía cạnh thực hiện việc phân tích một cách cụ thể và sâu sắc. Mức độ khía cạnh còn được gọi là mức độ tính năng. Thay vì nhìn vào các cấu trúc ngôn ngữ (tài liệu, đoạn văn, câu, mệnh đề, cụm từ), mức độ khía cạnh tiếp cận trực tiếp vào ý kiến của bản thân nó. Nó dựa vào nhận định về các ý kiến bao gồm cảm xúc (tích cực hay tiêu cực) và mục tiêu. Ta có ví dụ: “có thể cấu hình HP 450 chưa cao nhưng tôi vẫn rất hài lòng về nó”. Nếu nhìn vào đó ta thấy đây là một ý kiến tích cực tuy nhiên ta không thể kết luận nó hoàn toàn là ý kiến tích cực được, về mặt tổng quan người dùng đã hài lòng về laptop HP 450, tuy nhiên về mặt cấu hình, họ đã biểu lộ ý kiến chê cấu hình “chưa cao”. Trong nhiều ứng dụng cụ thể, việc tập

trung phân tích quan điểm còn chia ra nhiều khía cạnh khác nhau chứ không đơn thuần là một khía cạnh trong cùng một sản phẩm.

2.1.2. Các loại câu quan điểm chính

Các phân loại quan điểm chính bao gồm:

- Quan điểm thường: là ý kiến dựa trên thực thể đang xét, thường có 2 loại chính
 - + Quan điểm hiện (explicit opinion): trực tiếp đưa ra biểu lộ về sản phẩm hoặc khía cạnh cụ thể của sản phẩm. Ví dụ ta có nhận xét “nói về chụp hình chắc không loại nào qua nổi Samsung Galaxy Y ở phân khúc tầm trung”. Ở đây người viết đã nêu rõ về tính năng “camera” và của sản phẩm “Samsung Galaxy Y” .
 - + Quan điểm ẩn (implicit opinion): gián tiếp đưa ra biểu lộ về sản phẩm hoặc khía cạnh cụ thể của sản phẩm, người viết không ghi ra cụ thể nhưng người đọc có thể hiểu được khía cạnh mà người viết đang đề cập đến. Ví dụ ta có nhận xét “nhìn thẳng Dell này to quá, không hợp với tôi”. Khía cạnh được đề cập đến ở đây là “thiết kế”, tuy nhiên nó không được ghi trực tiếp ra trong nhận xét nhưng ta có thể hiểu được mục đích của người viết.
- Quan điểm so sánh: một quan điểm so sánh là một nhận xét biểu lộ một sự liên quan giống nhau hoặc khác nhau giữa hai hay nhiều thực thể. Ta có ví dụ: “Dùng Iphone thích thật, tôi thấy tốt hơn Samsung đó”. Bản thân các quan điểm so sánh này cũng có rất nhiều loại khác nhau và rất phức tạp, tùy thuộc vào hệ thống ngôn ngữ khác nhau.

2.1.3. Các thành phần cấu tạo của một câu quan điểm

Ta có thể hiểu quan điểm là một ý kiến, nhận định, thái độ, cảm xúc tích cực hoặc tiêu về một thực thể hay khía cạnh của một thực thể.

Một thực thể có thể là một sản phẩm, con người, sự kiện, tổ chức hoặc bất kì một chủ đề nào đó. Một quan điểm p thường được diễn tả bằng một bộ các thành phần:

$$p = (e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

Trong đó:

- e_i : là thực thể mục tiêu.
- a_{ij} : một khía cạnh của thực thể e_i
- s_{ijkl} : là giá trị quan điểm của người h_k trên khía cạnh a_{ij} của thực thể mục tiêu e_i . s_{ijkl} có thể là khẳng định, phủ định, tích cực, tiêu cực hoặc cũng có thể là chi tiết hơn nữa.
- h_k : là người đưa ra quan điểm p .
- t_l : thời gian đưa ra quan điểm.

Ở đây e được định nghĩa là một thực thể. Một thực thể có thể là: sản phẩm, dịch vụ, chủ đề, vấn đề, người tổ chức, sự kiện... Thực thể còn được mô tả $e(T, W)$, với T là hệ phân cấp của các thành phần, thành phần con.. W là tập các thuộc tính của e .

Cả bộ 5 ($e_j, a_{jk}, s_{ijkl}, h_i, t_l$) đều rất cần thiết đến sự chính xác và đưa ra nhận định của hệ thống. Bên cạnh 5 thành phần trên, người ta có thể bổ sung thêm các thành phần nếu thấy cần thiết cho quá trình phân tích. Để quá trình rút trích bộ 5 kể trên, người ta tiến hành đơn giản hóa thông tin. Tuy nhiên việc đơn giản hóa thông tin có thể dẫn đến việc mất mát thông tin không mong muốn trong quá trình thực thi. Mặc dù vậy, việc đơn giản hóa thông tin là cần thiết và việc chấp nhận sự mất mát không quá nhiều trong ứng dụng thực tế là điều khó tránh khỏi.

Ví dụ ta có đoạn thông tin được viết bằng ngôn ngữ tự nhiên:

“Trong quá trình công tác, tôi đã quyết định mua Iphone. Nó khá đẹp, kiểu dáng sang trọng và phù hợp với doanh nhân như tôi. Pin của nó lâu hơn pin của

con Samsung Z3 của tôi nhiều lắm. Tuy nhiên giá cả quá cao của Iphone là điều đã khiến tôi khá băn khoăn” - Tâm Nguyên 30/5/2016.

Quan điểm của đoạn thông tin trên được đơn giản hóa theo bộ 5 sau đây:

- (Iphone, thiết kế, positive, Tâm Nguyên, 30/5/2016)
- (Iphone, pin, positive, Tâm Nguyên, 30/5/2016)
- (Samsung Z3, pin, negative, Tâm Nguyên, 30/5/2016)
- (Iphone, giá cả, negative, Tâm Nguyên, 30/5/2016)

Một tài liệu sau khi đã được chuẩn hóa thành các bộ 5 như trên để có thể giải quyết vấn đề đơn giản hơn. Biến dữ liệu từ một tập văn bản truyền thống bằng ngôn ngữ tự nhiên, không có cấu trúc thành một tập tài liệu có cấu trúc. Bộ năm được rút trên là một cấu trúc dữ liệu cơ bản, đã được sắp xếp thứ tự, làm tiền đề để có thể đưa vào cấu trúc dữ liệu bảng truyền thống. Từ đó rút trích ra được xu hướng quan điểm dựa vào các câu truy vấn hoặc bằng bất cứ phương pháp nào.

2.2. Những kĩ thuật, phương pháp khai phá quan điểm đã được phát triển để phân loại quan điểm

Có 3 phương pháp chính để phân loại quan điểm:

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

2.2.1. Phân loại quan điểm dựa vào phương pháp phân lớp quan điểm

Đây là phương pháp đơn giản nhất để giải quyết các bài toán phân lớp quan điểm dựa vào chủ đề. Phân loại quan điểm được hiểu là một chương trình học có giám sát với hai lớp gán nhãn (tích cực và tiêu cực). Dữ liệu huấn luyện và kiểm tra đã sử dụng để nghiên cứu chủ yếu nhất là các đánh giá về sản phẩm. Từng quan điểm sẽ được gán nhãn phân loại (ví dụ 1-5 sao), những quan điểm với 4-5 sao là có chiều hướng tích cực và với những quan điểm với 1-2 sao là có chiều hướng tiêu cực. Phân loại quan điểm cũng có một số điểm giống và khác với

phân loại văn bản các chủ đề cơ bản được định nghĩa một số lớp chủ đề, chính trị, khoa học, thể thao, ... Trong phân loại theo chủ đề những từ liên quan đến chủ đề là rất quan trọng, trong phân loại quan điểm những từ liên quan đến chủ đề là không được quan tâm. Phân loại quan điểm chỉ tập trung vào những từ quan điểm để quyết định xem đánh giá đó tích cực hay tiêu cực. Ví dụ: *tốt, xấu, ngạc nhiên, kinh khủng, tồi tệ, ...*

Một số phương pháp phân loại dựa trên học có giám sát thường được áp dụng trong phân loại quan điểm là *naïve Bayesian*, *support vector machine (SVM)*. *Pang et al* đã áp dụng phương pháp để phân loại quan điểm phim thành hai lớp, tích cực và tiêu cực. Để áp dụng tốt *naïve Bayesian* và *support vector machine* thường sử dụng các từ, cụm từ như những đặc trưng trong phân loại. Những đánh giá trung lập không được sử dụng trong phần này, điều đó giúp cho chương trình trở lên đơn giản hơn.

2.2.2. Phân loại quan điểm dựa vào cụm từ thể hiện quan điểm

Kỹ thuật này sử dụng một phương pháp phổ biến trong xử lý ngôn ngữ tự nhiên, đó là tách từ và gán nhãn từ loại (part-of-speech), phân loại, gán nhãn một từ được đánh giá theo ngữ nghĩa của nó trong câu. Các loại nhãn từ loại dùng cho tiếng Anh như: danh từ, động từ, tính từ, trạng từ, đại từ, giới từ, ..

Number	Tag	Description	Number	Tag	Description
1.	CC	Coordinating conjunction	21.	RBR	Adverb, comparative
2.	CD	Cardinal number	22.	RBS	Adverb, superlative
3.	DT	Determiner	23.	RP	Particle
4.	EX	Existential <i>there</i>	24.	SYM	Symbol
5.	FW	Foreign word	25.	TO	<i>to</i>
6.	IN	Preposition or subordinating conjunction	26.	UH	Interjection
7.	JJ	Adjective	27.	VB	Verb, base form
8.	JJR	Adjective, comparative	28.	VBD	Verb, past tense
9.	JJS	Adjective, superlative	29.	VBG	Verb, gerund or present participle
10.	LS	List item marker	30.	VBN	Verb, past participle
11.	MD	Modal	31.	VBP	Verb, non-3rd person singular present
12.	NN	Noun, singular or mass	32.	VBZ	Verb, 3rd person singular present
13.	NNS	Noun, plural	33.	WDT	Wh-determiner
14.	NNP	Proper noun, singular	34.	WP	Wh-pronoun
15.	NNPS	Proper noun, plural	35.	WP\$	Possessive wh-pronoun
16.	PDT	Predeterminer	36.	WRB	Wh-adverb
17.	POS	Possessive ending			
18.	PRP	Personal pronoun			
19.	PRP\$	Possessive pronoun			
20.	RB	Adverb			

Hình 2.1: Bảng các nhãn từ loại của Penn Treebank

2.2.3. Phương pháp tính điểm dựa vào hàm số

Phương pháp phân lớp dựa vào hàm tính điểm số được Kushal Dave và cộng sự [8] đưa ra gồm 2 bước:

Bước 1: Tính điểm các từ trong văn bản của tập dữ liệu học theo công thức

$$score(t_i) = \frac{\Pr(t_i/C) - \Pr(t_i/C')}{\Pr(t_i/C) + \Pr(t_i/C')} \quad (1.4)$$

Trong đó:

- t_i : Từ cần được tính điểm.
- C : Một lớp quan điểm

- C' : Lớp phản bù của C (not C).
- $\mathbf{Pr}(t/C)$: Xác suất t xuất hiện ở lớp C , được tính bằng số lần xuất hiện của t trong lớp C .
- Điểm số chuẩn hóa trong khoảng $[-1, 1]$.

Bước 2: Một văn bản mới $d_i = t_1 \dots t_n$ sẽ được phân lớp theo công thức :

$$\mathbf{class}(d_i) = \begin{cases} C \text{ eval}(d_i) \\ C' \text{ eval}(d_i) \end{cases} \quad (1.5)$$

Với $\mathbf{eval}(d_i) = \sum_j \mathbf{score}(t_j)$

2.3. Kho ngữ liệu khai phá quan điểm

Việc thực hiện được bài toán khai phá quan điểm đặt ra, yêu cầu về một kho ngữ liệu thống kê (hay còn gọi là các từ điển quan điểm) chứa các từ quan điểm là cần thiết. Hiện nay cũng có một số kho ngữ liệu được biết đến như là WordNet, SentiWordNet cho một vài tiếng như: tiếng Anh, tiếng Ấn Độ... SentiNetWord được sử dụng vào khai phá quan điểm trên nhiều lĩnh vực khác nhau. Các nghiên cứu cũng đã cho thấy được khả năng ứng dụng của SentiNetWord là rất lớn và cần thiết.

2.3.1. Từ điển SentiWordNet

Các khái niệm trong một file SentiWordNet:

- Synset: là một dòng trong từ điển, được cấu tạo bởi 6 cột bao gồm POS, ID, PosScore, NegScore, SynsetTerms và các thành phần này cách nhau bởi 1 <Tab>.
 - + POS: từ loại của từ
 - + ID: mã số đại diện cho từ và là duy nhất cho một từ
 - + PosScore: trong số thể hiện sự tích cực của từ
 - + NegScore: trọng số thể hiện sự tiêu cực của từ
 - + SynsetTerms: là tập các từ đồng nghĩa trong 1 synnet

- Term: là những từ quan điểm trong synnet, một synnet có thể chứa nhiều tern khác nhau, tùy thuộc vào ngữ cảnh sử dụng từ đó. Chính vì chỉ giống nhau về hình thức nhưng ý nghĩa khác nhau nên cùng 1 term các trọng số về PosScore và NegScore cũng khác nhau. Trong SentiWordNet, các term sẽ được gán một nhãn riêng để phân biệt các từ đồng nghĩa với nhau. Ví dụ ta có từ beautiful có 2 ngữ cảnh beautiful#1 và beautiful#2. Ta có beautiful#1 có trọng số PosScore/NegScore là 0.75/0 và beautiful#2 có trọng số PosScore/NegScore là 0.625/0.
- Gloss: là cột giải nghĩa của 1 từ, ngoài ra còn ghi chú các ngữ cảnh sử dụng từ.

POS	ID	PosS	NegS	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to`) having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"

Hình 2.2: Một phần của file SentiWordNet

POS	ID	PosS core	NegS core	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to`) having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at

					last able to buy a car"; "able to get a grant for the project"
--	--	--	--	--	--

Bảng 2.1: Một ví dụ đầy đủ của 1 synnet bao gồm 6 cột

2.3.2. Từ điển Negdic

Là từ điển chứa các từ phủ định, trong tiếng Việt có một hệ thống các từ có ý nghĩa phủ định. Cấu trúc của từ điển Negdic gồm 2 phần: từ phủ định và trọng số.

word	score
define a filter	
k	(null)
ko	(null)
không	(null)
chẳng	(null)
chả	(null)
mất	(null)
cần	(null)
còn	(null)
phải chi	(null)
khỏi	(null)
già mà	(null)
thiếu	(null)

Hình 2.3: Các từ phủ định trong tiếng Việt và trọng số tương ứng

2.3.3. Từ điển thể hiện mức độ sắc thái:

Là loại từ điển chứa các từ gần như tương đương với trạng từ trong tiếng Anh, có tác dụng nhấn mạnh các sắc thái khác nhau của một tính từ hoặc động từ. Trong tiếng Việt có một số từ như vậy, ví dụ như: rất, lắm, quá...

Tuy nhiên cần phải nhấn mạnh rằng việc đánh giá trọng số cho các từ thể hiện sắc thái, quan điểm phụ thuộc vào cảm nhận và ý kiến chủ quan của người xây dựng từ điển. Hiện nay có một số phương pháp để tăng tính chính xác của trọng

số các từ sắc thái lên tuy nhiên vì giới hạn thời gian nên nhóm chúng em chưa thể thực hiện được, nếu tiếp tục phát triển khóa luận thì nhóm chúng em sẽ tập trung vào phần này để tăng độ chính xác của trọng số, từ đó có thể đưa ra các đánh giá cho mức độ câu, mức độ tài liệu chính xác hơn.

2.4. Ontology

2.4.1. Khái niệm Ontology

– Trong Triết học

Theo Aristelo – tác giả của cuốn *Metaphysics* (Siêu Hình) – định nghĩa : “ Ontology là một siêu hình học nghiên cứu về sự tồn tại và bản chất các sự vật trong thực tế ”. Ngoài ra, Ontology là một thuật ngữ có nguồn gốc từ triết học, tạm dịch là “ Bản thể học ” – bộ môn khoa học về nhận thức, cụ thể hơn là một nhánh của siêu hình học, nghiên cứu về tự nhiên và bản chất của thế giới, nhằm xem xét các vấn đề về sự tồn tại hay không tồn tại của các sự vật.

– Trong lĩnh vực Trí tuệ nhân tạo

Đối với lĩnh vực này có rất nhiều định nghĩa về Ontology, mỗi định nghĩa lại có cách nhìn nhận khác nhau và một phương pháp xây dựng Ontology riêng nhưng nhìn chung có thể mô tả khái quát như sau: “ Mỗi Ontology xác định một bộ từ vựng chung mô tả thông tin về một lĩnh vực cần chia sẻ, bao gồm định nghĩa các khái niệm cơ bản và sự liên quan giữa chúng giúp máy tính có thể hiểu được và dễ dàng xử lý ”. Đây là khái niệm được sử dụng trong đề tài này vì tính rõ ràng và minh bạch của nó.

Ngoài bộ từ vựng, Ontology còn cung cấp các ràng buộc, đôi khi các ràng buộc này cũng được coi như là các giả định cơ sở về định nghĩa mong muốn của bộ từ vựng, nó được sử dụng trong một lĩnh vực mà có thể được giao tiếp giữa người và các hệ thống ứng dụng phân tán khác.

2.4.2. Tính chất của Ontology

Một Ontology phải thỏa mãn các tính chất sau:

- Được sử dụng để mô tả một lĩnh vực ứng dụng cụ thể
- Các khái niệm và quan hệ phải được định nghĩa một cách rõ ràng trong phạm vi lĩnh vực đó
- Có cơ chế tổ chức khái niệm, thường là cơ chế phân cấp
- Có sự đồng thuận về mặt ý nghĩa của các khái niệm giữa những người sử dụng Ontology

2.4.3. Vai trò của Ontology

Ontology là một cơ chế nhằm cung cấp các chức năng sau:

- Chia sẻ sự hiểu biết chung giữa các ứng dụng và con người, hiểu biết về cấu trúc thông tin giữa con người và các tác tử.
- Cho phép tái sử dụng tri thức. Ví dụ, nếu một nhóm chúng em đã xây dựng Ontology về Hình Học, nhóm chúng em khác có thể sử dụng lại nó cho ứng dụng của họ.
- Làm rõ cho lĩnh vực cần quan tâm, đưa ra các giả thiết rõ ràng về miền: tạo điều kiện thay đổi khi tri thức về lĩnh vực thay đổi, các đặc tả rõ ràng về miền tri thức sẽ giúp cho người mới nghiên cứu về nó có thể dễ dàng tìm hiểu ngữ nghĩa của các từ trong lĩnh vực quan tâm
- Phân tách tri thức lĩnh vực với tri thức xử lý. Ví dụ, ta có tác vụ tạo một đối tượng Tam giác từ nhiều thành phần theo mô hình đặc tả cho Ontology Hình Học thì độc lập với chương trình ứng dụng làm nhiệm vụ này.
- Phân tích tri thức: Phân tích hình thức của các khái niệm, cần thiết cho việc tái sử dụng và mở rộng Ontology. Muốn kế thừa hay sử dụng một Ontology ta cần phải phân tích và tìm hiểu các khái niệm cũng như quan hệ giữa chúng trong Ontology đó.

Ontology cung cấp nguồn thông tin giàu ngữ nghĩa giúp cho các hệ thống thực hiện các tác vụ với kết quả tốt hơn. Nó không những phục vụ cho nhu cầu chia sẻ tri thức đơn thuần mà còn được áp dụng vào nhiều lĩnh vực khác nhau. Chính vì thế mà hiện nay Ontology được ứng dụng rộng rãi trong lĩnh vực Trí

tuệ nhân tạo, công nghệ Web ngữ nghĩa, các hệ thống kỹ thuật, kỹ thuật phần mềm, tin học y sinh và kiến trúc thông tin với vai trò là hình thức biểu diễn tri thức về thế giới hoặc một số lĩnh vực cụ thể.

2.4.4. Các thành phần chính của Ontology

Một Ontology bao gồm các thành phần sau:

- **Khái niệm – Concept:** Có nhiệm vụ mô tả các khái niệm trong miền lĩnh vực đang xét. Những khái niệm này được tổ chức phân loại để định nghĩa tập hợp thuộc tính (Property, Role hay Slot) và tập hợp các thao tác đặc trưng của bất kỳ thành phần nào của khái niệm. Các lớp thường được tổ chức phân cấp và áp dụng kỹ thuật thừa kế. Một lớp có thể có nhiều lớp con biểu diễn khái niệm cụ thể hơn so với lớp cha.
- **Quan hệ - Relation:** Là kiểu tương tác giữa cái khái niệm hay biểu diễn các kiểu quan hệ giữa các khái niệm. Các quan hệ nhị phân được sử dụng để biểu diễn thuộc tính. Tuy nhiên, giá trị của quan hệ khác với giá trị của thuộc tính ở chỗ giá trị của quan hệ là một khái niệm.
- **Hàm - Function:** Là các thao tác thực hiện trên Ontology. Nói cách khác, hàm là một loại thuộc tính hay quan hệ đặc biệt, trong đó, phần tử thứ n là duy nhất đối với n-1 phần tử còn lại.
- **Tiên đề - Axioms:** Biểu diễn các phát biểu luôn đúng mà không cần phải chứng minh hay giải thích. Tiên đề có thể được phân tích thành các luật thể hiện tri thức mang tính phổ quát trên các khái niệm hay các loại sự kiện khác nhau. Từ sự kiện đã có, mỗi luật cho ta một qui tắc suy luận để suy ra được các sự kiện mới. Cấu trúc của một luật bao gồm hai phần chính: phần giả thiết và phần kết luận (đều là tập hợp các sự kiện trên các đối tượng nhất định) . Mỗi luật được mô hình dưới dạng sau:

$$r: \{sk_1, sk_2, sk_3, ..., sk_n\} \Rightarrow \{sk_1, sk_2, sk_3, ..., sk_m\}$$

Các tiên đề được sử dụng để kiểm chứng sự nhất quán của Ontology. Cả hai thành phần là hàm và tiên đề đều góp phần tạo nên khả năng suy diễn trên Ontology.

- **Thể hiện - Instance:** Đại diện cho các phần tử riêng biệt của khái niệm (các thể hiện của lớp) hay các quan hệ. Mỗi thể hiện của lớp biểu diễn một sự cụ thể hóa của khái niệm đó.

2.4.5. Phân loại Ontology

Theo cách phân loại của D. Fensel thì Ontology gồm 7 loại sau:

Tên	Mô Tả	Đại Diện
Knowledge Representation Ontology (Ontology biểu diễn tri thức)	Dựa trên các cách biểu diễn tri thức truyền thống	Frame – Ontology, ...
General/Comon Ontology (Ontology tổng quát)	Bao gồm từ vựng liên quan tới sự vật, hiện tượng, thời gian, không gian, quan hệ nhân quả ... mang ý nghĩa chung chung, không chỉ riêng một lĩnh vực nào	WordNet, CYC, Ontology về bảng chuyển đổi giữa Meter và Inch, ...
Metadata Ontology	Định nghĩa các Ontology. Cung cấp bộ từ vựng dùng để mô tả nội dung của các nguồn thông tin trực tuyến	Registry Ontology (dùng để quản lý các Ontology khác), Dublin Core, ...
Domain Ontology (Ontology lĩnh vực)	Cung cấp từ vựng về các khái niệm và quan hệ giữa chúng trong một phạm vi lĩnh vực nào đó mà ta có thể tái sử dụng	Ontology về lý thuyết của một miền tri thức nào đó, Gene Ontology (Ontology về sinh học), ...

Task Ontology (Ontology tác vụ)	Cung cấp hệ thống các từ vựng của các thuật ngữ để giải quyết các vấn đề kết hợp, liên quan đến nhiệm vụ mà có thể cùng hoặc không cùng một phạm vi ứng dụng	Ontology về phân công công việc
Domain – Task Ontology (Ontology Lĩnh vực – Tác vụ)	Các Task Ontology được sử dụng lại trong một phạm vi ứng dụng cụ thể	Ontology về phân công công việc trong công ty
Application Ontology (Ontology Ứng dụng)	Bao gồm các tri thức cần thiết của một ứng dụng trong một lĩnh vực cụ thể	Ontology Giải tích, Điện xoay chiều, ...

Bảng 2.2: Phân loại Ontology

2.4.6. Ngôn ngữ Ontology

Ngôn ngữ Ontology là ngôn ngữ dùng để mô tả các thành phần của Ontology. Gồm 2 loại ngôn ngữ chung nhất:

- **Ngôn ngữ tự nhiên:** Mọi thứ của con người đều có thể được biểu diễn bằng một ngôn ngữ tự nhiên. Trong Ngôn ngữ học, một **ngôn ngữ tự nhiên** là bất kỳ ngôn ngữ nào phát sinh, không suy nghĩ trước trong não bộ của con người. Nó là một siêu ngôn ngữ có thể diễn đạt các ngôn ngữ khác như: ký số toán học, ngôn ngữ lập trình, ...
- **Ngôn ngữ dựa trên logic:** Mọi thứ được phát biểu một cách chính xác và thực sự rõ ràng. Bất cứ ngôn ngữ tự nhiên nào, nếu có thể nắm bắt được ý nghĩa của chúng một cách rõ ràng đều có thể biểu diễn được trong logic. Bất kỳ thứ gì có thể thực hiện trên máy tính bằng ngôn ngữ lập trình đều có thể biểu diễn được trong logic.

Các Ontology thường dùng ngôn ngữ dựa trên logic để biểu diễn các khái niệm.

Các ngôn ngữ thường được dùng để biểu diễn Ontology : Conceptual Graphs (CGs) –đồ thị khái niệm, Knowledge Interchange Format (KIF), Conceptual Graph Interchange Form (CGIF) là mức trung gian giữa CGs và KIF

Các ngôn ngữ biểu diễn Web ngữ nghĩa : XML, XML Schema, RDF, RDF Schema, OWL, OIL, SML + OIL .

2.5. Đồ thị khái niệm

2.5.1. Định nghĩa đồ thị khái niệm

Đồ thị khái niệm (Conceptual Graph) là một đồ thị hữu hạn, liên thông; các đỉnh được chia làm 2 loại đỉnh như sau:

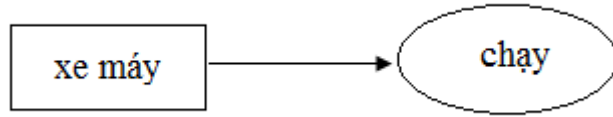
- Đỉnh khái niệm: Có đặc điểm là
 - + Biểu diễn các khái niệm cụ thể. Ví dụ: điện thoại, xe tải, cái nhà, ...
 - + Biểu diễn các khái niệm trừu tượng. Ví dụ: tình yêu, vẻ đẹp, ...
 - + Đỉnh loại này được biểu diễn bởi hình chữ nhật, có gán nhãn là khái niệm.
- Đỉnh quan hệ: Có đặc điểm là
 - + Biểu diễn quan hệ giữa các khái niệm có cung nối đến nó. Ví dụ: các mối quan hệ trong gia đình như cha mẹ và con cái, ông bà và cháu, ...
 - + Đỉnh loại này được biểu diễn bởi hình oval có gán nhãn quan hệ.
 - + Có thể là quan hệ một ngôi hay nhiều ngôi.

Các lưu ý về Đồ thị khái niệm:

- Chỉ có các đỉnh khác loại mới nối được với nhau (đỉnh khái niệm nối với đỉnh quan hệ). Dùng đỉnh quan hệ thì các cung không cần phải gán nhãn nữa.
- Mỗi đồ thị khái niệm biểu diễn một mệnh đề đơn.
- Tập hợp gồm các Đồ thị khái niệm tạo nên Cơ Sở Tri Thức.

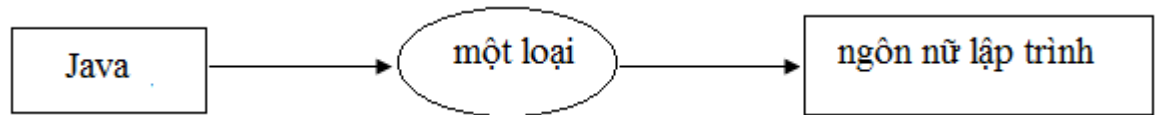
Ví dụ: Ta có các Đồ thị khái niệm sau:

- Biểu diễn loại quan hệ một ngôi:



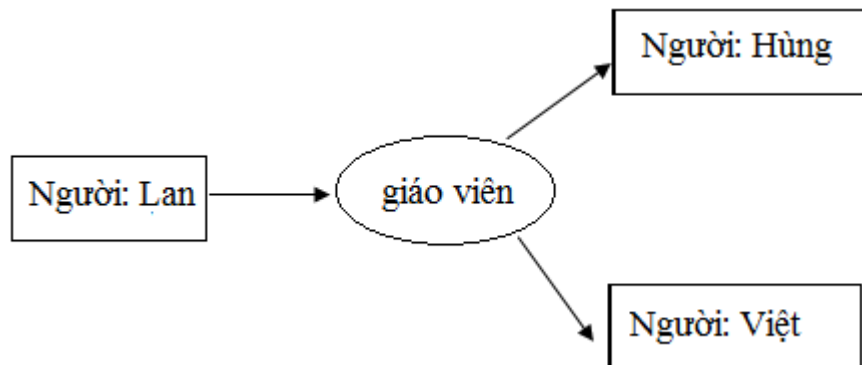
Hình 2.4: Ví dụ minh họa đồ thị biểu diễn quan hệ một ngôi

- Biểu diễn loại quan hệ hai ngôi:



Hình 2.5: Ví dụ minh họa đồ thị biểu diễn quan hệ hai ngôi

- Biểu diễn loại quan hệ ba ngôi:



Hình 2.6: Ví dụ minh họa đồ thị biểu diễn quan hệ ba ngôi

2.5.2. Loại, cá thể và tên

Để biểu diễn quan hệ giữa “loại” và “cá thể” thì mỗi đỉnh khái niệm được gán nhãn:

<loại>:<tên cá thể>

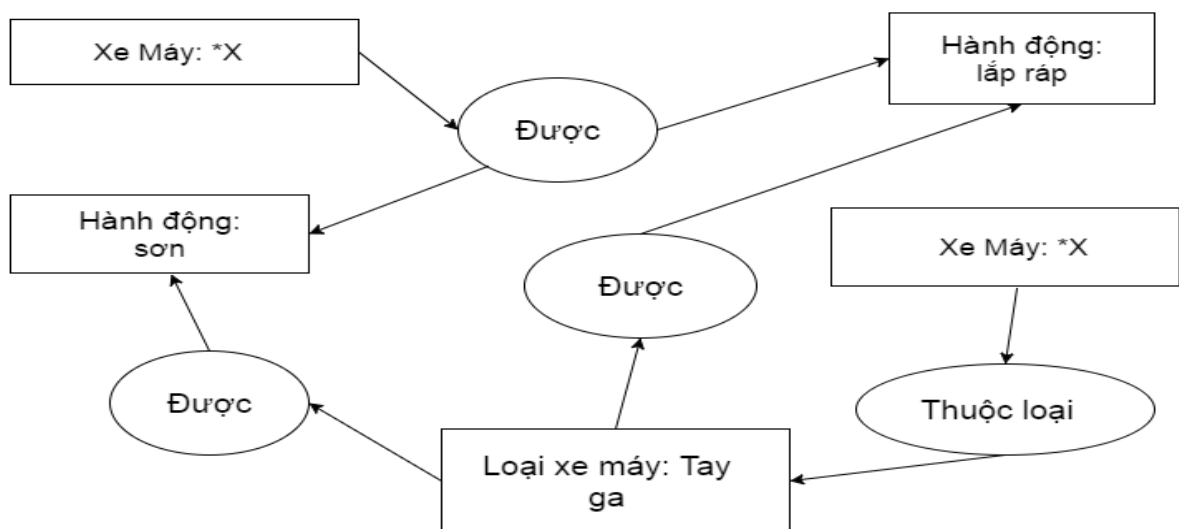
Tên cá thể có thể chia thành các loại sau:

- Một tên riêng nào đó. Ví dụ: Học sinh: Nam hoặc Giáo viên: Lan.
- Một khóa phân biệt, theo cú pháp #<khóa>. Ví dụ: Xe máy: #48478847.

- Một cá thể chưa xác định và được đại diện bởi dấu *. Trường hợp này, khái niệm được gọi là khái niệm tổng quát, còn 2 trường hợp ở trên được gọi là khái niệm cá thể. Ví dụ: Chuyến bay:* hay Sinh viên: T*, suy ra sinh viên này có tên bắt đầu bởi chữ “T”.

Ngoài ra, biến cũng có thể được dùng khi cần chỉ ra nhiều đỉnh khái niệm có tính đồng nhất với nhau.

Ví dụ: Đồ thị sau mô tả câu “Một chiếc xe tay ga đang được sơn và lắp ráp”.



Hình 2.7: Ví dụ minh họa dùng biến để biểu diễn đồ thị

2.5.3. Mô hình Đồ thị khái niệm cơ bản

Để tạo điều kiện thích hợp cho máy tính có thể tổ chức, lưu trữ và xử lý tri thức thì ứng với mỗi phương pháp biểu diễn tri thức phải có một mô hình tổ chức riêng, nhằm mã hóa tri thức ở dạng thức mà máy có thể hiểu được. Vậy mô hình cơ bản của một đồ thị khái niệm gồm những thành phần sau:

(V, G)

Trong đó:

- $V = (T_C, T_R, I)$: Bộ từ vựng của miền tri thức đang xét.
- $G = (C, R, E, I)$: Đồ thị biểu diễn các đỉnh khái niệm, đỉnh quan hệ và các cung nối đỉnh khái niệm với đỉnh quan hệ.

2.5.3.1. Bộ từ vựng V

Mô hình biểu diễn bộ từ vựng V như sau:

$$V = (T_C, T_R, I)$$

Trong đó:

- T_C : Tập các concept type (loại khái niệm);
- T_R : Tập các relation symbol (nhãn quan hệ) ;
- I : Tập các individual maker (đánh dấu cá thể hay tên cá thể);

Có những lưu ý về V cần biết là:

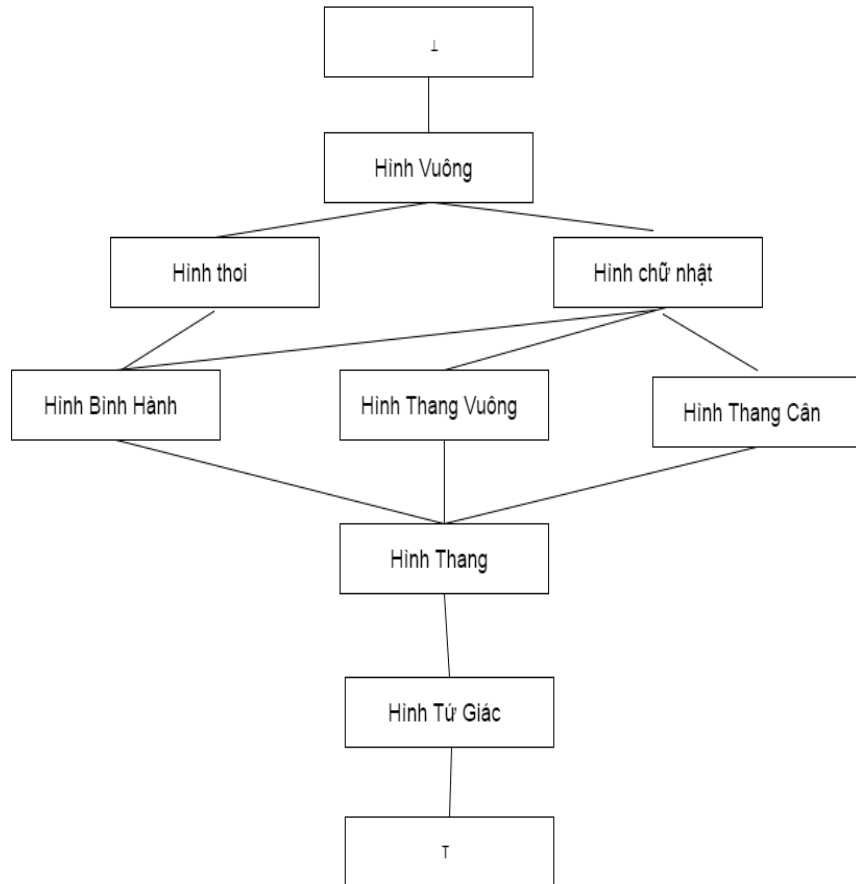
- T_C và T_R là một cặp hữu hạn phân chia thành từng đôi.
- T_C và T_R được biểu diễn bởi một cấu trúc phân cấp cụ thể. Ví dụ: cấu trúc cây, một lattice hoặc một semi – lattice.
- Một tập T_C (resp. T_R) của concept (resp. relation) types còn được gọi là hệ thống phân cấp (hierarchy) của concept (resp. relation) types đó.

2.5.3.2. Tập các concept type T_C

Tập T_C chứa các concept type (loại quan hệ), tập này một phần được sắp xếp bởi mối quan hệ “ \leq ” như sau:

- Với s và $t \in T_C$, nếu $s \leq t$ thì:
 - + s : Là subtype của t .
Ví dụ: Học sinh là subtype của Người, Giáo viên mầm non là subtype của Giáo viên.
 - + t : Là supertype của s .
Ví dụ: Động vật có xương sống là supertype của Bò sát.
- \top : Là supertype của mọi type.
- \perp : Là subtype của mọi type.

Ví dụ: Ta có một tập T_C như sau.



Hình 2.8: Sơ đồ phân cấp “Tứ giác”

2.5.3.3. Tập các relation symbol T_R

Tập T_R chứa các relation symbol (nhãn quan hệ), tập này một phần được sắp xếp bởi mỗi quan hệ “ \leq ” và có đặc điểm như sau:

- Phân thành các tập con $T_R^1, T_R^2, \dots, T_R^k$ của các *relation symbol* tương ứng với các arity 1, 2, ..., k.
- Arity (số ngôi) của một quan hệ r được ký hiệu là: $\text{arity}(r)$.
- Bất kỳ 2 quan hệ nào khác nhau arity thì không thể so sánh được.

Tương tự như tập T_C , tập T_R cũng được phân cấp như sau:

- Với s và $t \in T_R$, nếu $s \leq t$ thì:
 - + s : Là subsymbol của t .

Ví dụ: Quan hệ một phần “partOf” là subsymbol của quan hệ “relationWith”.

+ t: Là supersymbol của s.

Ví dụ: Quan hệ “relationWith” là supersymbol của quan hệ một dạng “kindOf”.

– T: Là supersymbol của mọi symbol.

– \perp : Là subsymbol của mọi symbol.

2.5.3.4. Tập các individual maker I

Tập I chứa các individual maker (đánh dấu cá thể hay tên cá thể) có những đặc điểm như sau:

– Dấu * đại diện cho các *generic maker* (các cá thể không xác định tên rõ ràng).

– $M = I \cup \{*\}$ đại diện cho các *maker*, được sắp xếp như sau:

+ * lớn hơn mọi thành phần thuộc I.

+ I gồm các cặp không thể so sánh được.

2.5.4. Đồ thị G

Mô hình biểu diễn đồ thị G như sau:

$$G = (C, R, E, I)$$

Trong đó:

– (C, R, E) : Một đa đồ thị hữu hạn, vô hướng, song phương;

– I: Hàm gán nhãn cho các đỉnh và các cạnh nối của đồ thị.

2.5.4.1. Đồ thị (C, R, E) :

Như đã nói ở trên, (C, R, E) là một đa đồ thị hữu hạn, vô hướng và song phương, được gọi là *underlying graph* của G, ký hiệu là $graph(G)$. Trong đó các thành phần của đồ thị được mô tả như sau:

– C: Tập các đỉnh khái niệm của đồ thị.

- **R**: Tập các đỉnh quan hệ của đồ thị.
- **N**: Tập các đỉnh của đồ thị, $N = C \cup R$.
- **E**: Tập các cạnh (cung) nối giữa các đỉnh thuộc C và các đỉnh thuộc R .

2.5.4.2. Hàm gán nhãn l

l là hàm gán nhãn cho các đỉnh và các cạnh nối của đồ thị, thỏa các điều kiện sau đây:

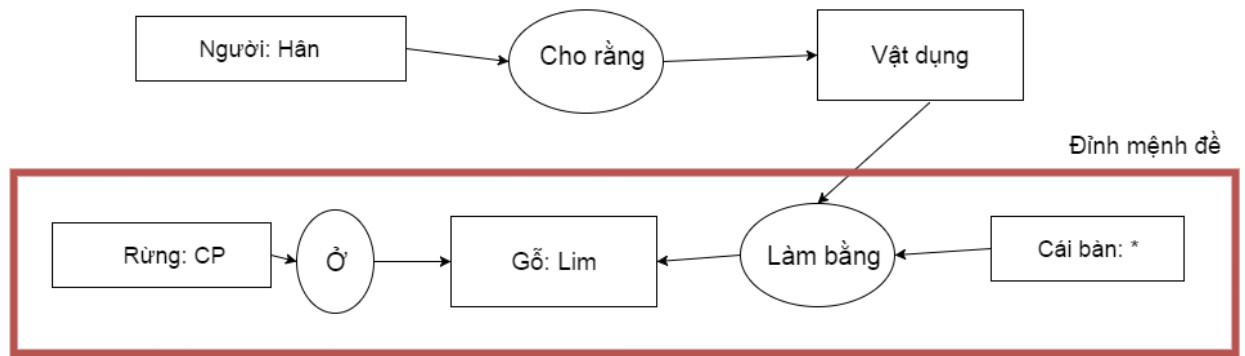
- Mỗi đỉnh khái niệm c được gán nhãn bởi cặp $(type(c), maker(c))$: $type(c) \in T_C$ và $maker(c) \in M$. Ví dụ: Học sinh: #12520855, Xe máy: Suzuki abc, ...
- Một đỉnh quan hệ r được gán nhãn bởi $l(r) \in T_R$ với $l(r)$: type hay symbol của r và còn được ký hiệu là $type(r)$.
- Bậc của đỉnh quan hệ r bằng arity (số ngôi) của $type(r)$, viết là: $arity(type(r))$.
- Một cạnh e nối với r thì đặt cố định và gán nhãn từ 1, ..., $arity(type(r))$.

2.5.5. Đỉnh mệnh đề trên Đồ thị khái niệm

Với Đồ thị khái niệm đã mở rộng có thể chứa cả mệnh đề trong một đỉnh khái niệm, từ đó ta có định nghĩa một đỉnh mệnh đề được phát biểu như sau:

Đỉnh mệnh đề là một đỉnh khái niệm có chứa một đồ thị khái niệm khác.

Ví dụ: Để biểu diễn câu “Hân cho rằng cái bàn đó được làm bằng gỗ Lim ở rừng CP” ta có đồ thị khái niệm sau:



Hình 2.9: Ví dụ minh họa đỉnh mệnh đề

Trong hình trên, phần đồ thị được khoanh đỏ chính là một đỉnh khái niệm.

2.5.6. Các phép toán cơ bản trên Đồ thị khái niệm

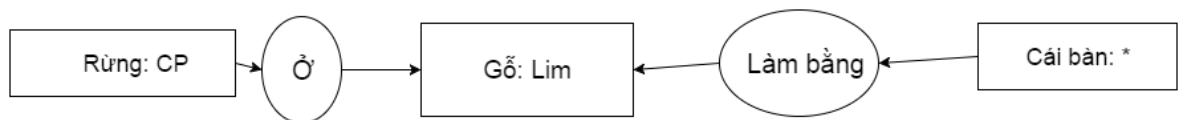
Trên Đồ thị khái niệm sẽ có các phép toán cơ bản thực hiện trên thành phần G như sau:

2.5.6.1. Phép Copy (nhân bản)

Ta có thể có được đồ thị B bằng cách nhân bản đồ thị A và gán cho B.

Ví dụ:

Đồ thị A:



Hình 2.10: Đồ thị A trong phép Copy

Đồ thị B có được nhờ phép copy A:



Hình 2.11: Đồ thị B trong phép Copy

2.5.6.2. Phép Restriction (giới hạn)

Phép Restriction giúp ta tạo ra một đồ thị mới từ đồ thị có sẵn, bằng cách thay một đỉnh khái niệm bằng đỉnh khác cụ thể hơn.

Các cách để thay một đỉnh khái niệm bằng đỉnh khác cụ thể hơn:

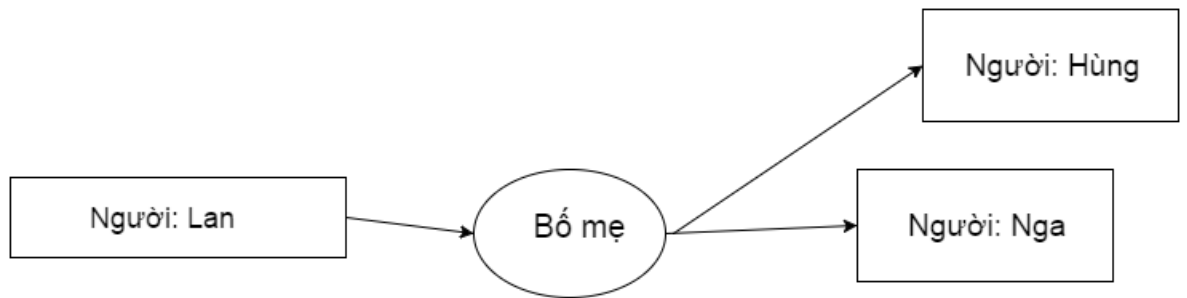
- Thay * bằng tên riêng hoặc một khóa.
Ví dụ: Thay đỉnh có nhãn “Học sinh: *” bằng “Học sinh: #12520001” hoặc “Học sinh: An”.
- Thay type bằng subtype của nó.
Ví dụ: Thay đỉnh “Người: A” bằng “Sinh viên: A”.

2.5.6.3. Phép Join (nối)

Phép Join giúp ta nối 2 đồ thị X và Y có sẵn thành một đồ thị mới, nếu có đỉnh khái niệm C xuất hiện trên cả hai đồ thị X và Y, thì chúng ta có thể nối hai đồ thị trên đỉnh chung C đó.

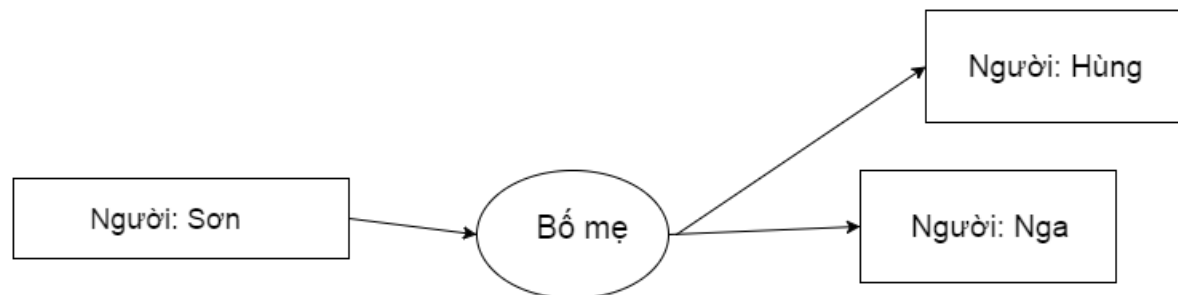
Ví dụ:

Đồ thị X biểu diễn “Lan là mẹ của Hùng và Nga”:



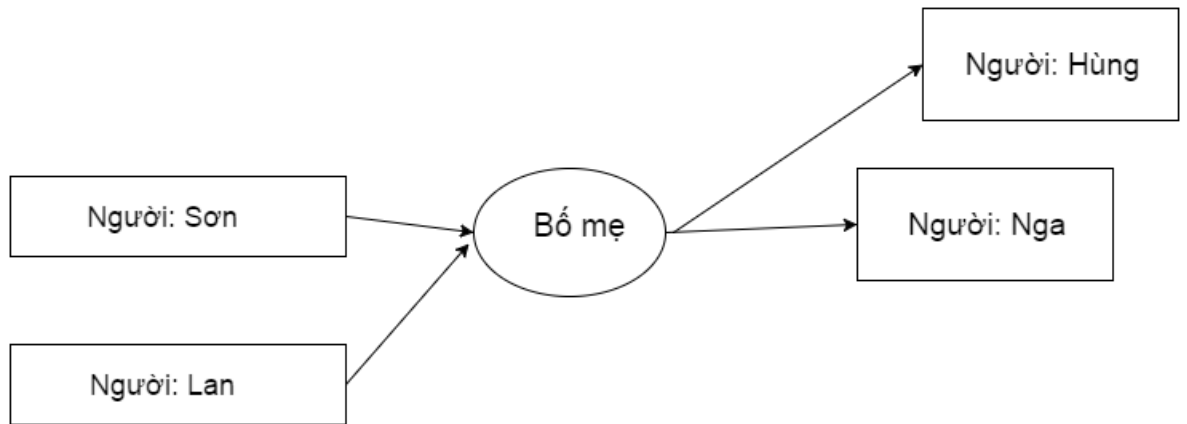
Hình 2.12: Đồ thị X trong phép Join

Đồ thị Y biểu diễn “Sơn là bố của Hùng và Nga”:



Hình 2.13: Đồ thị Y trong phép Join

Đồ thị Z được sinh ra bằng cách nối 2 đồ thị X và Y biểu diễn “Lan và Sơn là bố mẹ của Hùng và Nga”:



Hình 2.14: Đồ thị Z trong phép Join

Nhận xét: Phép Restriction và Join cho phép ta thực hiện tính kế thừa trên Đồ thị khái niệm như sau:

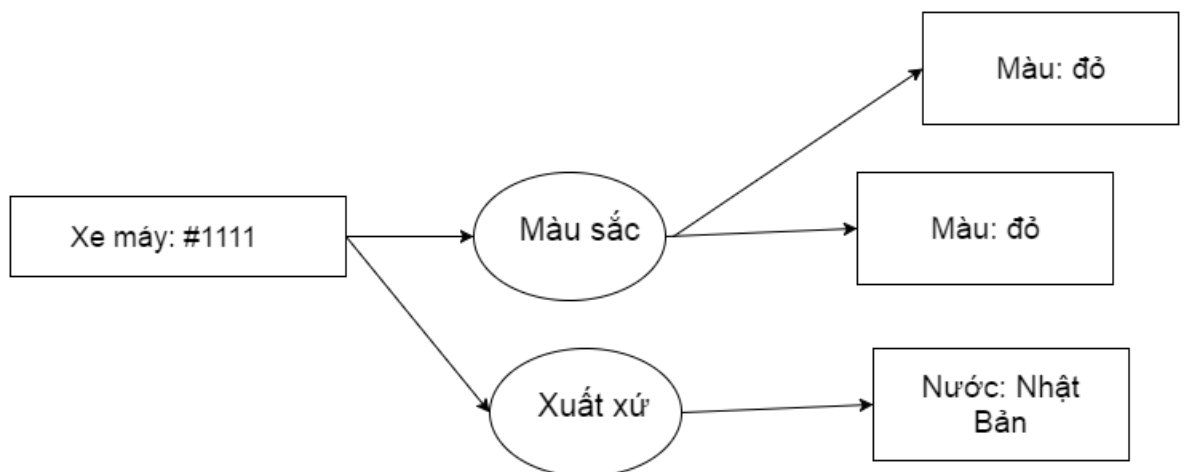
- Khi thay * bằng tên hoặc khóa: có thể kế thừa tính chất từ type của nó;
- Khi thay type bằng subtype: thể hiện sự kế thừa của subtype với type;
- Có đồ thị $Z = X \cup Y$ thì suy ra: X, Y là con của Z.

2.5.6.4. Phép Simplify (đơn giản)

Phép Simplify được phát biểu như sau: Nếu đồ thị X có 2 đồ thị con giống nhau thì lược bỏ một, ta được đồ thị mới có khả năng biểu diễn không đổi.

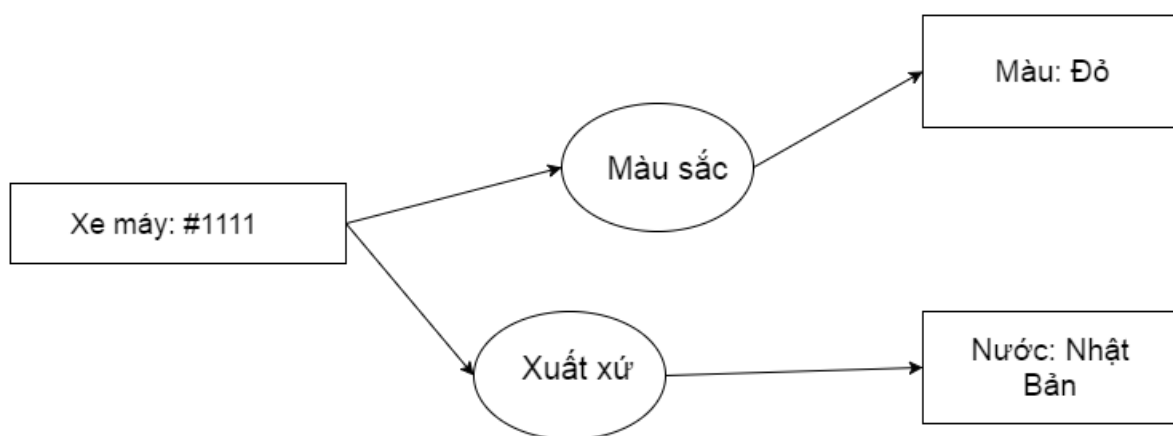
Ví dụ:

Đồ thị X:



Hình 2.15: Đồ thị X trong phép Simplify

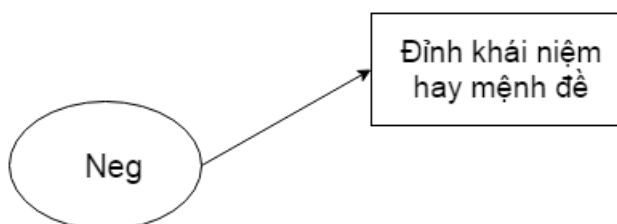
Đồ thị Y được tạo ra bằng cách đơn giản đồ thị X:



Hình 2.16: Đồ thị Y trong phép Simplify

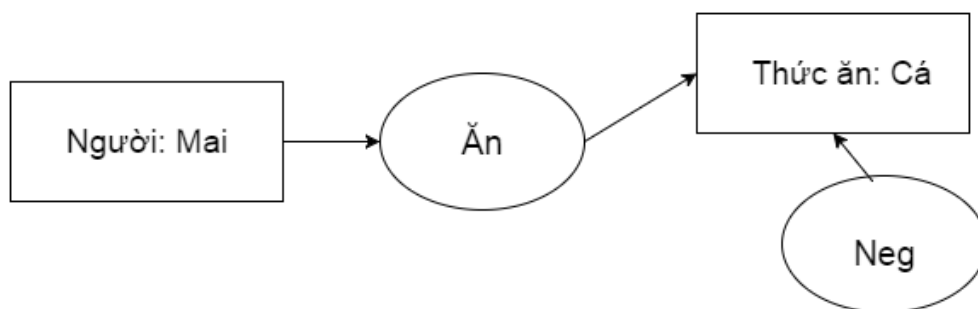
2.5.6.5. Phép Not (phủ định)

Ta thực hiện phép Not (phủ định) bằng cách đưa vào 1 đỉnh quan hệ có tên là neg (phủ định) để thể hiện khái niệm hay mệnh đề có tính phủ định:



Hình 2.17: Đỉnh phủ định neg

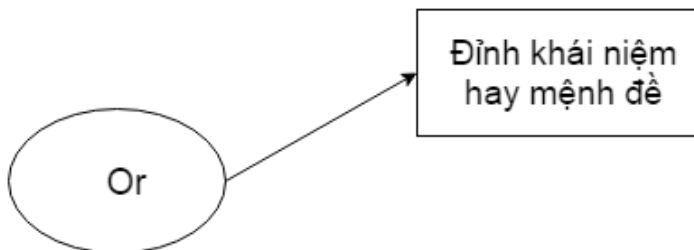
Ví dụ: Ta có đồ thị X biểu diễn câu “Mai không thích ăn cá” sau



Hình 2.18: Ví dụ về phép Not

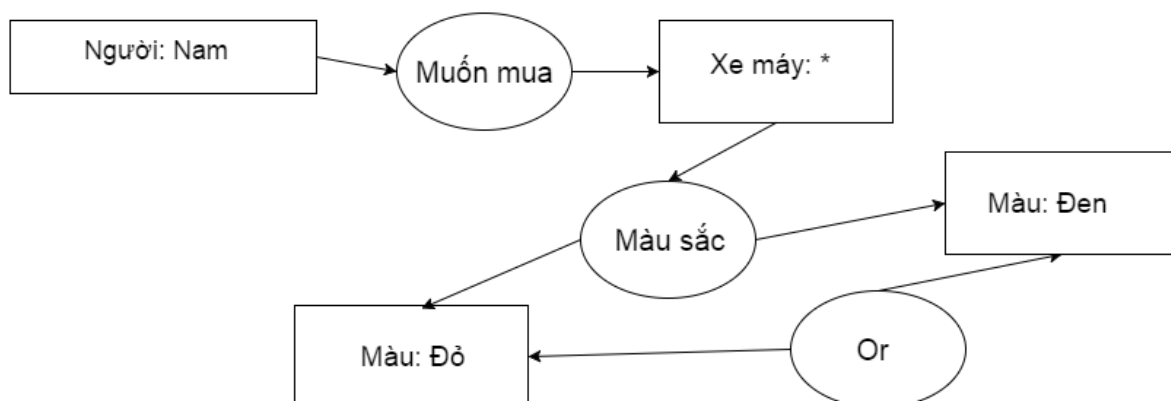
2.5.6.6. Phép Or (tuyển)

Phép Or (tuyển) thực hiện bằng cách đưa vào 1 đỉnh quan hệ có tên là or (tuyển) để thể hiện khái niệm hay mệnh đề như sau:



Hình 2.19: Đỉnh quan hệ Or

Ví dụ: Đồ thị biểu diễn câu “Nam muốn mua chiếc xe máy màu đỏ hoặc đen”



Hình 2.20: Ví dụ về phép Or

Chương 3. MÔ HÌNH VÀ GIẢI PHÁP

3.1. Phát biểu bài toán

Các mô hình và giải pháp được đề cập trong chương này nhằm tập trung giải quyết bài toán lớn gồm đầu vào và đầu ra như sau:

- Đầu vào (Input): Tên của một sản phẩm điện thoại di động mà người dùng quan tâm. Ví dụ: Bphone, Sony Xperia Z3, ...;
- Đầu ra (Output):
 - + Tập dữ liệu là các bình luận về sản phẩm đầu vào đã được phân tích và phân loại chiều hướng quan điểm (tích cực, tiêu cực, trung lập) theo trọng số tính được.
 - + Kết quả sau khi tổng hợp, thống kê các bình luận này theo từng tính năng của sản phẩm và tính điểm cho mỗi tính năng.

Muốn giải quyết được bài toán trên cần phải giải quyết các bài toán nhỏ sau:

- **Bài toán 1:** Tính trọng số cho từng câu trong mỗi bình luận của khách hàng về sản phẩm điện thoại di động. Để tính được trọng số này, ta cần phải tìm được ra cụm từ chỉ quan điểm và tính trọng số của cụm từ quan điểm. Từ đó tính trọng số của câu tương ứng chứa những cụm từ quan điểm vừa tìm được.
 - + Đầu vào (Input): Tập dữ liệu là các bình luận nguyên mẫu về sản phẩm điện thoại di động được lấy từ Web, chưa trải qua bước xử lý, phân loại và chứa rất nhiều thông tin gây nhiễu, không có ích.
 - + Đầu ra (Output): Tập dữ liệu là các bình luận về sản phẩm điện thoại di động đã được phân tích, rút trích cụm từ chỉ quan điểm và được tính trọng số.
- **Bài toán 2:** Phân loại chiều hướng quan điểm (tích cực, tiêu cực, trung lập) cho mỗi bình luận về sản phẩm điện thoại di động đã được phân tích, rút trích cụm từ chỉ quan điểm và tính trọng số.
 - + Đầu vào (Input): Kết quả đầu ra của bài toán 1.

- + Đầu ra (Output): Tập dữ liệu là các bình luận về sản phẩm điện thoại di động đã được phân loại chiều hướng quan điểm.
- **Bài toán 3:** Tổng hợp và tính điểm cho mỗi tính năng của sản phẩm điện thoại di động bởi trọng số của những cụm từ chỉ quan điểm nói về nó xuất hiện trong các câu bình luận đã được tính trọng số.
- + Đầu vào (Input): Kết quả đầu ra của bài toán 1.
- + Đầu ra (Output): Thống kê danh sách các tính năng đã được tính điểm của sản phẩm được yêu cầu.

3.2. Mô hình Ontology của ứng dụng

3.2.1. Mô hình Ontology cho lĩnh vực Sản Phẩm Điện Tử

Với kết quả sau khi thu thập và xử lý dữ liệu, dựa trên mô hình của một Ontology cơ bản, nhóm chúng em đề xuất mô hình cơ sở tri thức cho miền tri thức về Sản Phẩm Điện Tử gồm các thành phần như sau:

(Concepts, Relations, Rules)

Trong đó:

- **Concepts:** Tập hợp các lớp, khái niệm thuộc lĩnh vực;
- **Relations:** Tập hợp các biểu diễn của các kiểu quan hệ (khác lớp) giữa các thành phần thuộc Concepts;
- **Rules:** Tập hợp các luật xuất hiện trong miền tri thức.

3.2.1.1. Tập hợp các khái niệm Concepts

Một lớp, khái niệm (Concept) được mô tả bằng một bộ gồm các thành phần:

(Attributes, Label)

Trong đó:

- **Attributes:** Tập hợp các thuộc tính của mỗi khái niệm;

- **Label:** Nhãn phân loại của mỗi khái niệm. Ví dụ: Feature (tính năng) có nhãn là “ft”, Degree Word (từ chỉ mức độ) có nhãn là “dgw”, ...

Các Concept tồn tại trong cơ sở tri thức được mô tả chi tiết như sau:

STT	Tên	Mô tả	Nhãn	Danh sách các thuộc tính		
				Tên	Loại giá trị	Mô tả
1	Product	Sản phẩm	p	id	String	Mã sản phẩm
				name	String	Tên (đại diện) của sản phẩm
				list_name	String	Danh sách tên thay thế của sản phẩm
2	Feature	Tính năng của sản phẩm. Ví dụ: màn hình, giá, cấu hình, ...	ft	id	String	Mã tính năng
				name	String	Tên (đại diện) của

						tính năng
				list_indicator	String	Danh sách các dấu hiệu nhận biết (hiện)
				list_hidden_indicator	String	Danh sách các dấu hiệu nhận biết (ẩn)
3	SentiWord	Từ chỉ quan điểm. Ví dụ: đẹp, xấu, tốt, ồn, bắt mắt, ...	stw	id	String	Mã từ
				word	String	Nội dung từ
				feature	String	Tên tính năng mà từ này mô tả

				score	Float	Điểm của từ
4	DegreeWord	Từ chỉ mức độ. Ví dụ: rất, khá, hơi, ...	dgw	id	String	Mã từ
				word	String	Nội dung từ
				score	Float	Điểm của từ
5	DeniedWord	Từ phủ định. Ví dụ: không, chẳng, chả, ...	dnw	id	String	Mã từ
				word	String	Nội dung từ
				score	Float	Điểm của từ
6	Comparision Word	Từ dùng để thể hiện sự so sánh. Ví dụ: hơn, thua, bằng, ...	cw	id	String	Mã từ
				word	String	Nội dung từ
				symbol	String	Dấu đại diện. Ví dụ: “hơn” có symbol là “>”.

7	ReferWord	Từ kết nối giữa cụm từ chỉ quan điểm với tính năng, thường là động từ. Ví dụ: từ “về” trong câu “chuẩn về phong cách”.	rw	Id	String	Mã từ
				word	String	Nội dung từ

Bảng 3.1: Tập hợp Concepts của Ontology lĩnh vực Sản Phẩm Điện Tử

3.2.1.2. Tập hợp các quan hệ Relations

Tập Relations biểu diễn các kiểu quan hệ (khác lớp) giữa các đối tượng Concepts. Các kiểu quan hệ này đều là quan hệ hai ngôi giữa một đối tượng Concept với một tượng Concept hoặc giữa một đối tượng Concept với một nhóm các đối tượng Concepts kết hợp với nhau bằng. Mỗi đối tượng thuộc tập Relations được biểu diễn bởi một bộ gồm các thành phần sau:

(id, left_object, right_object, label)

Trong đó:

- **id**: Mã quan hệ;
- **left_object**: Nhãn của đối tượng 1;
- **right_object**: Nhãn của đối tượng 2 hoặc nhóm nhãn của các đối tượng nối với nhau bởi dấu “+”;
- **label**: Nhãn của quan hệ.

Các đối tượng trong tập Relations được mô tả chi tiết như sau:

Id	left_object	right_object	label	Mô tả
1	ft	P	PropertyOf	Left là thuộc tính của right
2	stw	Ft	Define	Left mô tả cho right
3	dgw	Stw	Modify	Left bổ nghĩa cho right
4	dnw	Stw	Deny	Left phủ định right
5	dgw	dnw + stw	Modify	Left bổ nghĩa cho right
6	dnw	dgw + stw	Modify	Left bổ nghĩa cho right

Bảng 3.2: Tập hợp Relations của Ontology lĩnh vực Sản Phẩm Điện Tử

3.2.1.3. Tập hợp các luật Rules

Mỗi thành phần $r \in \text{Rules}$ được mô tả bởi một bộ gồm các thành phần như sau:

(id, left_content, right_content)

Trong đó:

- **id**: Mã của luật;
- **left_content**: Phần giả thiết của luật, được mô tả bởi sự kiện kết hợp giữa các đối tượng có cú pháp:

<nhãn đối tượng 1> + <nhãn đối tượng 2> + ...

Các đối tượng này có thể là một đối tượng trong tập Concepts hoặc một đối tượng đánh dấu tên luật. Các sự kiện này được dùng để phân tích cấu trúc thông tin của các vế trong một câu của mỗi bình luận người dùng;

- **right_content**: Phần kết luận của luật, được mô tả bởi sự kiện đánh dấu tên của luật có cú pháp:

r<số thứ tự>

Ứng với mỗi sự kiện này, ta được một đồ thị riêng thể hiện cấu trúc thông tin của mỗi vế câu trong câu bình luận. Các đồ thị này được phân chia thành 5 nhóm, mỗi nhóm được ký hiệu như sau:

g<số thứ tự>

Cấu trúc của các đồ thị này sẽ được mô tả chi tiết trong mục ...

id	left_content	example	right_content	group
Click here to define a filter				
8 ft + r01		màn_hình đẹp	r08	g01
9 ft + r02		màn_hình không đẹp	r09	g02
10 ft + r03		màn_hình đẹp quá	r10	g03
11 ft + r04		màn_hình quá đẹp	r11	g03
12 ft + r05		màn_hình quá "là" không đẹp	r12	g04
13 ft + r06		màn_hình không đẹp quá	r13	g05

Hình 3.1: Ví dụ minh họa tập hợp Rules của Ontology

3.2.2. Quy trình xây dựng Ontology

3.2.2.1. Nguồn thu thập

Để tổng hợp dữ liệu cho các lớp trong Concepts, trước tiên phải thu thập các dữ liệu là các bình luận hay đánh giá của người dùng về một số sản phẩm điện tử nhất định, từ đó xây dựng nên tập mẫu để xử lý và rút trích các đối tượng Concepts mong muốn. Trong khóa luận này, nhóm chúng em đã chọn một số sản phẩm điện thoại di động đang được khách hàng quan tâm hiện nay và thực hiện lấy dữ liệu về (Bphone, Asus Zenfone 5, HTC One 10, ...). Để thuận tiện cho quá trình phân tích dữ liệu, các bình luận được thu về từ nguồn chính là mục “Số hóa” , trang điện tử VnExpress.Net (<http://sohoa.vnexpress.net/>). Nguyên nhân là do các bình luận trên mỗi bài báo trên trang điện tử này đều đã được kiểm duyệt bởi các biên tập viên trước khi được đăng, nên các bình luận chẳng những tập trung vào chủ đề mà còn tránh được phần nào các lỗi chính tả, cú pháp câu, ... Dữ liệu bình luận trên trang tin VnExpress.Net là nguồn giàu thông tin, với định dạng trả về là HTML thuần, không sử dụng JavaScript giúp dễ dàng trích xuất thông tin.

Ngoài việc phân tích các bình luận trên tập mẫu, cũng cần phải bổ sung dữ liệu từ các nguồn có liên quan khác như:

- Các chuyên gia trong lĩnh vực, các cá nhân và cơ quan có liên quan như: Xử lý ngôn ngữ tự nhiên, các sản phẩm điện tử, ...
- Các tài liệu, văn bản giấy chính qui như: Các tài liệu học tập (giáo trình, tài liệu tham khảo của ngành học), từ điển chuyên ngành, ...
- Các tài liệu từ Internet: Các trang Web của Bộ Ngành, Hiệp hội, các trang Web bách khoa toàn thư (Wikipedia tiếng Việt: <https://vi.wikipedia.org/>), từ điển chuyên ngành, cổng thông tin, diễn đàn, các trang báo điện tử chuyên về sản phẩm điện tử (<http://thegioididong.com/> , <http://tinhte.vn/> , ...), ...

3.2.2.2. Cách thức xây dựng

Sau khi thực hiện công việc thu thập dữ liệu, tách xử lý các file html được một tập dữ liệu gồm các bình luận có chứa và không chứa quan điểm của người dùng về các sản phẩm đã chọn. Để xây dựng tập mẫu, nhóm chúng em tiến hành loại bỏ các bình luận không chứa bất kỳ quan điểm nào. Kết quả thu thập các bình luận (đã loại bỏ bình luận không chứa quan điểm) cho tập mẫu như sau:

STT	Tên sản phẩm	Số bình luận mẫu
1	Bphone	53
2	SamSung Galaxy S7 Edge	60
3	Sony Xperia Z3	61
4	Asus Zenfone 5	51
5	Microsoft Lumia 950	54
6	Iphone SE	51

7	Iphone 6	51
8	HTC One 10	50

Bảng 3.3: Kết quả tập mẫu bình luận

3.2.2.2.1 Xây dựng tập sản phẩm Product

Dựa trên tập mẫu bình luận, ta xây dựng tập Product theo các bước sau:

- **Bước 1:** Khởi tạo các đối tượng trong tập Product có tên là tên của các sản phẩm trong tập mẫu: Bphone, SamSung Galaxy S7 Edge, Sony Xperia Z3, Asus Zenfone 5, Microsoft Lumia 950, Iphone SE, Iphone 6, HTC One 10.
- **Bước 2:** Duyệt trên từng tập mẫu ứng với mỗi tên sản phẩm lấy ra những cụm từ đại diện cho tên của sản phẩm đó. Ví dụ: Xét tập mẫu của dòng phẩm SamSung Galaxy S7 Edge, ta có câu bình luận sau: “Galaxy S7 edge là smartphone có camera tốt nhất, các smartphone cao cấp khác tất đài không cần bàn cãi.” Rút trích được cụm từ chỉ tên sản phẩm là “Galaxy S7 edge”.
- **Bước 3:** Bổ sung các đối tượng khác ngoài các sản phẩm mẫu cho tập Product bằng cách tìm thông tin về chúng trên các trang Web chuyên về lĩnh vực Sản Phẩm Điện Tử. Do thời gian có hạn nên bước này vẫn chưa được thực hiện trong khóa luận.

3.2.2.2.2 Xây dựng tập tính năng sản phẩm Feature

Các bước xây dựng tập tính năng sản phẩm Feature như sau:

- **Bước 1:** Chọn ra các tính năng sản phẩm mà người dùng hay quan tâm nhất và khởi tạo các đối tượng trong tập Feature với tên đại diện của các tính năng đã chọn. Trong khóa luận này nhóm chúng em chọn các tính năng sau: Thiết kế (Design), Giá (Price), Camera, Màn hình (Screen), Pin (Battery), Ứng dụng (Software), Cấu hình (Process & Memory).

- **Bước 2:** Đối với từng tính năng, duyệt mỗi bình luận mẫu chứa những tính năng đó để tìm ra các dấu hiệu nhận biết ở dạng hiện. Thường là các danh từ hay động từ ám chỉ tính năng, ví dụ:
 - + Tính năng Design thì có từ: kiểu dáng, bên ngoài, mẫu mã, ngoại hình, ...
 - + Tính năng Camera thì có các từ: hình ảnh, ảnh chụp, chụp, ...
- **Bước 3:** Đối với từng tính năng, duyệt mỗi bình luận mẫu chứa những tính năng đó để tìm ra các dấu hiệu nhận biết ở dạng ẩn. Thường là các tính từ đặc trưng duy nhất ám chỉ cho tính năng đó mà không có ở tính năng khác để tránh gây nhầm lẫn cho quá trình nhận dạng, ví dụ:
 - + Tính năng Design thì có từ: phong cách, lịch lãm, men, nữ tính, sang trọng, tinh tế, ...
 - + Tính năng Price thì có các từ: đắt, mắc, rẻ, ...

Kết quả xây dựng tập Feature như sau:

id	name	list_indicator	list_hidden_indicator
Click here to define a filter			
1	design	design, thiết kế, kiểu dáng, hình dáng, bề ngoài, dáng vẻ, hình trạng, mẫu mã, dáng dấp, nhìn, mặt, viền, nhìn, lưng, trông, dáng, hình thức, ngoại hình, kiểu cách	đẹp, xấu, xấu quắt, sang trọng, phong cách, tinh tế, thẩm mỹ, lịch lãm, thẩm mỹ, cá tính, gọn gàng, menly, nam tính, tối giản
2	screen	screen, màn hình, LCD, monitor, màn	
3	camera	cam, camera, ảnh, picture, hình ảnh, hình chụp, ảnh chụp, chụp, lấy nét, khung hình	trung thực
4	battery	battery, pin, sạc	
5	software	software, ứng dụng, phần mềm, app, apps, game, tiện ích	
6	price	price, money, giá, giá cả	đắt, rẻ, mắc
7	memory	memory, chip, ram, hiệu năng, tốc độ, xử lý, cấu hình, phần cứng, chạy, hệ điều hành, chip, vi xử lý, hiệu năng	mượt

Hình 3.2: Kết quả xây dựng tập Feature

3.2.2.2.3 Xây dựng tập từ chỉ quan điểm SentiWord

Tập SentiWord và Feature là hai tập Concepts quan trọng nhất, không thể thiếu, các tập này được dùng để nhận dạng và rút trích những cụm từ cần thiết chỉ tính năng và quan điểm. Thay vì sử dụng trực tiếp nội dung của bộ từ điển VietSentiWord do Vũ Xuân Sơn và các cộng sự xây dựng [13], trong khóa luận này nhóm chúng em đã xây dựng một tập từ chỉ

quan điểm SentiWord riêng dành cho lĩnh vực Sản Phẩm Điện Tử và chỉ sử dụng bộ từ điển VietSentiWord này cho việc đánh giá điểm trọng số cho từng từ trong tập SentiWord. Xuất phát từ nguyên nhân: bộ từ điển VietSentiWord là bộ từ điển chung không dành riêng cho lĩnh vực Sản Phẩm Điện Tử và còn hạn chế về số lượng các bộ từ. Việc xây dựng một từ điển dành riêng cho lĩnh vực đang xét sẽ làm tăng độ chính xác của việc phân tích và rút trích các cụm từ chỉ quan điểm trong mỗi bình luận.

Áp dụng các phương pháp xử lý ngôn ngữ tự nhiên kết hợp thống kê thủ công, tập SentiWord được xây dựng như sau:

- **Bước 1:** Duyệt mỗi đối tượng Feature lấy ra tập bình luận mẫu chứa đối tượng đó.
- **Bước 2:** Duyệt mỗi bình luận lấy ra được từ bước 1 và thực hiện các bước sau.
 - + **2.1:** Sử dụng công cụ **vnTokenizer** và **vnTagger** của tác giả Lê Hồng Phương để phân tách thành những từ, cụm từ và gán nhãn phân loại từ vựng cho mỗi bình luận.

Ví dụ: Xét tính năng Design, ta có câu: “Nhìn sản phẩm đẹp, cá tính, nhưng không biết chất lượng thế nào.” Kết quả phân tách từ và gán nhãn: “Nhìn/V sản_phẩm/N đẹp/A ./, cá_tính/A ./, nhưng/C không/R biết/V chất_lượng/N thế_nào/P ./.”
 - + **2.2:** Lấy những tính từ hoặc cụm tính từ có nhãn phân loại là “A” hoặc “A” + “V”. Ví dụ: cá_tính/A, lịch_lãm/A, ấn_tượng/A, nhanh/A hết/V, ...
 - + **2.3:** Loại bỏ các từ gây nhiễu (không chứa quan điểm hoặc có chứa nhưng không nói về Feature đang xét) bằng phương pháp thủ công.
- **Bước 3:** Tính điểm trọng số cho mỗi từ đã rút trích được ở bước 2. Duyệt từng từ và thực hiện các bước sau:

- + **3.1:** Duyệt toàn bộ các tập từ trong từ điển VietSentiWord, lấy ra các bộ từ chứa từ đang xét và lấy trọng số tích cực của bộ từ làm trọng số cho từ. Nếu xuất hiện nhiều bộ từ đồng thời thì chọn bộ từ có trọng số phù hợp nhất bằng kinh nghiệm của các chuyên gia trong lĩnh vực có liên quan;
- + **3.2:** Nếu không tìm thấy bất cứ bộ từ nào tiến hành khảo sát để lấy ý kiến của các chuyên gia trong lĩnh vực liên quan về độ tin cậy, từ này có phải là từ có chiều hướng tích cực?, có giá trị là: -1 (phủ định) hoặc 1 (khẳng định). Điểm của mỗi từ được tính bằng công thức sau:

$$\frac{\text{Tổng số ý kiến khẳng định}}{\text{Tổng số ý kiến}}$$

3.2.2.2.4 Xây dựng tập từ chỉ mức độ DegreeWord

Các bước xây dựng tập từ chỉ mức DegreeWord:

- **Bước 1:** Duyệt mỗi bình luận lấy ra từ tập mẫu và thực hiện các bước sau.
 - + **1.1:** Sử dụng công cụ **vnTokenizer** và **vnTagger** của tác giả Lê Hồng Phương để phân tách thành những từ, cụm từ và gán nhãn phân loại từ vựng cho mỗi bình luận.
Ví dụ: Xét tính năng Design, ta có câu: “Nhìn sản phẩm đẹp, cá tính, nhưng không biết chất lượng thế nào.” Kết quả phân tách từ và gán nhãn: “Nhìn/V rất/R đẹp/A ./, cá_tính/A./.”
 - + **1.2:** Lấy những trạng từ có nhãn phân loại là “R”, ví dụ: rất/R, quá/R, ... Các từ có nhãn “T” và “P” như: lắm/T, thế/P, ... Các tính từ có nhãn “A” không nằm trong tập SentiWord. Ví dụ: tương_đối/A, thật/A, có_vẻ/A, ...

- + **1.3:** Loại bỏ các từ gây nhiễu (không phải từ chỉ mức độ) bằng phương pháp thủ công theo ý kiến của các chuyên gia trong lĩnh vực xử lý ngôn ngữ tự nhiên.
- **Bước 2:** Bổ sung các đối tượng khác ngoài tập mẫu theo ý kiến các chuyên gia trong lĩnh vực xử lý ngôn ngữ tự nhiên và từ điển Tiếng Việt nói về từ chỉ mức độ.
- **Bước 3:** Tính điểm cho mỗi từ đã rút trích được ở bước 1 và 2 bằng phương pháp thống kê thủ công kết hợp ý kiến của chuyên gia trong lĩnh vực liên quan. Các từ này chỉ làm 2 nhóm:
 - + Nhóm dưới mức 1 điểm: Các từ được phân vào nhóm này là các từ khi bỏ nghĩa cho từ chỉ quan điểm sẽ làm giảm mức độ của từ đó. Ví dụ: cũng, hơi, có vẻ, tương đối, ...
 - + Nhóm trên mức 1 điểm: Các từ được phân vào nhóm này là các từ khi bỏ nghĩa cho từ chỉ quan điểm sẽ làm tăng mức độ của từ đó. Ví dụ: rất, quá, dữ dội, ...

Việc phân nhóm các từ này đều phải dựa trên ý kiến chuyên gia trong lĩnh vực. Từ đó ta có các bước áp dụng tính điểm cho từ chỉ mức độ như sau:

- + **3.1:** Phân các từ vào hai nhóm đã nêu trên bằng phương pháp thủ công;
- + **3.2:** Sắp xếp thứ tự theo chiều tăng dần mức độ cho danh sách từ ở mỗi nhóm. Ví dụ: kém → chút → cũng → có vẻ → hơi → tương đối. Sau đó phân nhóm các từ trong mỗi nhóm, các từ có cùng mức độ sẽ cùng một nhóm, ví dụ như: từ “rất” và “quá”.
- + **3.3:** Từ có mức độ nhỏ nhất ở 2 nhóm từ được cho điểm đầu tiên, bằng mức điểm tối thiểu (lớn hơn 0). Sau đó cho điểm các từ còn lại theo công thức:

Điểm của từ có mức độ nhỏ hơn liên kề trước nó +
Điểm tối đa mỗi nhóm từ – Điểm của phần tử nhỏ nhất nhóm
Tổng số từ nhóm từ trong nhóm – 1

Ví dụ: Đối với nhóm từ dưới 1, từ “kém” là từ có mức độ nhỏ nhất được khởi tạo bằng 0.1. Điểm tối đa cho nhóm là 0.9. Giả sử tổng số nhóm từ thuộc nhóm dưới 1 là 5, suy ra điểm của từ

$$\text{tiếp theo sau từ “kém” là: } 0.1 + \frac{0.9-0.1}{5-1} = 0.1 + \frac{0.8}{4} = 0.3$$

Mức điểm tối đa và tối thiểu ở mỗi nhóm được cho phải hợp lý, phụ thuộc vào số lượng nhóm từ trong nhóm.

id	word	score
Click here to define a filter		
1	quá	1.75
2	tương đối	0.8
3	thật	1.5
4	lắm	1.75
5	thế	1.25
6	rất	2
7	hơi	0.6
8	cũng	0.4
9	chút	0.3
12	nhất	3.25
13	khá	0.9
14	dữ dội	2.5
15	có vẻ	0.4
16	cực	2.25
17	cực kỳ	2.5
18	kém	0.1

Hình 3.3: Một mẫu của tập DegreeWord

3.2.2.2.5 Xây dựng tập từ phủ định DeniedWord

Xét tập quan hệ Relations xuất hiện 2 loại quan hệ có vẻ trái là “dnw”, thế nên tương ứng với mỗi loại quan hệ là có cách tính điểm từ phủ định riêng:

- Nhãn quan hệ “Deny”: Giá trị là -1;
- Nhãn quan hệ “Modify”: Giá trị được tính giống như giá trị của một từ chỉ mức độ, giá trị này chính là thuộc tính score của mỗi đối tượng trong tập DeniedWord.

Các bước xây dựng tập DeniedWord như sau:

- **Bước 1:** Duyệt mỗi bình luận lấy ra từ tập mẫu và thực hiện các bước sau.
 - + **1.1:** Sử dụng công cụ **vnTokenizer** và **vnTagger** của tác giả Lê Hồng Phương để phân tách thành những từ, cụm từ và gán nhãn phân loại từ vựng cho mỗi bình luận.
Ví dụ: Xét tính năng Design, ta có câu: “Nhìn sản phẩm đẹp, cá tính, nhưng không biết chất lượng thế nào.” Kết quả phân tách từ và gán nhãn: “Nhìn/V không/R đẹp/A ./, cá_tính/A ./.”
 - + **1.2:** Lấy những trạng từ có nhãn phân loại là “R”, ví dụ: không/R, ...
 - + **1.3:** Loại bỏ các từ gây nhiễu (không phải từ phủ định) bằng phương pháp thủ công theo ý kiến của các chuyên gia trong lĩnh vực xử lý ngôn ngữ tự nhiên.
- **Bước 2:** Bổ sung các đối tượng khác ngoài tập mẫu theo ý kiến các chuyên gia trong lĩnh vực xử lý ngôn ngữ tự nhiên và từ điển Tiếng Việt nói về từ phủ định.
- **Bước 3:** Tính điểm cho mỗi từ đã rút trích được ở bước 1 và 2 theo ý kiến chuyên gia bằng cách sắp xếp tập từ theo mức độ tăng dần và áp dụng công thức tính như đối với từ chỉ mức độ đã nhắc ở mục 3.2.2.2.4.

3.2.2.2.6 Xây dựng các tập từ so sánh, liên kết

- Đối với tập từ so sánh ComparisionWord:

- + **Bước 1:** Duyệt thủ công trên tập các bình luận mẫu và lấy ra những từ mang ý nghĩa so sánh.
- + **Bước 2:** Bổ sung những đối tượng nằm ngoài tập mẫu theo ý kiến các chuyên gia trong lĩnh vực xử lý ngôn ngữ tự nhiên.
- + **Bước 3:** Gán giá trị cho thuộc tính symbol theo ý kiến chuyên gia. Ví dụ: “hơn” gán symbol là “>”, “bằng” hoặc “như” gán symbol là “=”, ...
- Đối với tập từ liên kết, kết nối ReferWord:
 - + **Bước 1:** Duyệt mỗi bình luận lấy ra từ tập mẫu và thực hiện các bước sau.
 - **1.1:** Sử dụng công cụ **vnTokenizer** và **vnTagger** của tác giả Lê Hồng Phương để phân tách thành những từ, cụm từ và gán nhãn phân loại từ vựng cho mỗi bình luận.
 - **1.2:** Lấy những từ có nhãn “V”
 - **1.3:** Loại bỏ các từ không mang ý nghĩa liên kết bằng phương pháp thủ công.

3.2.2.2.7 Xây dựng tập luật Rules

Các bước xây dựng tập luật của Ontology như sau:

- **Bước 1:** Duyệt mỗi bình luận lấy ra từ tập mẫu và thực hiện các lấy ra những cấu trúc chỉ bao gồm các từ chỉ tính năng sản phẩm, từ chỉ quan điểm, mức độ, phủ định và liên kết. Các từ này được lấy ra theo thứ tự xuất hiện lần lượt trong câu và bắt buộc phải liên quan đến nhau theo các mối quan hệ có trong tập Relations.
Ví dụ: Xét câu: “Điện thoại này có cấu hình khủng”, ta được cụm từ “cấu hình khủng”. Hoặc xét câu: “Bphone chuẩn về phong cách” ta được cụm từ “chuẩn về phong cách”.
- **Bước 2:** Gán nhãn đối tượng cho từng đối tượng xuất hiện trong mỗi cụm từ. Tách lấy các nhãn và liên kết chúng bởi dấu “+”. Ví

dụ: Cụm từ “cấu hình khung” sau khi gán nhãn được “cấu_hình/ft khung/stw” và chuỗi “ft + stw”.

- **Bước 3:** Đánh dấu phân kết luận của mỗi luật bằng sự kiện đánh dấu tên luật theo thứ tự từ 01 trở đi cho mỗi chuỗi kết quả ở bước 2 sau khi xóa “ft” để tạo các luật cơ bản (chỉ dùng phân tích mà không dùng để tạo đồ thị cho câu bình luận). Ví dụ: Ta có luật cơ bản $stw = r01$, $dgw + stw = r02$, ...
- **Bước 4:** Tạo các luật đi kèm với “ft” kết hợp với một trong những luật đã xây dựng ở bước 3 và đánh dấu phân kết luận của mỗi luật bằng sự kiện đánh dấu tên luật theo thứ tự từ thứ tự cuối cùng trong tập luật cơ bản. Các luật này dùng để tạo đồ thị cho câu bình luận. Ví dụ: Giả sử ta có 2 luật cơ bản là $stw = r01$ và $dgw + stw = r02$. Ta có luật $ft + r01 = r03$ biểu diễn cho cấu trúc “ft + stw”.
- **Bước 5:** Khảo sát và dự kiến phân chia mỗi luật vào các nhóm đồ thị.

id	left_content	example	right_content	group
Click here to define a filter				
1	stw	đẹp	r01	(null)
2	dnw + r01	không đẹp	r02	(null)
3	r01 + dgw	đẹp quá	r03	(null)
4	dgw + r01	quá đẹp	r04	(null)
5	dgw + r02	quá không đẹp	r05	(null)
6	dnw + r03	không đẹp quá	r06	(null)
7	dnw + r04	không quá đẹp	r07	(null)

Hình 3.4: Các luật cơ bản trong tập Rules

3.3. Mô hình đồ thị khái niệm biểu diễn câu và vế câu trong một đánh giá của người dùng

Mỗi bình luận, đánh giá sẽ được phân tách thành tập các câu theo dấu hiệu kết thúc câu như: “.”, “!”, “?”, “\n”, ... Và mỗi câu lại được phân tách thành nhiều vế theo dấu hiệu phân tách vế như: “,”, “;”, “và”, “tuy”, “nhưng”, “không những”,

“mà còn”, ... Với mỗi câu và vế câu như vậy sẽ có một mô hình riêng để biểu diễn cấu trúc thông tin của chúng sau khi đã được phân tích. Những đồ thị này ngoài mục đích biểu diễn cấu trúc thông tin, nó còn tạo điều kiện thuận lợi để dễ dàng thực hiện các bước phân loại đánh giá người dùng theo các cụm từ chỉ quan điểm và tổng hợp kết quả để tính điểm cho từng tính năng của sản phẩm được yêu cầu.

3.3.1. Mô hình đồ thị khái niệm biểu diễn cho mức vế câu

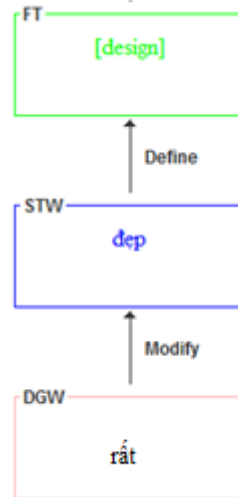
Mô hình đồ thị khái niệm biểu diễn cho mức vế câu gồm các thành phần sau:

(C, E, I)

Trong đó:

- **C**: Tập hợp bao gồm các đỉnh khái niệm của đồ thị biểu diễn cho một đối tượng trong tập Concepts, chỉ bao gồm các loại đối tượng: Feature, SentiWord, DegreeWord và DeniedWord. Vì xét ở mức độ vế câu nên đồ thị của ở mức vế chỉ chứa một đỉnh khái niệm loại Feature.
- **E**: Tập hợp cạnh nối các đỉnh của đồ thị trong tập C, để chỉ mối quan hệ giữa các đỉnh của đồ thị.
- **I** là hàm gán nhãn cho các đỉnh và cạnh của đồ thị sao cho:
 - + Một đỉnh $c \in C$ được gán nhãn bởi $\text{type}(c) \in \text{Labels}$ với Labels là tập các nhãn của các đối tượng trong tập Concepts.
 - + Một cạnh $e \in E$ được gán nhãn bởi $\text{type}(e) \in \text{TR}$ với TR là tập các ký hiệu hay nhãn cho loại quan hệ giữa các đối tượng Concepts đã được mô tả trong tập Relations theo thuộc tính label.

Ví dụ: Xét vế câu “nhìn rất đẹp” ta có đồ thị biểu diễn vế câu như sau:



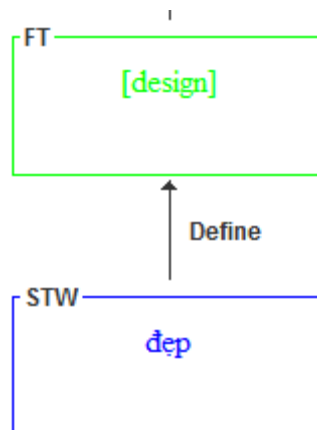
Hình 3.5 Ví dụ minh họa đồ thị về câu

Phân tích đồ thị trên ta có:

- Tập $C = \{c1 = "[design]", c2 = "đẹp", c3 = "rất"\}$.
- Tập $E = \{e1 = (c2, c1), e2 = (c3, c2)\}$.
- $l(c1) = "FT", l(c2) = "STW", l(c3) = "DGW", l(e1) = "Define", l(e2) = "Modify"$.

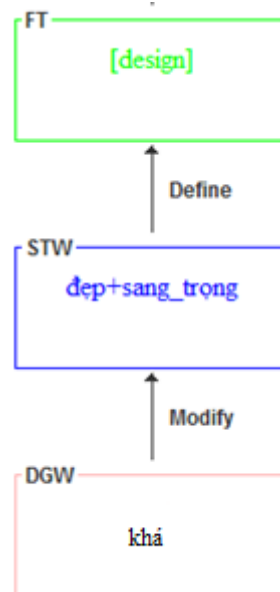
Đồ thị khái niệm biểu diễn về câu được chia làm 5 loại như sau:

- **Nhóm 1:** Tập C chỉ chứa 2 loại đối tượng là Feature và SentiWord.



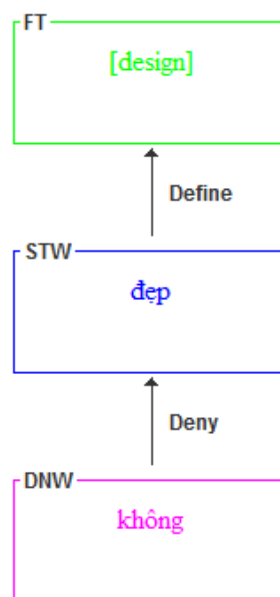
Hình 3.6: Ví dụ minh họa đồ thị về câu nhóm 1

- **Nhóm 2:** Tập C chỉ chứa các loại đối tượng là Feature, SentiWord và DegreeWord.



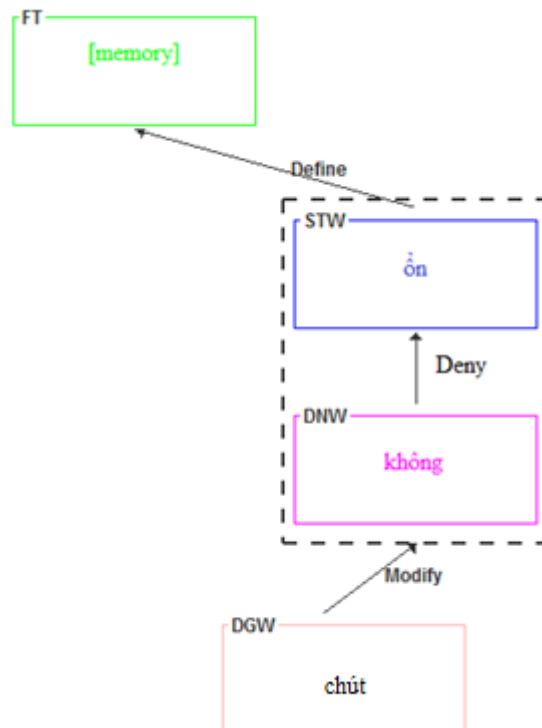
Hình 3.7: Ví dụ minh họa đồ thị về câu nhóm 2

- **Nhóm 3:** Tập C chỉ chứa các loại đối tượng là Feature, SentiWord và DeniedWord.



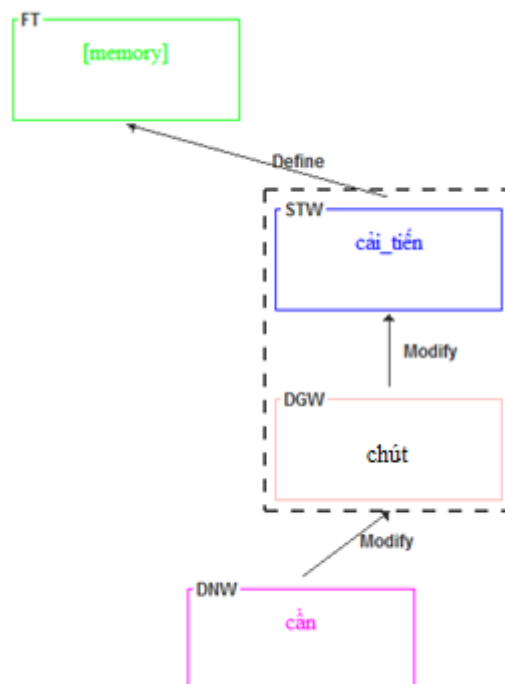
Hình 3.8: Ví dụ minh họa đồ thị về câu nhóm 3

- **Nhóm 4:** Tập C chứa đầy đủ 4 loại đối tượng Feature, SentiWord, DegreeWord và DeniedWord. Đỉnh DegreeWord bổ nghĩa cho đỉnh mệnh đề DeniedWord kết hợp với SentiWord.



Hình 3.9: Ví dụ minh họa đồ thị về câu nhóm 4

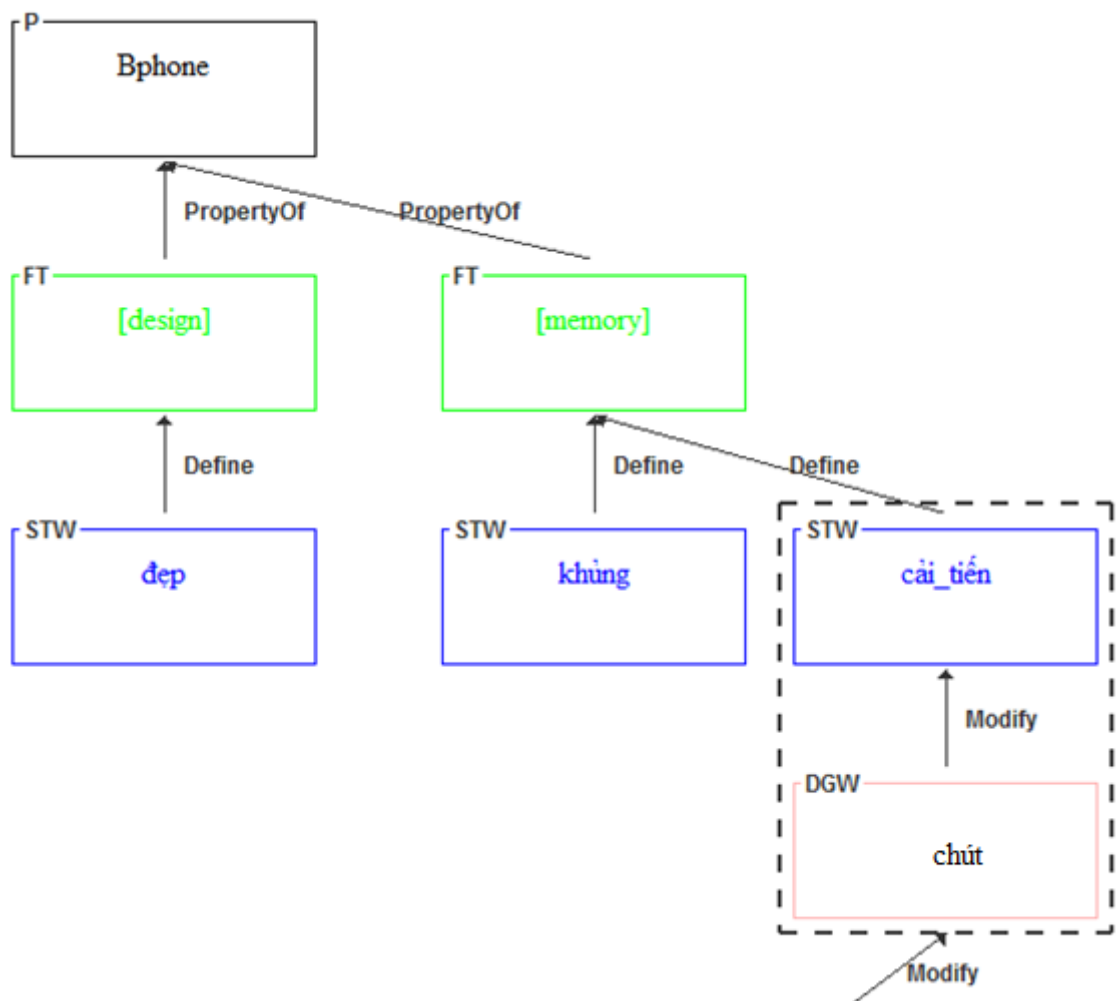
- **Nhóm 5:** Tập C chứa đầy đủ 4 loại đối tượng Feature, SentiWord, DegreeWord và DeniedWord. Đỉnh DeniedWord bỏ nghĩa đỉnh mệnh đề DegreeWord kết hợp với SentiWord.



Hình 3.10: Ví dụ minh họa đồ thị về câu nhóm 5

3.3.2. Mô hình đồ thị khái niệm biểu diễn cho mức câu

Đồ thị khái niệm mức câu được xây dựng bằng cách tổng hợp các đồ thị ở mức vế câu theo từng đỉnh đối tượng Feature xuất hiện trong câu. Vì thế mô hình của đồ thị khái niệm biểu diễn cho mức câu có các thành phần giống như đồ thị biểu diễn cho mức vế câu, chỉ khác ở chỗ: thành phần của tập C được bổ sung thêm đối tượng Product và có nhiều đối tượng Feature hơn.



Hình 3.11: Ví dụ minh họa đồ thị mức câu

- Cấu trúc dữ liệu của một đối tượng đồ thị ở mức câu:
 - + Vocabulary productNode: Đỉnh Product của đồ thị;
 - + ArrayList<FeatureNode> featureNodes: Danh sách đỉnh Feature của đồ thị.

Trong đó:

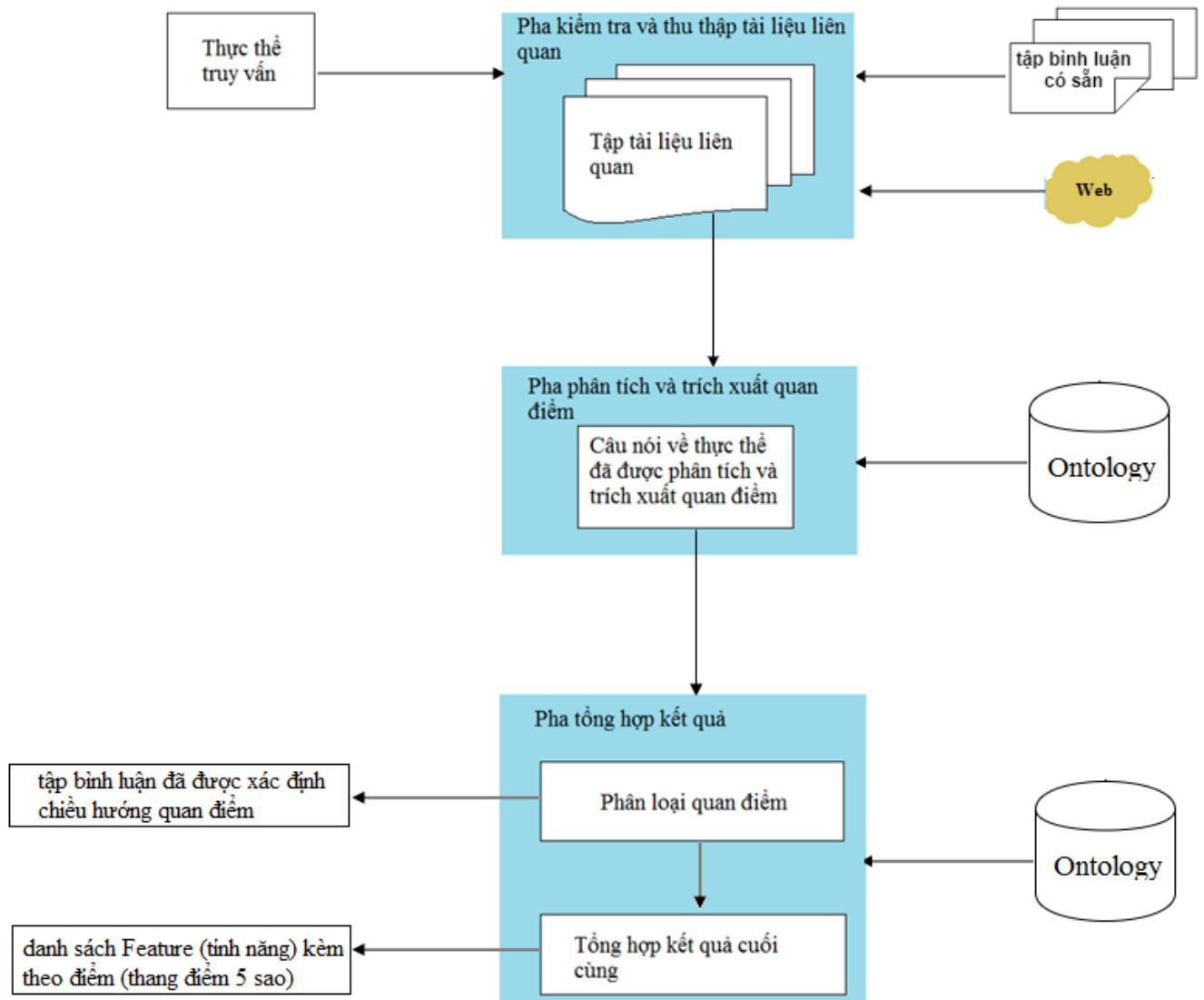
- + Vocabulary là đối tượng từ vựng của câu, bao gồm các thành phần:
 - String content: Nội dung từ vựng bao gồm cả phần từ lẫn phần nhãn;
 - String word: Phần từ của từ vựng;
 - String tag: Phần nhãn của từ vựng.
- + FeatureNode là đối tượng đỉnh Feature của đồ thị, bao gồm các thành phần:
 - Vocabulary ft: Phần từ vựng cho đỉnh;
 - ArrayList<SentiPhraseGraph> sps: Những đồ thị con chỉ chứa các đỉnh SentiWord, DegreeWord và DeniedWord.

Cấu trúc của một đối tượng SentiPhraseGraph như sau:

- Vocabulary stw: Phần từ vựng cho đỉnh SentiWord;
- Vocabulary dgw: Phần từ vựng cho đỉnh DegreeWord;
- Vocabulary dnw: Phần từ vựng cho đỉnh DeniedWord;
- String rule: Luật tương ứng với đồ thị.

3.4. Mô hình đề xuất

Sau khi phân tích các mô hình và phương có liên quan đã được đề cập trong chương 2, nhóm chúng em xin đề xuất mô hình cho bài toán phân tích, trích xuất và phân loại đánh giá người dùng về Sản Phẩm Điện Tử dựa trên mô hình thống kê kết hợp với bộ từ điển Ontology đã được xây dựng riêng cho lĩnh vực này. Mô hình đề xuất như sau:



Hình 3.12: Mô hình đề xuất

Trong đó các thành phần của mô hình được mô tả như sau:

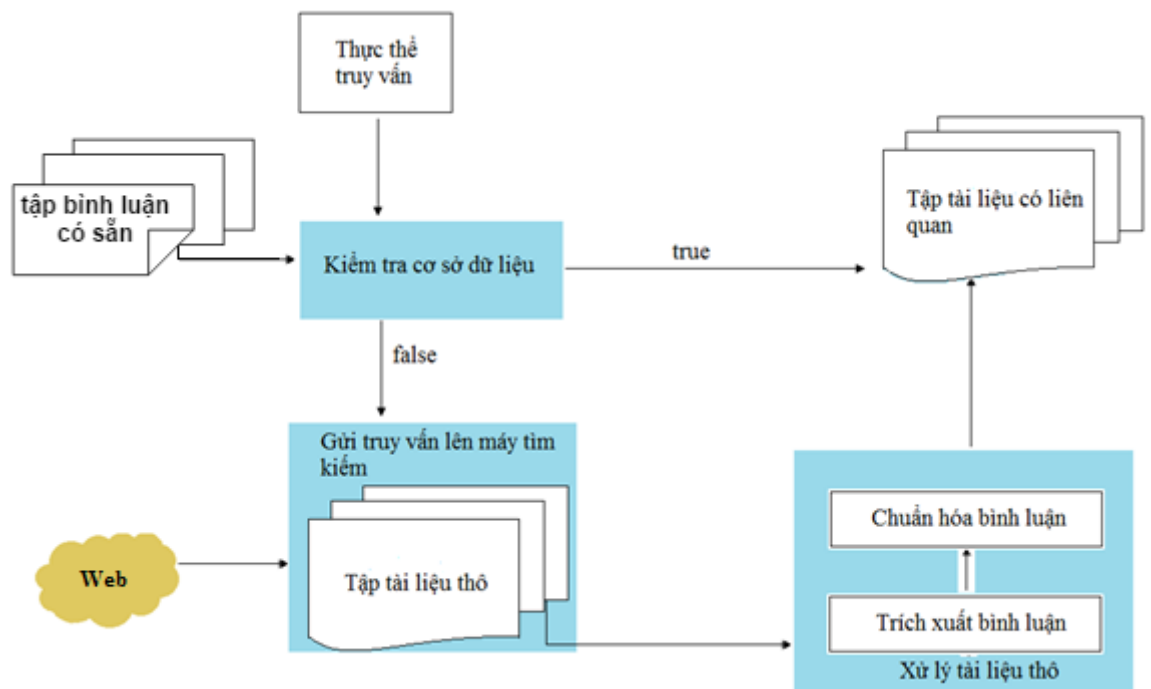
- **Thực thể truy vấn:** Là truy vấn đầu vào của người dùng, cụ thể là tên của một sản phẩm điện thoại di động mà người dùng quan tâm;
- **Pha kiểm tra và thu thập dữ liệu liên quan:** Pha này có nhiệm vụ truy tìm các tài liệu có liên quan đến thực thể truy vấn để làm dữ liệu đầu vào cho pha kế tiếp;
- **Pha phân tích và trích xuất quan điểm:** Pha này có nhiệm vụ phân mỗi đánh giá thành tập hợp các câu và phân tích mỗi câu để rút trích ra các cụm

từ chỉ quan điểm liên quan đến các tính năng của thực thể truy vấn được nhắc đến trong câu và xây dựng đồ thị khái niệm cho mỗi câu.

- **Pha tổng hợp kết quả:** Pha này có nhiệm vụ tính trọng số cho mỗi đánh giá dựa trên trọng số của các cụm từ chỉ quan điểm đã rút trích được, sau đó phân loại đánh giá thông qua trọng số đó. Cuối cùng tổng hợp để tính điểm cho từng tính năng của thực thể truy vấn dựa trên điểm của các cụm từ chỉ quan điểm liên quan đến nó và xuất kết quả tổng hợp được.

3.4.1. Pha kiểm tra và thu thập dữ liệu liên quan

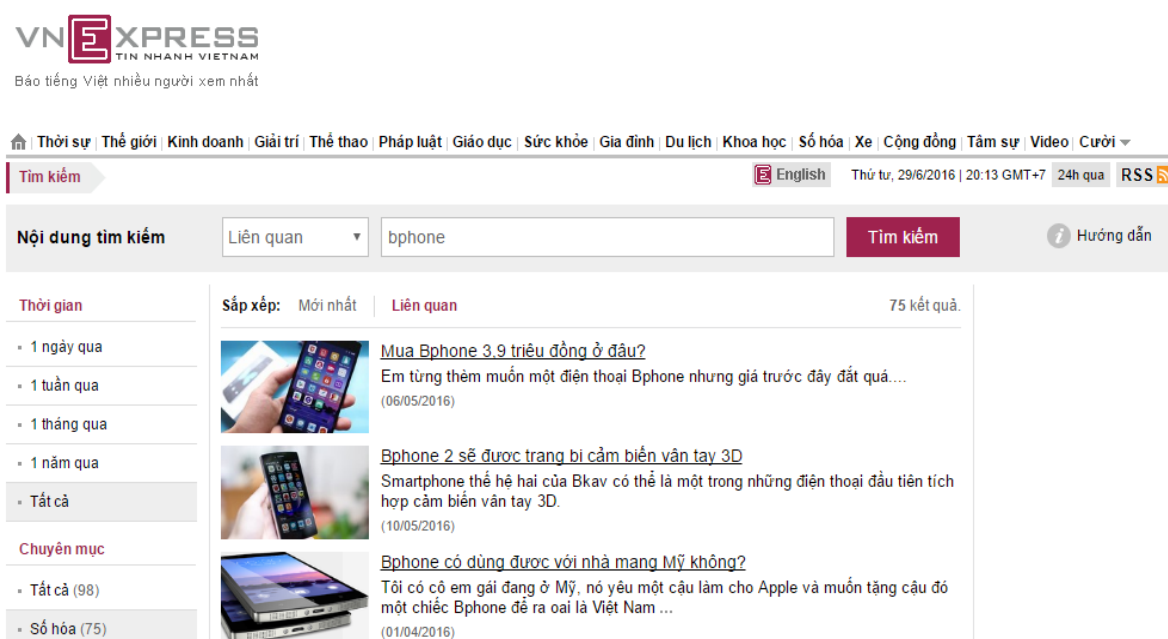
Mô hình của pha kiểm tra và thu thập dữ liệu liên quan:



Hình 3.13: Pha kiểm tra và thu thập tài liệu liên quan

Truy tìm các tài liệu có liên quan đến thực thể truy vấn bằng cách:

- Kiểm tra xem trong cơ sở dữ liệu đã xuất hiện các đánh giá liên quan đến thực thể truy vấn chưa? Nếu có thì không cần thu thập chuyển sang pha phân tích và trích xuất quan điểm, nếu chưa thì tiến hành gửi truy vấn lên máy tìm kiếm để thu thập các đánh giá của người dùng từ trang báo VnExpress.Net. Ở đây nhóm chúng em sử dụng chức năng tìm kiếm của chính trang báo điện tử này. Sau đó nhận được kết quả là một HTML có chứa danh sách mã bài và liên kết của những bài báo có liên quan đến truy vấn.



Hình 3.14: Kết quả tìm kiếm cho “Bphone” trên trang VnExpress.Net

- Duyệt mỗi phần tử trong danh sách và tiến hành tải các bình luận, đánh giá của người dùng ứng với mỗi bài báo từ liên kết sau:

“http://usi.saas.vnexpress.net/index/get?offset=0&limit=” + <số bình luận giới hạn> + “&frommobile=0&sort=like&objectid=” + <mã bài báo> + “&objecttype=1&siteid=1002592&categoryid=1002644”

Kết quả được một chuỗi JSON chứa các bình luận:

```
{
  "error": 0,
  "errorDescription": "",
  "data": {
    "total": 25,
    "totalitem": 32,
    "items": [
      {
        "comment_id": "16444148",
        "parent_id": "16444148",
        "article_id": "3398283",
        "content": "theo mình là thì đang có mấy em xiaomi giá cũng tầm này. lại có thương hiệu hơn nữa",
        "full_name": "anonym",
        "creation_time": "12:51 06/05",
        "userlike": 53,
        "like_ismember": false,
        "userid": null,
        "type": 4
      },
      {
        "comment_id": "16445110",
        "parent_id": "16445110",
        "article_id": "3398283",
        "content": "nếu tất cả mà tốt thì máy đó không lỗi, giá đó chấp nhận được.",
        "full_name": "Binh Minh",
        "creation_time": "12:51 06/05",
        "userlike": 20,
        "like_ismember": false,
        "userid": null,
        "type": 4
      },
      {
        "comment_id": "16450715",
        "parent_id": "16444071",
        "article_id": "3398283",
        "content": "ra đường nhìn thấy ai đó dùng Bphone trên tay , tới nay vẫn chưa biết Bphone như thế nào vư",
        "full_name": "coplambienhoa",
        "creation_time": "06/05",
        "userlike": 12,
        "like_ismember": false,
        "userid": "1003695292",
        "type": 8
      }
    ]
  }
}
```

Hình 3.15: Chuỗi kết quả rút trích bình luận

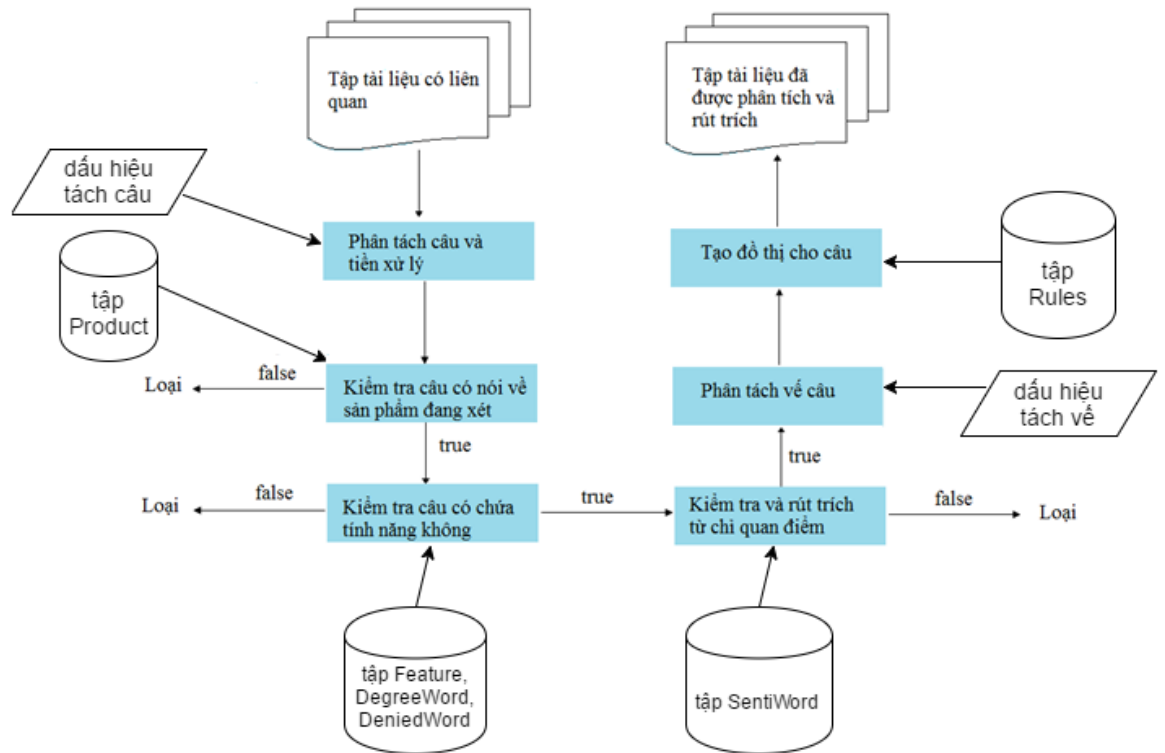
- Các bài toán khóa luận tập trung giải quyết chỉ quan tâm đến nội dung bình luận và bỏ qua các bài toán có liên quan đến tên người bình luận, thời gian đăng bình luận, ... Từ những chuỗi dữ liệu thô rút trích nội dung trong thẻ “content” để lấy nội dung của mỗi bình luận.
- Mỗi nội dung bình luận lấy được có thể chứa:
 - + Các thẻ HTML: ` `; `<`; `>`; `&`; `¢`; `£`; `¥`; `€`; `©`; `®`; `
`, `\`, ...
 - + Các liên kết URL: <http://www>, ...
 - + Tài khoản người dùng: `@Hoàng Nam`, ...

Để tránh gây nhiễu thông tin ta cần chuẩn hóa những thẻ HTML về dạng có thể hiển thị trên máy dùng để phân đoạn câu trong một bình luận hoặc có thể loại bỏ chúng nếu như không có tác dụng dùng để phân tách câu bình luận.

Ví dụ: Xét câu “@Hoàng Nam nếu tất cả mà tốt thì máy đó không lỗi, giá đó chấp nhận được.
Người đã dùng Bphone vì ủng hộ gà nhà.” Sau khi chuẩn hóa được: “nếu tất cả mà tốt thì máy đó không lỗi, giá đó chấp nhận được.\nNgười đã dùng Bphone vì ủng hộ gà nhà.”. Trong đó: “\n” là ký tự xuống dòng được dùng như một dấu hiệu phân tách câu.

3.4.2. Pha phân tích và trích xuất quan điểm

Các bước phân tích và trích xuất quan điểm được mô hình như sau:



Hình 3.16: Mô hình pha phân tích và rút trích quan điểm

Trong đó các thành phần trong mô hình có nhiệm vụ:

- **Phân tách câu và tiền xử lý:** Các bình luận, đánh giá sẽ được phân tách thành tập các câu theo dấu hiệu phân tách câu và mỗi câu sẽ trải qua các bước tiền xử lý.
- **Kiểm tra câu có nói về sản phẩm đang xét:** Thành phần này có nhiệm vụ kiểm tra từng câu trong mỗi bình luận xem câu đó có thật sự là đang nói về sản phẩm đang xét không? Nếu có chuyển qua bước tiếp theo, nếu không thì loại và kết luận câu không chứa quan điểm.
- **Kiểm tra câu có chứa tính năng không:** Thành phần này kiểm tra xem câu đang xét có chứa bất kỳ tính năng nào trong tập Feature không?. Nếu có thì thay thế chúng bằng cấu trúc nhãn và chuyển qua bước tiếp theo, nếu không thì loại và kết luận câu không chứa quan điểm.

- **Kiểm tra và rút trích từ chỉ quan điểm:** Thành phần này kiểm tra xem câu đang xét có chứa từ chỉ quan điểm (được giới hạn chỉ bao gồm những từ nói về các Features được nhắc đến trong câu). Nếu có thì thay thế chúng bằng cấu trúc nhãn và chuyển qua bước tiếp theo, nếu không thì loại và kết luận câu không chứa quan điểm.
- **Phân tách vế câu:** Thành phần này phân tách mỗi câu thành nhiều vế câu theo dấu hiệu tách vế và xử lý để phân tích cấu trúc thông tin.
- **Tạo đồ thị cho câu:** Xây dựng đồ thị cho câu bằng cách tổng hợp các đồ thị ở mức vế câu theo từng đỉnh đối tượng Feature xuất hiện trong câu.

3.4.2.1. Phân tách câu và tiền xử lý

Mỗi bình luận, đánh giá sẽ được phân tách thành tập các câu theo dấu hiệu kết thúc câu như: “.”, “!”, “?”, “\n”, ... Ví dụ: Xét bình luận: “Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay. Mặt dưới hơi cong nên có độ tì tốt khi dùng bằng 1 tay.”, được phân tách thành: “Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay”, “Mặt dưới hơi cong nên có độ tì tốt khi dùng bằng 1 tay”.

Từ kết quả phân tách tiến hành tiền xử lý từng câu bình luận. Các bước tiền xử lý bao gồm:

- **Bước 1:** Tìm và thay thế những từ nói về tên sản phẩm đang xét xuất hiện trong câu bằng tên đại diện. Ví dụ: Xét câu “Z3 rất đẹp” chuyển thành “Sony Xperia Z3 rất đẹp”.
- **Bước 2:** Tìm các từ so sánh xuất hiện trong câu và thay thế nội dung của chúng bằng thuộc symbol. Ví dụ: “Bphone đẹp hơn Iphone” chuyển thành “Bphone đẹp > Iphone”.
- **Bước 3:** Xóa các ký hiệu biểu lộ cảm xúc, tìm và thay thế các dấu hiệu phân tách vế câu bởi dấu “,”. Ví dụ: “Nhìn đẹp và sang trọng ☺” chuyển thành “Nhìn đẹp , sang trọng”.

- **Bước 4:** Phân tách câu thành những từ, cụm từ và gán nhãn phân loại từ vựng sử dụng vnTagger.
- **Bước 5:** Thay thế tên đại diện của sản phẩm đang xét bằng “tproduct”. Tìm và thay thế tên các sản phẩm khác xuất hiện trong câu bằng “oproduct”. Thay thế nhãn của các symbol so sánh bằng “/Cw”. Ví dụ: Ta có sản phẩm đang xét là Bphone, câu “Bphone nhìn đẹp hơn Iphone” chuyển thành “tproduct/N nhìn/V đẹp/A >/Cw oproduct/N”.

Một số ký hiệu biểu lộ cảm xúc thường xuất hiện như: ☺ , ☹ ;) :D ;) :x :”> :P =(x(☹ ...

3.4.2.2. Kiểm tra câu có nói về sản phẩm đang xét

Câu bình luận xuất hiện trong bài báo nói về sản phẩm đang xét có thể chưa chắc là đang nói về sản phẩm đó, nó có thể là nhắc đến sản phẩm khác hoặc đang so sánh sản phẩm đang xét với sản phẩm khác. Vì thế ta cần phải đảm bảo các câu trong bình luận phải nói về sản phẩm đó thì mới tính là câu chứa quan điểm.

Các bước để kiểm tra như sau:

- **Bước 1:** Kiểm tra xem câu có chứa “tproduct” không?
 - + **1.1:** Nếu có thì tiếp tục kiểm tra xem câu có chứa “oproduct” không?
 - Nếu có thì kiểm tra xem câu có chứa “>”, “<” hoặc “=” không?
 - Nếu có thì kiểm tra xem “tproduct”, “oproduct” và dấu so sánh xuất hiện theo thứ tự nào. Nếu xuất hiện trường hợp “oproduct”, dấu so sánh, “tproduct” thì đổi chỗ “oproduct” với “tproduct” cho nhau và đổi chiều dấu so sánh “>” thành “<” và ngược lại. Nếu dấu so sánh là dấu “<” thì thêm từ “không/R” vào sau “tproduct”. Cuối cùng trả về true.
 - Nếu không thì trả về false.
 - Nếu không thì trả về true.

- + **1.2:** Nếu không thì tiếp tục kiểm tra xem câu có chứa “oproduct” không?
 - Nếu có thì kiểm tra xem câu có chứa “>”, “<” hoặc “=” không?
 - Nếu có thì kiểm tra xem dấu so sánh có đứng trước “oproduct” không? Nếu không thì đổi chiều dấu so sánh. Nếu dấu so sánh là “<” thì thêm “không/R” vào đầu câu nếu từ “oproduct” đứng trước hoặc thay thế “oproduct” bằng “không/R” nếu “oproduct” đứng sau. Cuối cùng trả về true.
 - Nếu không thì trả về false.
 - Nếu không thì trả về true.

– **Bước 2:** Kết luận.

- + true: Khẳng định câu có nói về sản phẩm đang xét;
- + false: Khẳng định câu không nói về sản phẩm đang xét, đồng thời dừng phân tích.

Ví dụ: Xét câu đã được tiền xử lý “tproduct/N đẹp/A </Cw oproduct/N” sau khi kiểm tra và kết luận true thì câu trở thành “tproduct/N không/R đẹp/A </Cw oproduct/N”.

3.4.2.3. Kiểm tra câu có chứa tính năng không

Chỉ thực hiện bước này nếu bước kiểm tra sản phẩm có giá trị trả về là true. Để kiểm tra câu có chứa tính năng sản phẩm hay không ta làm như sau:

- **Bước 1:** Tìm và thay thế nhãn của các từ chỉ mức độ bằng “/Dgw”, các từ phủ định bằng “/Dnw” và các từ liên kết bằng “/Rw”;
- **Bước 2:** Duyệt từng phần tử trong tập Feature trong Ontology:
 - + **2.1:** Duyệt từng phần tử i trong list_indicator của f , kiểm tra xem câu có chứa i không?

Nếu có thì thay nhãn của i thành “/Ft” và i thành [<Tên đại diện của f >];

Thêm f vào danh sách feature của câu;

- + **2.2:** Chỉ thực hiện khi câu chứa “,” hoặc câu không chứa [<Tên đại diện của f>]:

Duyệt từng phần tử *i* trong *list_hidden_indicator* của *f*, kiểm tra xem câu có chứa *i* không?

Nếu có chứa thì thêm phần tử [<Tên đại diện của f>]/Ft vào câu ở vị trí trước *i* nếu trước *i* không có từ chỉ mức độ hay từ phủ định, vào vị trí trước từ cụm <từ chỉ mức độ> + *i* hoặc <từ phủ định> + <từ chỉ mức độ> + *i* hoặc <từ phủ định> + *i* nếu trước *i* có chứa các từ chỉ mức độ hay từ phủ định;

Thêm *f* vào danh sách feature của câu;

- **Bước 3:** Nếu danh sách Feature của câu khác rỗng thì trả về true, ngược lại trả về false;
- **Bước 4:** Kết luận.
 - + true: Khẳng định câu có chứa tính năng;
 - + false: Khẳng định câu không có chứa tính năng, đồng thời dừng phân tích.

Ví dụ: Xét câu “Bphone khá đẹp và sang trọng” qua bước kiểm tra này sẽ trở thành “tproduct/N [Design]/Ft khá/Dgw đẹp/A , sang_trọng/A” và trả về giá trị true.

3.4.2.4. Kiểm tra và rút trích từ chỉ quan điểm

Chỉ thực hiện bước này khi bước kiểm tra tính năng đúng. Kiểm tra và rút trích từ chỉ quan điểm trong câu được tiến hành như sau:

- **Bước 1:** Duyệt mỗi phần tử *s* của tập SentiWord trong Ontology:
Nếu câu chứa thuộc tính word của *s* và danh sách feature của câu chứa thuộc tính feature của *s* thì thực hiện thay nhãn của word trong câu bằng “/Stw”;
- **Bước 2:** Thay các cấu trúc <từ chỉ quan điểm 1>/Stw “,” <từ chỉ quan điểm 2>/Stw và [<tên đại diện feature 1>]/Ft “,” [<tên đại diện feature

2>]/Ft thành <từ chỉ quan điểm 1>+<từ chỉ quan điểm 2>/Stw và [<tên đại diện feature 1>+<tên đại diện feature 2>]/Ft;

- **Bước 3:** Nếu trong câu xuất hiện “/Stw” thì trả về true, ngược lại trả về false;
- **Bước 4:** Kết luận.
 - + true: Khẳng định câu có chứa từ quan điểm;
 - + false: Khẳng định câu không có chứa từ quan điểm, đồng thời dừng phân tích.

Ví dụ: Xét câu “Máy đẹp và cá tính” chuyển thành “Máy/N [Design]/Ft đẹp+cá_tính/Stw”.

3.4.2.5. Phân tách vế câu

Chỉ thực hiện bước này khi bước kiểm tra và rút trích từ chỉ quan điểm trả về true. Mỗi câu lại được phân tách thành nhiều vế theo dấu hiệu phân tách vế như: “,” , “;” , “và” , “tuy” , “nhưng” , “không những” , “mà còn” , ...

Xét tập vế câu ta tiến hành các bước xử lý sau:

- **Bước 1:** Kiểm tra từng vế câu xem có chứa “/Stw” không? Nếu không thì loại bỏ khỏi tập vế câu;
- **Bước 2:** Kiểm tra xem mỗi vế câu có nói về sản phẩm đang xét không? (thực hiện như ở bước kiểm tra của mức câu). Nếu không đúng thì loại vế câu khỏi tập vế câu;
- **Bước 3:** Duyệt mỗi vế câu lấy ra tập từ vựng là những từ có nhãn là: “Ft” , “Stw” , “Dgw” , “Dnw” và “Rw”;
- **Bước 4:** Nếu số phần tử của tập vế câu lớn hơn hoặc bằng 2 thì xét xem phần tử đầu tiên có chứa phần tử nhãn “Ft” không? Nếu không thì thêm phần tử nhãn “Ft” của phần tử vế câu thứ 2 vào tập từ vựng của phần tử đầu tiên đó;

- **Bước 5:** Duyệt mỗi về câu trong tập về câu, kiểm tra xem có chứa phần tử nhãn “Ft” không? Nếu không thì thêm phần tử nhãn “Ft” của phần tử về câu liền trước nó;
- **Bước 6:** Duyệt mỗi về câu tạo chuỗi có cấu trúc <nhãn từ vựng 1> + <nhãn từ vựng 2> + <nhãn từ vựng 3> + ... Với mỗi chuỗi vừa tạo xem có ứng với luật nào trong tập luật Rules của Ontology không để thay giả thiết của luật đó trong chuỗi bằng kết luận của luật và loại các thành phần thừa như: ft, dgw, dnw, rw, stw, các luật không dùng để tạo đồ thị ra khỏi chuỗi. Sau đó khởi tạo cho các thành phần đồ thị tương ứng với luật đó;

Ví dụ: Xét câu “Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay” có kết quả rút trích quan điểm là:

“Đã/R được/V cầm/V thử/V chiếc/Nc tproduct/N ./, [design]/Ft bên/N ngoài/N khá/Dgw bóng/Stw ./, [design]/Ft sang_trọng/Stw ./, cầm/V rất/Dgw chắc_tay/Stw ” và kết quả phân tích về câu như sau:

STT	Nội dung	Tập từ vựng	Chuỗi cấu trúc nhãn chưa áp luật	Chuỗi đã áp luật
1	[design]/Ft bên/N ngoài/N khá/Dgw bóng/Stw	[design]/Ft, khá/Dgw, bóng/Stw	Ft + Dgw + Stw	r11
2	[design]/Ft sang_trọng/Stw	[design]/Ft, sang_trọng/Stw	Ft + Stw	r08
3	[design]/Ft cầm/v rất/Dgw	[design]/Ft, rất/Dgw,	Ft + Dgw + Stw	r11

	chắc_tay/Stw	chắc_tay/Stw		
--	--------------	--------------	--	--

Bảng 3.4: Ví dụ minh họa bước phân tách vế câu

3.4.2.6. Tạo đồ thị cho câu

Chỉ thực hiện bước này nếu tập vế câu của câu không rỗng. Các bước tạo đồ thị cho câu như sau:

- **Bước 1:** Khởi tạo đỉnh Product cho đồ thị với nội dung là tên đại diện của sản phẩm đang xét;
- **Bước 2:** Duyệt mỗi phần tử f trong danh sách feature của câu để lấy ra những vế câu có chứa f và tạo một đồ thị con ứng với f chứa nhiều đồ thị con SentiPhraseWord;

3.4.3. Pha tổng hợp kết quả

Mô hình của pha tổng hợp kết quả như sau:

- **Bước 1:** Tính trọng số cho mỗi bình luận trong tập bình luận liên quan đến sản phẩm đang xét mà đã được phân tích và rút trích quan điểm;
- **Bước 2:** Phân loại chiều hướng quan điểm (tích cực, tiêu cực hay trung lập) cho mỗi bình luận;
- **Bước 3:** Tổng hợp, thống kê các bình luận này theo từng tính năng của sản phẩm và tính điểm cho mỗi tính năng.

3.4.3.1. Tính trọng số và phân loại chiều hướng quan điểm cho bình luận:

Sau khi phân tích và trích xuất quan điểm, ứng với mỗi vế câu trong từng câu của mỗi bình luận ta có cụm từ chỉ quan điểm của mỗi vế đó. Đó là một bộ gồm có từ chỉ quan điểm, từ chỉ mức độ và từ phủ định. Trong đó thành phần từ chỉ quan điểm luôn luôn phải tồn tại và các thành phần còn lại có thể có hoặc không tùy theo cấu trúc thông tin của câu. Và thứ tự xuất hiện các thành phần này trong cụm cũng ảnh hưởng đến cách tính trọng số cho cụm.

Ta có công thức tính điểm cho mỗi cụm từ quan điểm như sau:

$$ts = fp * s * fs$$

Trong đó:

- **ts**: Trọng số của cụm từ chỉ quan điểm;
- **s**: Trọng số từ chỉ quan điểm;
- **fs**: Trọng số từ chỉ mức độ;
- **fp**: Trọng số từ phủ định. Giá trị này phụ thuộc vào thứ tự xuất hiện của nó trong cụm: Nếu đứng trước từ quan điểm (cụm không chứa từ chỉ mức độ) thì giá trị là -1. Nếu đứng trước từ chỉ mức độ hay từ chỉ quan điểm + từ chỉ mức độ thì giá trị được lấy từ trong tập DeniedWord của Ontology).

Trọng số của mỗi vế câu cũng chính là trọng số của cụm từ chỉ quan điểm xuất hiện trong vế câu đó. Sau khi tính trọng số cho từng vế câu, ta tiến hành tính trọng số cho mức câu bằng công thức:

$$S_s = \sum ts_i$$

Trong đó:

- **S_s**: Trọng số của câu;
- **ts_i**: Trọng số của cụm từ quan điểm hay vế câu thứ i.

Ở mức bình luận, ta có công thức tính trọng số sau:

$$S_p = \sum S_{s_i}$$

Trong đó:

- **S_p**: Trọng số của bình luận;
- **S_{s_i}**: Trọng số của câu thứ i.

Xét trọng số của mỗi bình luận và xác định chiều hướng quan điểm theo các trường hợp sau:

- $S_p > 0.5$: Bình luận tích cực;
- $-0.5 \leq S_p \leq 0.5$: Bình luận trung lập;
- $S_p < -0.5$: Bình luận tiêu cực.

3.4.3.2. Thống kê kết quả theo tính năng sản phẩm:

Xét mỗi bình luận đã tính trọng số, tương ứng với mỗi tính năng sản phẩm xuất hiện trong câu bình luận ta tính trọng số cho mỗi tính năng bằng cách lấy ra những cụm từ chỉ quan điểm nói về tính năng đó trong câu và tính trọng số tính năng ở mức câu theo công thức:

$$S_{f_s} = \sum ts_i$$

Trong đó:

- S_{f_s} : Trọng số tính năng ở mức câu;
- ts_i : Trọng số của cụm từ chỉ quan điểm thứ i nói về tính năng trong câu.

Sau đó, ta tính trọng số cho tính năng đang xét ở mức bình luận theo công thức:

$$S_{f_p} = \sum S_{f_{s_i}}$$

Trong đó:

- S_{f_p} : Trọng số tính năng ở mức bình luận;
- $S_{f_{s_i}}$: Trọng số tính năng của câu thứ i có nói về tính năng đó.

Tương tự như phân loại chiều hướng quan điểm cho mỗi bình luận, ta cũng phân loại chiều hướng quan điểm cho tính năng đang xét ở mức bình luận. Tức là

xét một bình luận i có chứa tính năng j , xem j được nhắc đến với chiều hướng tích cực ($S_{fp} > 0.5$), tiêu cực ($S_{fp} < -0.5$) hay trung lập ($-0.5 \leq S_{fp} \leq 0.5$). Không dùng trực tiếp chiều hướng quan điểm đã xác định cho mỗi bình luận để làm chiều hướng quan điểm cho tính năng xuất hiện trong bình luận là do: một bình luận có thể nói về nhiều tính năng và từng tính năng lại có chiều hướng quan điểm khác nhau.

Từ kết quả tính trọng số và xác định chiều hướng quan điểm cho tính năng đang ở tất cả bình luận chứa nó, ta tính điểm cuối cùng cho tính năng đó và quy ra phần trăm để đánh số sao cho tính năng bằng công thức:

$$\frac{\text{Số câu chiều hướng tích cực} + \text{Số câu chiều hướng trung lập} / 2}{\text{Tổng số bình luận nhắc đến tính năng}}$$

Mỗi tính năng sẽ được tính tối đa 5 sao và cứ mỗi 20% thì tương ứng được 1 sao. Ví dụ: Xét tính năng Design của một sản phẩm nào đó, có số chiều hướng tích cực là 45, tổng số bình luận nhắc đến nó là 90 thì quy đổi ra điểm phần trăm là 50% tương ứng với 2.5 sao.

Chương 4. CÀI ĐẶT VÀ THỬ NGHIỆM

4.1. Ngôn ngữ sử dụng và công cụ hỗ trợ

4.1.1. Ngôn ngữ sử dụng

Ngôn ngữ sử dụng để cài đặt chương trình ứng dụng là ngôn ngữ lập trình Java.

Java là một ngôn ngữ lập trình hướng đối tượng (OOP) và dựa trên các lớp (class). Khác với phần lớn ngôn ngữ lập trình thông thường, thay vì biên dịch mã nguồn thành mã máy hoặc thông dịch mã nguồn khi chạy, Java được thiết kế để biên dịch mã nguồn thành bytecode, bytecode sau đó sẽ được môi trường thực thi (runtime environment) chạy.

Trước đây, Java chạy chậm hơn những ngôn ngữ dịch thẳng ra mã máy như C và C++, nhưng sau này nhờ công nghệ "biên dịch tại chỗ" - Just in time compilation, khoảng cách này đã được thu hẹp, và trong một số trường hợp đặc biệt Java có thể chạy nhanh hơn. Java chạy nhanh hơn những ngôn ngữ thông dịch như Python, Perl, PHP gấp nhiều lần. Java chạy tương đương so với C#, một ngôn ngữ khá tương đồng về mặt cú pháp và quá trình dịch/chạy.

Cú pháp Java được vay mượn nhiều từ C & C++ nhưng có cú pháp hướng đối tượng đơn giản hơn và ít tính năng xử lý cấp thấp hơn. Do đó việc viết một chương trình bằng Java dễ hơn, đơn giản hơn, đỡ tốn công sửa lỗi hơn.

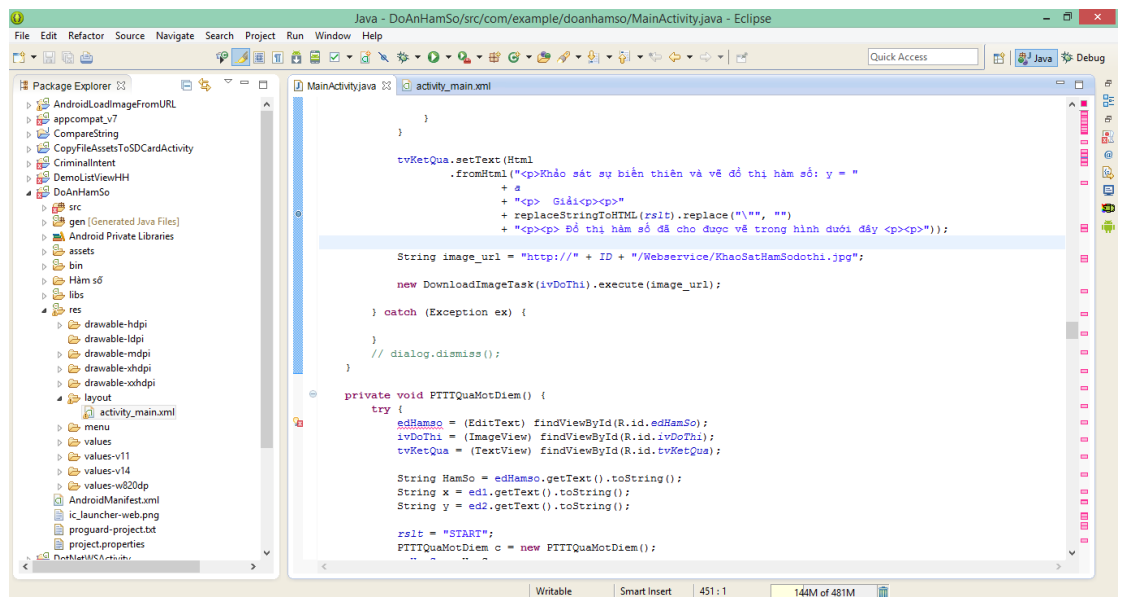
Trong Java, hiện tượng rò rỉ bộ nhớ hầu như không xảy ra do bộ nhớ được quản lý bởi Java Virtual Machine (JVM) bằng cách tự động "dọn dẹp rác". Người lập trình không phải quan tâm đến việc cấp phát và xóa bộ nhớ như C, C++. Tuy nhiên khi sử dụng những tài nguyên mạng, file IO, database (nằm ngoài kiểm soát của JVM) mà người lập trình không đóng (close) các streams thì rò rỉ dữ liệu vẫn có thể xảy ra.

4.1.2. Công cụ hỗ trợ

- **Công cụ lập trình: Eclipse IDE.**

Eclipse là một môi trường phát triển tích hợp cho Java, được phát triển ban đầu bởi IBM, và hiện nay bởi tổ chức Eclipse.

Ngoài Java, Eclipse còn hỗ trợ các ngôn ngữ lập trình khác như PHP, C, C++, C#, Python, HTML, XML, JavaScript khi dùng thêm trình bổ sung (*plug-in*).



Hình 4.1: Giao diện Eclipse

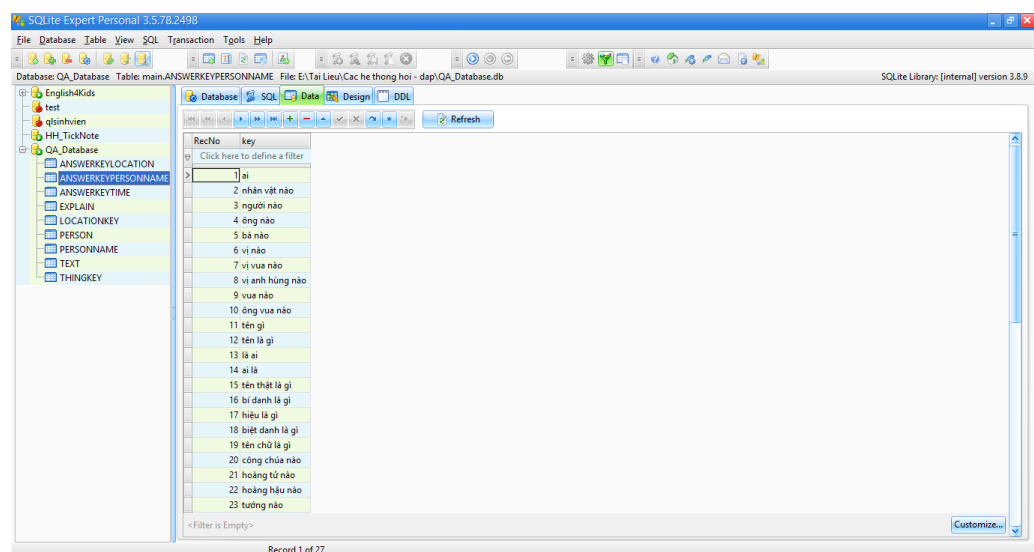
- **Công cụ quản lý và duy trì cơ sở dữ liệu: SQLite Expert Personal**

(version 3.5.78.2498). SQLite Expert Personal là chương trình giúp người dùng dễ dàng quản lý các cơ sở dữ liệu SQLite3. SQLite Expert tích hợp trình quản lý và duy trì cơ sở dữ liệu vào một môi trường chung duy nhất với giao diện đồ họa trực quan, dễ sử dụng với những đặc điểm sau:

- + Hiệu chỉnh bảng biểu và xem trực quan, không có một dòng văn bản của SQL. Một cách dễ dàng nhà trường, lập chỉ mục, khó khăn, triggers mà không bị mất dữ liệu đã tồn tại trong bảng..
- + Xây dựng và tạo ra các script SQL xem trực quan bằng cách sử dụng tích hợp Query Builder.

- + Tạo SQLite3 cơ sở dữ liệu, xem và thay đổi các tham số cơ sở dữ liệu, kiểm tra cơ sở dữ liệu và tính liên chân không (compact) cơ sở dữ liệu.
- + Dễ dàng chuyển dữ liệu giữa các cơ sở dữ liệu SQLite, nhập dữ liệu từ các tập lệnh SQL ADO hoặc nguồn dữ liệu, hoặc xuất khẩu sang các SQL script.
- + Hiển thị và chỉnh sửa các dữ liệu trong lưới điện, bao gồm cả hình ảnh và các lĩnh vực BLOB. Hiện nay hỗ trợ BMP, JPG và các định dạng hình ảnh PNG. BLOB lĩnh vực có thể được sửa đổi với việc tích hợp trình soạn thảo hex.
- + Thực hiện Xóa các truy vấn SQL và hiển thị kết quả trong các lưới điện hoặc là văn bản.

Giao diện chính của **SQLite Expert Personal (version 3.5.78.2498)**:



Hình 4.2: Giao diện SQLite Expert Personal (version 3.5.78.2498)

- **Công cụ dùng để phân đoạn từ trong Tiếng Việt: vnTokenizer của tác giả Lê Hồng Phương.** vnTokenizer là công cụ tự động phân đoạn văn bản thành các đơn vị từ và cụm từ trong Tiếng Việt. Nó được phát triển dưới dạng thư viện hỗ trợ viết bằng ngôn ngữ lập trình Java. Công cụ này có kết quả phân tách từ đạt độ chính xác trong khoảng từ 96% đến 98%.

- **Công cụ gán nhãn từ loại Tiếng Việt: vnTagger** của tác giả **Lê Hồng Phương**. **vnTagger** là công cụ tự động gán nhãn thẻ từ loại cho các từ vựng Tiếng Việt trong văn bản. Nó được phát triển dưới dạng thư viện hỗ trợ viết bằng ngôn ngữ lập trình Java. Kết quả chính xác nhất khi văn bản đã được phân đoạn từ vựng nhờ công cụ vnTokenizer. Độ chính xác của công cụ đạt từ 94% đến 95%.

4.2. Tổ chức lưu trữ Ontology và dữ liệu mẫu trên máy tính

4.2.1. Bảng Product

Bảng Product gồm các thuộc tính như sau: id (mã sản phẩm), name (tên đại diện) và list_name (danh sách tên thay thế).

Index	Name	Declared Type	Type	Size	Precision	Not Null
1	id	INT	INT	0	0	<input type="checkbox"/>
2	name	TEXT	TEXT	0	0	<input type="checkbox"/>
3	list_name	TEXT	TEXT	0	0	<input type="checkbox"/>

Hình 4.3: Bảng Product

4.2.2. Bảng Feature

Bảng Feature gồm các thuộc tính như sau: id (mã tính năng), name (tên đại diện), list_indicator (danh sách dấu hiệu nhận biết dạng hiện) và list_hidden_indicator (danh sách dấu hiệu nhận biết dạng ẩn).

Index	Name	Declared Type	Type	Size	Precision	Not Null
1	id	text	text	0	0	<input type="checkbox"/>
2	name	text	text	0	0	<input type="checkbox"/>
3	list_indicator	text	text	0	0	<input type="checkbox"/>
4	list_hidden_indicator	text	text	0	0	<input type="checkbox"/>

Hình 4.4: Bảng Feature

4.2.3. Bảng SentiWord

Bảng SentiWord gồm các thuộc tính như sau: id (mã từ), word (nội dung từ), feature (tên đại diện tính năng liên quan) và score (điểm).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	TEXT	TEXT	0	0	<input type="checkbox"/>
2	word	TEXT	TEXT	0	0	<input type="checkbox"/>
3	feature	TEXT	TEXT	0	0	<input type="checkbox"/>
4	score	FLOAT	FLOAT	0	0	<input type="checkbox"/>

Hình 4.5: Bảng SentiWord

4.2.4. Bảng DegreeWord

Bảng DegreeWord gồm các thuộc tính: id (mã từ), word (nội dung từ) và score (điểm).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	TEXT	TEXT	0	0	<input type="checkbox"/>
2	word	TEXT	TEXT	0	0	<input type="checkbox"/>
3	score	FLOAT	FLOAT	0	0	<input type="checkbox"/>

Hình 4.6: Bảng DegreeWord

4.2.5. Bảng DeniedWord

Bảng DeniedWord gồm các thuộc tính: id (mã từ), word (nội dung từ) và score (điểm).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	TEXT	TEXT	0	0	<input type="checkbox"/>
2	word	TEXT	TEXT	0	0	<input type="checkbox"/>
3	score	FLOAT	FLOAT	0	0	<input type="checkbox"/>

Hình 4.7: Bảng DeniedWord

4.2.6. Bảng ComparisionWord



Bảng ComparisionWord gồm các thuộc tính sau: id (mã từ), word (nội dung từ) và symbol (dấu biểu tượng).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	INT	INT	0	0	<input type="checkbox"/>
2	word	TEXT	TEXT	0	0	<input type="checkbox"/>
3	symbol	text	text	0	0	<input type="checkbox"/>

Hình 4.8: Bảng ComparisionWord

4.2.7. Bảng ReferWord


Bảng ReferWord gồm các thuộc tính sau: id (mã từ) và word (nội dung từ).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	TEXT	TEXT	0	0	
2	word	TEXT	TEXT	0	0	

Hình 4.9: Bảng ReferWord

4.2.8. Bảng Relation






Bảng Relation gồm các thuộc tính như sau: id (mã quan hệ), left_object (đối tượng 1), right_object (đối tượng 2), label (nhãn quan hệ) và explain (lời chú giải).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	TEXT	TEXT	0	0	
2	left_object	TEXT	TEXT	0	0	
3	right_object	TEXT	TEXT	0	0	
4	label	TEXT	TEXT	0	0	
5	explain	TEXT	TEXT	0	0	

Hình 4.10: Bảng Relation

4.2.9. Bảng Rule

Bảng Rule gồm các thuộc tính như sau: id (mã luật), left_content (giả thiết luật), right_content (kết luận luật), group (ký hiệu nhóm đồ thị) và example (ví dụ minh họa).

Index	Name	Declared Type	Type	Size	Precision	Not Null
> 1	id	int	int	0	0	
2	left_content	text	text	0	0	
3	right_content	text	text	0	0	
4	group	TEXT	TEXT	0	0	
> 5	example	text	text	0	0	

Hình 4.11: Bảng Rule

4.2.10. Bảng Comment

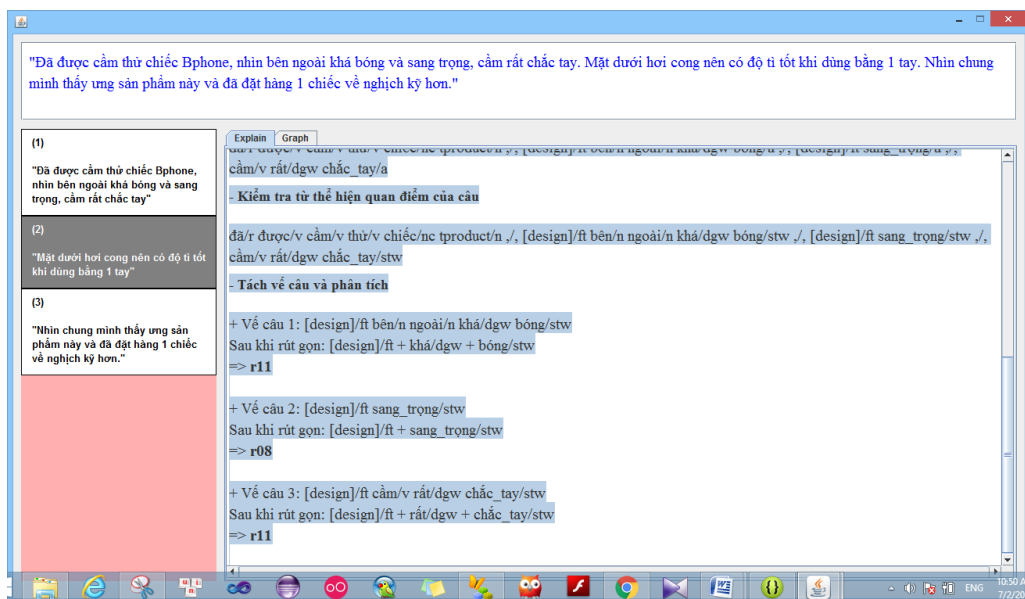
Bảng Comment chứa tập bình luận, đánh giá thử nghiệm về các dòng sản phẩm điện thoại di động, gồm các thuộc tính như sau: id (mã bình luận), product_name (tên đại diện sản phẩm), original_content (nội dung nguyên thủy), handle_type (chiều hướng quan điểm đúng của bình luận), option_type (chiều hướng quan điểm do ứng dụng đưa ra) và acticle_id (mã bài báo).

Index	Name	Declared Type	Type	Size	Precision	Not Null
1	id	int	int	0	0	<input type="checkbox"/>
2	product_name	text	text	0	0	<input type="checkbox"/>
3	original_content	text	text	0	0	<input type="checkbox"/>
4	handle_type	INT	INT	0	0	<input type="checkbox"/>
5	option_type	INT	INT	0	0	<input type="checkbox"/>
6	article_id	TEXT	TEXT	0	0	<input type="checkbox"/>

Hình 4.12: Bảng Comment

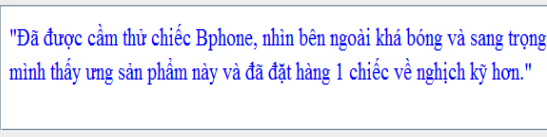
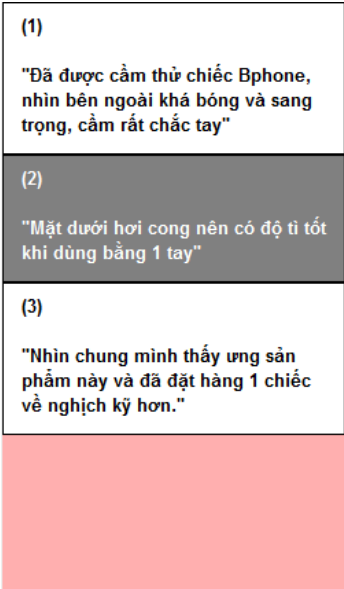
4.3. Giới thiệu giao diện chương trình ứng dụng

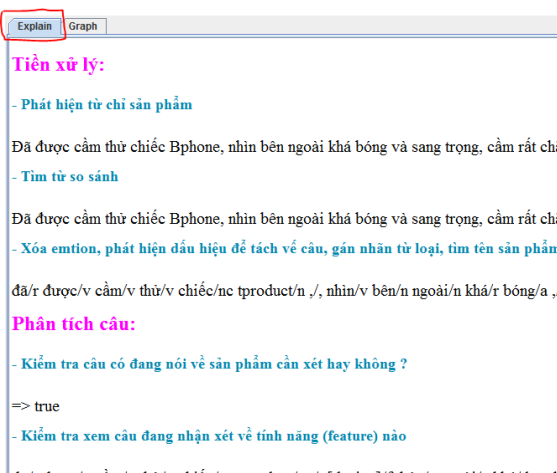
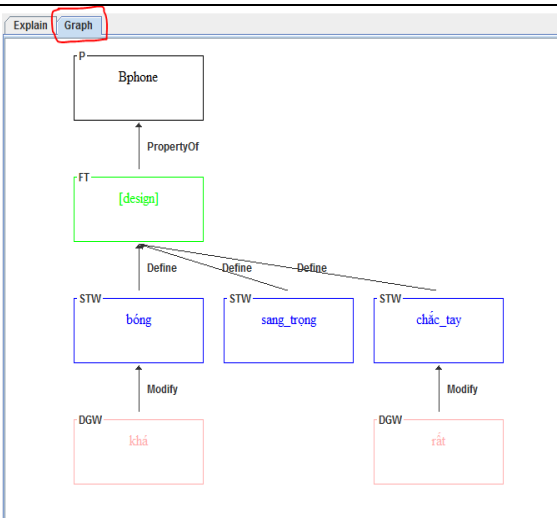
4.3.1. Frame hiển thị kết quả phân tích và rút trích quan điểm của một bình luận



Hình 4.13: Giao diện Frame hiển thị kết quả phân tích một bình luận

Sau đây là danh sách các thành phần của Frame và chức năng của chúng:

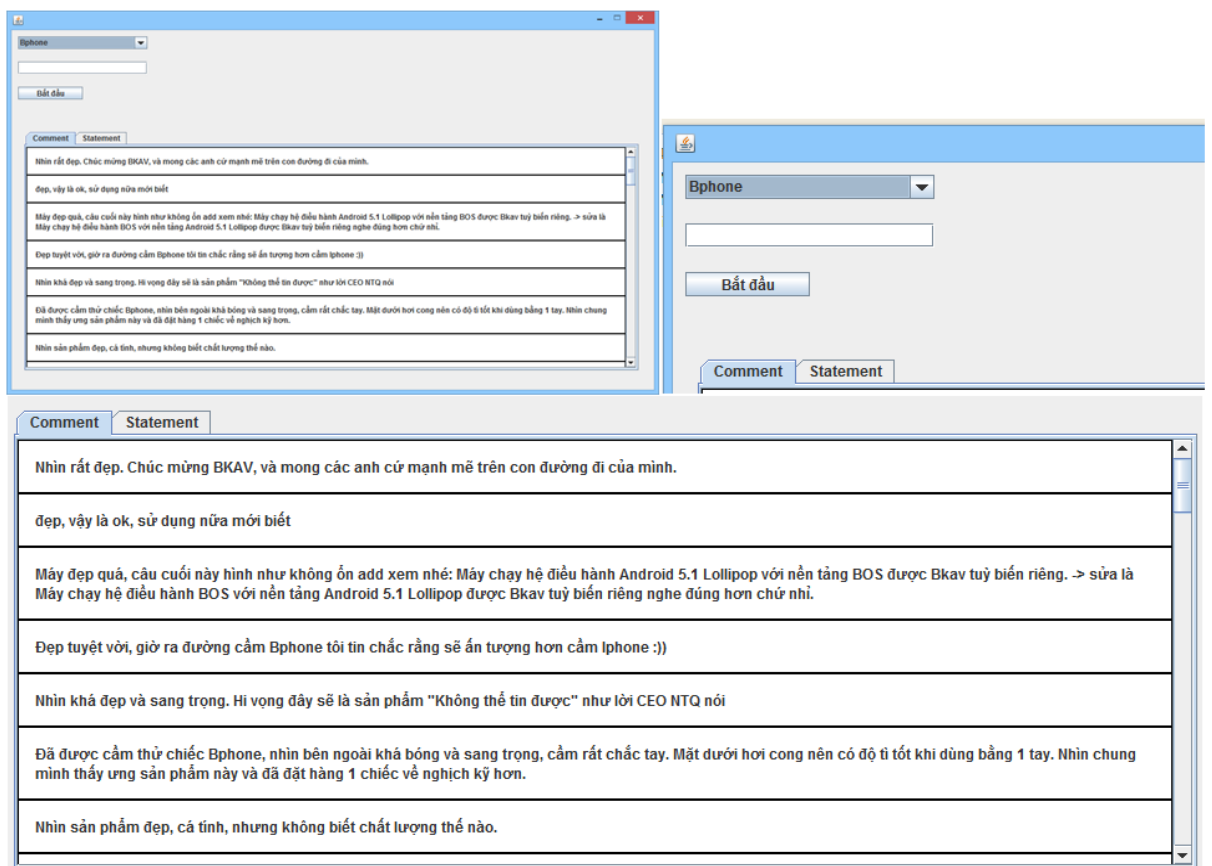
STT	Tên thành phần	Hình ảnh	Chức năng
1	JTextPane tpComment	 <p>Hình 4.14: JTextPane tpComment</p>	Hiển thị nội dung nguyên thủy của câu bình luận
2	JList lSentence	 <p>Hình 4.15: JList lSentence</p>	Hiển thị danh sách câu đã tách được của bình luận

3	JTextPane tpExplain	 <p>Tiền xử lý:</p> <ul style="list-style-type: none"> - Phát hiện từ chỉ sản phẩm - Tìm từ so sánh - Xóa emtion, phát hiện dấu hiệu để tách về câu, gán nhãn từ loại, tìm tên sản phẩm - Kiểm tra câu có đang nói về sản phẩm cần xét hay không ? <p>Đã được cảm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cảm rất chắc</p> <p>đã/r được/v cảm/v thử/v chiếc/nc tproduct/n ./, nhìn/v bên/n ngoài/n khá/r bóng/a ./</p> <p>Phân tích câu:</p> <p>=> true</p> <p>- Kiểm tra xem câu đang nhận xét về tính năng (feature) nào</p>	Hiển thị nội dung từng bước phân tích và rút trích quan điểm của từng câu trong bình luận
4	JPane pGraph		Hiển thị hình vẽ đồ thị của từng câu trong bình luận

Bảng 4.1: Các thành phần của Frame hiển thị kết quả phân tích và rút trích quan điểm của một bình luận

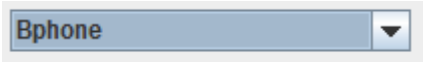
4.3.2. Frame hiển thị nội dung của pha tổng hợp kết quả


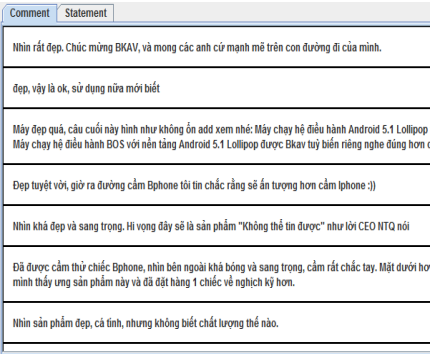
Đây là Frame tiếp nhận thực thể đầu vào và hiển thị nội dung của pha tổng hợp kết quả:

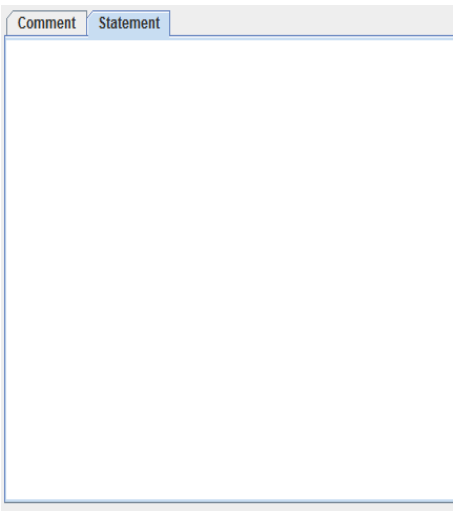


Hình 4.18: Giao diện Frame hiển thị nội dung của pha tổng hợp kết quả

Sau đây là danh sách các thành phần của Frame và chức năng của chúng:

STT	Tên thành phần	Hình ảnh	Chức năng
1	JComboBox cbbProductName	 <p>Hình 4.19: JComboBox cbbProductName</p>	ComboBox chứa các tên sản phẩm mẫu dùng để chọn và lấy ra những bình luận mẫu để tiến hành phân loại, đánh giá

2	JTextFieldProductname	 <p>Hình 4.20: JTextFieldProductname</p>	Khung nhập tên sản phẩm cần lấy ra những bình luận có liên quan để tiến hành phân loại, đánh giá
3	JButtonbtnStart	 <p>Hình 4.21: JButton btnStart</p>	Nút khởi động quá trình lấy ra danh sách bình luận sau khi nhập nội dung cho tfProductname và phân tích những bình luận
4	JList lCmts	 <p>Hình 4.22: JList lCmts</p>	Hiển thị danh sách bình luận lấy ra được

5	JTextPane tpStatement	 <p>Hình 4.23: JTextPane tpStatement</p>	Hiển thị kết quả tổng hợp điểm theo tính năng
---	-----------------------	--	---

Bảng 4.2: Các thành phần của Frame hiển thị nội dung của pha tổng hợp kết quả

4.4. Thử nghiệm

4.4.1. Pha phân tích một bình luận

Xét một câu bình luận về Bphone với nội dung: “Đã được cảm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay. Mặt dưới hơi cong nên có độ tì tốt khi dùng bằng 1 tay. Nhìn chung mình thấy ưng sản phẩm này và đã đặt hàng 1 chiếc về nghịch kỹ hơn.”, ta có kết quả phân tích như sau:

Danh sách tách câu:

(1)
"Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay"
(2)
"Mặt dưới hơi cong nên có độ ti tốt khi dùng bằng 1 tay"
(3)
"Nhìn chung mình thấy ụng sản phẩm này và đã đặt hàng 1 chiếc về nghịch kỹ hơn."

Hình 4.24: Kết quả phân tách câu của một bình luận

Kết quả phân tích rút trích quan điểm và đề thị câu 1:

– **Tiền xử lý:**

- + **Phát hiện từ chỉ sản phẩm:** “Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay”
- + **Tìm từ so sánh:** “Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay”
- + **Xóa emtion, phát hiện dấu hiệu để tách về câu, gán nhãn từ loại, tìm tên sản phẩm khác được nhắc tới:** “đã/r được/v cầm/v thử/v chiếc/nc tproduct/n ./, nhìn/v bên/n ngoài/n khá/r bóng/a ./, sang_trọng/a ./, cầm/v rất/r chắc_tay/a”

– **Phân tích câu:**

- + **Kiểm tra câu có đang nói về sản phẩm cần xét hay không?:**true
- + **Kiểm tra xem câu đang nhận xét về tính năng (feature) nào:** “đã/r được/v cầm/v thử/v chiếc/nc tproduct/n ./, [design]/ft bên/n ngoài/n khá/dgw bóng/a ./, [design]/ft sang_trọng/a ./, cầm/v rất/dgw chắc_tay/a”

- + **Kiểm tra từ thể hiện quan điểm của câu:** “đã/r được/v cầm/v thử/v chiếc/nc tproduct/n ./, [design]/ft bên/n ngoài/n khá/dgw bóng/stw ./, [design]/ft sang_trọng/stw ./, cầm/v rất/dgw chắc_tay/stw”
- + **Tách vế câu và phân tích:**
 - Vế câu 1: “[design]/ft bên/n ngoài/n khá/dgw bóng/stw”;

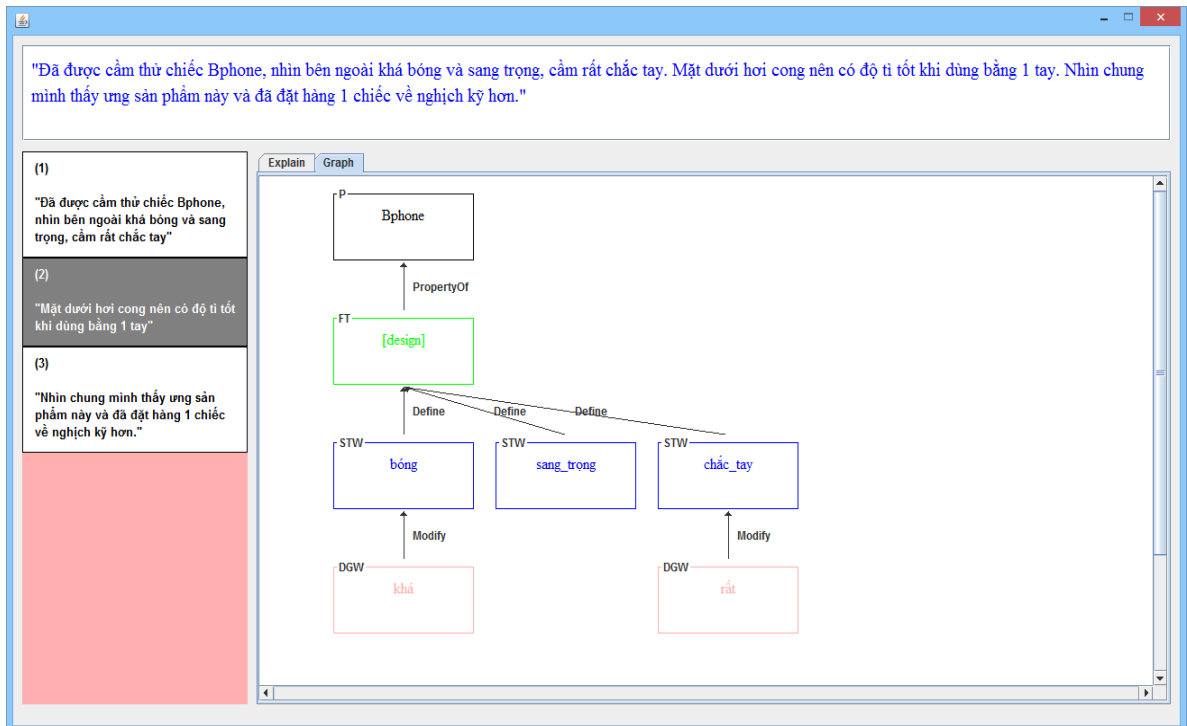
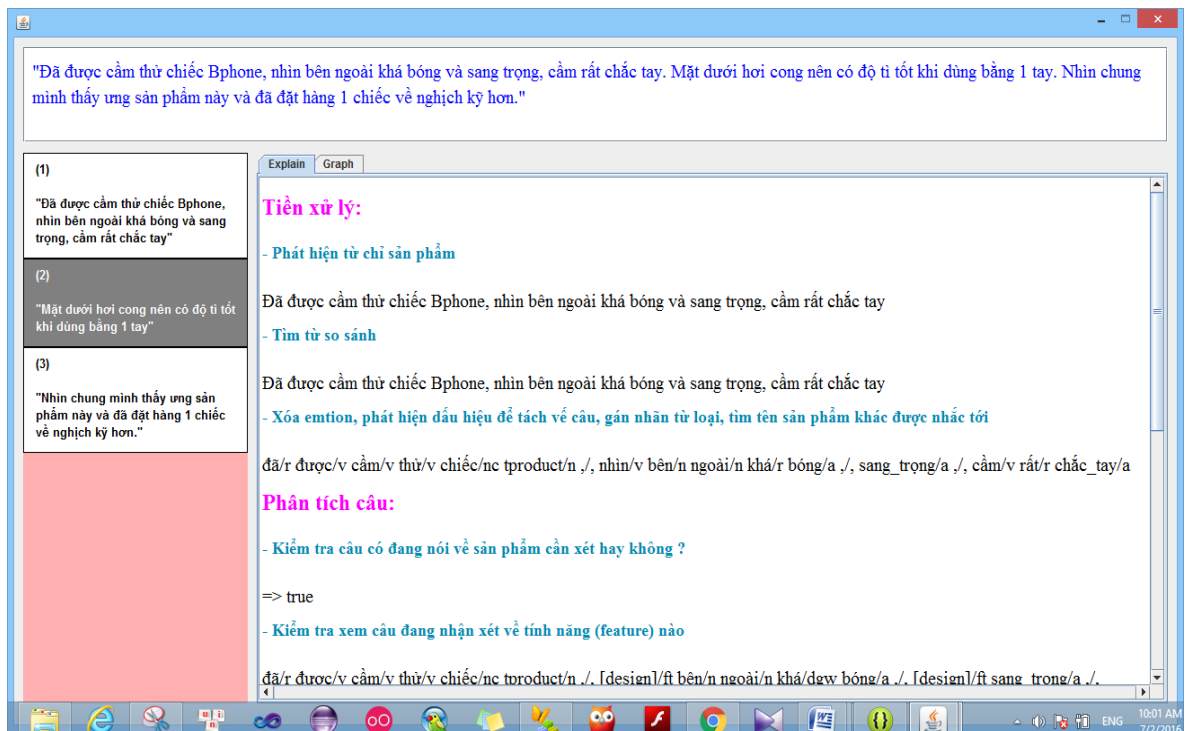
Sau khi rút gọn: “[design]/ft + khá/dgw + bóng/stw”;

Suy ra thuộc luật: **r11**.
 - Vế câu 2: “[design]/ft sang_trọng/stw”;

Sau khi rút gọn: “[design]/ft + sang_trọng/stw”;

Suy ra thuộc luật: **r08**.
 - Vế câu 3: “[design]/ft cầm/v rất/dgw chắc_tay/stw”;

Sau khi rút gọn: “[design]/ft + rất/dgw + chắc_tay/stw”;
Suy ra thuộc luật: **r11**.



Hình 4.25: Kết quả phân tích câu đầu tiên trong bình luận

Hình 4.26: Đồ thị của câu đầu tiên trong bình luận

4.4.2. Phương pháp chọn mẫu thử nghiệm

Trong khóa luận này nhóm chúng em đã tự phân loại khoảng 150 mẫu bình luận một cách chủ quan, hoàn toàn theo cảm tính của con người. Sau đó nhóm đã

lựa ra 120 câu bao gồm bình luận cho 2 loại sản phẩm điện thoại: Bphone và SamSung Galaxy S7 Edge. Vì đây là 2 loại sản phẩm có nhiều bình luận nhất vào thời điểm khảo sát nên sẽ đa dạng được các mẫu thử. Trong đó nhóm đã phân ra 4 nhóm sau: quan điểm tích cực (40 bình luận), quan điểm tiêu cực (40 bình luận), quan điểm trung lập (20 bình luận), câu không chứa quan điểm (20 bình luận).

* Nội dung cụ thể mỗi bình luận được trình bày trong phụ lục.

4.4.3. Đánh giá kết quả thử nghiệm

Kết quả khảo sát pha phân tích và rút trích quan điểm:

STT	Tên sản phẩm	Tổng số bình luận	Số bình luận đúng	Tỉ lệ đúng (%)
1	Bphone	60	50	83.33
2	SamSung Galaxy S7 Edge	60	53	88.33

Bảng 4.3: Kết quả đánh giá pha phân tích và rút trích “cụm quan điểm”

STT	Tên sản phẩm	Nhóm	Tổng số bình luận	Số bình luận đúng	Tỉ lệ đúng (%)	Tổng (%)
1	Bphone	Tích cực	20	17	85	78.33
		Tiêu cực	20	16	80	
		Trung lập	10	7	70	
		Không chứa quan điểm	10	7	70	
2	SamSung Galaxy S7 Edge	Tích cực	20	17	85	80
		Tiêu cực	20	15	75	
		Trung lập	10	8	80	
		Không chứa quan điểm	10	8	80	

Bảng 4.4: Kết quả đánh giá sau khi tính trọng số và phân loại chiều hướng quan điểm

Trong phạm vi tìm kiếm tài liệu tham khảo, nhóm chúng em chưa tìm thấy ứng dụng hay đề tài nào về khai phá quan điểm tiếng Việt trong miền tri thức các sản phẩm về điện thoại di động. Các đề tài tìm được chủ yếu nghiên cứu về lĩnh vực tin tức tiếng Việt [2], lĩnh vực tin tức tài chính [1].

Do đó nhóm chúng em chỉ dừng lại ở mức cải tiến những phần có thể, kế thừa một số phần đã có và đề xuất thêm một số phần. Về bộ dữ liệu mẫu đã áp dụng cho các đề tài trên nhóm cũng không tìm thấy nên không có điều kiện cài đặt để thử nghiệm lại.

Chương 5. TỔNG KẾT

5.1. Kết quả đạt được

Mục tiêu ban đầu của nhóm chúng em đã thực hiện được trong khóa luận là tìm hiểu được tổng quan của thuật ngữ khai phá quan điểm, một khái niệm chưa phải là đã được phổ biến, nhất là đối với ngôn ngữ tiếng Việt.

Từ cơ sở lý thuyết tìm hiểu được, nhóm chúng em đã xây dựng được một sản phẩm trực quan để thể hiện quá trình phân tích ở các mức độ tài liệu, câu, từ đáp ứng các yêu cầu:

- Thu thập dữ liệu mẫu, sau đó tiến hành phân tích thủ công những từ ngữ, cú pháp trong từng bình luận và rút trích ra được các dạng câu cơ bản cũng như một số trường hợp ngoại lệ. Từ đó tạo cơ sở cho việc xây dựng được mô hình Ontology riêng của mình cho việc phân tích.
- Sau khi đã có đủ các cơ sở tri thức nhóm chúng em đã xây dựng được mô hình để xử lý phân tích các mẫu thử như:
 - + Kiểm tra và thu thập các tài liệu liên quan, chuẩn bị cho quá trình phân tích.
 - + Phân tích và rút trích quan điểm thông qua các “cụm quan điểm”, tính năng người viết đề cập đến như: “màn hình”, “pin”, “thiết kế”, “giá cả”, “camera”, “cấu hình”, “ứng dụng”.
 - + Các bình luận về sản phẩm đầu vào đã được phân tích và phân loại chiều hướng quan điểm (tích cực, tiêu cực, trung lập) theo trọng số tính được.
 - + Thống kê các bình luận theo từng tính năng của sản phẩm và tính điểm cho mỗi tính năng.
- Đưa ra đồ thị khái niệm biểu diễn cấu trúc thông tin của câu bình luận sau khi đã được đơn giản hóa một số thành phần không cần thiết.

5.2. Hạn chế

Bên cạnh đó, khóa luận lần này cũng còn một số hạn chế:

- Ứng dụng chưa phân tích dữ liệu theo thời gian thực mà dữ liệu chỉ là được lấy cứng từ trên trang mạng và lưu về local.
- Việc trích xuất thông tin còn cứng nhắc và thủ công, thêm vào đó là chưa áp dụng các giải pháp máy học cho việc cho điểm trọng số các cụm từ quan điểm để tăng độ chính xác.
- Việc tiền xử lý dữ liệu chỉ dừng ở mức đơn giản, chưa xử lý được nhiều các câu sai lỗi chính tả, sai ngữ pháp hoặc xử dụng từ lóng, từ địa phương.

5.3. Hướng phát triển

- Nếu tiếp tục thực hiện và nâng cấp khóa luận, nhóm chúng em sẽ cố gắng áp dụng những phương pháp khác nhau trong việc tìm ra cụm từ quan điểm.
- Cố gắng nâng cao tỷ lệ chính xác của các trọng số trong các cụm từ quan điểm bằng một số phương pháp máy học.
- Tiếp tục mở rộng phạm vi trên nhiều lĩnh vực khác nhau còn thiếu trong việc khai phá dữ liệu tiếng Việt, thay vì gói gọn trong miền lĩnh vực các sản phẩm điện thoại thông minh.

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Việt

- [1] Lê Thu Hà, *Phân lớp quan điểm theo chủ đề dựa vào chuỗi con và cây con phụ thuộc trên miền tin tức tài chính*, khóa luận tốt nghiệp đại học chính quy, trường đại học Công Nghệ - ĐHQG Hà Nội, 2011.
- [2] Vũ Xuân Sơn, *Tổng hợp quan điểm dựa trên mô hình thống kê và ứng dụng vào khai phá quan điểm trong văn bản tin tức tiếng Việt*, khóa luận tốt nghiệp đại học chính quy, trường đại học Công Nghệ - ĐHQG Hà Nội, 2011.

Tài liệu Tiếng Anh

- [3] Alec Go, Richa Bhayani, and Lei Huang, <http://sentiment140.com>.
- [4] B.Pang, dL.Lee, *Thumbs up?Sentiment classification using machine learning techniques*, 2002, pp.1-8.
- [5] Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012.
- [6] Binh Thanh Kieu and Son Bao Pham, *Sentiment Analysis for Vietnamese*, 2010
- [7] Bo Pang and Lillian Lee, *Opinion Mining and Sentiment Analysis*, Foundations and Trends R in Information Retrieval, 2, 1–2 ,2008,pp. 1–135.
- [8] Healey và Ramaswamy,
https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/ .
- [9] Liu et al., “*Low-quality product review detection in opinion summarization*”, MNLP, 2007
- [10] Nabeela Altrabsheh et al., *SA-E: Sentiment Analysis for Education*, 2013.

- [11] Nguyen Thi Duyen et al, *An Empirical Study on Sentiment Analysis for Vietnamese*, 2014
- [12] Sunghwan Mac Kim et al., *Sentiment Analysis in Student Experiences of Learning*, The 3rd International Conference on Educational Data Mining, Pittsburgh, PA, USA, 2010.
- [13] Xuan-Son Vu, Hyun-Je Song and Seong-Bae Park, *Building a VietNameese SentiWordNet Using VietNameese Electronic Dictionary and String Kernel*, 2014.

PHỤ LỤC

Danh mục các bình luận đã được chọn để thử nghiệm:

STT	Tên sản phẩm	Nội dung	Nhóm
1	Bphone	Nhìn rất đẹp. Chúc mừng BKAV, và mong các anh cứ mạnh mẽ trên con đường đi của mình.	Tích cực
2	Bphone	đẹp, vậy là ok, sử dụng nữa mới biết	Tích cực
3	Bphone	Máy đẹp quá, câu cuối này hình như không ổn add xem nhé: Máy chạy hệ điều hành Android 5.1 Lollipop với nền tảng BOS được Bkav tùy biến riêng. -> sửa là Máy chạy hệ điều hành BOS với nền tảng Android 5.1 Lollipop được Bkav tùy biến riêng nghe đúng hơn chứ nhỉ.	Tích cực
4	Bphone	Đẹp tuyệt vời, giờ ra đường cầm Bphone tôi tin chắc rằng sẽ ấn tượng hơn cầm Iphone :))	Tích cực
5	Bphone	Nhìn khá đẹp và sang trọng. Hi vọng đây sẽ là sản phẩm "Không thể tin được" như lời CEO NTQ nói	Tích cực
6	Bphone	Đã được cầm thử chiếc Bphone, nhìn bên ngoài khá bóng và sang trọng, cầm rất chắc tay. Mặt dưới hơi cong nên có độ tì tốt khi dùng bằng 1 tay. Nhìn chung mình thấy ưng sản phẩm này và đã đặt hàng 1 chiếc về nghịch kỹ hơn.	Tích cực

7	Bphone	Nhìn sản phẩm đẹp, cá tính, nhưng không biết chất lượng thế nào.	Tích cực
8	Bphone	cấu hình như vậy có thể nói là rất tốt, còn lại chỉ là vấn đề tương thích giữa phần cứng và phần mềm và các ứng dụng khi sử dụng. điều đó mới quyết định máy mượt hay không. mong là ấn tượng như lời anh Quảng.	Tích cực
9	Bphone	Quá đẹp.	Tích cực
10	Bphone	Đẹp, nhưng xem thôi, con mình uống sữa quan trọng hơn.	Tích cực
11	Bphone	Quá đẹp. Chúc mừng anh Quảng và BKAV.	Tích cực
12	Bphone	đẹp quá mình sẽ mua 1000 chiếc	Tích cực
13	Bphone	mình dùng rồi, cảm giác máy rất mượt và sang trọng. máy không nóng mấy khi truy cập web với thời gian dài.	Tích cực
14	Bphone	Thật tuyệt vời ! và tự hào ! không phải vì nó là dt đẹp nhất thế giới hay cấu hình mạnh nhất. Mà vì nó sp dt đầu tay của người việt do chính người việt thiết kế và sản xuất !	Tích cực
15	Bphone	Nhái Iphone nhưng mà so về độ sang trọng thì thua xa	Tiêu cực
16	Bphone	Quá chuẩn về phong cách, quá đẹp về ngoại hình!!! Gặp em Bphone này thì nhất	Tích cực

		<p>định phải xin cưới luôn!:D:D:D</p> <p>Tôi yêu Bphone vì từ nay ra thế giới chúng ta không còn phải quá ngậm ngùi như xưa. Cần nữa, nhiều nữa, nhiều mãi những trí tuệ Việt, những sản phẩm Việt như thế này...</p>	
17	Bphone	camera kém quá!	Tiêu cực
18	Bphone	Xấu quắt, thấy được cái khe giữa kính và thân rồi nhé hoàn thiện còn kém làm chưa xứng tầm Sam và Apple đâu nhé, với số tiền trên mua con MI4 6.5 chính hãng cấu hình y chang hệ điều hành mượt hơn đư cả khối tiền.	Tiêu cực
19	Bphone	Đẹp và sang trọng!	Tích cực
20	Bphone	Nhìn giống Oppo lai Iphone lai Sony, Không có cá tính thiết kế riêng, vượt trội.....Cấu hình này so với tầm giá thì không ổn chút nào...Các Rivewer đâu rồi, hãy thể hiện những bài so sánh với các sản phẩm khác xem	Tiêu cực
21	Bphone	Quan trọng là chất lượng và kiểu dáng máy quá đẹp và sang trọng kìa.	Tích cực
22	Bphone	Khá đẹp, nếu có tiền cũng mua 1 em về dùng	Tích cực
23	Bphone	nhìn thô gần chết mà cũng khen đẹp được	Tiêu cực
24	Bphone	Không đẹp	Tiêu cực

25	Bphone	Gớm thật	Tiêu cực
26	Bphone	giá thì đắt mà chắc gì dùng được 1 năm, haha	Tiêu cực
27	Bphone	cho em xin, nhái của Iphone hả anh Quảng	Không chứa quan điểm
28	Bphone	nhìn xàm xàm	Tiêu cực
29	Bphone	anh Quảng được cái nỏ to	Không chứa quan điểm
30	Bphone	Kiểu dáng nữ tính quá. Thích vuông vắn của SONY hơn	Tiêu cực
31	Bphone	Chụp hình hơi mờ	Tiêu cực
32	Bphone	đừng vội chê những cái gì mình chưa thấy chưa cảm nhận được và ăn theo phong trào như thế!tôi thực sự ấn tượng trước những gì mình thấy được,cộng đồng mạng chỉ ghét cái tính nỏ của ông Quảng thôi!	Không chứa quan điểm
33	Bphone	Quá đẹp, rất thời thượng!!mua zô mua zô	Tích cực
34	Bphone	Rất đẹp, tinh xảo, chúc mừng bphone. Mình sẽ mua một em	Tích cực
35	Bphone	:D	Không chứa quan điểm
36	Bphone	giá đắt quá	Tiêu cực
37	Bphone	chơi game chậm như rùa	Tiêu cực
38	Bphone	9tr cho cho cái điện thoại này hả, đắt vãi	Tiêu cực

39	Bphone	anh Quảng nỏ to quá, cấu hình yếu như rùa	Tiêu cực
40	Bphone	Nói thật chứ nhìn chả ưa tí nào!	Tiêu cực
41	Bphone	Giá cao quá ko đủ thóc :(Tiêu cực
42	Bphone	Xấu toàn tập	Tiêu cực
43	Bphone	Giá cao trên trời lấy đâu mà mua	Tiêu cực
44	Bphone	chụp hình chả đâu vào đâu, xấu !	Tiêu cực
45	Bphone	hahah, Việt Nam cũng sản xuất được smartphone chứ đùa à	Không chứa quan điểm
46	Bphone	cố lên, ủng hộ hàng VN	Không chứa quan điểm
47	Bphone	Chả hiểu cứ của Việt Nam là bị đim đến chết	Không chứa quan điểm
48	Bphone	Thôi cho em xin	Không chứa quan điểm
49	Bphone	wtf, có cần PR vậy không	Không chứa quan điểm
50	Bphone	Hết bà một tháng lương cho cái đt này hả	Không chứa quan điểm
51	Bphone	Cấu hình có vẻ tốt đây, có điều mức giá cao quá	Trung lập
52	Bphone	mình thích cái thiết kế này, khá giống Iphone nhưng giá hơi cao so với thị trg chung	Trung lập
53	Bphone	Máy đẹp, cấu hình khủng nhưng pk đi kèm, tại nghe dạng thường, ko phù hợp với	Trung lập

		sản phẩm, cần có chút cải tiến. ^^	
54	Bphone	Ứng hộ hàng Việt Nam, nhưng nếu giá giảm tí nữa thì mình mua. Máy đẹp !	Trung lập
55	Bphone	Xài rồi, chụp hình xấu, được cái thiết kế bắt mắt với hàng VN nên cũng muốn dùng thử	Trung lập
56	Bphone	Camera chuối quá, được mỗi cái pin trâu	Trung lập
57	Bphone	Giá này phải kèm điều kiện mới mua được. Thôi 3tr đi mua Zenfone sài đi bạn, cấu hình ổn, thiết kế cũng ngon lành.	Trung lập
58	Bphone	Không ngờ anh Quảng chém gió vậy mà cũng sản xuất được con điện thoại oách ra phết	Trung lập
59	Bphone	Lúc đầu định mua rồi, nhưng sau nghĩ lại cầm 9tr ra mua con iphone 6 xách tay còn hơn, tuy nhiên bạn nào có tinh thần yêu nước thì cứ mua nhé, cũng ko đến nỗi đâu !	Trung lập
60	Bphone	So với Iphone tuổi gì, tuy nhiên hàng VN thiết kế vậy cũng khá ok rồi	Trung lập
61	SamSungS7	Samsung đang dần khẳng định đẳng cấp. đẹp thật.	Tích cực
62	SamSungS7	Đẹp thật, nhưng nên thêm màu xanh ngọc lục bảo cũng rất đẹp.	Tích cực
63	SamSungS7	Con này cài mấy game khủng vào là tạch	Tiêu cực

64	SamSungS7	thảm họa!	Tiêu cực
65	SamSungS7	màu đẹp quá, nhưng đợi Ip7 ra màu gì rồi S8 edge ra màu đỏ nhé.	Tích cực
66	SamSungS7	Ko đúng, S7 chụp ảo lòi. Không trung thực. Chụp tối phải tối thui như iphone mới đẹp. Vote iphone chụp trung thực tối đen.	Tiêu cực
67	SamSungS7	Galaxy S7 edge là smartphone có camera tốt nhất, các smartphone cao cấp khác tất đài không cần bàn cãi.	Tích cực
68	SamSungS7	Xanh đỏ tím hồng đẹp không các tím yêu Sam :)	Không chứa quan điểm
69	SamSungS7	Độc, lạ, tiện dụng lên ngôi. Tiếc là mình không đủ tiền!	Tích cực
70	SamSungS7	đẹp quá, đúng là chỉ dành cho dân chơi	Tích cực
71	SamSungS7	không phải anti fan samsung, nhưng tôi chẳng thấy đẹp chỗ nào, cộng với khí hậu nắng nóng, nắng mưa thất thường, mồ hôi bám vào thì chắc là mùi kinh khủng lắm.	Tiêu cực
72	SamSungS7	Đẹp quá chất quá mê em này rồi 6s nhìn chán chán xấu xấu dc mỗi thương hiệu	Tích cực
73	SamSungS7	Chạy mượt lắm các bác ạ.	Tích cực
74	SamSungS7	Pin 3600 vẫn độ phân giải nv mà pin lại kém con note 5 của mình là sao	Tiêu cực
75	SamSungS7	Vòng vo chức năng cũng chỉ nhieu đó, chạy	Tiêu cực

		theo mấy thứ này có nước nghèo luôn.	
76	SamSungS7	hao Pin quá vậy lúc 19:19 98% pin đến lúc 19:34 còn có 96% Pin ah.	Tiêu cực
77	SamSungS7	fan Sam đâu rồi, go	Không chứa quan điểm
78	SamSungS7	500 anh em đâu, mau ra rước em nó về nào	Không chứa quan điểm
79	SamSungS7	đây mới là siêu phẩm của năm	Không chứa quan điểm
80	SamSungS7	iphone tất cả các dòng Trung Quốc đều có thể làm giả, duy chỉ có dòng Edge của nhà Samsung thì đồ làm giả được :)))	Không chứa quan điểm
81	SamSungS7	GATO từ trong giấc ngủ	Không chứa quan điểm
82	SamSungS7	Ồi. Ước ao	Không chứa quan điểm
83	SamSungS7	Nhái của anh Quảng à =)) Cái góc bo kiểu này thì em thua rồi	Không chứa quan điểm
84	SamSungS7		Không chứa quan điểm
85	SamSungS7	Nhìn thì đẹp đó mà sao cái màn hình tối thui vậy	Trung lập
86	SamSungS7	Rước em nó về rồi, chụp đẹp, pin hơi kém	Trung lập
87	SamSungS7	Màu Blue Coral trên hình đẹp!. Khắc phục được nhược điểm pin yếu nữa thì hay	Trung lập

88	SamSungS7	mua con này thả rước 3 em xiaomi về cho cháu nó chơi game còn hay hơn, màn hình thì xấu, pin thì yếu. Được cái sang trọng, đem đi lòe mấy bạn nữ được :))	Trung lập
89	SamSungS7	Đẹp vậy chứ có đẹp hơn nữa tui cũng chẳng bao giờ nghĩ tới một ngày nào đó lại dùng điện thoại SAMESAME cả!	Trung lập
90	SamSungS7	Vừa mua hôm qua, phải nói quả ảnh hiện ra ảo lòi. Được cái rẻ trong tầm giá	Trung lập
91	SamSungS7	S7 edge đứng đầu bảng smartphone chụp ảnh đẹp nhất hiện nay rồi. Theo đánh giá từ trang dxomark, trang web chuyên đánh giá camera thì 6s chụp thua cả Note 4 và LG G4. Những Smartphone chụp đẹp đầu bảng là S7 edge, S6 edge plus, Xperia Z5, Note 5, S6 Edge và cách đó rất xa là Iphone 6s plus	Trung lập
92	SamSungS7	sao không nhìn mượt mà gì hết dù là thiết kế cũn khá nam tính...	Trung lập
93	SamSungS7	Mê 2 em này rồi. Đẹp, cấu hình mạnh, chống nước, có thể nhớ và mình lại thích số 7. Tích cóp tiền từ giờ sang năm mua, giá chất quá :((Trung lập
94	SamSungS7	Thất vọng, vẫn chưa có thêm công nghệ nào thật sự đột phá so với Galaxy S5 của mình. Nguyên khối, màn hình cong? Không cần thiết (tay chân ướt mà sờ vào	Tiêu cực

		máy đang sạc thì hơi ngán); camera F1.7? Tốt nhưng nó vẫn cái điện thoại; Hiệu năng nhanh, mạnh? cài softs/games vào nhiều máy rồi cũng chậm. Xem ra Samsung đang loay hoay cải tiến những công nghệ họ đang có từ S5. Cần thêm nữa những đột phá về công nghệ không chỉ đi thay cái vỏ và chăm chút cái màn hình	
95	SamSungS7	Đang tết không muốn "rầy". Thiết kế hàng rẻ tiền mà nhiều người có làm cả năm còn chưa chắc mua được mà to miệng chê bai. Thùng rỗng kêu to mà.	Không chứa quan điểm
96	SamSungS7	Năm ngoái giữ quả pin lại làm của, năm nay làm cái thiết kế nhìn như hàng rẻ tiền há! :))	Tiêu cực
97	SamSungS7	Xấu dã man	Tiêu cực
98	SamSungS7	Quá đẹp	Tích cực
99	SamSungS7	quá xấu! trông đợi ở Note 6!	Tiêu cực
100	SamSungS7	thằng sam đẹp dùm mình cái loa luôn đi, nghe như loa tàu	Tiêu cực
101	SamSungS7	Nhìn xấu thật..	Tiêu cực
102	SamSungS7	Cái xấu luôn đeo bám điện thoại Samsung. Nhìn rẻ tiền thế không biết.	Tiêu cực
103	SamSungS7	Cá nhân mình thấy máy rất đẹp.Tuy rằng mình là fan LG,và có lẽ đợi LG G5 có gì không để nâng cấp hoặc chuyển qua s7 của	Tích cực

		samsung	
104	SamSungS7	Tương đối đẹp.	Tích cực
105	SamSungS7	Điểm Antutu với cấu hình như vậy thì không có gì là khủng vì HTC One M9 điểm Antutu cũng đã hơn 118.000 rồi	Tiêu cực
106	SamSungS7	SS đã trang bị pin khủng, 1 tín hiệu tích cực mà người dùng đang chờ đợi. Sẽ knockout apple.	Tích cực
107	SamSungS7	Để xem sung rụng lần này làm được cái trò j. Kiểu dáng không thay đổi. S6 s7 y trang. Chỉ khác cái khe cắm thẻ nhớ.	Không chứa quan điểm
108	SamSungS7	Nhìn xấu và rẻ tiền, không so được với iPhone.	Tiêu cực
109	SamSungS7	Viền trên và dưới cũng còn dày quá, nhất là viền trên quá dày luôn.	Tiêu cực
110	SamSungS7	Kể từ năm 2015 thì bộ đôi Samsung Galaxy S6 và S6 Edge đã đánh bật Iphone bằng doanh thu tăng cao và vượt Apple.Năm 2016 khả năng kỉ nguyên người người,nhà nhà Iphone sẽ chấm dứt. Cú bật lớn nhất của Sam là những thế hệ Galaxy S và Note đẹp về thiết kế và độc về màn hình cong khiến ai cũng thích thú	Tiêu cực
111	SamSungS7	Không nói về cấu hình...cái này chỉ khác cái s6 là camera làm không bị lỗi lên	Tích cực
112	SamSungS7	Nhìn đẹp và sang trọng quá. Thích thiết kế	Tích cực

		sang trọng và dễ cầm ôm tay như vậy	
113	SamSungS7	Công nhận s7 chụp đẹp , phải nói còn đẹp hơn cả ngoài đời thật luôn, người ta còn hay gọi là "áo cái chảo"	Tích cực
114	SamSungS7	Đẹp nhưng không thật! :))	Tích cực
115	SamSungS7	Nhìn màu SS luôn ảo như cái chảo	Tiêu cực
116	SamSungS7	Wao, Ấn tượng quá! tấm đầu tiên Cam thể hiện khả năng bắt chuyển động rất tốt. Hầu hết các tấm có độ chi tiết cao.	Tích cực
117	SamSungS7	khaoduong: Chụp đẹp thật	Tích cực
118	SamSungS7	loa vẫn xấu như con gấu	Tiêu cực
119	SamSungS7	dòng android chạy vậy là nhanh rồi, mấy máy khác lag như gì !	Tích cực
120	SamSungS7	Sao thẳng sam nó ko bo lại cái góc chút nhỉ, pin thì trâu đó nhưng cái thiết kế có quả góc tròn tròn nhìn thô quá	Trung lập