ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

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 HỆ CHÍNH QUI
Ngành: Công nghệ thông tin

HÀ NỘI - 2011

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

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 HỆ CHÍNH QUI
Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: ThS. Nguyễn Thu Trang

HÀ NỘI - 2011

Lời cảm ơn

Trước tiên, em muốn gửi lời cảm ơn sâu sắc nhất đến thầy giáo, Phó Giáo sư Tiến sĩ. Hà Quang Thụy, Thạc sỹ Nguyễn Thu Trang và Cử nhân Lê Hoàng Quỳnh, những người đã tận tình hướng dẫn em trong suốt quá trình nghiên cứu Khoa học và làm khóa luận tốt nghiệp.

Em xin bày tỏ lời cảm ơn sâu sắc đến những thầy cô giáo đã giảng dạy em trong bốn năm qua, những kiến thức mà em nhận được trên giảng đường đại học sẽ là hành trang giúp em vững bước trong tương lai.

Em cũng muốn gửi lời cảm ơn đến các anh chị và các bạn trong nhóm seminar về "Khai phá dữ liệu" đã giúp đỡ và cho em những lời khuyên bổ ích về chuyên môn trong quá trình nghiên cứu.

Cuối cùng, em muốn gửi lời cảm ơn sâu sắc đến tất cả bạn bè, và đặc biệt là cha mẹ và anh trai, những người luôn kịp thời động viên và giúp đỡ em vượt qua những khó khăn trong cuộc sống.

Sinh Viên

Lê Thu Hà

Tóm tắt

Phân lớp quan điểm là một bài toán quan trọng trong khai phá quan điểm. Bài toán phân tích các đánh giá cho một chủ đề nhất định, hoặc sự kiện , sản phẩm để tự động phân loại đánh giá theo hai hướng tích cực hay tiêu cực của quan điểm. Với sự phát triển nhanh chóng của các ứng dụng internet, phân lớp quan điểm cần thiết để giúp người dùng và nhà sản xuất nhanh chóng xác định quan điểm của khách hàng từ thông tin bình luận.

Có rất nhiều phương pháp phân lớp quan điểm nhưng chủ yếu theo hai hướng chính : phương pháp học máy và phương pháp hướng ngữ nghĩa dựa vào độ đo thông tin (PMI). Khóa luận này trình bày phương pháp tiếp cận học máy bằng cách sử dụng các mối quan hệ cú pháp giữa từ trong câu cho phân lớp quan điểm. Phương pháp sử dụng tần suất của chuỗi từ con và cây con phụ thuộc làm đặc trưng của máy hỗ trợ vector(SVM). Thực nghiệm trên dữ liệu miền tin tức tài chính với 312 bình luận trên 180 bài báo cho độ chính xác cao nhất là 72%.

Lời cam đoan

Tôi xin cam đoan khóa luận "Phân lớp quan điểm theo chủ đề dựa vào chuỗi con và cây con phụ thuộc trên miền tin tức tài chính " dưới sự hướng dẫn của Thạc sỹ Nguyễn Thu Trang và cử nhân Lê Hoàng Quỳnh là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả được trình bày trong khóa luận là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác.

Tôi đã trích dẫn đầy đủ các tài liệu tham khảo, công trình nghiên cứu liên quan ở trong nước và quốc tế. Ngoại trừ các tài liệu tham khảo này, khóa luận hoàn toàn là công việc của riêng tôi.

Khóa luận được hoàn thành trong thời gian tôi làm Sinh viên tại Bộ môn Các hệ thống thông tin, Khoa Công nghệ Thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Hà Nội, ngày 25 tháng 05 năm 2011

Sinh viên

Lê Thu Hà

Mục lục

Danh sách bảng biểu

Danh sách hình vẽ

Danh sách từ viết tắt

Từ và cụm từ	Viết tắt
A Library for Support Vector Machines LibSVM	LibSVM
Support vector machine SVM	SVM
Frequent tree miner	FREQT
Sequential pattern minier	PrefixSpan

Mở đầu

Trong các nghiên cứu mới nhất của phân lớp quan điểm, phân lớp dựa vào kỹ thuật học máy [5][13][17] đạt được nhiều thành công chứng tỏ hiệu quả hơn phân lớp dựa vào quy tắc phân lớp. Bo Pang và các cộng sự [6] đạt được độ chính xác 87% trong phân lớp quan điểm đánh giá bình luận phim thông qua sử dụng mô hình n-gram như đặc trưng cho máy hỗ trợ vector .

Theo nghiên cứu của Shotaro Matsumoto và cộng sự [21] về phân lớp quan điểm, một tài liệu được biểu diễn như một tập hợp từ (bag-of-words) . Nhược điểm của cách biểu diễn này là họ không quan tâm tới trật tự và quan hệ ngữ nghĩa của các từ xuất hiện trong một câu. Tuy nhiên, trật tự từ và quan hệ ngữ nghĩa giữa các từ trong một câu được đánh giá là quan trọng và hữu ích trong phân lớp quan điểm. Vì vậy, để kết hợp các thông tin đó vào phân lớp quan điểm, họ đề xuất phương pháp sử dụng một chuỗi từ và cây con phụ thuộc như thể hiện cho một câu và tính tần suất của các mẫu con của câu trong tài liệu làm đặc trưng cho phân lớp quan điểm. Khóa luận với tên đề tài "Phân lớp quan điểm theo chủ đề dựa vào chuỗi con và cây con phụ thuộc trên miền tin tức tài chính" đưa ra phương pháp phân lớp quan điểm dựa trên trích chọn đặc trưng chuỗi con và cây con phụ thuộc kết hợp với mô hình học máy SVM.

Khóa luận được tổ chức thành 4 chương:

- *Chương 1* giới thiệu về khai phá quan điểm, và ứng dụng trong thực tế.
- *Chương 2* được chia thành hai phần :

- Phần một giới thiệu bài toán phân lớp quan điểm, và giới thiệu một số hướng tiếp cận nhằm giải quyết bài toán phân lớp quan điểm
- Phần hai trình bày hai thuật toán tính tần suất khai phá mẫu: prefixspan và freqt của Kenji Abe và các cộng sự [10, 12, 20].

- *Chương 3* trình bày mô hình đề xuất giải quyết bài toán phân lớp quan điểm. Mô hình gồm ba bước : phân tích chủ đề, trích chọn đặc trưng và phân lớp dựa vào học máy SVM.

- *Chương 4* trình bày nội dung thực nghiệm mô hình đề xuất trên dữ liệu tin tức tài chính và đánh giá kết quả .

Chương 1. Giới thiệu khai phá quan điểm

1.1 Giới thiệu khai phá quan điểm

1.1.1 Giới thiệu bài toán khai phá quan điểm

Theo Bo Pang và Lillian Lee, 2008 [7], thuật ngữ ***khai phá quan điểm*** (Opinion Mining) xuất hiện lần đầu tiên trong bài báo của Kushal Dave và cộng sự [13] được công bố ở hội nghị WWW năm 2003. Thuật ngữ này có thể giải thích theo nhiều cách khác nhau, liên quan chặt chẽ với tìm kiếm và trích chọn thông tin trên Web. Theo Kushal Dave và cộng sự, công cụ khai phá quan điểm lí tưởng sẽ "xử lí một tập các kết quả tìm kiếm cho một đối tượng nhất định, tạo ra một danh sách thuộc tính (chất lượng, đặc trưng…) và tổng hợp quan điểm cho mỗi thuộc tính đó (xấu, tốt,..)". Tuy nhiên, đến khi nghiên cứu của Bing Liu [4] được công bố năm 2006 thì thuật ngữ khai phá quan điểm mới được đưa ra một cách rõ ràng hơn.

Theo Bing Liu, các quan điểm có thể thể hiện về bất cứ điều gì, ví dụ như một sản phẩm , một cá nhân, một tổ chức, một chủ đề,…Ông coi các thực thể được nhận xét là đối tượng(object). Đối tượng này gồm một tập hợp các thành phần (components) và một tập các thuộc tính(attributes). Vì vậy, một đối tượng có thể được phân ra theo thành phần của mối quan hệ, tức là mỗi thành phần cũng có thể là thành phần con của nó.Ví dụ: một sản phẩm (ví dụ: ô tô, máy ảnh…) có thể có nhiều thành phần khác nhau, một sự kiện có thể có nhiều sự kiện con, một chủ đề có thể là chủ đề con của chủ đề nào đó… Theo đó, tác giả đưa ra một số khái niệm trong khai phá quan điểm:

❖ Đối tượng (object):

Một đối tượng O là một thực thể (người, sản phẩm, sự kiện…) được đánh giá. O được kết hợp bởi một cặp , O: (T,A). Trong đó, T là một cấu trúc phân cấp thành phần cha, thành phần con, A là tập các thuộc tính của thành phần O. Mỗi thành phần có tập các thành phần con và thuộc tính của nó.

Ví dụ: Xe đạp có một tập các thành phần : bánh xe, yên xe, .. và các thuộc tính : màu sắc, nhà sản xuất… Thành phần bánh xe có các thành phần con : nan hoa, trục xoay …

Về bản chất, một đối tượng được thể hiện như một cây. Gốc là đối tượng của nó . Các nút không phải là gốc là một thành phần hoặc thành phần con của một đối tượng. Mỗi

nhánh thể hiện một liên kết giữa các thành phần. Mỗi nút cũng được thể hiện bởi một tập các thuộc tính. Một quan điểm được thể hiện trên bất kì nút nào và bất kỳ thuộc tính nào của nút.

❖ Các đặc trưng hiện và ẩn:

Với mỗi một đánh giá r bao gồm một tập các câu $r = \{s_1, s_2, \ldots, s_m\}$. Nếu đặc trưng *f* xuất hiện trong *r*, ta nói *f* là đặc trưng hiện (explicit feature). Ngược lại, ta nói, *f* là đặc trưng ẩn (implicit feature).

Ví dụ:

+ *"Khung xe đạp này rất chắc chắn"*: đặc trưng "*khung xe*" là đặc trưng hiện.

+ "*Xe này đi chậm quá* ": đặc trưng "*tốc độ*" là đặc trưng ẩn

❖ Đoạn đánh giá (opinion passage) về một đặc trưng:

Đoạn đánh giá về một đặc trưng *f* của đối tượng *O* trong *r* là một tập các câu liên tiếp trong *r* diễn tả quan điểm tích cực hay tiêu cực về đặc trưng *f*.

Thông thường, đoạn đánh giá là chuỗi các câu trong văn bản thể hiện một quan điểm cho một đối tượng nào đó hoặc một đặc trưng của đối tượng. Nó cũng có thể là một câu đơn thể hiện quan điểm cho nhiều hơn một thuộc tính.

Ví dụ: "Kiểu dáng xe đẹp nhưng xe không được bền lắm"

Đoạn đánh giá bao gồm tối thiểu ít nhất một câu. Hầu hết các nghiên cứu hiện tại tập trung vào mức câu: mỗi một đoạn bao gồm một câu. Khái niệm đoạn và câu được dùng tương đương về ngữ nghĩa trong ngữ cảnh này.

❖ Quan điểm hiện, ẩn:

Quan điểm hiện (explicit opinion) về một đặc trưng f là một câu thể hiện quan điểm mang tính chủ quan, diễn trả trực tiếp quan điểm tích cực hay tiêu cực của tác giả. Quan điểm ẩn (implicit opinion) về một đặc trưng *f* là câu thể hiện quan điểm tích cực hay tiêu cực một cách không tường minh.

Ví dụ:

+ Quan điểm hiện : *"Xe đạp này bền đấy*"

+ Quan điểm ẩn : "*Tai nghe mới mua mà đã hỏng*"

❖ Người đánh giá (opinion holder):

Là người hay tổ chức cụ thể đưa ra lời đánh giá. Trong trường hợp, đánh giá sản phẩm, forum, blogs người đánh giá luôn là các tác giả của đánh giá hay bài viết đó.

1.1.2 Các bài toán điển hình trong khai phá quan điểm

Theo Bing Liu [4] có ba bài toán chính trong khai phá quan điểm:

❖ Phân lớp quan điểm:

Với bài toán này có thể coi khai phá quan điểm như bài toán phân lớp văn bản. Bài toán phân lớp một văn bản đánh giá là tích cực hay tiêu cực.Ví dụ: với một đánh giá sản phẩm, hệ thống xác định xem nhận xét về sản phẩm ấy là tốt hay xấu. Phân lớp này thường là phân lớp ở mức tài liệu. Thông tin được phát hiện không mô tả chi tiết về những gì mọi người thích hay không thích.

❖ Khai phá và tổng hợp quan điểm dựa trên đặc trưng :

Bài toán đi sâu vào phát hiện quan điểm ở mức câu, tức là tìm hiểu các khía cạnh của đối tượng mà người đánh giá thích hay không thích. Một đối tượng có thể là một sản phẩm, một dịch vụ, một chủ đề, cá nhân hay một tổ chức… Ví dụ, trong một bài đánh giá về một sản phẩm, bài toán phải xác định các đặc trưng của sản phẩm mà người dùng bình luận và xác định các ý kiến là tích cực hay tiêu cực.

Ví dụ: Trong một câu, "tuổi thọ của pin này quá ngắn".

Người dùng nhận xét về "tuổi thọ của pin" và ý kiến là tiêu cực.

Kết quả, bài toán đưa ra một bản đánh giá tổng hợp quan điểm về đối tượng được đề cập .

❖ Khai phá quan hệ (so sánh câu):

So sánh là một loại đánh giá khác, tức là so sánh trực tiếp một đối tượng đối với một hoặc nhiều đối tượng khác tương tự. Ví dụ: câu sau so sánh hai máy ảnh:

"Tuổi thọ pin của máy ảnh này ngắn hơn nhiều so với máy ảnh B". Chúng ta muốn xác định câu đó và trích xuất các mối quan hệ so sánh thể hiện trong đó.

Trong giới hạn của khóa luận, bài toán phân lớp quan điểm: Coi khai phá quan điểm như là phân lớp văn bản được đề cập : mỗi văn bản thể hiện một quan điểm và quá trình phân lớp quan điểm chính là phân lớp văn bản. Các quan điểm được phân vào hai lớp

tích cực (tốt) và tiêu cực (xấu), và không quan tâm tới lớp trung lập (neutral) bởi những nhận định mang tính trung lập không ảnh hưởng tới kết quả tổng hợp quan điểm.

1.2 Ý nghĩa và ứng dụng của bài toán khai phá quan điểm

1.2.1 So sánh sản phẩm

Đó là một thực tế phổ biến cho các dịch vụ thương mại trực tuyến để yêu cầu khách hàng của họ đưa ra ý kiến về sản phẩm mà họ đã mua. Vì các doanh nghiệp luôn muốn biết ý kiến của người dùng về sản phẩm và dịch vụ của họ sau khi sản xuất. Song song với điều đó, khách hàng trước khi mua một sản phẩm hay sử dụng một dịch vụ nào đó , đều muốn biết ý kiến của người sử dụng trước đó. Hầu hết các nghiên cứu về những đánh giá đã được tập trung vào tự động phân loại các sản phẩm vào "bình luận" hoặc “"không bình luận". Nhưng mỗi sản phẩm có nhiều tính năng, trong đó có thể chỉ là một phần trong số chúng được người dùng quan tâm. Như vậy, phân tích đánh giá trực tuyến giúp cho khách hàng và doanh nghiệp thấy rõ những lợi thế và điểm yếu của sản phẩm.Với khách hàng, họ có thể dễ dàng đưa ra quyết định mua sản phẩm. Đối với nhà sản xuất, so sánh giúp họ nhanh chóng tiếp nhận thông tin tiếp thị để đưa ra chiến lược phát triển sản phẩm hiệu quả.

1.2.2 Tổng hợp quan điểm

Một vấn đề với các đánh giá sản phẩm trực tuyến là số lượng lớn đặc biệt là các sản phẩm phổ biến. Hơn nữa, nhiều đánh giá dài nhưng chỉ có một số câu chứa quan điểm đánh giá sản phẩm. Điều này gây khó khăn cho khách hàng khi quyết định mua một sản phẩm .Số lượng lớn các đánh giá cũng làm cho các nhà sản xuất khó theo dõi ý kiến khách hàng của họ. Tổng hợp quan điểm là bản đánh giá xác định quan điểm phân cực, mức độ và liên quan của các sự kiện. Với tổng hợp quan điểm, người dùng dễ dàng xác định ý kiến của khách hàng hiện tại. Các nhà sản xuất có được những lý do giải thích tại sao người dùng thích hay không thích sản phẩm của mình .

1.3 Những khó khăn trong bài toán khai phá quan điểm tiếng Việt

Ta có thể thấy ba khó khăn cơ bản trong bài toán khai phá quan điểm trên miền dữ liệu tiếng Việt như sau:

Thứ nhất, chưa có một nghiên cứu về khai phá quan điểm tiếng Việt nào được công bố.

Thứ hai, trong các kỹ thuật khai phá quan điểm cần có một bộ từ vựng lớn. Với tiếng Anh, có Sentiwordnet và Wordnet hỗ trợ. Với tiếng Việt chưa có một bộ từ điển như vậy hỗ trợ cho khai phá quan điểm.

Thứ ba, dữ liệu bình luận thường ít, có nhiều từ lóng, thiếu dấu câu…gây khó khăn trong bước tiền xử lí dữ liệu.

Tổng kết chương

Chương này khóa luận trình bày tổng quát khai phá quan điểm. Một số khái niệm, các bài toán chính và ứng dụng của khai phá quan điểm trong thực tế. Khai phá quan điểm là bài toán gốc, quan trọng trước khi đi sâu vào bài toán con. Ở chương tiếp theo, khóa luận sẽ trình bày nội dung bài toán phân lớp quan điểm, một lớp bài toán chính trong khai phá quan điểm.

Chương 2. Phân lớp quan điểm

2.1 Giới thiệu phân lớp quan điểm

2.1.1 Khái niệm phân lớp quan điểm

Theo Huifeng Tang và cộng sự [9], phân lớp quan điểm bao gồm hai dạng phân lớp: phân lớp quan điểm nhị phân và phân lớp quan điểm đa lớp. Cho một tập văn bản cần đánh giá D ={d1, ... ,dn} và một tập đánh giá được xác định trước C = {tích cực(positive), tiêu cực(negative)}. Phân lớp quan điểm nhị phân là phân loại mỗi tài liệu di $\subset$ D vào một trong hai lớp: tích cực và tiêu cực. Nếu d thuộc lớp tích cực có nghĩa là tài liệu d thể hiện quan điểm tích cực. Ngược lại, d thuộc tiêu cực có nghĩa tài liệu d thể hiện quan điểm tiêu cực.

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

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

2.1.2 Một số phương pháp phân lớp quan điểm

Trong [4], Bing Liu đưa ra ba phương pháp chính để phân lớp quan điểm.

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

2.1.2.1 Phân lớp dựa vào cụm từ thể hiện quan điểm

Phương pháp phân lớp dựa vào từ thể hiện quan điểm tích cực hay tiêu cực trong mỗi văn bản đánh giá. Thuật toán mô tả dựa trên nghiên cứu của Turney[15], được thiết kế để phân loại đánh giá của khách hàng. Thuật toán này sử dụng kỹ thuật xử lí ngôn ngữ tự nhiên gọi là gán nhãn từ loại (part-of-speech). Đánh dấu cho một từ được xác định bởi cú pháp ngữ nghĩa của nó. Các loại nhãn chung cho ngữ pháp tiếng Anh là : danh từ, động từ, tính từ, trạng từ, đại từ, giới từ, từ chuyển tiếp và thán từ. Có nhiều loại sinh ra từ

các kiểu khác nhau của các loại này. Ở đây, ta có thể sử dụng bảng các nhãn từ loại của Pennn Treebank bảng 1

Thẻ	**Mô tả**	**Thẻ**	**Mô tả**
CC	Từ nối	PRP$	Đại từ sở hữu
CD	Số đếm	RB	Trạng từ
DT	Từ hạn định	RBR	Trạng từ, so sánh hơn
EX	Từ chỉ sự tồn tại	RBS	Trạng từ, so sánh hơn nhất
FW	Từ nước ngoài	RP	Tiền tố , hậu tố
IN	Giới từ, hoặc từ nối ngoài	SYM	Từ đại diện
JJ	Tính từ	TO	Trong
JJR	Tính từ, so sánh hơn	UH	Thán từ
JJS	Tính từ, so sánh hơn nhất	VB	Động từ, từ nguyên thể
LS	Danh sách mục đánh dâu	VBD	Động từ thì quá khứ
MD	Từ chỉ cách thức	VBG	Động từ, danh động từ, hiện tại hoàn thành
NNS	Danh từ, từ số nhiều	VBP	Động từ chia ở thời hiện tại thứ ba số nhiều
NNP	Đại từ, từ số ít	VBZ	Động từ chia ở thời hiện tại thứ ba số ít
NNPS	Đại từ số nhiều	WDT	Từ hạn định bắt đầu bằng Wh

PDT	Từ hạn định	WP	Đại từ bắt đầu bằng Wh
POS	Từ sở hữu	WP$	Đại từ sở hữu bắt đầu bằng Wh
PRP	Đại từ chỉ người	WRB	Trạng từ bắt đầu bằng Wh
NN	Danh từ, từ số ít	VBN	Động từ, từ quá khứ

Bảng 1. Bảng các nhãn từ loại của Pennn Treebank

Thuật toán được trình bày rõ trong [13] được chia làm ba phần:

Bước 1:

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

Ví dụ: "không đoán trước được".

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

Trong câu "hắn ta không đoán trước được âm mưu " thì mang hướng tích cực

Do đó, thuật toán trích chọn hai từ liên tiếp trong đó một từ là tính từ hoặc trạng từ , từ kia thể hiện ngữ cảnh. Hai từ được trích chọn nếu nhãn của chúng phù hợp với bất kì các mẫu nào trong bảng 2

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

Bảng 2.Nhãn của mẫu cho trích chọn với cụm có hai từ

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

Bước 2:

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

Độ tương đồng ngữ nghĩa giữa hai cụm từ tính theo công thức (a)

$$PMI(term_1, term_2) = \log 2\left(\frac{\Pr(term_1 \Lambda term_2)}{\Pr(term_1)\Pr(term_2)}\right) \quad \text{(a)}$$

Trong đó:

- Pr(term1 ^ term2): xác suất đồng xuất hiện của term1 và term2
- Pr(term1), Pr(term2) : xác suất mà term1, term2 xuất hiện khi thống kê chúng riêng rẽ.
- Log của tỉ lệ trên là lượng thông tin mà ta có được về sự hiện diện của một term khi ta quan sát term kia.

Xu hướng ngữ nghĩa hay quan điểm (SO) của một từ/cụm từ được tính dựa trên việc tính toán độ đo PMI của từ/cụm đó với 2 từ *"excellent"* và *"poor"* theo công thức (b)

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$ (b)

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

- Pr(term): số kết quả trả về (hits) của máy tìm kiếm khi truy vấn là *term.* Thêm 0,01 vào hits **tránh trường hợp chia cho 0**
- Pr(term1^term2): số kết quả trả về khi sử dụng máy tìm kiếm Alta Vista sử thêm toán tử NEAR.

$$SO(phrase) = \log_2\left(\frac{hits(pharseNEAR"excellent")hits("poor")}{hits(phraseNEAR"poor")hits("excellent")}\right)$$ (c)

Bước 3:

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

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

Kết quả: Hệ thống thu được độ chính xác thay đổi theo các miền ứng dụng khác nhau

- 84% với các đánh giá về ôtô
- 66% với các đánh giá về phim

2.1.2.2 Phân lớp dựa vào phương pháp phân lớp văn bản

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

Cách tiếp cận này được thử nghiệm với Bo Pang và các cộng sự [5] áp dụng để đánh giá xem phim thành hai lớp tích cực hay tiêu cực. Bài toán chỉ ra việc sử dụng unigram trong phân lớp cho kết quả thực nghiệm tốt khi sử dụng Bayesian hoặc SVM.

Kết quả thực nghiệm qua sử dụng 700 đánh giá tiêu cực và 700 đánh giá tích cực cho thấy các thuật toán phân lớp đạt độ chính xác là 81% và 82% tương ứng với hai

thuật toán Baysie hoặc SVM. Tuy nhiên đánh giá trung lập không được đề cập trong bài báo .

2.1.2.3 Phân lớp dựa vào hàm tính điểm số

Phương pháp phân lớp dựa vào tính điểm được Kushal Dave và cộng sự [13] đưa ra gồm hai bước sau:

Bước 1:

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

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

Trong đó:

- t_i là từ cần được tính điểm
- C là một lớp quan điểm; C' là lớp phần bù của C (not C)
- Pr(t|C): xác suất t xuất hiện ở lớp C, được tính bằng số lần xuất hiện của t trong lớp C
- Điểm số được chuẩn hóa trong khoảng [-1, 1]

Bước 2:

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

$$class(d_i) = \begin{cases} C & eval(d_i) > 0 \\ C' & eval(d_i) \leq 0 \end{cases} \quad \text{(e)}$$

với

$$eval(d_i) = \sum_j score(t_j)$$

Kết quả:

- Kiểm thử trên 13000 đánh giá của 7 sản phẩm
- bigrams và trigrams cho kết quả chính xác cao nhất, từ 84.6% tới 88.3%
- Không loại bỏ từ dừng…

2.1.3 Phân lớp dựa vào kỹ thuật học máy

Có rất nhiều phương pháp có thể xác định tính phân cực của một tài liệu. Trong [9], Huifeng Tang và cộng sự đề cập phương pháp sử dung kỹ thuật học máy để xác định tích cực hay tiêu cực của bình luận với việc chuẩn bị dữ liệu học bằng tay. Các tác giả cũng đưa ra hai vấn đề quan trọng khi phân lớp quan điểm dựa vào kỹ thuật học máy : trích chọn đặc trưng và huấn luyện bộ phân lớp.

2.1.3.1 Trích chọn đặc trưng

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

❖ Dựa vào từ vựng :

Có hai phương pháp dựa vào từ vựng được sử dụng :

Thứ nhất là dựa trên từ điển WordNet. WordNet thay thế các từ trong đánh giá bởi một tập từ đồng nghĩa chung với nó có trong WordNet. Bởi vì các từ trong các bình luận có thể không phải là từ phổ biến trong đánh giá. Có rât nhiều nghiên cứu sử dụng kỹ thuật này, nghiên cứu gần đây nhất là của Taboada [16]. Nghiên cứu mô tả và so sánh các phương pháp để tạo bộ từ điển tương ứng với ngữ nghĩa định hướng của nó(SO). Qua thực nghiệm, tác giả cho thấy hiệu quả của việc sử dụng từ điển để xác định hướng ngữ nghĩa cho văn bản. Để trích chọn riêng mỗi từ, tác giả sử dụng một phương pháp dựa trên độ đo thông tin lẫn nhau (PMI). Thông tin lẫn nhau giữa một tập từ mồi và tập từ mục tiêu được tính toán bằng cách sử dụng hai phương pháp khác nhau: một là tìm kiếm NEAR dựa vào kỹ thuật tìm kiếm Altavista, hai là AND tìm kiếm trên Google. Hai tập từ điển được thử nghiệm so tập từ điển được gán nhãn tích cực, tiêu cực bằng tay. Kết quả của ba phương pháp khá gần nhau và không một phương pháp nào trong số họ thực sự có hiệu quả đặc biệt . Bài báo cũng chỉ ra hướng nghiên tiềm năng trong việc tính toán độ đo thông tin lẫn nhau PMI bằng cách sử dụng Google.

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

❖ Đánh giá tính từ:

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

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

Beineke và cộng sự [3] mở rộng phương pháp này bằng cách trích chọn tập các đặc trưng được kết hợp tuyến tính để xác định hướng quan điểm. Với phương pháp này có thể nâng cao kết quả so với phương pháp gốc, chủ yếu thông qua hai chiến lược : tích hợp các đặc trưng bổ sung vào mô hình và có thể sử dụng gán nhãn dữ liệu để đánh giá ảnh hưởng của chúng trong ngữ cảnh. Đặc biệt, khóa luận tập trung vào phương pháp của Takamura và cộng sự [21] sử dụng kỹ thuật trích chọn đặc trưng chuỗi từ con và cây con phụ thuộc của câu dựa trên thuật toán tính tần suất khai phá mẫu và kết hợp sử dụng máy hỗ trợ vecto. Các thuật toán này sẽ được khóa luận trình bày kỹ hơn trong mục 2.2

2.1.3.2 Huấn luyện bộ phân lớp SVM nhị phân

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

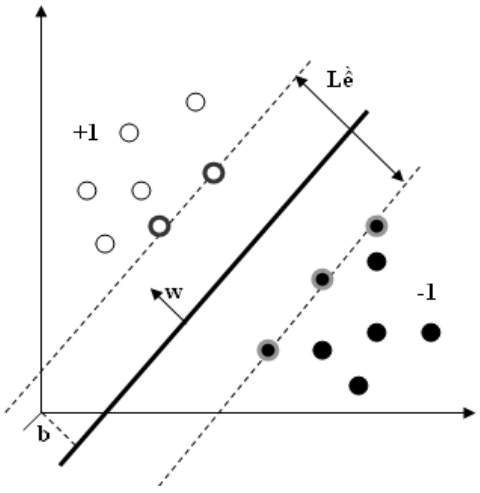
Mặt khác, trong số các công cụ trên , SVM được chứng minh là công cụ phân lớp mạnh, hiệu quả hơn phân lớp văn bản truyền thống như Naïve Bayes [22]. Thêm vào đó, B.Pang và các cộng sự [5] áp dụng kỹ thuật Naïve Bayes , maximum entropy và SVM để

xác định hướng quan điểm phân cực trong bình luận về phim. Kết quả phân lớp sử dụng mô hình unigram và phân lớp SVM đạt kết quả cao nhất 82.9% . Điều đó cho ta thấy rằng SVM vẫn là một công cụ hiệu quả cho phân lớp quan điểm.

SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis [8] xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dựng trong thực tiễn. Tư tưởng chung của học máy SVM như sau:

- Giai đoạn xây dựng mô hình: Cho một tập mẫu dữ liệu huấn luyện đã được gán nhãn lớp, như vậy có một tập nhãn lớp tương ứng xác định tên tập mẫu. Mỗi mẫu dữ liệu được biểu diễn dưới dạng một vector đặc trưng. Dựa vào vector đặc trưng của các mẫu dữ liệu huấn luyện, mô hình máy vector hỗ trợ sẽ được xây dựng để phân tách các mẫu học. Trong trường hợp khả tách tuyến tính, nó là một siêu phẳng (hyperplane) trong không gian dùng để phân tách tuyến tính các mẫu thuộc các nhãn lớp khác nhau với khoảng cách lớn nhất có thể. Trong trường hợp không khả tách tuyến tính, chúng ta có thể sử dụng lề mềm (soft margin) để phân tách mẫu học, hay sử dụng ánh xạ phi tuyến để chuyển không gian ban đầu sang không gian mới có số chiều lớn hơn mà ở đó các mẫu học có khả năng phân tách tuyến tính.
- Giai đoạn sử dụng mô hình: Mô hình đã xây dựng sẽ được sử dụng để gán nhãn lớp cho các mẫu dữ liệu mới.

a. Trường hợp khả tách tuyến tính



Hình 1. Mô hình máy vector hỗ trợ khả tách tuyến tính

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

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

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

Bộ phân lớp SVM được định nghĩa như sau:

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

trong đó

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

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

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

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

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

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

Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán:

Cực đại hóa:

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

với các ràng buộc:

$$\alpha_i \geq 0 \quad \text{(k)}$$

$$\text{và } \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{(l)}$$

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

b. Trường hợp không khả tách tuyến tính

Có thể giải quyết theo 2 phương pháp sau:

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

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

$$w^T x_i + b \geq +1 - \xi \text{ nếu } y_i = +1 \quad \text{(n)}$$

$$w^T x_i + b \leq -1 + \xi \text{ nếu } y_i = -1 \quad \text{(o)}$$

$$\xi \geq 0 \quad \text{(p)}$$

Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán:

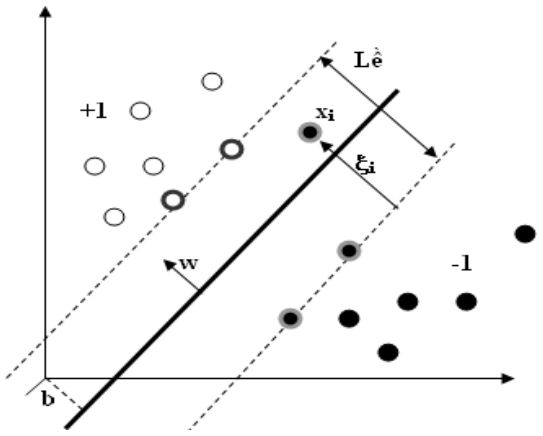
Cực đại hóa:

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

với các ràng buộc

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (r)$$

$$0 \leq \alpha_i \leq C \qquad (s)$$



Hình 2.Phương pháp lề mềm

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

$$\Phi : R^d \rightarrow R^D \ (D >> d)$$

$$x \rightarrow \Phi(x)$$

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

Cực đại hóa:

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

với các ràng buộc:

$$0 \leq \alpha_i \leq C \qquad (u)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (v)$$

trong đó k là một hàm nhân thoản mãn:

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

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

- Đa thức: $k(x_i, x_j) = (gamma.x_i^T x_j + coef0)^{degree}$
- Hàm vòng RBF (Radial Basic Function): $k(x_i, x_j) = \exp(-gamma.|x_i - x_j|^2)$
- Hàm chữ S Sigmoid: $k(x_i, x_j) = \tanh(gamma.x_i^T.x_j + coef0)$

trong đó gamma, coef0 và degree là các tham số nhân.

2.1.4 Các công trình nghiên cứu liên quan

Có ba mức phân lớp quan điểm : mức từ, mức câu và mức tài liệu.

2.1.4.1 Phân lớp quan điểm mức từ

Trong [23], Hatzivassiloglou và các cộng sự sử dụng các biểu thức kết nối như " nhanh và đẹp", " nhanh nhưng không chính xác"… để trích xuất ra các từ thể hiện quan điểm phân cực.

Cùng là phân lớp mức từ , nhưng Turney [15] sử dụng phương pháp xác định độ tương đồng giữa hai từ bằng việc đếm số kết quả trả về từ web tìm kiếm. Quan hệ giữa từ phân cực chưa biết và một tập seed được lựa chọn để phân lớp từ chưa biết vào lớp tích cực hay tiêu cực. Thuật toán đạt độ chính xác trung bình là 74% với 410 bình luận từ Epinions.

2.1.4.2 Phân lớp quan điểm mức câu

Trong [18], Taka Kudo sử dụng cây con của cây phụ thuộc như đặc trưng cho một câu phân lớp quan điểm. Các tác giả sử dụng thuật toán boosting với cây con quyết định như việc học yếu. Các tác giả cũng đưa ra quan hệ giữa các thuật toán với mô hình SVM và cây nhân (tree kernel). Hai thực nghiệm phân lớp quan điểm chứng tỏ đặc trưng của cây con là quan trọng.

2.1.4.3 Phân lớp quan điểm mức tài liệu

B.Pang và cộng sự [5] tiến hành thực nghiệm phân lớp quan điểm trên dữ liệu các bình luận về phim. Các tác giả áp dụng phân lớp quan điểm ở mức tài liệu sử dụng kỹ

thuật học máy giám sát để phân lớp tài liệu. Để trích chọn đặc trưng, họ sử dụng mô hình n-gram trong tập dữ liệu xem như đặc trưng bag-of-word để phân lớp. Một từ n-gram là một tập n từ liên tiếp trích xuất từ một câu. Kết quả tốt nhất từ mô hình dựa trên unigram chạy qua SVM , với độ chính xác 82.9%.

Bo Pang [6] đã cải tiến phương pháp phân lớp học máy bằng cách chỉ sử dụng các câu thể hiện quan điểm chủ quan trong bình luận. Nhưng độ chính xác của phương pháp này thấp hơn so với phân lớp bình luận đầy đủ được giới thiệu trong nghiên cứu trước đây của B. Pang [5].

Kushal Dave và cộng sự [13] sử dụng phương pháp học máy để phân lớp bình luận trên một số loại sản phẩm. Không giống như nghiên cứu của Bo Pang, các tác giả thu được kết quả tốt nhất với mô hình phân lớp dựa vào từ bigram trên dữ liệu của họ. Kết quả này đã chỉ ra rằng mô hình dựa trên unigram không phải luôn luôn tốt nhất và cách phân lớp tốt nhất là dựa vào dữ liệu.

Để sử dụng những tri thức có sẵn trong tài liệu, Mullen và Collier [17] áp dụng định nghĩa hướng ngữ nghĩa của từ bởi Peter Turney [15] và một số loại thông tin trên Internet và lí thuyết. Các tác giả đánh giá trên tập dữ liệu của Bo Pang [5] và đạt độ chính xác lên đến 84.6% với mô hình n-gram và hướng ngữ nghĩa của từ.

2.2 Thuật toán tính tần suất mẫu

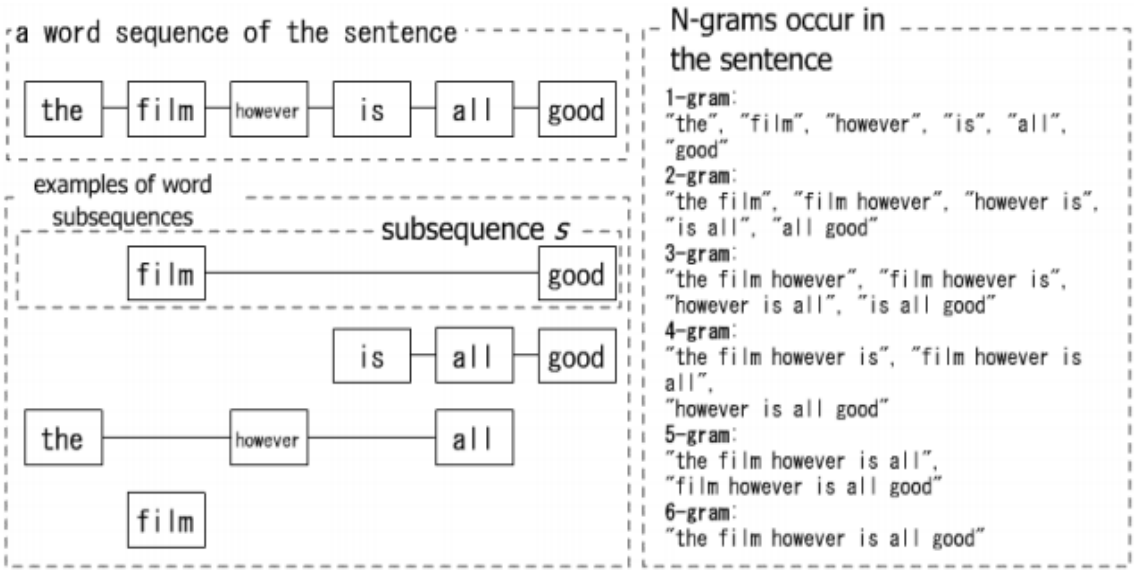
2.2.1 Chuỗi từ con

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

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

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

Ví dụ: n-grams không thể hiện được sự đồng xuất hiện của "film" và "good", khi một từ khác xuất hiện giữa hai từ như trong hình 3. Ngược lại, với chuỗi con luôn chứa mẫu "film-good", được chú ý bởi s trong hình.



Hình 3.Một ví dụ chuỗi con trong câu " The film however is all good"

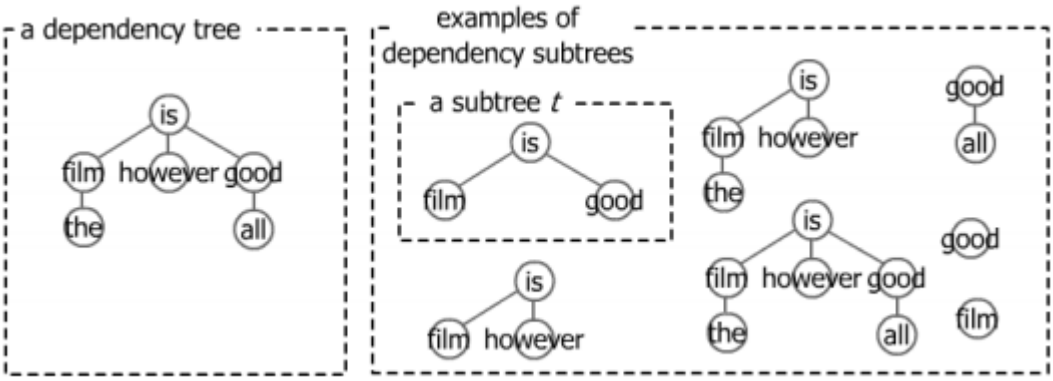
2.2.2 Cây con phụ thuộc

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

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

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

Ví dụ : trong hình 4, thể hiện quan hệ giữa các từ "good" và "film" , sự phụ thuộc cây con t (được chú ý như là is((film)(good))) không chỉ thấy được sự đồng xuất hiện của từ "good" và "film", mà còn bảo đảm "good" và "film" được kết nối cú pháp với nhau qua từ "is".



Hình 4.: Một ví dụ cây con phụ thuộc trong câu " The film however is all good"

2.2.3 Thuật toán tính tần suất mẫu

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

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

2.2.3.1 Tần suất khai phá chuỗi con. Thuật toán PrefixSpan

❖ Một số định nghĩa

- Định nghĩa 1(prefix, projection, postfix) :

  Giả sử có tất cả các items của một thành phần được sắp xếp theo thứ tự a,b,c. Với một chuỗi $\alpha = \langle e_1 e_2 e_3 \ldots\ldots e_n \rangle$ và một chuỗi $\beta = \langle e'_1 e'_2 e'_3 \ldots e'_m \rangle$ ($m \leq n$) là tiền tố (prefix) của α khi và chỉ khi:

  - $e_i' = e_i$ for($i \leq m-1$)
  - $e_{m'} \subseteq e_m$
  - Tất cả các items trong $(e_m - e'_m)$ được sắp xếp sau e'_m

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

- α' có tiền tố β
- Không tồn tại chuỗi α'' nào là tiền tố của β mà lớn hơn α'

Với a'= $\langle e_1 e_2 e_3 ..... e_n \rangle$ là hình chiếu(projection) của α tương ứng với tiền tố β = $\langle e_1 e_2 e_3 ..... e_{m-1} e'_m \rangle$

Chuỗi con $\gamma = (e''_m e_{m+1} ... e_n)$ gọi là hậu tố của α tương ứng với tiền tố của β khi $\gamma = \alpha / \beta$ với $e''_m = (e_m - e'_m)^2$.

Ví dụ : cho một dãy α= <a(abc)(ac)d(cf)>

- <a>, <aa>, <a(ab)> và <a(abc)> là tiền tố α
- <(abc)(ac)d(cf)> là hậu tố chuỗi α tương ứng tiền tố <a>
- <(_bc)(ac)d(cf)> là hậu tốt tương ứng tiền tố <aa>
- <(_c)(ac)d(cf)> là hậu tố tương ứng la tiền tố <ab>

- Định nghĩa 2:

Cho α là một chuỗi trong dữ liệu S. α-projected database kí hiệu là $S|_\alpha$, là tất cả các postfixes của các chuỗi trong S tương ứng là tiền tố của α

- Định nghĩa 3:

Cho chuỗi α trong chuỗi dữ liệu S và β là một chuỗi có tiền tố là α. Độ hỗ trợ(support count)của β trong α-projected database $S|_\alpha$, kí hiệu support $S|_\alpha(\beta)$, là số lượng chuỗi γ trong $S|_\alpha$, ký hiệu

Prefix	Projected database	Sequential patterns
<a>	<(abc)(ac)d(cf)>,<(_d)c(bc)(ae)>,<(_b)(df)cb>, <(_f)cbc>	<a>, <aa>,<ab>, <a(bc)>, <a(bc)a>, <aba>, <abc>, <(ab)>,<(ab)c>, <(ab)d>,
<b>	<(_c)(ac)d(cf)>,<(_c)(ae)>,<(af)cb>,<c>	<b>,<ba>,<bc>,<(bc)>,<(bc)a>,<bd>,<bdc><bf>
<c>	<(ac)d(cf)>,<(bc)(ae),<b>,<bc>	<c>,<ca>,<cb>,<cc>
<d>	<(cf)>,<c(bc)(ae)>,<(_f)cb>	<d>,<db>,<dc>,<dcb>
<e>	<(_f)(ab)(df)cb>,<(af)cb>	<e>,<ea>,<eab>,<eac>,<eacb>,<eb>, <ebc>,<ec>, <ecb>,<ef>,<efb>,<efc>,<efcb>
<f>	<(ab)(af)cb>,<cbc>	<f>,<fb>,<fbc>,<fc>,<fcb>

Bảng 3.prefix, postfix và các mẫu tuần tự tương ứng

❖ Thuật toán PrefixSpan

Thuật toán Prefixspan [10] tính tần suất của tất cả các chuỗi con trong tập dữ liệu của câu. Đầu tiên, thuật toán bắt đầu với một tập hợp tần suất của các chuỗi con gồm các từ đơn(single items). Sau đó, thuật toán được mở rộng , với mỗi chuỗi con có kích thước k gắn thêm một từ mới để tính được tần suất của dãy con có kích thước k+1. Thuật toán tính được tất cả tần suất của chuỗi con thông qua lặp đệ quy.

Tuy nhiên, việc mở rộng chuỗi con bằng cách thêm một nút mới vào bất kì vị trí của lá có thể dẫn đến tình trạng trùng lặp các cây con mới được sinh ra. Để tránh điều này, thuật toán hạn chế vị trí để đính kèm một nút mới vào cuối cây con mới theo thứ tự từ trái sang phải.

Thuật toán prefixspan được mô tả chi tiết như sau:

a. Đầu vào: Một chuỗi dữ liệu S, và độ ngưỡng hộ trợ min_sup.

b. Đầu ra : Tập các mẫu liên tiếp sinh ra từ chuỗi dữ liệu ban đầu.

c. Hàm : PrefixSpan(α, l, $S|_\alpha$)

d. Tham số: α: là mẫu liên tục ; l: độ dài của α; $S|_\alpha$: α-projected database, nếu α # $\langle\rangle$; và ngược lại, chuỗi dữ liệu S

e. Phương thức :

   a. Quét $S|_\alpha$ một lần, tìm tập các tần suất items b như sau:

      i. b có thể được thêm vào phần tử cuối của α để thành một mẫu tuần tự hoặc

      ii. <b> có thể thêm vào α mẫu tuần tự

   b. Lặp với mỗi item thường xuyên b, thêm nó vào chuỗi α để tạo thành chuỗi mới α', in ra α'.

   c. Với mỗi α, sinh ra α'-projected database $S|_{\alpha'}$, và gọi lại hàm PrefixSpan(α', l+1, $S|_{\alpha'}$)

❖ Đánh giá :

- PrefixSpan chỉ tăng số lượng các mẫu tuần tự dài hơn từ các mẫu ngắn hơn của nó. Thuật toán không tự tạo ra và cũng không kiểm tra được bất kì các chuỗi ứng viên(candidate) nào không tồn tại trong cơ sở dữ liệu dự kiến. So với thuật toán GSP, quá trình sinh và kiểm tra một số lượng lớn các của các chuỗi, PrefixSpan có không gian tìm kiếm nhỏ hơn.
- Một cơ sở dữ liệu dự kiến thường nhỏ hơn cơ sở dữ liệu gốc bởi vì chỉ có các chuỗi con hậu tố của các tiền tố thường xuyên mới được đưa vào cơ sở dữ liệu dự kiến.
- Chi phí chính của PrefixSpan là xây dựng cơ sở dữ liệu dự kiến. Trong trường hợp xấu nhất, PrefixSpan xây dựng mỗi cơ sở dữ liệu dự kiến cho mỗi mẫu tuần tự .

2.2.3.2. Tần suất khai phá cây con. Thuật toán Freqt

Thuật toán freqt tính tần suất của tất cả các cây con trong một cây được Kenji Abe và cộng sự mô tả chi tiết trong [12][20]. Đầu tiên, thuật toán bắt đầu với một tập hợp tần suất của các cây con gồm các từ đơn(single node). Sau đó, thuật toán được mở rộng , với mỗi cây con có kích thước k gắn thêm một từ mới để tính được tần suất của cây con có kích thước k+1. Thuật toán tính được tất cả tần suất của chuỗi con thông qua lặp đệ quy.

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

Đầu vào : tập nhãn L của cây cấu trúc D và độ hộ trợ nhỏ nhất $0 < \sigma \leq 1$

Đầu ra : Tập F của tất cả các mẫu có $\sigma - frequent$ trong D

- Gán tập $C_1 = F_1$ của 1 mẫu và tập RMO_1 của các nhánh đồng xuất hiện bên phải , bằng cách quét toàn bộ tập D, gán k = 2
- Trong khi $F_{k-1} \neq \varnothing$ lặp:
  - $< C_k, RMO_k > := \mathrm{Ex}pand - Trees(C_{k-1}, RMO_{k-1}); F_k := \varnothing$
  - Với mỗi mẫu T , $T \in C_k$, thực hiện các bước sau :
    - ➔ tính lại $freqt_D(T)$ từ $RMO_k(T)$. Nếu $freqt_D(T) \geq \sigma$ thì gán $F_k = F_k \cup T$
- Trả lại F = $F_1 \cup ..... \cup F_{k-1}$

Trong đó:

- Occ(T) là tập các gốc đồng xuất hiện của T trong D, $Occ(T) = \{\mathrm{Roo}t(\varphi)\}$, φ là một hàm chuyển của cây T sang tập D
- $freq_D(T)$ xác định bởi số lượng các nút gốc khác nhau của T trên tổng số nút trong D, $freq_D(T) = \#Occ(T)/|D|$

- Minimum support : σ của mẫu T trong D với $0 < \sigma \leq 1$, khi đó một mẫu T là σ-frequent trong D nếu $freq(T) \geq \sigma$

❖ Đánh giá thuật toán:

Vấn đề phát hiện các mẫu $\sigma - frequent$ là khó để áp dụng vào thực tế. Tuy nhiên, phương pháp này chỉ hiệu quả để giải quyết những bài toán khai phá dữ liệu phức tạp hơn như phát hiện mẫu thường xuyên với số lượng tài liệu chính là tập đầu vào của cây và tìm ra mô hình tối ưu hóa phương pháp thống kê như sử dụng độ đo thông tin entropy.

Tổng kết chương

Chương này giới thiêu khái niệm và một số phương pháp giải quyết bài toán phân lớp quan điểm. Trong đó, khóa luận tập trung vào phương pháp phân lớp dựa vào kỹ thuật học máy. Bên cạnh đó, khóa luận cũng trình bày hai thuật toán tính tần suất khai phá mẫu: prefixspan và freqt . Từ so sánh với mô hình n-gram, ta thấy chuỗi con và cây con phụ thuộc cho thông tin giàu ngữ nghĩa hơn n-gram. Đây là lí do, khóa luận dựa vào tần suất mẫu để trích trọn chuỗi con và cây con phụ thuộc làm đặc trưng phân lớp quan điểm trong mô hình ở chương sau.

Chương 3. Mô hình đề xuất bài toán phân lớp quan điểm theo chủ đề trên miền tin tức tài chính

3.1 Phân lớp quan điểm trên miền tài chính

Trong nền kinh tế phát triển ngày nay, thông tin có tác động rất lớn đến thị trường đặc biệt là thông tin tài chính. Những bản tin tài chính , những đánh giá của các chuyên gia có thể gây biến động về giá vàng, giá cổ phiếu, khối lượng giao dịch, thậm chí là mức thu nhập của một công ty, doanh nghiệp nào đó trong tương lai. Từ nhận định hay đánh giá về thị trường, nhà đầu tư quyết định mua vào hoặc bán ra một loại cổ phiếu. Trong nghiên cứu của mình, Robert Engle[2] đã mô tả toán học về sự tác động không đối xứng về mặt tin tức giá cả. Tác giả cho rằng với những tin tức tốt thường liên quan đến những thay đổi lớn về giá nhưng chỉ trong một thời gian ngắn, ngược lại những ảnh hưởng của tin tức tiêu cực về giá dẫn đến khối lượng giao dịch kéo dài hơn. Vậy làm sao, có thể xác định được hướng quan điểm của các bản tin tài chính, để từ đó các công ty, nhà đầu tư nhanh chóng đưa ra quyết định đầu tư vào thị trường. Đây là một bài toán khó, thách thức lớn cho những nhà nghiên cứu, chuyên gia về lĩnh vực này. Bởi lẽ, dữ liệu về tin tức tài chính chiếm số lượng lớn và thường xuyên được cập nhật. Yêu cầu đặt ra cho bài toán là tự động phân lớp hướng quan điểm của tin tức tài chính thông qua phân tích các đánh giá về tin tức đó.

Trên thế giới có rất nhiều nghiên cứu về lĩnh vực này, Ahmad và các cộng sự [11] đã nghiên cứu phương pháp xác định thông tin tích cực và tiêu cực trong luồng thông tin và xác định sự ảnh hưởng của các tin tức . Các tác giả xác định một sự kiện tin tức gây tranh cãi có thể dẫn đến những quan điểm khác nhau và sử dụng các tin tức tiếp theo như là ngữ liệu(corpus).Kết quả thực nghiệm đạt độ chính xác lên đến 70%. Tuy nhiên, các tác giả phân tích dữ liệu duy nhất trên bản tin mà chưa phân tích đến các đánh giá của người bình luận. Trong [14], O'Hare và cộng sự chỉ ra rằng dữ liệu tin tức nói chung thường thể hiện quan điểm khách quan và không phải là nguồn thông tin lý tưởng cho khai phá quan điểm. Mặt khác, xác định được những khó khăn với trên nguồn dữ liệu tin tức tài chính tiếng Việt như dữ liệu đánh giá ít và có độ nhiễu cao, mục đích của khóa luận là làm sao có thể xác định hướng quan điểm từ các bình luận của độc giả về tin tức tài chính đó.

Có rất nhiều kỹ thuật trích chọn đặc trưng phân lớp, trong đó mô hình n-gram là điển hình. B.Pang [6] tiến hành thực nghiệm phân lớp quan điểm trên dữ liệu các bình

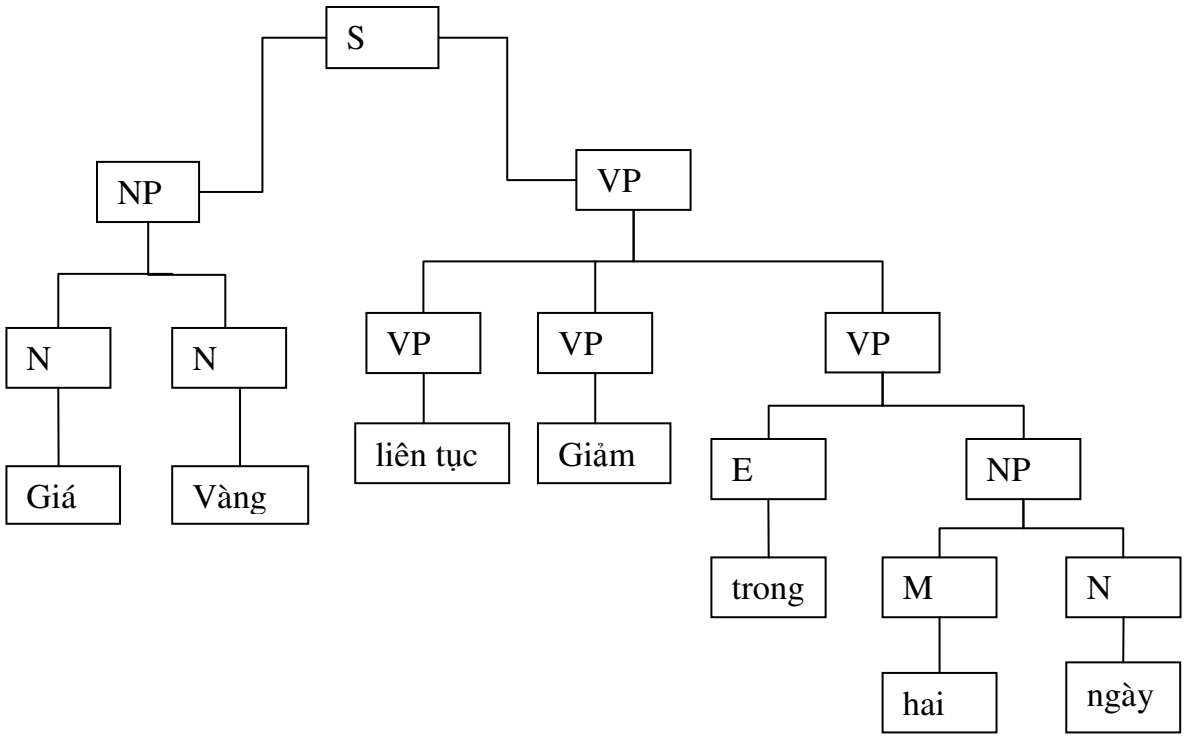
luận về phim. Kết quả tốt nhất từ mô hình dựa trên unigram kết hợp SVM , với độ chính xác 82.9%. Không giống như kết quả nghiên cứu của Pang, Kushal Dave và cộng sự [13] sử dụng phương pháp học máy để phân lớp đánh giá sản phẩm. Các tác giả thu được kết quả tốt nhất với mô hình phân lớp dựa vào bigram. Kết quả này đã chỉ ra rằng mô hình dựa trên unigram không phải luôn luôn tốt nhất và cách phân lớp tốt nhất là dựa vào dữ liệu . Đây là nhận định của Shotaro Matsumoto và cộng sự [21]. Từ đó, các tác giả đưa ra phương pháp trích chọn đặc trưng dựa vào tần suất của chuỗi con và cây con phụ thuộc của câu. Vì chuỗi từ và cây con phụ thuộc giữ được trật tự và thể hiện được mối quan hệ giữa các từ trong câu, mang lại thông tin giàu ngữ nghĩa hơn n-gram. Qua kết quả thực nghiệm, các tác giả đã chứng minh mô hình đưa ra khả thi. Trong [10], Huifeng Tang và cộng sự cũng đánh giá mô hình của các tác giả [21] là một trong những phương pháp học máy phân lớp quan điểm hiệu quả. Dựa vào đánh giá này và kết quả thực nghiệm, em chọn phương pháp của Shotaro Matsumoto và cộng sự làm tiền đề để xây dựng mô hình phân lớp quan điểm trên miền tin tức tài chính.

3.2. Cây phân tích cú pháp tiếng Việt

Theo một số nghiên cứu trong tài liệu [1], các tác giả chỉ ra rằng cây con phụ thuộc là cây được trích xuất từ cây phân tích cú pháp đầy đủ. Cây con phụ thuộc được sinh ra bằng cách sử dụng thông tin về các cụm từ trung tâm dựa vào phân tích cú pháp và liên kết tất cả các thành phần của cụm từ tới từ trung tâm của cụm từ đó.

Theo [24], với tập đầu vào là một chuỗi các từ tố (là kết quả của quá trình phân tích từ tố, thông thường đối với xử lý ngôn ngữ là các từ), tiến hành phân tích cú pháp (parsing hay syntatic analys) là quá trình phân tích nhằm đưa ra cấu trúc ngữ pháp của chuỗi từ đó dựa vào một văn phạm nào đó. Thông thường cấu trúc ngữ pháp được là ở dạng cây, bởi thông qua dạng này sự phụ thuộc của các thành phần là trực quan. Cây này được gọi là cây phân tích cú pháp.

Ví dụ : "Giá vàng liên tục giảm trong hai ngày "



Hình 5.Ví dụ cây phân tích cú pháp

Cấu trúc của cây cú pháp như sau:

- Nút gốc thể hiện loại câu (trần thuật, nghi vấn, cảm thán, cầu khiến)
- Các nút lá biểu diễn các từ trong câu
- Nút cha của các nút lá này biểu diễn nhãn từ loại tương ứng của nút con.
- Các nút trung gian còn lại thể hiện chức năng ngữ pháp (cụm danh từ, cụm động từ, bổ ngữ …)

3.3 Mô hình phân lớp quan điểm

Mục tiêu của bài toán phân lớp quan điểm theo chủ đề là xác định hướng quan điểm phân cực của các bình luận cho một nội dung tin tức tài chính nào đó

- Dữ liệu đầu vào:

  Nội dung bài viết và tập các bình luận đi kèm .

- Kết quả đầu ra:

  Xác định hướng quan điểm(tiêu cực, tích cực) của mỗi bình luận cho nội dung tin tức đưa ra.

Để xác định hướng quan điểm phân cực của mỗi bình luận, ta tập trung giải quyết bài toán phân lớp quan điểm. Dựa vào phương pháp trích chọn đặc trưng chuỗi con và cây con phụ thuộc kết hợp học máy SVM của Shotaro Matsumoto và cộng sự [21], khóa luận xây dựng được pha chính phân lớp quan điểm gồm 2 modul chính : trích chọn đặc trưng và sử dụng bộ phân lớp SVM.

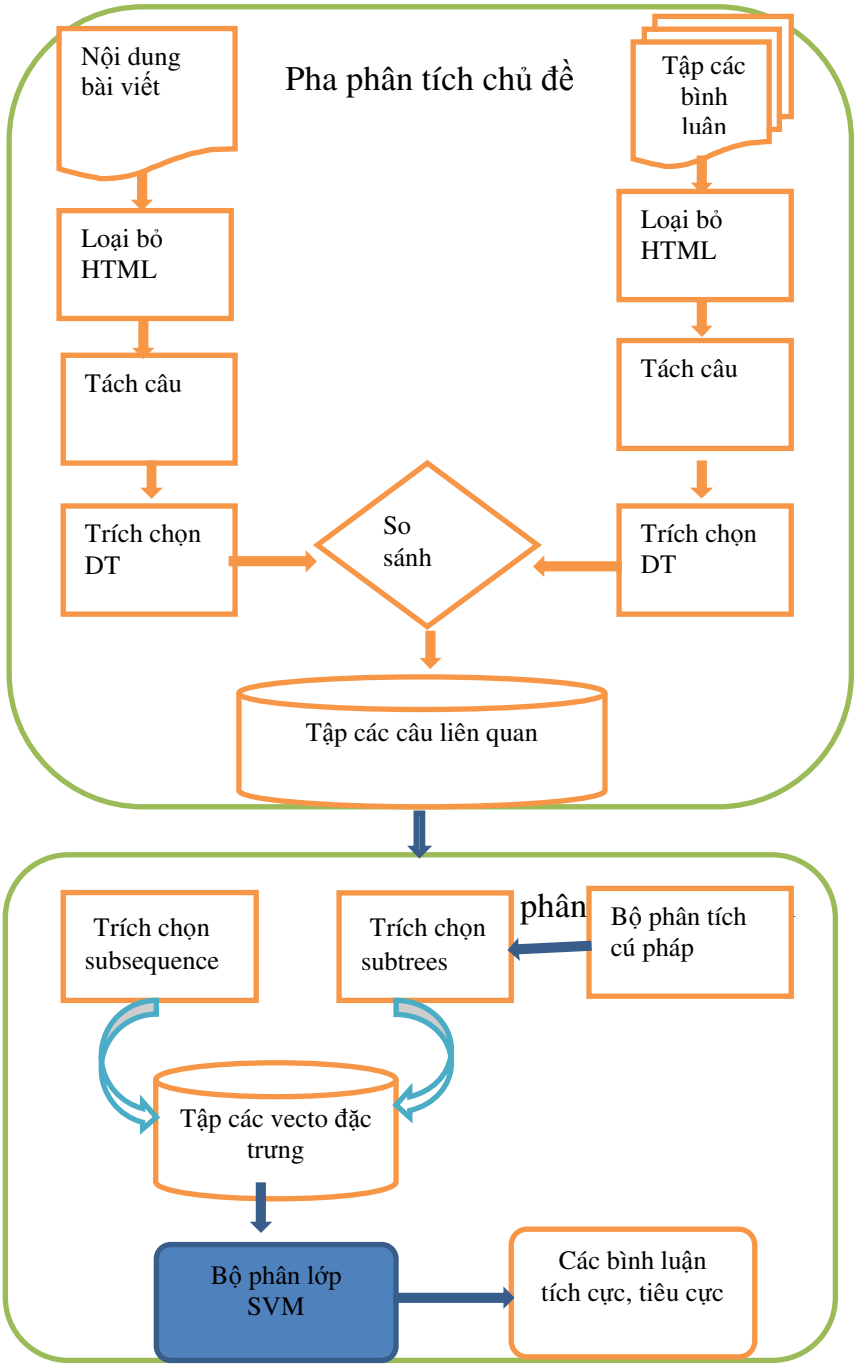
Mặt khác, khảo sát qua các bình luận trong bài báo , ta thấy một số lượng lớn các bình luận không đề cập đến nội dung bài báo đưa ra, coi đây là các dữ liệu nhiễu . Nhằm loại bỏ những bình luận này, ta xây dựng pha hai : phân tích chủ đề.

Mỗi pha được mô tả chi tiết trong mô hình như sau:

❖ Phân tích chủ đề : trích chọn các câu trong đoạn bình luận có liên quan đến chủ đề bài viết

❖ Phân lớp quan điểm : Trong pha này gồm 2 module:

- Trích chọn đặc trưng: trích chọn tần suất của chuỗi từ con, cây con phụ thuộc làm đặc trưng phân lớp.
- Bộ phân lớp SVM : phân lớp ở mức tài liệu. Số chiều của vecto tương ứng vơi số đặc trưng của tài liệu



Hình 6.Mô hình giải quyết bài toán

3.4 Phân tích các thành phần

3.4.1 Phân tích chủ đề

Trong pha này, nhận đầu vào là nội dung bài viết và một tập các bình luận trên trang tin tức tài chính vneconomy.vn.

Lần lượt từng trang sẽ được loại bỏ các thẻ html.Tiến hành tách câu tách từ, gán nhãn từ loại cho nội dung bài viết và tập các bình luận sử dụng bộ công cụ JvnTextPro[25].

Bởi vì mục tiêu là xác định hướng quan điểm của vấn đề mà bài báo đề cập. Tuy nhiên, qua khảo sát nội dung các bình luận cho thấy : các bình luận không chỉ đánh giá hướng quan điểm của vấn đề mà bài báo đề cập mà còn bày tỏ quan điểm đồng ý với những người bình luận trước họ, hoặc bàn về những vấn đề không liên quan đến nội dung bài báo. Ta coi những nội dung không liên quan, hoặc bày tỏ quan điểm với những người bình luận là dữ liệu nhiễu .

Để loại bỏ dữ liệu nhiễu này, coi mỗi danh từ trong bài viết như một đối tượng.Tiến hành thực hiện các bước sau :

+ Trích chọn các danh từ trong các câu bình luận .

+ So sánh danh từ trong từng câu bình luận với danh từ trong nội dung bài viết. Nếu câu nào chứa danh từ trùng với danh từ trong bài viết thì trích chọn câu đó. Khi đó với mỗi bài viết, ta trích chọn được một tập các câu bình luận liên quan đến chủ đề bài báo.

3.4.2 Trích chọn đặc trưng

Để sinh ra vecto đặc trưng của phân lớp, khóa luận kết hợp 4 đặc trưng : unigram, bigram, subsequences, subtrees. Với các đặc trưng được xác định cụ thể như sau:

a. Unigram là các từ đơn trong câu

Trong phạm vi bài toán , chỉ lấy các unigram xuất hiện ít nhất bốn lần trong tài liệu.

Ví dụ : Vàng |tăng |giá| liên |tục

b. Bigram là hai từ đơn , xuất hiện liên tiếp nhau trong câu.

Trong phạm vi bài toán, chỉ lấy các bigram xuất hiện ít nhất bốn lần trong tài liệu.

Ví dụ : Vàng tăng giá liên tục => vàng tăng| tăng giá| giá liên | liên tục

c. Tần suất các chuỗi con(seq)

Với tập các câu bình luận và tham số đầu vào ngưỡng hỗ trợ, ta sử dụng thuật toán prefixspan thu được các chuỗi con tương ứng. Trong phạm vi bài toán, chỉ lấy các chuỗi con có kích thước hai từ trở lên và lấy ngưỡng hộ trợ là 10 . Ở đây, nếu lấy ngưỡng hỗ trợ 10 tức là tần suất xuất hiện của các chuỗi con ít nhất là 10 trong tập dữ liệu.

So sánh với mô hình n-gram ta thấy rằng, trong khi n-grams chỉ thể hiện sự đồng xuất hiện của n từ liên tục trong một câu thì chuỗi từ con thể hiện sự đồng xuất hiện của một số lượng bất kì các từ liên tục hoặc không liên tục. Vì vậy, cùng với một câu đầu vào thì chuỗi con thu được thông tin giàu ý nghĩa hơn.

Ví dụ : với câu đầu vào là "Giá xăng tăng cao"

Chuỗi con	**n-grams**
Cao	Giá|xăng|tăng|cao
Giá cao|tăng cao| xăng cao|	Giá xăng| xăng tăng| tăng cao
Giá tăng cao|giá xăng cao|xăng tăng cao	Giá xăng tăng|xăng tăng cao|
Giá xăng tăng cao	Giá xăng tăng cao

Bảng 4.Bảng ví dụ chuỗi con

Trong một đoạn bình luận gồm nhiều câu, mỗi câu sinh ra nhiều chuỗi con. Vì vậy , ta phải đặt ra một ngưỡng hỗ trợ(tần suất xuất hiện của các chuỗi) để thu được những đặc trưng tốt. Ta có thể mô tả qua ví dụ sau:

- Đầu vào : gồm hai câu:
  - Giá xăng tăng cao
  - Giá cả cũng tăng theo
- Đầu ra
  - Với ngưỡng hỗ trợ là 1 , ta có tập các chuỗi con

```
cao/1
cả/1 theo/1
cả/1 tăng/1 theo/1
giá/2 cao/1
giá/2 cả/1 theo/1
giá/2 cả/1 tăng/1 theo/1
giá/2 theo/1
giá/2 tăng/2 cao/1
giá/2 tăng/2 theo/1
giá/2 xăng/1 cao/1
giá/2 xăng/1 tăng/1 cao/1
theo/1
tăng/2 cao/1
tăng/2 theo/1
xăng/1 cao/1
xăng/1 tăng/1 cao/1
```

- Với ngưỡng hỗ trợ là 2, ta có tập các chuỗi con

```
giá/2 tăng/2
tăng/2
```

Kết quả trên có nghĩa là:

giá là chuỗi có tần suất xuất hiện là 2

giá tăng là chuỗi có tần suất xuất hiện là 2

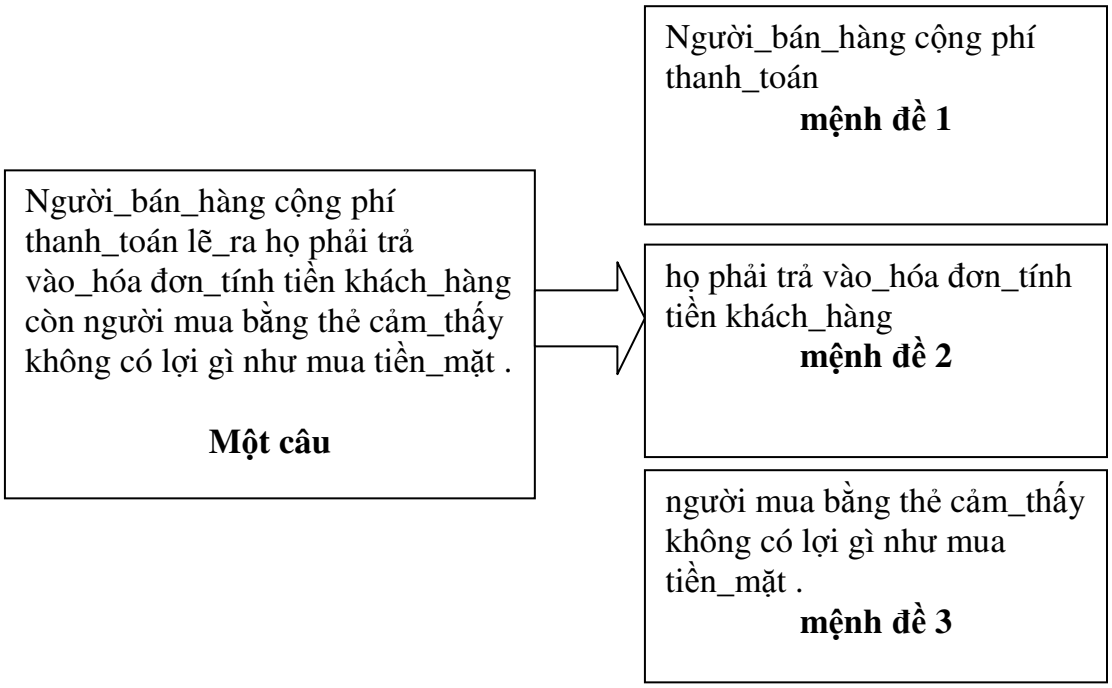
tăng là chuỗi có tần suất xuất hiện là 2.

Từ đây, ta thu được ba đặc trưng : *giá* , *giá tăng* và *tăng*. Đây là ba đặc trưng tốt cho phân lớp.

Mặt khác , theo thuật toán prefixspan [10], một chuỗi con được sinh ra bằng cách gắn một hoặc nhiều từ vào chuỗi hiện tại . Vì vậy, với câu càng dài, thì số lượng chuỗi càng lớn. Theo nhận định của Takamura và cộng sự [21], số lượng của các chuỗi con thường tăng theo cấp số nhân với độ dài của chuỗi được đánh giá . Khóa luận không sử dụng tất cả các câu mà coi các mệnh đề ngắn như các câu.Ví dụ trong hình 7, ta coi mỗi mệnh đề của câu như một chuỗi từ. Để xử lí những câu dài, ta chia nhỏ câu thành nhiều mệnh đề với các nút được gán nhãn SBAR (là các nút gốc của mệnh đề phụ) trong cấu trúc cây của câu . Thêm vào đó, nhằm rút gọn chuỗi, ta tiếp tục loại bỏ dấu câu và gán

nhãn từ , với các nhãn như bảng 5. Các từ được gán nhãn thường không phải là thành phần thể hiện quan điểm chủ quan.

Ví dụ : một câu có nhiều mệnh đề được chia nhỏ như hình 7



Hình 7.Các mệnh đề thu được chia ra từ một câu

Thẻ	Mô tả	Thẻ	Mô tả
CC	Từ nối	PRP$	Đại từ sở hữu
CD	Số đếm	RP	Tiền tố , hậu tố
DT	Từ hạn định	POS	Từ sở hữu
EX	Từ chỉ sự tồn tại	NNPS	Đại từ số nhiều
FW	Từ nước ngoài	RP	Tiền tố , hậu tố
IN	Giới từ, hoặc từ nối ngoài	SYM	Từ đại diện
LS	Danh sách mục đánh dâu	TO	Trong
NNP	Đại từ số ít	WDT	Từ hạn định bắt đầu bằng Wh
AUX	Trợ động từ	PRP	Đại từ chỉ người

Bảng 5.Danh sách các tag để loại bỏ các từ trong chuỗi của một câu

d. Tần suất các cây con(deq)

❖ Sinh cây phân tích cú pháp

Sử dụng bộ phân tích cú pháp coltechParser của Nguyễn Phương Thái và cộng sự [24] thu được nhãn của cây phân tích cú pháp tương ứng với từng câu thu được ở pha 1.

Ví dụ : Kết_nối là một chuyện , còn cơ_chế khuyến_khích cho người_dùng thẻ và người_bán_hàng_hóa là chuyện quan_trọng hơn.

=> ((S (NP[-H] (NP[-H] (N[-H] Kết_nối))) (VP (V[-H] là) (NP (NP[-H] (NP[-H] (M một) (N[-H] chuyện) (VP (, ,) (V[-H] còn) (NP (NP[-H] (N[-H] cơ_chế) (VP (V[-H] khuyến_khích) (PP (E[-H] cho) (NP (NP[-H] (N[-H] người_dùng) (N thẻ))))))))))) (C và) (NP (NP[-H] (N[-H] người_bán_hàng_hóa))) (C là) (NP (NP[-H] (N[-H] chuyện) (AP (A[-H] quan_trọng) (R hơn)))))) (. .)))

❖ Tính tần suất các cây con

Tần suất của cây con phụ thuộc là các mẫu có số nút lớn hơn hoặc bằng 2. Qua khảo sát dữ liệu, ta chọn ngưỡng hộ trợ : 10. Để tránh phân tích các mẫu nhiễu, ta loại bỏ các dấu chấm câu. Các mẫu này được trích xuất từ thuật toán Freqt nêu ở trên, trong trường hợp này được mô tả như sau:

- Đầu vào: tập nhãn L của các cây phân tích cú pháp D (thu được từ bước trên), và ngưỡng hộ trợ (minimum support)
- Đầu ra: tập các cây con phụ thuộc có tần suất lớn hơn hoặc bằng độ minimum support

3.4.3 Phân lớp sử dụng kỹ thuật học máy SVM

Sau khi trích chọn đặc trưng, thực hiện phân lớp nhị phân sử dụng mô hình học máy SVM.

❖ Sinh vecto đặc trưng:

Thông qua tính tần suất xuất hiện của bốn đặc trưng :unigram, bigram, subsequences, subtrees ở phần 3.4.2 , ta thu được các đặc trưng của tài liệu. Coi mỗi vecto đặc trưng là thể hiện của mỗi bình luận, trong đó mỗi chiều tương ứng với một đặc trưng. Như vậy:

- Số chiều của vecto V_i trong tập vecto $V = \{V_1, V_2, ... V_n\}$ của bình luận D_i trong tập bình luận $D = \{D_1, D_2, ... D_n\}$ là số đặc trưng xuất hiện trong D_i .Số đặc trưng của vecto là tổng tất cả các đặc trưng của unigram, bigram, substree, subsequences : $V_i = \{\{u_i\}, \{b_i\}, \{seq_i\}, \{deq_i\}\}$
    - Mỗi unigram là một đặc trưng của vecto nếu có tần suất xuất hiện trong tài liệu >= 4
    - Mỗi bigram là một đặc trưng của vecto nếu có tần suất xuất hiện trong tài liệu >= 4

- Mỗi substree là một đặc trưng của vecto nếu có tần suất xuất hiện trong tài liệu >= 10
- Mỗi subsequences là một đặc trưng của vecto nếu có tần suất xuất hiện trong tài liệu >= 10

- Giả sử tổng số lượng đặc trưng của tập các bình luận là m. Mỗi bình luận D_i được biểu diễn bằng một vecto $\overrightarrow{v_i} = (f_1, f_2, ... f_m)$

- f_j = 1 nếu đặc trưng thứ j xuất hiện trong D_i
- f_j = 0 nếu đặc trưng thứ j không xuất hiện trong D_i

❖ Học và phân lớp SVM, sử dụng một trong số các hàm nhân sau :

- Hàm tuyến tính
- Hàm đa thức
- Hàm vòng RBF
- Hàm chữ S Sigmoid

Tổng kết chương 3

Trong chương này, dựa trên những phân tích đặc trưng ngữ nghĩa của dữ liệu tiếng Việt, nhằm loại bỏ những dữ liệu nhiễu, những đánh giá không liên quan đến chủ đề bài báo, khóa luận đưa ra một phương pháp phân tích chủ đề và mô hình trích chọn đặc trưng chuỗi từ con và cây con phụ thuộc. Kết quả thực nghiệm ở chương sau cho thấy mô hình đề ra là hoàn toàn khả thi.

Chương 4. Thực nghiệm và đánh giá

4.1 Môi trường thực nghiệm

4.1.1 Cấu hình phần cứng

Thành phần	Chỉ số
CPU	2.2 GHz Core Duo Intel
RAM	1GB
OS	Windows7
Bộ nhớ ngoài	160GB

Bảng 6.Cấu hình hệ thống thử nghiệm

4.1.2 Công cụ phần mềm

STT	Tên phần mềm	Tác giả	Nguồn
1	Eclipse-SDK-3.5-win32		http://www.eclipse.org/downloads
2	Coltech parser	Nguyễn Phương Thái	http://vlsp.vietlp.org:8080/demo/?page=home
3	LibSVM	C. Chang, C.-J. Lin	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
4	PrefixSpan	Taku Kudo	http://www.chasen.org/~taku/software/
5	Freqt	Taku Kudo	http://www.chasen.org/~taku/software/freqt/
6	N-gram Extraction		http://homepages.inf.ed.ac.uk/lzhang10/ngram.html#intro
7	JvnTextPro	Nguyễn Cẩm Tú	http://jvntextpro.sourceforge.net/

Bảng 7.Công cụ phần mềm sử dụng

4.2 Dữ liệu thực nghiệm

Khóa luận thực nghiệm trên miền dữ liệu tin tức tài chính. Trong mục tài chính trên trang thời báo kinh tê Việt Nam, http://vneconomy.vn/p0c6/tai-chinh.htm, ta thu về nội dung bài báo và các bình luận đi kèm bài báo đó.

Trong 400 bài báo thu về, ta thu được 180 bài báo có bình luận đi kèm. Tổng số các bình luận của tất cả 180 bài báo này là 312. Trong 312 bình luận thì có 140 bình luận tích cực và 172 bình luận tiêu cực. Với mỗi bài báo, ta chia dữ liệu thành hai phần : một phần là nội dung bài báo, phần còn lại là tập các bình luận đi kèm bài báo đó.

4.3 Quá trình thực nghiệm

4.3.1. Phân tích chủ đề

Sử dung IDM Grabber, tiến hành clawer dữ liêu từ trang http://vneconomy.vn/p0c6/tai-chinh.htm

Thực hiện tách thẻ HTML.

Qua bước phân tích chủ đề, ta loại bỏ được một số các bình luận không đề cập đến vấn đề mà nội dung tin tức đưa ra, thông qua so sánh đối tượng mà bài viết nói đến và đối tượng độc giả bình luận .

Tuy nhiên, ta cũng thấy rằng, phương pháp phân tích chủ đề mà khóa luận đề xuất chưa thực sự hiệu quả. Khóa luận mới chỉ quan tâm đến đối tượng có được đề cập trong câu bình luận không mà chưa xét đến ngữ nghĩa của câu. Với những trường hợp, cả câu bình luận và nội dung bài viết đều đề cập đến đối tượng nhưng ở hai vấn đề khác nhau thì khóa luận chưa giải quyết được. Ví dụ : cùng đều đề cập đến USD

*1- Tỷ giá **USD** kịch trần biên độ*

2-** Ford thua lỗ hàng chục tỷ **USD

4.3.2 Trích chọn đặc trưng

4.3.2.1. Trích chọn tần suất của chuỗi con

Qua bước phân tích chủ đề, ta thu được tập các bình luận mới. Với mỗi bình luận, tiến hành chia nhỏ các mệnh đề trong một câu. Trong mỗi mệnh đề, tiếp tục loại bỏ các từ được gán nhãn như ở bảng 5.

Sử dụng công cụ PrefixSpan để sinh và xác định tần suất của các các chuỗi con.

- Đầu vào : là tập các mệnh đề, mỗi dòng là một mệnh đề. Các từ trên mỗi dòng được cách đều nhau bởi một dấu cách. File đầu vào của thuật toán PrefixSpan có cấu trúc như ví dụ dưới đây. Giả sử đoạn bình luận chỉ có ba mệnh đề (*)
  giá xăng tăng cao

giá cả tăng theo

giá cả làm đời sống người dân càng chật vật

- Đầu ra : là tập các chuỗi từ và tần suất tương ứng của chuỗi. File đầu ra có cấu trúc như sau:

  item/freq. item/freq. ...

  item/freq. item/freq. ...

Với ví dụ ở trên file in put ở trên có kết quả đầu ra như sau :

cả/2

giá/3 cả/2

giá/3 tăng/2

tăng/2

Kết quả trên có nghĩa

Mẫu tuần tự	**Tần suất**
Cả	2
Giá	3
giá ->cả	2
giá ->tăng	2
Tăng	2

Bảng 8.Ví dụ tần suất của các chuỗi con

PrefixSpan bổ sung thêm tham số ngưỡng hỗ trợ -m, để chỉ in ra các chuỗi có tần suất lớn hơn hoặc bằng m. Xét ví dụ(*), với m = 2, ta có kết quả như ở trên. Với m = 3, ta có kết quả như dưới đây: *giá/3*

4.3.2.2. Trích chọn tần suất của cây con

❖ Sinh nhãn cây phân cú pháp

Sử dụng bộ công cụ Coltech Parse để sinh ra nhãn các cây phân tích cú pháp của từng câu. File đầu vào là tập các câu, được bao trong cặp dấu ngoặc đơn có cấu trúc như ví dụ sau

(giá xăng tăng lên.)

➔ ((S (NP[-H] (NP[-H] (N[-H] giá) (N xăng))) (VP (V[-H] tăng) (R lên))))(**)

❖ Trích chọn các cây con:

Sử dụng bộ công cụ Freqt để sinh và tính xác suất của các cây con. Nhằm loại bỏ các dữ liệu nhiễu và giảm phức tạp khi trích chọn, ta loại bỏ tất cả các dấu câu, trước khi đưa vào Freqt

- *Đầu vào*: tập nhãn của các cây phân tích cú pháp và ngưỡng hỗ trợ , có cấu trúc như (**)
- *Đầu ra* : các cây con có tần suất xuất hiện lớn hơn hoặc bằng ngưỡng hỗ trợ có cấu trúc như ví dụ mẫu ở hình 10. Trong đó mỗi dòng tương ứng là là tần suất của các cây con.

```
2    (NP(NP(NP[-H])))
3    (NP(NP[-H]))
2    (NP(NP[-H])(Cvà))
2    (NP(NP[-H])(Cvà)(NP))
2    (NP(NP[-H])(Cvà)(NP(NP[-H])))
2    (NP(NP[-H])(NP))
2    NP(NP[-H])(NP(NP[-H])))
2    (NP(NP[-H](NP[-H])))
2    (NP(NP[-H](NP[-H]))(Cvà))
2    (NP(NP[-H](NP[-H]))(Cvà)(NP))
2    (NP(NP[-H](NP[-H]))(Cvà)(NP(NP[-H])))
2    (NP(NP[-H](NP[-H]))(NP))
2    (NP(NP[-H](NP[-H]))(NP(NP[-H])))
2    (NP(NP[-H](Pgì)))
2    (NP(NP[-H](VP)))
3    (NP[-H])
3    (NP[-H](NP[-H]))
```

Hình 8.Ví dụ kết quả đầu ra của thuật toán freqt

Trong bước này, ta trích chọn được tổng 919 đặc trưng cây con

4.3.2.3 Trích chọn bigram, unigram

Trích chọn các bigram, unigram với điều kiện tần suất xuất hiện lớn hơn hoặc bằng 4. Ta tiến hành loại bỏ các bigram chứa các loại dấu câu (., ?)

+ Tổng đặc trưng thu được của bigram là 500 đặc trưng.

+ Tổng đặc trưng thu được của unigram là 600 đặc trưng.

4.3.3 Phân lớp

- Gọi a là % dữ liệu huấn luyện

- Gọi b là % dữ liệu test

❖ Sinh vecto đặc trưng:

Thực hiện gán nhãn bằng tay toàn bộ các tập vecto thành hai lớp Pos(tích cực) và Neg(tiêu cực) .

Chia tập vecto thành a% làm dữ liệu huấn luyện, b% làm dữ liệu test.

❖ Học và phân lớp

Tiếp theo, ta phân lớp nhị phân sử dụng mô hình học máy SVM.

Lấy a% dữ liệu được gán nhãn cho vào bộ phân lớp SVM để sinh ra mô hình phân lớp. Tiến hành huấn luyện mô hình SVM với tập dữ liệu huấn luyện này.

Lấy b% dữ liệu đã có nhãn cho vào bộ phân lớp. Sau khi phân lớp , ta đánh giá độ chính xác của mô hình bằng cách so sánh kết quả của việc gán nhãn qua bộ phân lớp với nhãn đã được gán bằng tay.

4.4.Đánh giá

Sử dụng phần mềm LibSVM tiến hành phân lớp nhị phân với một trong 3 tham số hàm nhân t = 0 , t =1, t = 2. Áp dụng công thức (v) tính được độ chính xác của phân lớp :

$$Accuracy = \frac{correct}{total} * 100\% \quad \text{(v)}$$

Trong đó :

- correct là tổng số vecto mô hình gán nhãn đúng
- total là tổng số vecto của b% dữ liệu đưa vào test

❖ Với a = 60%, b=40% ta có kết quả phân lớp như bảng 9

Đặc trưng	**Độ chính xác**		
	t= 0	t= 1	t= 2
Unigram	57%	62%	62%
Bigram	52%	60.2%	57%
Unigram+bigram=bow	64.6%	67%	65.3%
Bow+substree	69%	71.6%	70%

Bảng 9.Bảng kết quả phân lớp lần 1

❖ Với a = 70%m b=30%, kết quả phân lớp như bảng 10

Đặc trưng	**Độ chính xác**		
	t= 0	t= 1	t= 2
Unigram	60%	63%	63%
Bigram	56%	62%	58%
Unigram+bigram=bow	66%	68%	65.7%
Bow+substree	70%	72%	70.6%

Bảng 10.Bảng kết quả phân lớp lần 2

Nhận xét:

So sánh cả hai bảng kết quả , ta thấy với tham số t= 1 , cho kết quả phân lớp tốt nhất.

Vì số lượng đặc trưng của mỗi bình luận nhiều nên cần dữ liệu học lớn, so sánh kết quả ở bảng 10 và bảng 9 cho ta thấy , với tỉ lệ tỉ lệ huấn luyện và test 70/30 cho kết quả tốt hơn 60/40.

Phân lớp đạt kết quả tốt nhất với đặc trưng bow kết hợp với đặc trưng substree.

Kết quả thực nghiệm một lần nữa khẳng định cơ sở lý thuyết và mô hình khóa luận đưa ra là hoàn toàn khả thi.

Kết luận

Xác định hướng quan điểm từ đánh giá của người dùng trên miền tin tức tài chính là bài toán có ý nghĩa và thực tiễn trong thời đại kinh tế ngày nay. Từ việc nghiên cứu bài toán phân lớp quan điểm , khóa luận đã đề xuất mô hình giải quyết bài toán nói trên . Qua những kết quả thực nghiệm đạt được cho thấy mô hình đưa ra là hoàn toàn khả thi và có thể áp dụng được .

Khóa luận đã đạt được những kết quả :

- Giới thiệu khai phá quan điểm và ứng dụng khai phá quan điểm trong thực tế
- Tìm hiểu và phân tích các phương pháp phân lớp quan điểm, trong đó tập trung vào phương pháp phân lớp sử dụng kỹ thuật học máy SVM.
- Áp dụng hai thuật toán tính tần suất mẫu : PrefixSpan và Freqt để trích chọn chuỗi con và cây con phụ thuộc làm đặc trưng cho máy hỗ trợ vecto
- Khóa luận đã thực nghiệm với phương pháp trích chọn đặc trưng cây con phụ thuộc kết hợp với hai đặc trưng unigram và bigram cho kết quả khá tốt.

Bên cạnh những kết quả đạt được , do hạn chế về mặt thời gian và kiến thức , khóa luận còn một số hạn chế sau:

- Khóa luận chưa thực nghiệm được với đặc trưng chuỗi con
- Dữ liệu thực nghiệm còn ít dẫn đến số lượng vecto đặc trưng chưa nhiều mà số chiều của vecto khá lớn, đây là một trong những nguyên nhân dẫn đến kết quả thực nghiệm chưa đạt được như mong muốn.

Về định hướng nghiên cứu trong tương lai , khóa luận sẽ phát triển theo các hướng sau:

- Bổ sung thực nghiệm kết hợp giữa bốn đặc trưng chuỗi con , cây con phụ thuộc, unigram và bigram.
- Cải thiện phương pháp phân tích chủ đề, để việc loại bỏ dữ liệu nhiễu tốt hơn.
- Thực nghiệm trên miền dữ liệu lớn hơn, đủ đề học và phân lớp với số lượng đặc trưng lớn hơn nữa

Tài liệu tham khảo

Tiếng Việt

[1]. Nguyễn Tiến Thanh, *Trích chọn quan hệ thực thể trên Wikipedia tiếng Việt dựa vào cây phân tích cú pháp,* Khóa luận tốt nghiệp Trường Đại học Công Nghệ, Đại học Quốc Gia Hà Nội , 2010, tr. 32-33.

Tiếng Anh

[2]. A. Devitt and K. Ahmad, "Sentiment polarity identication in financial news: A cohesion-based approach", *In Annual Meeting of the Association of Computational Linguistics*, volume 45,2007, pp 984–991,.

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

[4]. Bing Liu*, Web data mining; Exploring hyperlinks, contents, and usage data*, 2006, chapter 11: Opinion Mining. Springer.

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

[6]. Bo Pang, Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts" *Proc. of 42nd ACL*, 2004,pp. 271-278.

[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]. Corinna Cortes, Vladimir Vapnik (1995). Support-Vector Networks, *Machine Learning, 20(3): 273-297.*

[9]. Huifeng Tang, Songbo Tan *, Xueqi Cheng, " A survey on sentiment detection of reviews", *Journal Expert Systems with Applications: An International Journal archive* ,Volume 36 Issue 7, September, 2009, pp.10760- 10773.

[10]. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu, "Prefixspan: Mining Sequential Patterns Efficiently by Prefixspan-Projected Pattern Growth", *Proc. of 17th ICDE*, 2001, pp.215-224.

[11]. K. Ahmad, D. Cheng, and Y. Almas. "Multilingual sentiment analysis of financial news streams". *In Proceedings of the 1st International Conference on Grid in Finance, Palermo*, 2006, pp .1-8.

[12]. Kenji Abe, Shinji Kawasoe, Tatsuya Asai, and Hiroki Arimura, Setsuo Arikawa, "Optimized substructure discovery for semi-structured data", *Proc. of 6th PKDD*,2002, pp.1-14.

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

[14]. O'Hare, Neil and Davy, Michael and Bermingham, Adam and Ferguson, Paul and Sheridan, Páraic and Gurrin, Cathal and Smeaton, Alan F,"Topic-dependent sentiment analysis of financial blog", *In: TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, 6 November ,2009, pp.1-8.

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

[16]. Taboada, M., Caroline A., & Kimberly V, "Creating semantic orientation Dictionaries", *in Proceedings of 5th international conference on language resources and evaluation (LREC), Genoa, Italy*, 2006

[17]. Tony Mullen and Nigel Collier, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources", *Proc. of 9th EMNLP*, 2004, pp.412-418.

[18]. Taku Kudo and Yuji Matsumoto, " A Boosting Algorithm for Classification of Semi Structured Text", *Proc. of 9th EMNLP,* 2004, pp.301-308.

[19]. Tanasanee Phienthrakul, Boonserm Kijsirikul, Hiroya Takamura, Manabu Okumura, "Sentiment Classification with Support Vector Machines and Multiple Kernel Functions", 2009*, ICONIP* (2) pp. 583-592.

[20]. T.Asai, K.Abe, S.Kawasoe, H.Arimura, H.Sakamoto, and S.Arikawa, "Efficient substructure discovery from large semi-structured data ", *in Proc. The 2nd SIAM Int'l Conf on Data Mining (SDM2002)*, 2002, pp.158-170.

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

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

[23]. Vasileios Hatzivassiloglou and Kathleen McKeown, "Predicting the Semantic Orientation of Adjectives", *Proc. of 35th ACL and 8th EACL*,1997, pp. 174-181.

[24]. Đề tài KC01.01/06-10 **"Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP).** http://vlsp.vietlp.org:8080/demo/?page=parser

[25]. http://jvntextpro.sourceforge.net/