BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

ISO 9001:2000

ĐỒ ÁN TỐT NGHIỆP

NGHÀNH CÔNG NGHỆ THÔNG TIN

HẢI PHÒNG 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

PHÁT HIỆN TỪ QUAN ĐIỂM MỚI CHO PHÂN TÍCH CẢM XÚC

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Nghành : Công Nghệ Thông Tin

Hải Phòng 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

PHÁT HIỆN TỪ QUAN ĐIỂM MỚI CHO PHÂN TÍCH CẢM XÚC

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Nghành : Công Nghệ Thông Tin

Sinh viên thực hiện : Nguyễn Danh Long

Giáo viên hướng dẫn : Ths. Nguyễn Thị Xuân Hương

Mã số sinh viên : 1413101003

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o-------

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

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

Sinh viên : Nguyễn Danh Long Mã số : 1413101003

Lớp: CLT 801 Nghành: Công Nghệ Thông Tin

Tên đề tài : Phát hiện từ quan điểm mới cho phân tích cảm xúc

NHIỆM VỤ ĐỀ TÀI

1. Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp

a. Nội dung

b. Các yêu cầu cần giải quyết

2. Các số liệu cần thiết để thiết kế , tình toán

3. Địa điểm thực tập

CÁN BỘ HƯỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

Người hướng dẫn thứ nhất:

Họ và tên:.....

Học hàm, học vị:.....

Cơ quan công tác:.....

Nội dung hướng dẫn: ……

……

……

……

……

Người hướng dẫn thứ hai:

Họ và tên: ……

Học hàm, học vị……

Cơ quan công tác: ……

Nội dung hướng dẫn: …….....

……

……

……

……

Đề tài tốt nghiệp được giao ngày 18 tháng 04năm 2016

Yêu cầu phải hoàn thành trước ngày 9 tháng 07 năm 2016

Đã nhận nhiệm vụ: Đ.T.T.N	Đã nhận nhiệm vụ: Đ.T.T.N
Sinh viên	Cán bộ hướng dẫn Đ.T.T.N

Hải Phòng, ngày .....tháng.....năm 2016

HIỆU TRƯỞNG

GS.TS.NGƯT Trần Hữu Nghị

PHẦN NHẬN XÉT TÓM TẮT CỦA CÁN BỘ HƯỚNG DẪN

1. Tinh thần thái độ của sinh viên trong quá trình làm đề tài tốt nghiệp:

.....

.....

.....

.....

.....

.....

.....

.....

2. Đánh giá chất lượng của đề tài tốt nghiệp (so với nội dung yêu cầu đã đề ra trong nhiệm vụ đề tài tốt nghiệp)

.....

.....

.....

.....

.....

.....

.....

.....

3. Cho điểm của cán bộ hướng dẫn:

(Điểm ghi bằng số và chữ)

.....

.....

Ngày.....tháng.....năm 2016

Cán bộ hướng dẫn chính

(Ký, ghi rõ họ tên)

PHẦN NHẬN XÉT ĐÁNH GIÁ CỦA CÁN BỘ CHẤM PHẢN BIỆN ĐỀ TÀI TỐT NGHIỆP

1. **Đánh giá chất lượng đề tài tốt nghiệp (về các mặt như cơ sở lý luận, thuyết minh chương trình, giá trị thực tế, ...)**

2. Cho điểm của cán bộ phản biện

(*Điểm ghi bằng số và chữ*)

.....

.....

Ngày.....tháng.....năm 2016

Cán bộ chấm phản biện

(Ký, ghi rõ họ tên)

MỤC LỤC

DANH MỤC BẢNG

LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành và biết ơn sâu sắc nhất tới Cô Nguyễn Thị Xuân Hương, Trường Đại học Dân lập Hải Phòng đã chỉ bảo và hướng dẫn tận tình cho em trong suốt quá trình tìm hiểu và thực hiện khóa luận này.

Em xin chân thành cảm ơn các Thầy, Cô trong Khoa Công nghệ Thông tin đã tận tình giảng dạy và truyền cho em những kiến thức quý báu cho em trong suốt quá trình học tập và làm luận văn tốt nghiệp

Em xin chân thành cảm ơn tới các Thầy, Cô và các Cán bộ, Nhân viên của trường Đại học Dân Lập Hải Phòng đã tạo cho em những điều kiện thuận lợi để học tập và nghiên cứu.

Cuối cùng em muốn gửi lời cảm ơn tới gia đình và bạn bè những người thân yêu đã luôn bên cạnh động viên trong suốt quá trình học tập và làm khóa luận tốt nghiệp.

Mặc dù em đã rất cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự cảm thông và tận tình chỉ bảo, góp ý của quý Thầy Cô và các bạn.

Em xin chân thành cảm ơn!

Hải Phòng , ngày….. tháng….. năm…….

Sinh viên

Nguyễn Danh Long

LỜI NÓI ĐẦU

Ngày nay, với sự phát triển mạnh mẽ của Internet, các hình thức kết nối và chia sẻ thông tin trong cộng đồng mạng ngày càng phát triển đã thu hút một lượng lớn người dùng tham gia. Qua đó, họ có thể dễ dàng trao đổi, chia sẻ thông tin, thảo luận các vấn đề và sở thích cùng quan tâm. Một số mạng xã hội phổ biến trên thế giới như: facebook, twitter và ở Việt Nam như Zing có số lượng người tham gia ngày càng đông đảo. Các bài nhận xét thảo luân trên các trang mạng tin tức, dịch vụ hay các diễn đàn cũng là một hình thức thể hiện khác rất phát triển.

Các thông tin được chia sẻ và thảo luận trên các trang mạng xã hội, trên các diễn đàn thuộc rất nhiều chủ đề từ các lĩnh vực kinh tế, chính trị, xã hội … đến các vấn đề kỹ thuật, dịch vụ, cuộc sống hàng ngày.... Từ đó hình thành nên các xu hướng, quan điểm của cộng đồng đối với việc đánh giá nhận xét một vấn đề, một đối tượng, một sản phẩm hay một hiện tượng nào đó. Các quan điểm, xu hướng này sẽ có tác động mạnh mẽ đến định hướng, quan điểm của người dùng khác.

Người dùng, hay chính các nhà cung cấp sản phẩm, dịch vụ cũng có xu hướng khai thác ý kiến đánh giá của người khác để sử dụng vào nhiều mục đích khác nhau. Người dùng cần biết nhận xét về sản phẩm, dịch vụ cho lựa chọn của mình, còn các nhà sản xuất, cung cấp dịch vụ thì thu thập thông tin để phân tích nhu cầu, thị hiếu của khách hàng, phân tích các đối thủ cạnh tranh để từ đó có chiến lược nâng cao chất lượng sản phẩm và dịch vụ của mình… Và với lượng dữ liệu khổng lồ được tạo ra mỗi ngày thì nhu cầu cần có các hệ thống khai thác và phân tích quan điểm một cách tự động là rất cần thiết.

Để phân tích quan điểm ta cần phải trích các thông tin chứa quan điểm để phân loại có quan điểm hay không. Các thông tin chưa quan điểm có thể là từ hoặc một cụm từ có chứa quan điểm cảm xúc tích cực, tiêu cực, trung lập. Trong khi người dùng cập nhật và chia sẻ thông tin của họ trên các trang web

xã hội họ thường sử dụng lối hành văn tự do theo phong cách ngôn ngữ riêng của họ. Do đó xuất hiện các từ mới thể hiện quan điểm trong các lĩnh vực. Việc xuất hiện ngày càng nhiều các từ quan điểm mới nên việc nghiên cứu các phương pháp trích rút tự động các từ quan điểm mới là rất cần thiết.

Chính vì lý do trên, em đã chọn đề tài "Phát hiện từ quan điểm mới cho phân tích cảm xúc" cho đồ án tốt nghiệp của mình.

Nội dung đồ án bao gồm 3 chương :

Chương 1 : Tổng quan về phân tích quan điểm

Chương 2 : Phát hiện từ mới cho phân tích quan điểm

Chương 3 : Ứng dụng tìm từ quan điểm mới cho dữ liệu Tiếng Việt

Cuối cùng là phần kết luận.

CHƯƠNG 1 : TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM

1.1 Nhu cầu về thông tin quan điểm và nhận xét

Những thông tin nhận xét góp ý đã luôn luôn là một phần quan trọng trong việc cung cấp thông tin cho quá trình ra quyết định của hầu hết chúng ta. Trước khi Internet trở lên phổ biến, chúng ta thường yêu cầu bạn bè hay người thân giới thiệu một thợ cơ khí tự động hoặc yêu cầu tài liệu tham khảo liên quan đến xin việc từ các đồng nghiệp, hoặc tư vấn tiêu dùng. Ngày nay, Internet và Web đã giúp cho chúng ta có thể dễ dàng tiếp cận các ý kiến và kinh nghiệm của những người khác mà không nhất thiết phải là những người quen biết cá nhân, không phải là các nhà phê bình chuyên nghiệp nổi tiếng, những người mà chúng ta chưa bao giờ nghe nói tới trong không gian rộng lớn. Và ngược lại, ngày càng nhiều và nhiều hơn nữa những người sẵn sàng cung cấp các ý kiến của mình cho những người khác qua Internet.

Theo hai cuộc khảo sát của hơn 2000 người Mỹ trưởng thành mỗi: 81% người dùng Internet (hoặc 60% người Mỹ) đã thực hiện nghiên cứu trực tuyến về một sản phẩm ít nhất một lần; 20% (15% của tất cả các người Mỹ) làm như vậy trong một ngày. Trong số các độc giả đánh giá trực tuyến của nhà hàng, khách sạn, và các dịch vụ khác nhau (ví dụ như, các cơ quan du lịch hoặc bác sĩ), giữa 73% và 87% báo cáo đánh giá đã có một ảnh hưởng đáng kể mua hàng của họ. Người tiêu dùng sẵn sàng trả từ 20% đến 99% một mục được đánh giá 5 sao cao hơn so với một mục đánh giá 4 sao, 32% đã cung cấp một đánh giá về một sản phẩm, dịch vụ thông qua một hệ thống xếp hạng trực tuyến, trong đó có 18% của công dân trực tuyến cao cấp, có đăng một bình luận trực tuyến hoặc xem xét về một sản phẩm hay dịch vụ.

Thống kê nhanh chỉ ra rằng việc tiêu thụ hàng hóa và dịch vụ không phải là động cơ duy nhất khi người dùng tìm kiếm hoặc thể hiện ý kiến trực tuyến. Sự cần thiết của những thông tin chính trị cũng là một yếu tố quan trọng. Ví dụ, trong một cuộc khảo sát hơn 2500 người Mỹ trưởng thành, Rainie và Horrigan nghiên cứu có 31% người Mỹ - trên 60 triệu người - 2006 người dùng Internet vận động tranh cử, là những người thu thập thông tin về

cuộc bầu cử năm 2006 trực tuyến và trao đổi nhận xét thông qua email. Trong số này:

• 28% nói rằng nguyên nhân chính cho các hoạt động trực tuyến này để thu nhận được quan điểm từ bên trong cộng đồng của họ, và 34% cho biết một lý do chính là để nhận được quan điểm từ bên ngoài cộng đồng của họ.

• 27% đã xem đánh giá trực tuyến cho sự tán thành hoặc xếp hạng của các tổ chức bên ngoài.

• 28% cho biết rằng hầu hết các trang web mà họ sử dụng để chia sẻ quan điểm, nhưng 29% nói rằng phần lớn các trang web mà họ sử dụng thách thức quan điểm của họ, chỉ ra rằng nhiều người không chỉ đơn giản là tìm kiếm để xác nhận các quan điểm có trước của họ.

• 8% đăng bình luận trực tuyến bình luận chính trị riêng của họ.

Đối với người dùng tìm kiếm sự tin cậy trong những lời khuyên và tư vấn trực tuyến quan tâm đến việc xây dựng một hệ thống mới để xử lý trực tiếp các quan điểm trước tiên là phân loại chúng. Theo Horrigan thống kê rằng trong khi đa số người sử dụng internet của Mỹ cho rằng kinh nghiệm tích cực trong nghiên cứu sản phẩm trực tuyến, 58% cho rằng thông tin trực tuyến là thiếu, khó tìm, khó hiểu và hoặc quá nhiều. Vì vậy, nhu cầu có một hệ thống để hỗ trợ người tiêu dùng tìm kiếm thông tin là rất cần thiết.

Các nhà cung cấp sản phẩm ngày càng chú ý hơn đến sự quan tâm mà người dùng cá nhân thể hiện trong các nhận xét trực tuyến về sản phẩm và dịch vụ, và sự ảnh hưởng như xu thế sử dụng.

Với sự bùng nổ của nền tảng Web 2.0 như các blog, diễn đàn thảo luận, peer-to-peer mạng, và các loại khác nhau của các mạng xã hội...

• Thống kê của Facebook: có hơn 500 triệu người dùng ở trạng thái hoạt động (active) mỗi người có trung bình 130 bạn (friends), trao đổi qua lại trên 900 triệu đối tượng.

• Twitter (5/2011): có hơn 200 triệu người dùng. Một ngày có hơn 300 nghìn tài khoản mới, trung bình hơn 190 triệu tin nhắn, xử lý trung bình khoảng 1,6 tỷ câu hỏi.

• Ở Việt Nam: các mạng xã hội zing.vn, go.vn … thu hút được đông đảo người dùng tham gia.

Một lượng đông đảo người dùng gia tăng chưa từng có và có quyền chia sẻ kinh nghiệm và nhận xét của riêng họ về bất kỳ sản phẩm hoặc dịch vụ, là tích cực hay tiêu cực. Khi các công ty lớn đang ngày càng nhận ra, những tiếng nói của người tiêu dùng có thể vận dụng rất lớn ảnh hưởng trong việc hình thành nhận xét của người tiêu dùng khác, cuối cùng để trung thành với thương hiệu của họ, họ quyết định mua, và vận động cho chính thương hiệu của họ... Công ty có thể đáp ứng với những hiểu biết của người tiêu dùng mà họ tạo ra thông qua điều khiển phương tiện truyền thông xã hội và phân tích các thông điệp maketing của họ, định vị thương hiệu, phát triển sản phẩm và các hoạt động phù hợp khác.

Tuy nhiên, các nhà phân tích ngành công nghiệp lưu ý rằng việc tận dụng các phương tiện truyền thông mới cho mục đích theo dõi hình ảnh sản phẩm đòi hỏi cần phải có công nghệ mới.

Các nhà tiếp thị luôn luôn cần giám sát các phương tiện truyền thông cho thông tin liên quan đến thương hiệu của mình - cho dù đó là đối với các hoạt động quan hệ công chúng, vi phạm gian lận, hoặc tình báo cạnh tranh. Nhưng phân mảnh các phương tiện truyền thông và thay đổi hành vi của người tiêu dùng đã loại trừ các phương pháp giám sát truyền thống. Technorati ước tính rằng 75. 000 blog mới được tạo ra mỗi ngày, cùng với 1, 2 triệu bài viết mỗi ngày, trong đó có nhiều nhận xét người tiêu dùng thảo luận về sản phẩm và dịch vụ.

Vì vậy, không chỉ có cá nhân, mà các công ty, các tổ chức đều quan tâm đến một hệ thống có khả năng tự động phân tích quan điểm của người tiêu dùng.

1.2 Lịch sử của phân tích quan điểm và khai thác quan điểm

Lĩnh vực phân tích quan điểm (sentiment analysis) hay khai thác quan điểm (opinion mining) gần đây đã thu hút được sự quan tâm rộng rãi của các nhà nghiên cứu. Năm 2001 bắt đầu đánh dấu sự lan rộng nhận thức về các vấn đề nghiên cứu và cơ hội nâng cao phân tích tình cảm và khai thác quan điểm.

Các nhân tố được nghiên cứu gồm:

• Sự gia tăng của các phương pháp học máy, xử lý ngôn ngữ tự nhiên và khôi phục thông tin.

• Sự sẵn có của các tập dữ liệu đào tạo cho các thuật toán học máy, sự phát triển của Internet, cụ thể là sự phát triển của tập hợp các trang Web thu thập các ý kiến và quan điểm.

• Thực hiện những thách thức trí tuệ, thương mại và các ứng dụng thông minh trong lĩnh vực này.

Thuật ngữ khai thác quan điểm (Dave et al. 2003) là các công cụ khai thác quan điểm sẽ xử lý một tập hợp các kết quả tìm kiếm cho một đối tượng nhất định, sinh ra một danh sách các thuộc tính sản phẩm (chất lượng, đặc trưng, vv) và các quan điểm tổng hợp về chúng (kém, bình thường, tốt).

“Phân tích quan điểm” là cụm từ song song của "khai thác quan điểm" ở những khía cạnh nhất định (Das và Chen Tong, 2001). “Phân tích quan điểm" và "khai thác quan điểm" biểu thị cùng một lĩnh vực nghiên cứu.

1.3 Một số nghiên cứu trong phân tích quan điểm

Gần đây, khai thác quan điểm đã trở thành chủ đề nóng giữa các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin. Có khá nhiều các bài báo được xuất bản và những ứng dụng khác nhau có sử dụng hệ thống đánh giá quan điểm được phát triển và đưa vào trong hoạt động thương mại. Các tiếp cận chủ yếu với bài toán này là:

- ✓ *Phân lớp quan điểm thông qua việc xác định từ, cụm từ chỉ quan điểm*
- ✓ *Xác định quan điểm với các thể hiện trong từng thuộc tính của đối tượng cần tìm kiếm quan điểm.*

1.3.1 Xác định cụm từ, quan điểm

Những từ, cụm từ chỉ quan điểm là những từ ngữ được sử dụng để diễn tả cảm xúc, ý kiến người viết, những quan điểm chủ quan đó dựa trên những vấn đề mà anh ta hay cô ta đang tranh luận. Việc rút ra những từ, cụm từ chỉ quan điểm là giai đoạn đầu tiên trong hệ thống đánh giá quan điểm, vì những

từ, cụm từ này là những chìa khóa cho công việc nhận biết và phân loại tài liệu sau đó.

Ứng dụng dựa trên hệ thống đánh giá quan điểm hiện nay tập trung vào các từ chỉ nội dung câu: danh từ, động từ, tính từ và phó từ. Phần lớn công việc sử dụng từ loại để rút chúng ra (Hu và Liu, 2004, Turney, 2002). Việc gán nhãn từ loại cũng được sử dụng trong công việc này, điều này có thể giúp cho việc nhận biết xu hướng quan điểm trong giai đoạn tiếp theo. Những kĩ thuật phân tích ngôn ngữ tự nhiên khác như xóa: *stopwords*, *stemming* cũng được sử dụng trong giai đoạn tiền xử lý để rút ra từ, cụm từ chỉ quan điểm

1.3.2 Sử dụng tính từ và phó từ

Những hệ thống hiện tại dùng để nhận biết những từ chỉ quan điểm hay xu hướng quan điểm tập trung chủ yếu vào các tính từ và phó từ vì chúng được xem là sự biểu lộ rõ ràng nhất của tính chủ quan (Hatzivassiloglou and McKeown, 1997, Wiebe and Bruce, 1999).

Hu và Liu (2004) áp dụng việc gán nhãn từ loại và kĩ thuật xử lý ngôn ngữ tự nhiên nhằm rút ra những tính từ cũng như những từ chỉ quan điểm. Phương pháp của họ dựa vào việc phân loại dựa trên dấu hiệu quan điểm về sản phẩm:

- Định nghĩa một câu mà chứa một hay nhiều dấu hiệu sản phẩm và từ chỉ quan điểm được xem là một câu chỉ quan điểm.
- Với mỗi câu trong dữ liệu chỉ quan điểm, rút ra tất cả những tính từ được coi là những từ chỉ quan điểm.
- Kết quả thực nghiệm việc rút ra những câu đánh giá quan điểm có độ chính xác (*precision*) khoảng 64.2% và *recall* là 69.3%.
- Sử dụng WordNet (Fellbaum, 1998) để xác định các tính từ được rút ra mang chiều hướng tích cực (*positive*) hay tiêu cực (*negative*).

Trong WordNet, các tính từ được tổ chức thành các cụm từ lưỡng cực, nửa cụm thứ hai phần đầu là từ trái nghĩa của cụm thứ nhất. Mỗi nửa cụm là phần đầu của tập từ đồng nghĩa chính, tiếp theo là tập từ đồng nghĩa kèm theo, đại diện cho ngữ nghĩa tương tự như những tính từ quan trọng. Ngược với cách tiếp cận dựa trên từ điển, họ sử dụng định hướng quan điểm của những từ đồng nghĩa và từ trái nghĩa để dự đoán định hướng của các tính từ. Họ bắt

đầu với một danh sách khởi đầu gồm 30 tính từ thông dụng được chọn thủ công (bằng tay). Sau đó sử dụng WordNet để dự đoán định hướng của tất cả các tính từ trong danh sách từ quan điểm được rút ra bằng cách tìm kiếm qua cụm lưỡng cực để tìm ra liệu các từ đồng nghĩa hay trái nghĩa có trong danh sách khởi đầu hay không. Khi định hướng của tính từ được dự đoán, nó sẽ được bổ sung vào danh sách khởi đầu và có thể được sử dụng để xác định định hướng của các tính từ khác. Trong phương pháp này, danh sách khởi đầu sẽ dần tăng lên khi sự định hướng của các tính từ được nhận dạng, và khi nó ngừng gia tăng, tức qui mô của danh sách khởi đầu trùng với qui mô của danh sách từ chỉ quan điểm, thì tất cả định hướng của các tính từ đã được nhận biết và quá trình này kết thúc.

Những từ quan điểm thường tập trung chủ yếu vào hai từ loại: tính từ và phó từ vì vậy càng nhận dạng chính xác được nhiều hai loại từ này hệ thống càng có độ chính xác cao

1.3.3 Sử dụng các động từ

Các tính từ và phó từ đóng một vai trò quan trọng trong việc phân tích quan điểm và là các loại từ có lợi thế trong việc nhận biết định hướng và rút ra các từ chỉ quan điểm trong các nghiên cứu hiện nay. Tuy nhiên, các loại từ khác, ví dụ như động từ cũng được sử dụng để diễn tả cảm xúc hay ý kiến trong các bài viết.

Nasukawa và Yi (2003) xem xét rằng bên cạnh các tính từ và phó từ, thì các động từ cũng có thể diễn tả quan điểm trong hệ thống đánh giá quan điểm của họ. Họ phân loại các động từ có liên quan đến quan điểm thành 2 loại. Loại thứ nhất trực tiếp thể hiện quan điểm tích cực hay tiêu cực, theo lý giải của họ thì “beat” trong “X beats Y”. Loại thứ hai không thể hiện quan điểm trực tiếp nhưng dẫn đến những quan điểm, giống như “is” trong “X is good”.

Họ sử dụng gán nhãn từ loại dựa trên mô hình Markov (HMM) (Manning and Schutze, 1999) và phân tích cú pháp nông dựa trên luật (Neff et al., 2003) cho bước tiền xử lý. Sau đó họ phân tích tính phụ thuộc về mặt cú pháp giữa các cụm từ và tìm kiếm các cụm từ có một từ chỉ quan điểm mà nó bổ nghĩa hoặc được bổ nghĩa bởi một thuật ngữ chủ thể.

1.3.4 Xác định chiều hướng, cụm từ, quan điểm

Trong phân tích quan điểm, xu hướng của những từ, cụm từ trực tiếp thể hiện quan điểm, cảm xúc của người viết bài. Phương pháp chính để nhận biết xu hướng quan điểm của những từ, cụm từ chỉ cảm nghĩ là dựa trên thống kê hoặc dựa trên từ vựng.

1.4. Nhiệm vụ của phân tích quan điểm

Phân tích quan điểm là những nghiên cứu nhằm phát hiện ra quan điểm hay xu hướng của người dùng dựa trên các kỹ thuật liên quan đến vấn đề xử lý ngôn ngữ tự nhiên. Có hai hướng tiếp cận chính cho bài toán này là : Phân lớp quan điểm (*Sentiment Classification*) và trích quan điểm (*Sentiment Extraction*).

Trích quan điểm: bao gồm 3 nhiệm vụ chính là:

- *Trích các đặc trưng đối tượng có nhận xét trong mỗi quan điểm.*
- *Xác định có hay không các quan điểm trong các đặc trưng là positive, negative hay neutral (phụ thuộc vào định dạng của các quan điểm)*
- *Nhóm các cụm từ cùng nghĩa đặc trưng*

1.5. Bài toán phân lớp quan điểm

Phân lớp là quá trình "nhóm" các đối tượng "giống" nhau vào "một lớp" dựa trên các đặc trưng dữ liệu của chúng. Tuy nhiên, phân lớp là một hoạt động tiềm ẩn trong tư duy con người khi nhận dạng thế giới thực, đóng vai trò quan trọng làm cơ sở đưa ra các dự báo, các quyết định. Phân lớp và cách mô tả các lớp giúp cho tri thức được định dạng và lưu trữ trong đó.

Khi nghiên cứu một đối tượng, hiện tượng, chúng ta chỉ có thể dựa vào một số hữu hạn các đặc trưng của chúng. Nói cách khác, ta chỉ xem xét biểu diễn của đối tượng, hiện tượng trong một không gian hữu hạn chiều, mỗi chiều ứng với một đặc trưng được lựa chọn. Khi đó, phân lớp dữ liệu trở thành phân hoạch tập dữ liệu thành các tập con theo một tiêu chuẩn nhận dạng được.

Nhiệm vụ phân lớp quan điểm được xem xét với hai tiếp cận chính là:

- Phân lớp câu chứa quan điểm

- Phân lớp tài liệu chứa quan điểm.

Phân lớp câu/tài liệu chứa quan điểm có thể được phát biểu như sau: Cho một câu hay một tài liệu chứa quan điểm, hãy phân loại xem câu hay tài liệu đó thể hiện quan điểm mang xu hướng tích cực(*positive*) hay tiêu cực (*negative*), hoặc trung lập (*neutral*).

Theo Bo Pang và Lillian Lee (2002) phân lớp câu/tài liệu chỉ quan điểm không có sự nhận biết của mỗi từ/ cụm từ chỉ quan điểm. Họ sử dụng học máy có giám sát để phân loại những nhận xét về phim ảnh. Không cần phải phân lớp các từ hay cụm từ chỉ quan điểm, họ rút ra những đặc điểm khác nhau của các quan điểm và sử dụng thuật toán Naïve Bayes (NB), Maximum Entropy (ME) và Support Vector Machine (SVM) để phân lớp quan điểm. Phương pháp này đạt độ chính xác từ 78, 7% đến 82, 9%.

Input: Cho một tập các văn bản chứa các ý kiến đánh giá về một đối tượng nào đó.

Output: Mỗi văn bản được chia vào một lớp theo mức độ phân cực (*polarity*) về tiếp cận ngữ nghĩa nào đó (tích cực, tiêu cực hay trung lập).

Phân lớp tài liệu theo hướng quan điểm thật sự là vấn đề thách thức và khó khăn trong lĩnh vự xử lý ngôn ngữ đó chính là bản chất phức tạp của ngôn ngữ của con người, đặc biệt là sự đa nghĩa và nhập nhằng nghĩa của ngôn ngữ. Sự nhập nhằng này rõ ràng sẽ ảnh hưởng đến độ chính xác bộ phân lớp của chúng ta một mức độ nhất định. Một khía cạnh thách thức của vấn đề này dường như là phân biệt nó với việc phân loại chủ đề theo truyền thống đó là trong khi những chủ đề này được nhận dạng bởi những từ khóa đứng một mình, quan điểm có thể diễn tả một cách tinh tế hơn. Ví dụ câu sau: "*Làm thế nào để ai đó có thể ngồi xem hết bộ phim này ?*"không chứa ý có nghĩa duy nhất mà rõ ràng là nghĩa tiêu cực. Theo đó, quan điểm dường như đòi hỏi sự hiểu biết nhiều hơn, tinh tế hơn.

1.5.1 Phân cực quan điểm và mức độ phân cực

- Mức độ phân cực: *positive*/*negative*/*neutral*
- Nhận xét về sản phẩm, dịch vụ: Like/ dislike/ So so
- Nhận xét về phim ảnh thumbs up/ thumbs down

- Nhận xét về quan điểm chính trị:like to win/ unlike to win Liberal/conservative
- Phân loại bài báo là good new/ bad new.

Các bài toán liên quan đến phân lớp phân cực quan điểm:

- Xác định sự phân cực của văn bản (tài liệu/câu) chứa quan điểm: tích cực, tiêu cực hay trung tính.

  VD: Thông qua nhận xét: "This laptop is great".

- Xác định một đoạn thông tin "khách quan" là tốt hoặc xấu =>thách thức liên quan đến phân tích quan điểm.

  VD: "The stock prise rose"

- Phân biệt giữa câu "chủ quan"và "khách quan"

Rating inference (*ordinal regression*): Sắp xếp các quan điểm theo nhiều mức:

- Sắp xếp các đánh giá từ theo nhiều mức: VD: 1 sao đến 5 sao. Hay theo mức độ phân cực: rất thích, thích, bình thường, không thích,…
- Khi phân loại vào 3 lớp: *positive, negative, neutral*: neutral được coi là giá trị trung bình giữa *positive* và *negative*.
- Nhãn "*neutral*": một số được sử dụng như là lớp khách quan(thiếu quan điểm).
- Theo Cabral và Hortacsu, 2006: nhãn *neutral* có thể gần *negative* hơn vì con người có xu hướng phản ứng mạnh với nhận xét *negative*: 40% so với nhận xét neutral là 10%.

1.5.2 Nhiệm vụ của bài toán phân lớp quan điểm

Bài toán phân lớp quan điểm được biết đến như là bài toán phân lớp tài liệu với mục tiêu là phân loại các tài liệu theo định hướng quan điểm.

Đã có rất nhiều tiếp cận khác nhau được nghiên cứu để giải quyết cho loại bài toán này. Để thực hiện, về cơ bản có thể chia thành hai nhiệm vụ chính như sau:

- *Trích các đặc trưng nhằm khai thác các thông tin chỉ quan điểm để phục vụ mục đích phân loại tài liệu theo định hướng ngữ nghĩa.*
- *Xây dựng mô hình để phân lớp các tài liệu*.

CHƯƠNG 2 : PHÁT HIỆN TỪ MỚI CHO PHÂN TÍCH QUAN ĐIỂM

Tự động trích các từ mới là sự cần thiết trước tiên cho một số nhiệm vụ xử lý ngôn ngữ tự nhiên như tách từ (ngôn ngữ Tiếng Trung, Tiếng Việt), trích tên của thực thể và phân tích tình cảm, phân tích quan điểm.

Minlie Hoang et al., 2014 đã giới thiệu một phương pháp để trích các từ quan điểm mới từ dữ liệu lớn do người dùng tạo ra. Trong đó, các tác giả đã giới thiệu một phương pháp học hoàn toàn không giám sát và một Framwork hoàn toàn dựa trên dữ liệu cho trích từ quan điểm mới và ứng dụng nó trong bài toán phân tích quan điểm. Các tác giả đã thiết lập các độ đo thống kê để xác định tính phù hợp của một mẫu từ vựng và đo khả năng một từ là từ mới.

Phương pháp này chỉ sử dụng rất ít các thông tin ngôn ngữ (gần với các nguồn tài nguyên ngôn ngữ tự do) ở đây chỉ sử dụng thông tin gán nhãn từ loại POS, và không yêu cầu phải xây dựng quy tắc ngôn ngữ. Các tác giả đã chứng minh từ cảm xúc (quan điểm) mới sẽ mang lại lợi ích trong phân tích cảm xúc hay phân tích quan điểm. Các kết quả thực nghiệm chứng minh tính hiệu quả của phương pháp đề xuất.

2.1. Giới thiệu phương pháp

Các từ mới trên Internet xuất hiện ngày càng nhiều, đặc biệt là trong nội dung do người dùng tạo ra. Người dùng muốn cập nhật và chia sẻ thông tin của họ trên các trang web xã hội với phong cách ngôn ngữ riêng của họ, trong đó xuất hiện các từ mới trong các lĩnh vực như chính trị, kinh tế, văn hóa, xã hội.

Tuy nhiên, những từ mới như vậy đã tạo ra những thách thức cho một số nhiệm vụ xử lý trong ngôn ngữ tự nhiên. Việc tự động trích các từ mới là không thể thiếu đối với nhiều công việc như phân đoạn từ (Tiếng Trung hay Tiếng Việt..), dịch máy, trích xuất thực thể có tên, hỏi và trả lời, và phân tích cảm xúc, phân tích quan điểm. Phát hiện từ mới là một trong những vấn đề quan trọng nhất trong tách từ Tiếng Trung. Các nghiên cứu gần đây (Sproat và Emerson, 2003) (Chen, 2003) đã chỉ ra rằng hơn 60% tách từ từ kết quả của từ

mới. Thống kê cho thấy đã có hơn 1.000 từ mới tiếng Trung xuất hiện mỗi năm (Trung tâm Nghiên cứu Thesaurus, 2003). Các từ này là những thuật ngữ kỹ thuật chủ yếu và là các từ nhạy cảm với thời gian trong văn hóa, xã hội, chính trị. Phần lớn các từ này chưa được nhận dạng chính xác bởi các thuật toán tách từ, và nó nằm ngoài các miền từ vựng.

Phát hiện từ mới cũng rất quan trọng để phân tích cảm xúc như cụm trích các cụm từ chứa quan điểm và phân loại mức độ quan điểm (tích cực, tiêu cực hay trung lập). Một cụm từ cảm xúc với đầy đủ ý nghĩa có một ranh giới chính xác, tuy nhiên, các ký tự trong một từ mới có thể được chia nhỏ.

Ví dụ, trong một câu " 表 演/ n 非 常/ adv 给/ v 力/ n（artists' performance is very impressive – Biểu diễn của các nghệ sĩ rất ấn tượng)" hai ký tự “给/v 力/n (cool; powerful – tuyệt vời, mạnh mẽ) nên được trích cùng nhau. Trong phân loại mức độ cảm xúc, các từ mới có thể là các đặc trưng thông tin cho các mô hình phân loại. Trong ví dụ trước " 给 力 (cool; powerful – tuyệt vời, mạnh mẽ) là đặc trưng mạnh cho các mô hình phân loại trong khi mỗi một từ đơn thì ngược lại. Việc thêm các từ mới như là một đặc trưng trong mô hình phân loại sẽ cải thiện hiệu suất của phân loại quan điểm.

Trong bài báo này, các tác giả đặc biệt quan tâm đến trích xuất từ cảm xúc mới có thể bày tỏ ý kiến hay cảm xúc, các từ này có giá trị cao cho bài toán phân tích quan điểm.

Từ cảm xúc mới, như được minh họa trong Bảng 1, là một tập con của các thể hiện nhiều từ chính là một chuỗi của các từ láng giềng "có ý nghĩa chính xác và rõ ràng hay hàm ý không thể được bắt nguồn từ ý nghĩa hay hàm ý của các thành phần của nó" (Choueka, 1988). Các từ mới như vậy không thể được xác định trực tiếp bằng sử dụng quy tắc ngữ pháp, nó đặt ra một thách thức lớn trong phân tích tự động. Hơn nữa, nguồn lực từ vựng hiện có không bao giờ đảm bảo đầy đủ và kịp thời khi từ mới xuất hiện liên tục. Do đó người dùng đến các phương pháp thống kê như (Pointwise Mutual Information) (Church và Hanks, 1990), Xác suất có điều kiện (da Silva and Lopes, 1999), Kỳ vọng tương hỗ (Mutual Expectation) (Dias et al., 2000), Thông tin tương hỗ tăng cường (Enhanced Mutual Information) (Zhang et al.,

2009), và Khoảng cách thể hiện giữa nhiều từ (Multiword Expression Distance) (Bu etal., 2010).

New word	English Translation	Polarity
口爱	Lovely	positive
杯具	tragic/tragedy	negative
给力	very cool; powerful	positive
坑爹	reverse one's expectation	negative

Bảng 1 : Các ví dụ của từ quan điểm mới

Ý tưởng chính cho mới phát hiện từ cảm xúc như sau:

Bắt đầu từ rất ít từ hạt giống (ví dụ, chỉ một từ hạt giống), chúng ta có thể trích xuất các mẫu từ vựng có kết hợp thống kê mạnh mẽ với những từ hạt giống; các mẫu từ vựng trích xuất có thể được tiếp tục sử dụng trong việc tìm kiếm nhiều từ mới, và các từ mới có khả năng nhất có thể được thêm vào từ hạt giống cần thiết cho sự tương tác tiếp theo.

Quá trình này có thể được lặp đi lặp lại cho đến khi gặp điều kiện dừng đã được thiết lập. Các vấn đề chính là làm thế nào để đo sự hữu ích của một mẫu và định lượng khả năng của một từ là một từ mới.

Những đóng góp chính của bài báo như sau:

• Đề xuất một framework mới để phát hiện từ mới từ dữ liệu lớn người dùng tạo ra. Framework này là hoàn toàn không có giám sát và hoàn toàn do dữ liệu điều khiển, và chỉ cần nghuồn ngôn ngữ rất ít thông tin (ví dụ, chỉ gồm thông tin gán nhãn từ loại - POS’s tag).

• Thiết kế các độ đo thống kê để xác định độ hữu ích của một mẫu và định lượng khả năng của một từ là một từ mới, tương ứng. Không sử dụng quy tắc ngôn ngữ cần thiết để lọc các kết quả không mong muốn. Đặc trưng này có thể cho phép tiếp cận áp dụng được cho các ngôn ngữ khác.

• Phân tích các vấn đề của dự đoán mức độ quan điểm của một từ cảm xúc mới và chứng minh sự bao gồm của từ quan điểm mới mang lại lợi ích cho các bài toán phân tích quan điểm.

2.2 Phương pháp

2.2.1. Các định nghĩa

Định nghĩa 2.2.1.1. (Phó từ - Adverbial word) Các từ mà chỉ dùng để bổ nghĩa cho động từ hoặc tính từ như

" 太 (too - quá) ", "非常 (very – rất) ", "十分 (very – rất) ", and " 特 别 (specially – đặc biệt) ".

Định nghĩa 2.2.1.2 (trợ từ). Các từ là trợ từ, hoặc dấu ngắt câu bao gồm: “,. ! ? ; :”

Định nghĩa 3.3 (Mẫu từ vựng). Một mẫu từ vựng là một bộ ba gồm <AD, *, AU>, trong đó AD là một phó từ, các ký tự đại diện * là một số tùy ý các từ ngữ, và tất cả các biểu thị một từ phụ trợ.

Bảng 2 đưa ra một số ví dụ về các mẫu từ vựng. Để có được mô hình từ vựng, chúng ta có thể xác định biểu thức thông thường với các nhãn từ loại và áp dụng các biểu thức thông thường trên các văn bản được gán nhãn từ loại. Khi các nhãn của phó từ và trợ từ là quan hệ tĩnh và có thể dễ dàng được xác định, phương pháp này có thể đạt được các mẫu từ vựng một cách an toàn.

Pattern	Frequency
<"都",*,"了">	562,057
<"都",*,"的">	387,649
<"太",*,"了">	380,470
<"不",*,", ">	369,702

Bảng 2: Tần xuất của một mẫu từ vựng trên các bình luận của mạng Weibo.

2.2.2. Tổng quan thuật toán

Thuật toán làm việc như sau:

Bắt đầu từ một tập rất nhỏ các từ giống (ví dụ như bảng 1), thuật toán có thể tìm các mẫu có kết hợp thống kê với các từ giống khi sử dụng Kiểm tra tỷ lệ thích hợp (likelihood ratio test – LRT) để xác định mức độ của sự kết hợp. Sau đó, các mẫu từ vựng đã trích có thể tiếp tục sử dụng trong việc tìm kiếm nhiều từ mới.

Một số độ đo được thiết kế để định lượng khả năng của một từ ứng cử viên là một từ mới, và những các từ đứng đầu xếp hạng sẽ được thêm vào tập các từ hạt giống cho tương tác tiếp theo. Quá trình này có thể được lặp lại cho đến khi gặp điều kiện dừng được thiết lập.

Các mẫu (P) ở mỗi lần lặp sẽ không được tăng lên, thay vào đó, việc cố định một số lượng nhỏ các mẫu trong thời gian lặp lại sẽ cho các kết quả tối ưu.

Từ khía cạnh ngôn ngữ, các từ cảm xúc mới thường được bổ nghĩa bởi các phó từ và do đó có thể được trích bằng các mẫu từ vựng. Đây là lý do tại sao sử dụng thuật toán này. Thuật toán này dựa trên ý tưởng của của thuật toán lan truyền kép (Qiu et al, 2011), tuy nhiên sự khác biệt là ở chỗ: trước hết, nó sử dụng rất ít thông tin ngôn ngữ (chỉ sử dụng thông tin gán nhãn từ loại); Thứ hai, các đóng góp chính là để đề xuất các độ đo thống kê để giải quyết các vấn đề quan trọng sau đây: thứ nhất là đo tính hữu ích của các mẫu từ vựng; thứ hai là đo khả năng của một từ ứng cử viên là một từ mới.

2.2.3. Độ đo tính hữu ích của một mẫu

Vấn đề mấu chốt đầu tiên là đo khả năng thích hợp của một mẫu trong mỗi bước lặp. Điều này có thể thực hiện được bởi sự kết hợp của một mẫu với tập từ hiện tại được sử dụng trong thuật toán. Sử dụng Kiểm tra tỷ lệ thích hợp (Dunning, 1993) cho mục đích này. Đây mà mô hình kết hợp thường được sử dụng giữa các từ mục tiêu quan điểm của (Hai et al., 2012). Likelihood Ratio Test (LRT) nổi tiếng không chỉ dựa chủ yếu vào các giả định của trạng thái bình thường, mà nó sử dụng các giả định tiệm cận của các tỷ lệ khả năng khái quát hóa. Trong thực tế, việc sử dụng các tỷ lệ thích hợp có xu hướng mang lại những cải thiện đáng kể trong hiệu xuất phân tích văn bản.

Ở đây, kiểm tra tỷ lệ thích hợp (LRT) tính một bảng dự phòng của một mẫu p và một từ w, xuất phát từ các thống kê ngữ liệu, được cho trong Bảng

3, khi $k_1(w,p)$ là số tài liệu w phù hợp với mẫu p, $k_2(w,\bar{p})$ là số tài liệu w xuất hiện mà p thì không tồn tại, $k_3(\bar{w},p)$ là số lượng tài liệu mà p xuất hiện trong khi w không có, và $k_4(\bar{w},\bar{p})$ là số tài liệu w và mẫu p đều không có.

Statistics	p	$\bar{p}$
w	$k_1(w,p)$	$k_2(w,\bar{p})$
$\bar{w}$	$k_3(\bar{w},p)$	$k_4(\bar{w},\bar{p})$

Bảng 3: Bảng ngẫu nhiên kiểm tra tỷ lệ thích hợp.

Dựa trên các thống kê được đưa ra trong bảng 3, các kiểm tra tỷ lệ thích hợp, mô hình thu được kết hợp thống kê giữa một mẫu p và một từ w bằng việc thực hiện công thức sau:

$$LRT(w,p) = log\frac{L(p_1,k_1,l_1)*L(p_2,k_2,l_2)}{L(p,k_1,l_1)*L(p,k_2,l_2)} (1)$$

Trong đó:

$L(p,k,n) = p^k * (1-p)^{n-k}$;

$n_1 = k_1 + k_3$;

$n_2 = k_2 + k_4$;

$p_1 = k_1/n_1$; $p_2 = k_2/n_2$;

$p = (k_1+k_2)/(n_1+n_2)$.

Khi đó sự thích hợp của một mẫu có thể được tính như sau:

$$U(p) = \sum_{w_i \in W} LPT(p,w_i)(2)$$

Trong đó: W là tập các từ giống hiện tại sử dụng trong thuật toán.

Thuật toán 1: Thuật toán phát hiện từ mới

Input:

D: là một tập lớn chứa các bình luận được gán nhãn

W_s: là một tập các từ giống

k_p: số các mẫu được chọn cho mỗi lần lặp

k_c: số các mẫu trong tập mẫu ứng cử

k_w: số các từ được thêm vào trong mỗi lần lặp.

K: số các từ được trả về.

Output: Một danh sách các từ mới W

1. Lấy tất cả các mẫu từ vựng sử dụng biểu thức thông thường trên D
2. Đếm tần xuất xuất hiện của mỗi mẫu từ vựng và trích các từ phù hợp với mỗi mẫu;
3. Lấy k_c mẫu có tần xuất cao nhất là tập các mẫu ứng cử cho tập P_c và 5000 từ có tần xuất cao nhất là tập các từ ứng cử W_c;
4. $P = \emptyset; W = W_s; t = 0;$
5. $for\ |W| < K\ do$
6. Sử dụng W để tính điểm với mỗi mẫu trong P_c với U(p);
7. P = {top k_p các mẫu}
8. Sử dụng P để trích các từ mới và nếu các từ này nằm trong W_c, tính điểm chúng với F(w);
9. $W = W \cup \{k\ các\ từ\ có\ giá\ trị\ cao\ nhất\}$
10. $W_c = W_c - W$;
11. Sắp xếp các từ trong W với F(w);
12. Kết quả là danh sách các từ được xếp trong W;

2.3 Độ đo khả năng để là các từ mới.

Một vấn đề quan trọng trong thuật toán được giới thiệu là độ đo khả năng của một từ ứng cử là từ mới.

2.3.1. Kiểm tra tỷ lệ thích hợp (LRT)

Tương tự như độ đo thích hợp mẫu, LRT có thể được sử dụng để tính sự kết hợp của một từ ứng cử viên với một tập mẫu đã cho như sau:

$$LRT(w) = \sum_{w_i \in W} LPT(w, p_i) (3)$$

Trong đó: P là tập mẫu hiện tại được sử dụng trong thuật toán 1, và p_i là một mẫu từ vựng.

Độ đo này chỉ định lượng sự kết hợp của từ mới ứng cử và một tập mẫu đã cho. Nó không cho biết khả năng từ này có phải là từ mới hay không, tuy nhiên một từ quan điểm mới phải có sự kết hợp chặt chẽ với các mẫu từ vựng. Điều này có giải thích ngôn ngữ vì những từ cảm xúc (quan điểm) mới thường được bổ nghĩa bởi từ phó từ và do đó cần phải có liên kết chặt chẽ với các mẫu từ vựng. Độ đo này được chứng minh là một yếu tố có ảnh hưởng bởi các thực nghiệm.

2.3.2. Entropy mẫu bên trái (Left pattern Entropy)

Nếu một từ ứng cử là một từ mới, nó sẽ thường được sử dụng với các mẫu từ vựng đa dạng khi không thể cấu thành của một từ mới có nghĩa là từ này có thế được sử dụng trong nhiều kịch bản ngôn ngữ khác nhau. Điều này có thể được đo bằng entropy thông tin như sau:

$$LPE(w) = -\sum_{l_i \in L(p_c,w)} \frac{c(l_i,w)}{N(w)} * log \frac{c(l_i,w)}{N(w)} \quad (4)$$

Trong đó $L(p_c, w)$ là tập các từ bên trái của tất cả các mẫu với mỗi từ w có thể ghép phù hợp trong p_c

$c(l_i, w)$ là số các từ w có thể ghép phù hợp bằng các mẫu mà từ bên trái là l_i

$N(w)$ là số các từ w có thể ghép phù hợp bởi các mẫu trong p_c

Sử dụng p_c thay cho P vì một tập sau đó là rất nhỉ trong khi tính toán entropy cần một số lớn các mẫu.

2.3.3. Xác xuất của một từ mới

Một số từ xuất hiện rất thường xuyên và có thể được ghép mở rộng bởi các mẫu từ vựng, nhưng nó không phải là các từ mới. Ví dụ: "..(love to eat – thích ăn)" và ".. (love to talk – thích nói) có thể ghép phù hợp bởi một số mẫu. Tuy nhiên, chúng không phải là các từ mới trong khi thiếu nguyên tắc không

cấu thành (non-compositionality). Trong trường hợp này, mỗi ký tự đơn lẻ có xác xuất cao có thể là từ mới. Do đó, cần sử dụng độ đo sau cho quan sát này:

$$NWP(w) = \prod_{i=1}^{n} \frac{p(w_i)}{1 - p(w_i)} \quad (5)$$

Trong đó, w= $w_1w_2…w_n$, mỗi w_i là một ký tự đơn lẻ, và $p(w_i)$là xác xuất để một ký tự (w_i) trở thành một từ, và được tính như sau:

$$p(w_i) = \frac{all(w_i) - s(w_i)}{all(w_i)}$$

Trong đó:

$all(w_i)$ là tổng số lần xuất hiện của w_i

$s(w_i)$ là tổng số lần xuất hiện của w_i là một ký tự từ đơn. Trước đó, chúng ta sử dụng một số công cụ tách từ tiếng Trung.

2.3.4. Các độ đo nguyên tắc không cấu thành.

Các từ mới thường được biểu hiện bằng nhiều từ, trong khi đó, chúng ta có rất nhiều độ đo thống kê được giới thiệu nhằm phát hiện các biểu hiện nhiều từ này. Do đó, các độ đo này có thể được kết hợp một cách tự nhiên trong thuật toán của chúng tôi.

Độ đo đầu tiên là thông tin tương hỗ tăng cường (EMI - enhanced mutual information) (Zhang et al., 2009):

$$EMI(w) = log_2 \frac{F/_N}{\prod_{i=1}^{n} \frac{F_i - F}{N}} \quad (6)$$

Trong đó: F là số các bình luận trong mỗi biểu hiện của nhiều từ $w = w_1w_{2…}w_n$ xuất hiện F_i là số các bình luận mà w_i xuất hiện.

N là tổng số các bình luận.

Ý tưởng chính của EMI là để đo các cặp từ độc lập là tỷ lệ của xác xuất nó là nhiều từ cùng xuất hiện và xác xuất nó không cùng xuất hiện. Giá trị này lớn thể hiện khả năng thể hiện sẽ là nhiều từ.

Độ đo thứ 2 chúng tôi dùng để tính khoảng cách chuẩn giữa nhiều từ (Bu et al., 2010), được giới thiệu để đo độ không cấu thành của các thể hiện nhiều từ.

$$NMED(w) = \frac{\log|\mu(w)| - \log|\emptyset(w)|}{logN - \log|\emptyset(w)|} \quad (7)$$

Trong đó $\mu(w)$ là tập các tài liệu mà trong đó tất cả các từ đơn trong $w = w_1 w_{2\ldots} w_n$ cùng xuất hiện đồng thời

$\emptyset(w)$ là tập các tài liệu mà trong đó mỗi từ w xuất hiện toàn bộ,

N: là tổng số tài liệu

Khác với EMI, độ đo này là độ đo khoảng cách chặt, nghĩa là giá trị này nhỏ sẽ cho biết khả năng lớn hơn nó là thể hiện nhiều từ. Chúng ta có thể thấy trong công thức, ý tưởng chính của độ đo này là để tính tỷ lệ của việc xuất hiện đồng thời của tất cả các từ trong các thể hiện nhiều từ và sự xuất hiện toàn bộ của một giải thích.

2.3.5 Cấu hình để kết hợp các yếu tố khác nhau

Để kết hợp các cách tính trên, chúng tôi đã có các thiết lập để tính điểm cho một từ mới, như sau:

$$F_{LRT}(w) = LRT(w) (8)$$

$$F_{LRT}(w) = LRT(w) * LPE(w) (9)$$

$$F_{LWP}(w) = LRT(w) * LPE(w) * LWP(w) (10)$$

$$F_{EMI}(w) = LRT(w) * EMI(w) (11)$$

$$F_{NMED}(w) = \frac{LRT(w) * LPE(w)}{NMED(w)} (12)$$

2.4. Thực nghiệm

Các tác giả đã thực hiện các thực nghiệm sau:

Trước tiên, họ so sánh phương pháp đề xuất với một số phương pháp có bảns và thực hiện điều chỉnh tham số trong quá trình thực nghiệm.

Tiếp theo là thực hiện phân lớp mức độ cảm xúc của các từ quan điểm mới (sử dụng hai phương pháp).

Cuối cùng, là phần chứng minh là các từ quan điểm mới mang lại lợi ích cho bài toán phân tích cảm xúc.

2.4.1 Chuẩn bị dữ liệu

Các tác giả crawled 237,108,977 bình luận trên trang mạng xã hội Weibo từ trang http://Weibo.com, đây là trang mạng xã hội lớn nhất Trung Quốc. Các bình luận này sẽ được gán nhãn từ loại sử dụng công cụ tách từ tiếng Trung là ICTCLAS (Zhang et al., 2003).

Họ sử dụng hai người gán nhãn cho 5000 từ tuần tự đã trích được bằng các mẫu từ vựng được mô tả trong thuật toán 1.

Người gán nhãn được yêu cầu lựa chọn có hay không một từ ứng cử là một từ mới, và cũng quyết định mức độ cảm xúc của một từ mới (tích cực, tiêu cực hay trung lập). Nếu không có sự thống nhất giữa hai công việc này, họ sẽ thảo luận để đưa ra quyết định. Người gán nhãn đã thực hiện 323 từ mới, trong đó có 116 từ tích cực, 112 từ tiêu cực và 95 từ trung lập.

2.4.2. Các độ đo

Như đề cập ở phần giới thiệu, đầu ra của thuật toán là danh sách các từ được xếp hạng, áp dụng độ chính xác trung bình để đánh giá hiệu suất của việc phát hiện từ cảm xúc mới. Các độ đo được tính như sau:

$$AP(K) = \frac{\sum_{k=1}^{K} P(k) * rel(k)}{\sum_{k=1}^{K} rel(k)}$$

Trong đó: $P(k)$ là độ chính xác tại lát cắt k, $rel(k)$là 1 nếu từ tại vị trí k mà một từ mới và là 0 trong trường hợp ngược lại.

K là số các từ trong danh sách đã xếp hạng. Một danh sách hoàn hảo (tất cả K từ là đúng) có một giá trị AP bằng 1.0.

2.4.3. Đánh giá các độ đo và so sánh với các phương pháp có bản

Trước tiên là xem xét độ trơn của đánh giá tỷ lệ thích hợp, các độ đo sự kết hợp cảu một từ trong một tập mẫu. Mô hình kết hợp (LRT) tăng cường đáng kể hiệu xuất của việc phát hiện một từ mới, điều này cho thấy LRT là một nhân tố quan trọng cho việc trích từ quan điểm mới. Từ quan điểm về ngôn ngữ, các từ quan điểm mới thường được bổ nghĩa bởi các phó từ, do đó nó có mối quan hệ kết hợp với các mẫu từ vựng.

Tiếp theo là phần so sánh các thiết lập của phương pháp này so với phương pháp cơ bản.

Phương pháp có bản thứ nhất là sử dụng thông tin tương hỗ tăng cường (EMI). Chúng tôi thiết lập $F(w) = EMI(w)$

Phương pháp có bản thứ hai là khoảng cách thể hiện nhiều từ chuẩn hóa (NMED) (Bu et al., 2010), chúng tôi thiết lập $F(w) = NMED(w)$

Kết quả ở hình 1 cho thấy, các thiết lập về độ đo của các tác giả đều cho kết quả tốt hơn so với các phương pháp có bản. F_{NMED} cho kết quả tốt nhất.

Việc thêm giá trị NMED hoặc EMI mang lại những cải tiến đáng chú ý nhờ khả năng đo sự không cấu thành của các từ mới. Chỉ sử dụng LRT có thể có được một kết quả khá tốt khi K là nhỏ, tuy nhiên, hiệu suất giảm mạnh bởi vì nó không thể xác định sự không cấu thành.

So sánh giữa LRT + LPE (hoặc LRT + LPE + NWP) và LRT cho thấy các mô hình trái entropy cũng tăng cường hiệu suất rõ ràng. Tuy nhiên, xác suất từ mới (NWP) chỉ đóng góp biên cho việc cải thiện này.

Trong các thực nghiệm trên, các tác giả thiết lập k_p = 5 (số lượng các mẫu được lựa chọn tại mỗi lần lặp) và k_w = 10 (số từ bổ sung ở mỗi lần lặp), đó là các thiết lập tối ưu và sẽ được thảo luận trong phần tiếp theo. Và chỉ có một từ hạt giống "^ H; (ngược lại kỳ vọng của một)" được sử dụng.

2.4.4 Điều chỉnh tham số

Các tham số được lựa chọn cho mô hình:

K_p= 5; k_w=10; $|P_c|$ = 100.

2.4.5. Dự đoán mức độ cảm xúc của các từ quan điểm mới.

Trong phần này, các tác giả cố gắng phân loại mức độ cảm xúc cho 323 từ mới đã được gán nhãn.

Hai phương pháp được lựa chọn cho nhiệm vụ này. Đầu tiên là phương pháp bỏ phiếu đa số (MV), và thứ hai là thông tin tương hỗ, tương tự như (Turney và Liftman, 2003).

Phương pháp bỏ phiếu đa số được xây dựng như sau:

$$MV(w) = \sum_{w_{p \in PW}} \frac{\#(w, w_p)}{|PW|} - \sum_{w_p \in NW} \frac{\#(w, w_p)}{|NW|}$$

Trong đó PW và NW và tập các biểu tượng cảm xúc tích cực và tiêu cực tương ứng (hoặc tập các từ giống)

$\#(w, w_p)$ là đếm sự xuất hiện đồng thời của từ vào w và một mục w_p.

Mức độ cảm xúc được quyết định dựa vào luật sau: nếu MV(w) > th_1, thì từ w là tích cực; nếu MV(w) < - th_1 thì từ w là tiêu cực và ngược lại là neutral. Ngưỡng th_1 được điều chỉnh bằng tay.

PMI được tính toán như sau:

$$PMI(w) = \sum_{w_p \in PW} \frac{PMI(w, w_p)}{|PW|} - \sum_{w_n \in NW} \frac{PMI(w, w_n)}{|NW|}$$

Trong đó:

$$PMI(x, y) = log_2(\frac{\Pr(x, y)}{\Pr(x) * \Pr(y)},$$

Pr(.) là xác xuất.

Mức độ cảm xúc được quyết định dựa vào luật sau:

Nếu PMI(w) > th_2 thì w là tích cực, nếu PMI(w) < -th_2 thì w là tiêu cực và ngược lại là trung lập. Ngưỡng th_2 được thiết lập bằng tay.

2.4.6. Ứng dụng của các từ quan điểm mới cho phân tích cảm xúc.

Trong phần này, các tác giả minh họa cho việc có hay không các từ quan điểm mới mang lại lợi ích cho phân tích quan điểm.

Lấy mẫu ngẫu nhiên 4500 bình luận trên Weibo chứa ít nhất một từ quan điểm trong tập các từ quan điểm Hownet và các từ quan điểm mới đã gán nhãn. Họ áp dụng hai mô hình để phân lớp mức độ quan điểm.

Mô hình thứ nhất là dựa trên từ vựng, trong đó đếm số từ tích cực và tiêu cực trong mỗi bình luận tương ứng và phân lớp các bình luận là tích cực nếu nó chứa nhiều từ tích cực hơn và là tiêu cực trong trường hợp ngược lại.

Mô hình thứ 2 là phân lớp dựa trên SVM với các từ quan điểm là các đặc trưng. Sử dụng đánh giá chéo với 5-fold dữ liệu.

Thử nghiệm với các thiết lập khác nhau của các nguồn tài nguyên từ vựng Hownet:

• Các từ quan điểm Hownet (biểu thị bằng Hownet): Sau khi loại bỏ một số từ không thích hợp, các từ trái có 627 từ quan điểm tích cực và 1.038 từ tiêu cực, tương ứng.

• Các từ quan điểm nhỏ Hownet (ký hiệu là cptHownet): chúng tôi đếm tần xuất của các từ quan điểm trên trên các dữ liệu đào tạo và xóa các từ có tần xuất ít hơn 2. Kết quả cho 138 từ tích cực và 125 từ tiêu cực.

Sau đó, thêm vào các nguồn tài nguyên trên các từ mới đã được gán nhãn mức độ quan điểm (ký hiệu là NW, bao gồm 116 từ tích cực và 112 từ tiêu cực) và 100 từ đầu tiên được trích ra bởi thuật toán (ký hiệu T100). Lưu ý rằng các mô hình dựa trên từ vựng đòi hỏi phải có sự định hướng cảm xúc của mỗi mục từ điển, chúng ta gán nhãn bằng tay cho 100 từ đầu tiên. Kết quả trả về là 52 từ tích cực và 34 từ tiêu cực.

	#Pos/Neg	Lexincon	SVM
Hownet	627/1,038	0.737	0.756
Hownet+NW	743/1,150	0.770	0.779
Hownet+T100	679/1,172	0.761	0.774
cptHownet	138/125	0.738	0.758
cptHownet+NW	254/237	0.774	0.782
cptHownet+T100	190/159	0.764	0.775

Bảng 4: Các kết quả với việc có sử và không sử dụng đánh giá độ đo phù hợp.

Kết quả ở bảng 4 cho thấy thêm các từ mới trong cả hai mô hình đều cải thiện hiệu suất một cách đáng kể. Trong các thiết lập của từ vựng gốc (Hownet), cả hai mô hình tăng 2-3% độ chính xác khi thêm các từ mới, tương tự trong thiết lập các từ vựng nhỏ cũng cho cải thiện tương. Lưu ý rằng, T100 từ được tự động lấy từ thuật toán 1 để nó có thể chứa các từ mà không phải là từ cảm xúc mới, nhưng cũng cải thiện đáng kể hiệu xuất của các nguồn tài nguyên.

CHƯƠNG 3 : ỨNG DỤNG TÌM TỪ QUAN ĐIỂM MỚI CHO DỮ LIỆU TIẾNG VIỆT

3.1. Đặt vấn đề

Ngày nay, cùng với sự phát triển bùng nổ của mạng xã hội trên Internet, người dùng đã và đang tạo ra một lượng dữ liệu rất lớn trong đó thể hiện các quan điểm, nhận xét, đánh giá, cảm xúc của họ về các sản phẩm, dịch vụ, các vấn đề về kinh tế chính trị, xã hội, văn hóa… Các văn bản dạng này thường thể hiện lối viết khá tự do và người dùng thường mong muốn diễn đạt theo cách riêng của mình. Do đó, có rất nhiều từ mới thể hiện quan điểm do người dùng tạo ra.

Ví dụ, khi đánh giá về một sản phẩm điện thoại di động:

"Pin của chiếc điện thoại này rất trâu!"

Hay "Điện thoại này trông ngầu quá!"

Ở đây, rõ ràng là từ "trâu" không phải là cách biểu đạt trong văn bản chính quy khi nhận xét về một chiếc điện thoại, nhưng khi đọc thì người ta có thể hiểu rằng nó ngụ ý cho ta thấy cái điện thoại này pin rất bền.

Hay "ngầu" là từ thường dùng để mô tả cho con người và không phải là từ thông thường khi đánh giá về hình thức của điện thoại. Vậy rõ ràng, tập các từ như vậy có thể coi là các từ cảm xúc do người dùng tạo ra khi nhận xét, đánh giá về sản phẩm, dịch vụ hay các vấn đề khác.

Đã có một số phương pháp đề xuất để phát hiện các từ quan điểm cho các bài toán phân tích quan điểm Tiếng Việt cho kết quả khá tốt. Tuy nhiên việc áp dụng một phương pháp học hoàn toàn không giám sát để phát hiện các từ quan điểm mới có ý nghĩa quan trọng nhằm phát hiện ra các tập từ quan điểm hữu ích cho các bài toán phân tích quan điểm. Do đó trong phần này, chúng tôi áp dụng thuật toán do Minlie Huang và các cộng sự đã đề xuất ở trên để tìm các từ quan điểm mới.

Trong thuật toán, chúng tôi cũng đề xuất một số điều chỉnh cho phù hợp với dữ liệu tiếng Việt.

3.2. Phương pháp

Các từ loại trong gán nhãn từ loại cho tiếng Việt:

Theo Nguyễn Phương Thái và các cộng sự (2009) các nhãn từ loại sử dụng cho Tiếng Việt bao gồm:

STT	Nhãn	Tên	Ví dụ
1	N	Danh từ	Tiếng, nước, thủ đô, nhân dân, đồ đạc, cây cối, chim muông
2	Np	Danh từ riêng	Nguyễn Du, Việt Nam, Hải Phòng, Trường Đại học Bách khoa Hà Nội, Mộc tỉnh, Hóa tỉnh, Phật, Đạo Phật
3	Nc	Danh từ chỉ loại	Con, cái, đứa, bức
4	Nu	Danh từ đơn vị	Mét, cân, giờ, nắm, nhúm, hào, xu, đồng
5	V	Động từ	Ngủ, ngồi, cười, đọc, viết, đá, đặt ; thích, yêu, ghét, giống, muốn
6	A	Tính từ	Tốt, xấu, đẹp, cao, thấp, rộng
7	P	Đại từ	Tôi, chúng tôi, hắn, nó, y, đại nhân, đại ca, huynh, đệ
8	L	Định từ	Mỗi, từng, mọi, cái ; các, những, mấy
9	M	Số từ	Một, mười, mười ba ; dăm, vài, mươi, nửa, rưỡi
10	R	Phó từ	Đã, sẽ, đang, vừa, mới, từng, xong, rồi ; rất, hơi, quá
11	E	Giới từ (kết từ chính phụ)	Trên, dưới, trong, ngoài ; của, trừ, ngoài, khỏi, ở

12	C	Liên từ (kết từ đẳng lập)	Và, với, cùng, vì vậy, tuy nhiên, ngược lại
13	I	Thán từ	Ôi, chao, a ha
14	T	Trợ từ, tình thái từ (tiểu từ)	À, a, á, ạ, ấy, chắc, chăng, cho, chứ
15	B	Từ tiếng nước ngoài (hay từ vay mượn)	Internet, email, video, chat
16	Y	Từ viết tắt	OPEC, WTO, HIV
17	S	Yếu tố cấu tạo tư	Bắt, vô
18	X	Các từ không phân loại được	

Bảng 5: Các nhãn từ loại Tiếng Việt.

Về chức năng ngữ pháp:

Chức năng ngữ pháp của Phó từ trong tiếng Việt cũng thường dùng để bổ nghĩa cho động từ hoặc tính từ. Đây là các từ thường được dùng để nhấn mạnh diễn đạt quan điểm hay cảm xúc. Một số phó từ thường được sử dụng như: đã, sẽ, rất, hơi, quá,...

Tuy nhiên, trong Tiếng Việt không sử dụng các trợ động từ và trong một số trường hợp các trợ động từ này cũng chính là các phó từ như: đã, sẽ, đang,…

Do đó, chúng ta sẽ giải thích lại các định nghĩa như sau:

Định nghĩa 1’: (Phó từ - Adverbial word) Các từ mà chỉ dùng để bổ nghĩa cho động từ hoặc tính từ như: khá, sẽ, rất, hơi, quá,...

Định nghĩa 2': (trợ từ). Các từ là trợ từ hay các tình thái từ, hoặc dấu ngắt câu bao gồm: "á, chăng, chớ,ấy, chắc,. ! ? ; :"

Định nghĩa 3-1 (Mẫu từ vựng). Chúng ta sử dụng mẫu từ vựng để từ đó phát hiện ra các từ quan điểm mới. Xét theo khía cạnh ngôn ngữ Tiếng Việt, thì một phó từ có thể bổ nghĩa cho một động từ hoặc tính từ. Đây cũng là các cụm từ thường được sử dụng để diễn đạt quan điểm hay cảm xúc.

Một mẫu từ vựng là một bộ ba gồm <AD, *, AU>, trong đó AD là một phó từ, các ký tự đại diện * là một số tùy ý các từ ngữ, và tất cả các biểu thị một từ phụ trợ.

Áp dụng thuật toán 1 để tìm các từ mới như sau:

Input: D: là một tập lớn chứa các bình luận được gán nhãn

W_s: là một tập các từ giống

k_p: số các mẫu được chọn cho mỗi lần lặp

k_c: số các mẫu trong tập mẫu ứng cử

k_w: số các từ được thêm vào trong mỗi lần lặp.

K: số các từ được trả về.

Output: Một danh sách các từ mới W

Lấy tất cả các mẫu từ vựng sử dụng biểu thức thông thường trên D

Đếm tần xuất xuất hiện của mỗi mẫu từ vựng và trích các từ phù hợp với mỗi mẫu;

Lấy k_c mẫu có tần xuất cao nhất là tập các mẫu ứng cử cho tập P_c và 1000 từ có tần xuất cao nhất là tập các từ ứng cử W_c;

1. $P = \emptyset; W = W_s; t = 0;$
2. $for\ |W| < K\ do$
3. Sử dụng W để tính điểm với mỗi mẫu trong P_c với U(p);
4. P = {top k_p các mẫu}

5. Sử dụng P để trích các từ mới và nếu các từ này nằm trong W_c, tính điểm chúng với F(w);
6. $W = W \cup \{k \text{ các từ có giá trị cao nhất}\}$
7. $W_c = W_c - W$;
8. Sắp xếp các từ trong W với F(w);
9. Kết quả là danh sách các từ được xếp trong W;

Một đặc điểm tương đồng của tiếng Việt và Tiếng Trung đó là một từ có thể là từ chứa một âm tiết hay nhiều âm tiết. Do đó, thuật toán cũng sử dụng các độ đo đã được giới thiệu trong chương 2 để đánh giá sự thích hợp của mẫu từ vựng và xác định khả năng một từ là từ mới bằng phương pháp như các tác giả đã đề xuất.

3.3. Thực nghiệm

3.3.1. Dữ liệu

Sử dụng dữ liệu đánh giá sản phẩm kỹ thuật (điện thoại di động, máy tính xách tay, tablet, máy ảnh, truyền hình) được thu thập từ một số diễn đàn kỹ thuật Việt Nam như tinhte.vn, voz.vn, thegioididong.com.

Tiền xử lý dữ liệu: loại bỏ những từ không có dấu, các lỗi chính tả trong các bình luận.

Dữ liệu bao gồm 6000 bình luận. Sử dụng các công cụ tách từ tiếng Việt, gán nhãn từ loại để làm dữ liệu vào cho thuật toán.

Một số bình luận:

“Khi mua điện thoại phổ thông cho những người lớn tuổi (cô, bác, bố, mẹ) hoặc cho trẻ con (cháu, em) thì mình vẫn chọn Nokia. Nhưng mua Smartphone thì Nokia không còn là một sự lựa chọn nữa, kể từ khi mình bị Nokia cho nếm mấy "trái đắng" là N96, N97 và N900. N8 sau đó cũng không cứu vãn được.”

“đã mua con này từ cái thời còn 2tr190k giờ nó giảm còn có ~800k theo nhận xét của em là con này (~800k chứ lúc em mua là ~2m2 thì chát quá) wifi nhà ổn, xài viettel gprs ~5-10kb/s, còn 3g thì ~45kb/s nói chung là chậm như bác chủ đã nói. Về độ bền thì em đã xài hơn 1 năm, làm rớt cũng 3-4 lần từ

trên bàn, giường. nhưng mà vẫn ổn. còn cái sim1 bị lock mạng mobi thì sau 1 năm xài tự nhiên nó tự unlock chứ chả hiểu thế nào nên giờ đang xài sim 1 là viettel còn con sim 3g vứt xó."

"Cái đó thì em bó tay,máy cũng bị tình trạng tương tự.mà em thấy để cũng chẳng ảnh hưởng gì nên chẳng quan tâm lắm.máy vẫn chạy bình thường mà."

"Nokia đang trên đường lấy lại thị phần, nhiều con khả thi ghớm."

"mãi yêu nokia dù không còn đỉnh cao như xưa nhưng vẫn trung thành, khổng thể gió chiều nào hùa theo chiều đó."

"Đáp ứng nhu cầu cơ bản của người dùng. Nhưng Nokia dạo này ra em nào cũng khá nhiều màu. Mùa hè nhìn chắc nóng lắm đây."

"Umk mình cũng không hiểu tại sao lại vậy! Chiếc 201 thực tế là bản sao của 200 lại hỗ trợ, phải chăng tùy vào thị trường cung ứng.! haiz chán Nokia chưa thấy sản phẩm nào thực sự hoàn hảo của Nokia có tính năng 2 Sim cả, vẫn có thiếu sót! A em cố gắng góp ý nhiệt tình rồi phản hồi với Nokia VN!"

"mình mới rước em này về từ nguyễnkim rất hài lòng với em nó, nhưng khi mình kết nối wifi máy báo tốt nhưng vao opera chạy được 50% báo 'không tìm thấy máy chủ' hic bác nào biết giúp minh với. có ai biết up rom con nay không?"

"nói chung là em vẫn thích con FPT này hơn: 1-nhìn nó sang hơn con avio 2- có wifi, em toàn nghe online với down nhạc qua wifi không àh"

3.3.2. Xử lý dữ liệu

Ở đây , ta xử dụng các công cụ tách từ Tiếng Việt và gán nhãn từ loại để xử lý dữ liệu bình luận ở trên

Dữ liệu lấy được sau khi sử dụng công cụ tách từ vntokenizer :

"Khi mua điện_thoại phổ_thông cho những người lớn_tuổi (cô , bác , bố , mẹ) hoặc cho trẻ_con (cháu , em) thì mình vẫn chọn Nokia . Nhưng mua Smartphone thì Nokia không còn là một sự lựa_chọn nữa , kể từ khi mình

bị Nokia cho nếm mấy " trái đắng " là N96 , N97 và N900 . N8 sau đó cũng không cứu_vãn được .

đã mua con này từ cái thời còn 2tr190k giờ nó giảm còn có ~ 800k theo nhận_xét của em là con này (~ 800k chứ lúc em mua là ~ 2m2 thì chát quá) wifi nhà ổn , xài viettel gprs ~ 5-10kb / s , còn 3g thì ~ 45kb / s nói_chung là chậm như bác chủ đã nói . Về độ bền thì em đã xài hơn 1 năm , làm rớt cũng 3-4 lần từ trên bàn , giường . nhưng_mà vẫn ổn . còn cái sim 1 bị lock mạng mobi thì sau 1 năm xài tự_nhiên nó tự unlock chứ chả hiểu thế_nào nên giờ đang xài sim 1 là viettel còn con sim 3g vứt xó .

Cái đó thì em bó_tay , máy cũng bị tình_trạng tương_tự . mà em thấy để cũng chẳng ảnh_hưởng gì nên chẳng quan_tâm lắm . máy vẫn chạy bình_thường mà .

Nokia đang trên đường lấy lại thị_phần , nhiều con khả_thi ghớm .

mãi yêu nokia dù không còn đỉnh_cao như xưa nhưng vẫn trung_thành , khổng thể gió chiều nào hùa theo chiều đó .

Đáp_ứng nhu_cầu cơ_bản của người dùng . Nhưng_Nokia dạo này ra em nào cũng khá nhiều màu . Mùa_hè nhìn chắc nóng lắm đây .

Umk mình cũng không hiểu tại_sao lại vậy ! Chiếc 201 thực_tế là bản_sao của 200 lại hỗ_trợ , phải_chăng tùy vào thị_trường cung_ứng . ! haiz chán Nokia chưa thấy sản_phẩm nào thực_sự hoàn_hảo của Nokia có tính_năng 2 Sim cả , vẫn có thiếu_sót ! A em cố_gắng góp_ý nhiệt_tình rồi phản_hồi với Nokia VN !

mình mới rước em này về từ nguyễnkim rất hài_lòng với em nó , nhưng khi mình kết_nối wifi máy báo tốt nhưng vao opera chạy được 50% báo ' không tìm thấy máy_chủ ' hic bác nào biết giúp minh với . có ai biết up rom con nay không ?

nói_chung là em vẫn thích con FPT này hơn : 1-nhìn nó sang hơn con avio 2 - có wifi , em toàn nghe online với down nhạc qua wifi không àh”

Dữ liệu lấy được sau khi sử dụng công cụ gán nhãn từ loại vntagger :

Khi/N mua/V điện/N _/M thoại/N phổ/V _/M thông/N cho/E những/L người_lớn/N _/M tuổi/N (/M cô/N ,/, bác/N ,/, bố/N ,/, mẹ/N)/V hoặc/CC cho/V trẻ/A _/M con/Nc (/N cháu/N ,/, em/N)/A thì/C mình/P vẫn/R chọn/V Nokia/Np ./. Nhưng/C mua/V Smartphone/Np thì/C Nokia/Np không/R còn/R là/V một/M sự/Nc lựa/V _/A chọn/V nữa/R ,/, kể/V từ/E khi/N mình/P bị/V Nokia/Np cho/V nếm/V mấy/L "/" trái/N đắng/A "/" là/V N96/Np ,/, N97/Np và/CC N900/Np ./. N8/Np sau/E đó/P cũng/R không/R cứu/V _/N vãn/V được/R ./.

đã/R mua/V con/N này/P từ/E cái/Nc thời/N còn/R 2tr190k/M giờ/Nu nó/P giảm/V còn/R có/V ~/N 800k/M theo/E nhận/V _/N xét/V của/E em/N là/V con/N này/P (/M ~/Nu 800k/M chứ/C lúc/N em/N mua/V là/V ~/M 2m2/M thì/C chát/A quá/R)/A wifi/V nhà/N ổn/A ,/, xài/V viettel/N gprs/V ~/N 5-10kb/M //X s/A ,/, còn/C 3g/M thì/C ~/M 45kb/M //X s/N nói/V _/N chung/A là/C chậm/A như/C bác/Nc chủ/N đã/R nói/V ./. Về/V độ/N bền/N thì/C em/N đã/R xài/V hơn/R 1/M năm/N ,/, làm/V rớt/V cũng/R 3-4/M lần/N từ/E trên/E bàn/N ,/, giường/N ./. nhưng/C _/A mà/C vẫn/R ổn/A ./. còn/R cái/Nc sim/N 1/M bị/V lock/V mạng/N mobi/V thì/C sau/N 1/M năm/N xài/V tự/P _/M nhiên/N nó/P tự/P unlock/V chứ/C chả/R hiểu/V thế/P _/M nào/P nên/C giờ/N đang/R xài/V sim/N 1/M là/V viettel/N còn/V con/N sim/N 3g/M vứt/V xó/N ./.

Cái/Nc đó/P thì/C em/N bó/V _/M tay/N ,/, máy/N cũng/R bị/V tình/N _/V trạng/N tương/N _/M tự/P ./. mà/C em/N thấy/V để/E cũng/R chẳng/R ảnh/N _/M hưởng/V gì/P nên/C chẳng/R quan/N _/M tâm/N lắm/R ./. máy/N vẫn/R chạy/V bình/N _/M thường/R mà/T ./.

Nokia/Np đang/R trên/A đường/N lấy/V lại/R thị/N _/M phần/N ,/, nhiều/A con/Nc khả/N _/M thi/V ghớm/V ./.

mãi/R yêu/V nokia/N dù/C không/R còn/V đỉnh/N _/N cao/A như/C xưa/N nhưng/C vẫn/R trung/V _/M thành/N ,/, khổng/N thể/R gió/N chiều/N nào/P hùa/V theo/E chiều/N đó/P ./.

Đáp/V _/N ứng/V nhu/N _/M cầu/N cơ/N _/M bản/N của/E người/N dùng/V ./. Nhưng/C _/M Nokia/Np dạo/V này/P ra/R em/N nào/P cũng/R

khá/R nhiều/A màu/N ./. Mùa/N _/M hè/N nhìn/V chắc/A nóng/A lắm/R đây/P ./.

Umk/N mình/P cũng/R không/R hiểu/V tại/E _/M sao/N lại/R vậy/P !/! Chiếc/Nc 201/M thực/N _/Np tế/V là/V bản/N _/M sao/N của/E 200/M lại/R hỗ/V _/N trợ/V ,/, phải/V _/V chăng/R tùy/V vào/E thị/N _/M trường/N cung/N _/M ứng/V ./. !/! haiz/A chán/A Nokia/Np chưa/R thấy/V sản/N _/M phẩm/N nào/P thực/V _/M sự/Nc hoàn/V _/A hảo/V của/E Nokia/Np có/V tính/N _/A năng/R 2/M Sim/N cả/T ,/, vẫn/R có/V thiếu/A _/A sót/V !/! A/I em/N cố/V _/M gắng/N góp/V _/M ý/N nhiệt/N _/M tình/N rồi/C phản/V _/M hồi/N với/E Nokia/Np VN/Np !/!

mình/P mới/R rước/V em/N này/P về/V từ/E nguyễnkim/N rất/R hài/A _/M lòng/N với/E em/N nó/P ,/, nhưng/C khi/N mình/P kết/V _/N nối/V wifi/N máy/N báo/V tốt/A nhưng/C vao/V opera/N chạy/V được/R 50%/M báo/N '/V không/R tìm/V thấy/V máy/N _/M chủ/N '/N hic/V bác/N nào/P biết/V giúp/V minh/N với/E ./. có/V ai/P biết/V up/V rom/N con/N nay/P không/R ?/?

nói/V _/N chung/A là/C em/N vẫn/R thích/V con/Nc FPT/Np này/P hơn/A :/: 1-nhìn/M nó/P sang/V hơn/R con/N avio/N 2/M -/- có/V wifi/N ,/, em/N toàn/R nghe/V online/V với/E down/N nhạc/N qua/V wifi/N không/R àh/V

KẾT LUẬN

Đồ án đã đạt được một số kết quả như sau

- Tìm hiểu tổng quan về phân tích quan điểm hay khai thác quan điểm và các vấn đề đặt ra với bài toán này.
- Tìm hiểu về phương pháp trích từ quan điểm mới trên dữ liệu, ứng dụng vào bài toán phân tích quan điểm
- Tìm hiểu về gán nhãn từ loại cho Tiếng Việt và một số đặc điểm ngôn ngữ tiếng Việt để từ đó lựa chọn đề xuất cho ứng dụng tìm từ quan điểm mới cho dữ liệu Tiếng việt.
- Phân tích dữ liệu thu thập từ các bình luận trên các trang mạng xã hội, tiền xử lý dữ liệu, tách và gán nhãn từ loại để chuẩn bị dữ liệu cho thực nghiệm.

Đề tài với những nội dung kiến thức hoàn toàn mới đối với em nên việc đọc tài liệu, trình bày và tìm hiểu ứng dụng là một thách thức không nhỏ. Do đó với một khoảng thời gian ngắn được phép thực hiện đề tài, em chưa hoàn thành được chương trình cho ứng dụng. Trong thời gian tới, em sẽ tiếp tục phát triển đề tài, đánh giá kết quả thực nghiệm của phương pháp để từ đó có những điều chỉnh và đề xuất mở rộng phù hợp với ngữ liệu.

Em cũng sẽ tiếp tục thử nghiệm với ngữ liệu đủ lớn để đánh giá kết quả của phương pháp.

Trong một khoảng thời gian có hạn, nên việc phát triển trình bày vấn đề em đã nghiên cứu được không tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu của thầy cô và các bạn

Em xin thân thành cảm ơn !

PHỤ LỤC

Công cụ tách từ vntokenizer

I) TỔNG QUAN

Chương trình vnTokenizer được sử dụng để tách từ các văn bản tiếng Việt (mã hóa bằng bảng mã Unicode UTF-8). Chương trình chạy dưới dạng dòng lệnh:

- vnTokenizer.sh nếu chạy trên các hệ điều hành Linux/Unix/Mac OS

- vnTokenizer.bat nếu chạy trên các hệ điều hành MS Windows

Yêu cầu: Máy cần cài JRE (Java Runtime Environment) phiên bản 1.6. JRE có thể tải về từ địa chỉ website Java của Sun Microsystems: http://java.sun.com/

II) DỮ LIỆU

Trong một lần chạy vnTokenizer có thể tách từ một tệp hoặc đồng thời nhiều tệp nằm trong cùng một thư mục.

1) Tách từ một tệp:

Dữ liệu cần cung cấp cho chương trình gồm 1 tệp văn bản tiếng Việt, dạng thô (ví dụ như tệp README.txt này).

Kết quả: Một tệp văn bản kết quả tách từ được ghi dưới định dạng đơn giản hoặc định dạng XML, tùy theo lựa chọn của người sử dụng (xem ví dụ dưới đây).

2) Tách từ nhiều tệp nằm trong một thư mục:

Dữ liệu cần cung cấp gồm một thư mục chứa các tệp văn bản thô cần tách từ (thư mục input) và một thư mục trống (thư mục output) để chứa kết quả tách từ.

Mặc định, chương trình sẽ tự động quét toàn bộ thư mục input và lọc ra tất cả các tệp có đuôi là ".txt". Người sử dụng có thể thay đổi đuôi mặc định

này thành đuôi bất kì, ví dụ ".seg" bằng tùy chọn -e của dòng lệnh (xem ví dụ dưới đây).

Kết quả: Tập các tệp kết quả tách từ trong thư mục output, các tệp này có cùng tên với tệp input tương ứng, tức là tệp input/abc.txt sẽ có kết quả là tệp output/abc.txt.

III) CHẠY CHƯƠNG TRÌNH

1) Tách từ một tệp: vnTokenizer.sh -i <tệp-input> -o <tệp-output> [<các-tùy-chọn>]

Hai tùy chọn -i và -o là bắt buộc. Ngoài ra, người dùng có thể cung cấp các tùy chọn không bắt buộc sau đây:

-xo: dùng định dạng XML để biểu diễn kết quả thay vì định dạng mặc định là văn bản thô.

-nu: không sử dụng dấu gạch dưới (no underscore) khi ghi kết quả. Nếu tùy chọn này được sử dụng thì trong kết quả, các âm tiết không được nối với nhau bằng ký tự gạch dưới, mà bằng ký tự trắng.

-sd: sử dụng mô-đun tách câu trước khi thực hiện tách từ. Nếu tùy chọn này được sử dụng thì trước tiên vnTokenizer thực hiện tách văn bản input thành một tập các câu, sau đó thực hiện tách từ từng câu một. Mặc định thì mô-đun tách câu không được sử dụng, vnTokenizer thực hiện tách từ trên toàn bộ văn bản. Các tùy chọn này có thể được phối hợp đồng thời với nhau để cho ra kết quả mong muốn.

Ví dụ:

a) vnTokenizer.sh -i samples/test0.txt -o samples/test0.tok.txt

Tách từ tệp samples/test0.txt và ghi kết quả vào tệp samples/test0.tok.txt

b) vnTokenizer.sh -i samples/test0.txt -o samples/test0.tok.xml -xo

Tương tự như a), tuy nhiên tệp kết quả samples/test0.tok.xml sẽ có định dạng XML.

c) vnTokenizer.sh -i samples/test0.txt -o samples/test0.tok.txt -sd

Tương tự như a) và sử dụng mô-đun tách câu trước khi tách từ.

2) Tách từ một thư mục: Ngoài các tùy chọn như ở trên, khi tách từ thư mục, chương trình cung cấp thêm tùy chọn không bắt buộc

-e : chỉ định phần mở rộng của các tệp cần tách. Ví dụ:

a) vnTokenizer.sh -i samples/input -o samples/output Thực hiện tách từ tất cả các tệp samples/input/*.txt, ghi kết quả ra thư mục samples/output.

b) vnTokenizer.sh -i samples/input -o samples/output -e.xyz

Thực hiện tách từ tất cả các tệp samples/input/*.xyz, ghi kết quả ra thư mục samples/output

Công cụ gán nhãn vntagger

I) TỔNG QUAN

Chương trình vnTagger là công cụ gán nhãn văn bản tiếng việt có độ chính xác khá cao khoảng 96%

Các thư viện được cung cấp trong tệp tin jars của thư mục lib

Chương trình sử dụng 18 nhãn từ loại như đã nêu trong nội dụng

II) YÊU CẦU

Máy cần cài jre 6.0 trở lên. JRE có thể tải về từ địa chỉ website Java của Sun Microsystems: http://java.sun.com/

III) CHẠY CHƯƠNG TRÌNH

- vnTagger.sh nếu chạy trên các hệ điều hành Linux/Unix/Mac OS

- vnTagger.bat nếu chạy trên các hệ điều hành MS Windows

Chương trình chính không có giao diện người dùng. nếu bạn muốn sử dụng phiên bản có giao diện người dùng, bạn nên tải vnToolkit

Cách gán nhãn một tệp văn bản

Bạn nên cung cấp 2 tham số cho chương trình : tệp văn bản đầu vào để gán nhán (lựa chọn tham số -i) và tệp văn bản đầu ra thể hiện kết quả của chương trình (lựa chọn tham số -o)

Ví dụ : ./vnTagger.sh -i samples/0.txt -o samples/0.tagged.xml

File “0.txt” chứa văn bản tiếng việt có sử dụng mã UTF-8. file "0.tagged.xml" được tạo bởi chương trình và nó cũng có mã UTF-8. Theo mặc định các từ ghép được tách với nhau bằng dấu cách, bạn có thể sử dụng –u để tách chúng bằng dấu gạch dưới. Nếu bạn muốn tệp kết quả là một tệp văn bản đơn giản thay vì là tệp XML, sử dụng lựa chọn -p

Do đó lênh :./vnTagger.sh -i samples/0.txt -o samples/0.tagged.xml -u

Sẽ xuất ra với các âm tiết được tách bởi dấu gạch dưới

Do đó lênh :./vnTagger.sh -i samples/0.txt -o samples/0.tagged.xml –u -p

Sẽ xuất ra với các âm tiết được tách bởi dấu gạch dưới và xuất ra tệp đơn giản thay vì tệp XML

Cách kiểm tra tệp đã gán nhãn

Nếu bạn muốn kiểm tra độ chính xác của việc gán nhãn, sử dụng tham số -t trên tệp cần kiểm tra

Ví dụ : ./vnTagger.sh -t samples/1.tagged.txt

Kết quả kiểm tra sẽ được xuất ra giao diện điều khiển chuẩn.

IV) SỬ DỤNG THE API

Lớp chính của chương trình là vn.hus.nlp.tagger.VietnameseMaxentTagger. lớp cung cấp 3 phương thức gán nhãn sau :

+ public String tagText(String text)

Gán nhãn một văn bản và kết quả là một chuỗi

+ public void tagFile(String inputFile, String outputFile, IOutputer outputer)

Gán nhãn một tệp văn bản và kết quả được xuất ra một tệp

+ public void tagFile(String inputFile, String outputFile)

Gán nhãn một tệp văn bản và kết quả được xuất ra một tệp, sử dụng một tệp đơn giản mặc định

Và một phương thức để kiểm tra tệp đã gán :

+ public void testFile(String filename)

TÀI LIỆU THAM KHẢO

[1]. Phạm Văn Sơn. Tìm hiểu về support vector machine cho bài toán phân lớp quan điểm. Đồ án tốt nghiệp ngành Công nghệ Thông tin, trường ĐHDL Hải Phòng, 2012.

[2]. Lê Hồng Phương. Tài liệu hướng dẫn sử dụng công cụ tách từ Tiếng Việt vnTokenizer, version 4.1.1

[3]. Lê Hồng Phương. Tài liệu hướng dẫn sử dụng công cụ gán nhãn từ loại Tiếng Việt vnTagger, version 4.1.1

[4]. Nguyễn Phương Thái, Nguyễn Lương, Nguyễn Thị Minh Huyền. Tài liệu hướng dẫn gán nhãn từ loại tiếng Việt.

[5]. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinionwordexpansionandtargetextraction through double propagation. Computational linguistics,37(1):9--27.

[6]. MinlieHuang, BoruiYe, YichenWang, HaiqiangChen, JunjunCheng, XiaoyanZhu. 2014. In Proceedings of the Association for Computational Linguistics, 2014.