BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

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

ISO 9001:2000

ISO 9001 : 2008

ĐỒ ÁN TỐT NGHIỆP

NGÀNH CÔNG NGHỆ THÔNG TIN

HẢI PHÒNG 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

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

TÌM HIỂU PHƯƠNG PHÁP TRÍCH VÀ SẮP XẾP CÁC ĐẶC TRƯNG THỂ HIỆN QUAN ĐIỂM

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

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

TÌM HIỂU PHƯƠNG PHÁP TRÍCH VÀ SẮP XẾP CÁC ĐẶC TRƯNG THỂ HIỆN QUAN ĐIỂM

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

Sinh viên thực hiện: Nguyễn Tiến Dũng

Giáo viên hướng dẫn: Ths. Nguyễn Thị Xuân Hương

Mã số sinh viên: 1413101001

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

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

-------*o0o*-------

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

Sinh viên: Nguyễn Tiến Dũng Mã số: 1413101001

Lớp: CTL 801 Ngành: Công nghệ Thông tin

Tên đề tài: Tìm hiểu phương pháp trích và sắp xếp các đặc trưng thể hiện quan điểm

NHIỆM VỤ ĐỀ TÀI

1. Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp
a. Nội dung:
- Đọc tài liệu Tiếng Anh
- Tìm hiểu phương pháp
- Tìm hiểu ngữ liệu
- Cài đặt phương pháp

b. Các yêu cầu cần giải quyết
- Trình bày và giải thích được các yêu cầu của phương pháp, việc áp dụng phương pháp trên dữ liệu tìm hiểu
- Cài đặt thử nghiệm thuật toán

2. Các số liệu cần thiết để thiết kế, tính 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

MỤC LỤC 1

LỜI CẢM ƠN 12

LỜI NÓI ĐẦU 13

CHƯƠNG 1 : TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM – PHÂN TÍCH CẢM XÚC 16

1.1. Sự kiện (Facts) và quan điểm (Opinions) 16

1.2 Lịch sử của phân tích cảm xúc và khai thác quan điểm 19

1.3. Khai thác quan điểm - sự trừu tượng hoá 20

1.3.1. Các thành phần cơ bản của quan điểm: 20

1.3.2. Biểu diễn của đối tượng (Object)/ thực thể (entity): 21

1.3.3. Mô hình của một bình luận cho đối tượng: 21

1.4. Một số nghiên cứu trong phân tích quan điểm 22

1.4.1. Xác định cụm từ, quan điểm 23

1.4.2. Xác định chiều hướng, cụm từ, quan điểm 25

1.5. Bài toán phân lớp quan điểm 28

CHƯƠNG 2: PHƯƠNG PHÁP XẾP HẠNG CÁC ĐẶC TRƯNG SẢN PHẨM CHO XẾP HẠNG CÁC SẢN PHẨM 31

2.1. Giới thiệu 31

2.2. Định hướng xếp hạng dựa trên đặc trưng của các sản phẩm 32

2.2.1 Các thực nghiệm 38

2.2.2. Các kết quả 39

2.3. Tổng kết 41
CHƯƠNG 3: THỬ NGHIỆM TRÊN DỮ LIỆU 43
3.1. Dữ liệu thử nghiệm cho đồ án 43
3.2. Phương pháp 46
3.3. Giới thiệu công cụ JFSA 46
KẾT LUẬN 49
TÀI LIỆU THAM KHẢO 50

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 08 tháng 07 năm 2016

Sinh viên

Nguyễn Tiến Dũng

LỜI NÓI ĐẦU

Cộng đồng người dùng Internet ngày càng phát triển phong phú với nhiều hình thức kết nối, chia sẻ đa dạng như các diễn đàn, trang tin tức, trang thương mại, mạng xã hội như facebook, twitter... Sự phát triển này kéo theo một hình thức mới trong trao đổi thông tin, đó là việc cộng đồng mạng tăng cường chia sẻ cảm nghĩ, nhận xét, đánh giá, nói chung là quan điểm của mỗi người đối với các vấn đề, sự kiện xã hội, kinh tế, chính trị hay kinh nghiệm về một sản phẩm, dịch vụ mà mình từng sử dụng.

Các thông tin thể hiện đánh giá, quan điểm, nhận xét của người dùng đối với các sản phẩm, dịch vụ trên mạng đang trở nên rất hữu ích và có ý nghĩa quan trọng đối với người dùng mới, cũng như đối với các nhà sản xuất, cung cấp dịch vụ. Trước đó, một người dùng khi muốn mua một sản phẩm hay sử dụng dịch vụ nào đó thường có xu hướng tìm hiểu thông tin qua những người xung quanh. Nhưng với sự phát triển của Internet như hiện nay, họ lại thường tìm hiểu thông tin qua mạng. Ví dụ:

- Một người trước khi mua một chiếc điện thoại di động sẽ lên mạng tìm hiểu bình luận (khen, chê) của những người đã sử dụng chiếc điện thoại này, hay xem xu hướng mọi người cộng đồng hay sử dụng loại sản phẩm nào. Một người đi du lịch sẽ chọn khách sạn có các tiêu chí quan tâm được cộng đồng đánh giá tích cực.

- Các thông tin được chia sẻ và thảo luận thông quan mạng xã hội thuộc rất nhiều chủ đề trong các lĩnh vực kinh tế, chính trị, xã hội. Từ đó hình thành nên xu hướng, quan điểm của cộng đồng đối với việc đánh giá một vấn đề, hay một sản phẩm, dịch vụ nào đó. Các quan điểm, xu hướng này sẽ có tác động mạnh mẽ đến định hướng, quan điểm của người dùng khác.

Mặt khác, đối với các nhà sản xuất, các nhà cung cấp dịch vụ để tìm

hiểu các đánh giá của người dùng về sản phẩm và dịch vụ của mình, thay vì phải lấy phiếu điều tra cho sản phẩm một cách thủ công, họ có thể thu thập các thông tin thống kê quan điểm, xu hướng người dùng thông qua các trang mạng. Từ đó sẽ giúp các nhà sản xuất, các nhà cung cấp dịch vụ hoạch định các chính sách cần thiết để phát triển sản phẩm và đáp ứng phù hợp nhu cầu của thị trường.

Để có thể khai thác được các thông tin quan điểm của người dùng, việc tìm kiếm, trích các thông tin có liên quan đến các sản phẩm, dịch vụ có ý nghĩa quan trọng phục vụ cho hệ thống xử lý, đánh giá các quan điểm về sản phẩm dịch vụ mà người dùng hay nhà sản xuất quan tâm.

Với việc mở rộng nhanh chóng của thương mại điện tử trong vòng 15 năm qua, các sản phẩm được bán ngày càng nhiều hơn trên các trang Web và ngày càng có nhiều người dùng đang mua sản phẩm trực tuyến. Để nâng cao kinh nghiệm mua sắm của khách hàng, các trang Web cho phép khách hàng của họ để viết nhận xét về sản phẩm mà họ đã mua. Một số sản phẩm phổ biến có thể nhận được hàng trăm, hàng ngàn ý kiến khác nhau. Từ quan điểm của thương mại điện tử, việc tiếp nhận thông tin phản hồi của người dùng có thể cải thiện chiến lược và phát triển các sản phẩm cho các doanh nghiệp. Vậy làm thế nào để biết được sản phẩm nào được đánh giá tốt, các tính năng (đặc trưng) của sản phẩm nào đang được người dùng quan tâm nhiều hơn và mang yếu tố sống còn cho sản phẩm?

Đã có các tiếp cận khác nhau sử dụng các phương pháp khai phá quan điểm để xếp thứ hạng cho các sản phẩm. Việc xếp hạng từng đặc trưng cụ thể bằng những biểu hiện cụ thể cho đặc trưng đó của sản phẩm rồi kết hợp các xếp hạng cho từng đặc trưng sẽ cho chúng ta xếp hạng của sản phẩm đó. Các thứ hạng của đặc trưng có thể được sử dụng để xác định ảnh hưởng của một đặc trưng trên bảng xếp hạng tổng thể.

Cũng vì lý do đó, trong đồ án này, em nghiên cứu về phương pháp trích và sắp xếp các đặc trưng của sản phẩm, từ đó có đưa ra thứ hạng của từng sản phẩm trong bài toán xếp hạng sản phẩm.

Nội dung đồ án bao gồm 3 chương

Chương 1: *Giới thiệu về bài toán phân tích quan điểm*

Chương 2: *Một số phương pháp trích và sắp xếp đặc trưng*

Chương 3: *Dữ liệu thực nghiệm và kết quả*

Cuối cùng là phần kết luận

CHƯƠNG 1 : TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM – PHÂN TÍCH CẢM XÚC

1.1. Sự kiện (Facts) và quan điểm (Opinions)

Thông tin dạng văn bản có thể chia thành 2 loại chính:

- Sự kiện: là những biểu hiện khách quan về các thực thể, các sự kiện và các thuộc tính của chúng.
  Ví dụ về câu chứa thông tin khách quan:
  "C*hiếc điện thoại này có màu xanh"*
- Quan điểm: là những biểu hiện chủ quan mô tả tình cảm, đánh giá hay cảm xúc của con người đối với các thực thể, sự kiện và thuộc tính của chúng: thể hiện dạng tích cực, tiêu cực hay trung lập.
  Ví dụ câu thể hiện quan điểm:
  "*Chiếc điện thoại này rất mượt"*

Những thông tin nhận xét góp ý hay những thông tin chủ quan chứa quan điểm 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 cảm xúc và khai thác quan điểm

Lĩnh vực phân tích cảm xúc (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.

Hai tiếp cận chính trong phân tích quan điểm: sentiment classification và opinion extraction.

- Sentiment classification: khai thác các kỹ thuật để phân loại các văn bản hoặc thông qua tiếp cận semantic/sentiment như postive, negative [Dave *et al., 2003;* Pang and Lee, 2004; Turney, 2002, etc.].

- Opinion extraction: trích rút các quan điểm bao gồm các thông tin về các nhân tố hướng ngữ nghĩa trong dạng cấu trúc từ văn bản không có cấu trúc, đang được cộng đồng nghiên cứu quan tâm. [Hu and Liu, 2004; Kanayama and Nasukawa, 2004; Popescu and Etzioni, 2005, etc.].

1.3. Khai thác quan điểm - sự trừu tượng hoá

1.3.1. Các thành phần cơ bản của quan điểm:

Quan điểm của một người dùng về một đối tượng có thể được thể hiện bằng các thành phần sau:

- Opinion holder: cá nhân, hoặc tổ chức nắm giữ quan điểm về đối tượng

- Object: đối tượng chứa quan điểm được thể hiện.

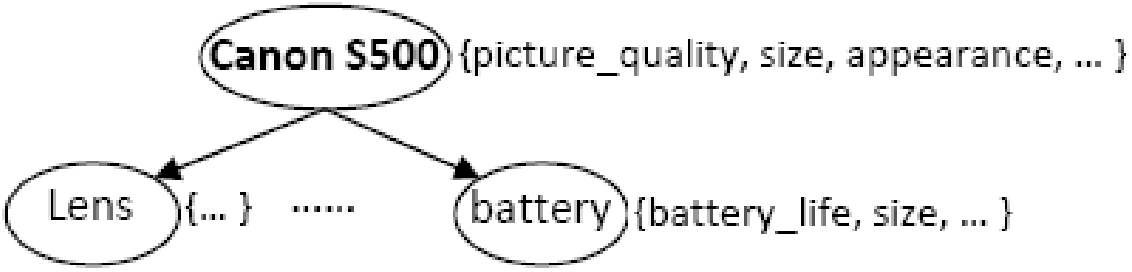
- Opinion: nhận xét, thái độ, đánh giá về đối tượng từ opinion holder.

1.3.2. Biểu diễn của đối tượng (Object)/ thực thể (entity):

Chúng ta có thể biểu diễn thông tin của đối tượng hay thực thể được đánh giá, nhận xét như sau:

- Đối tượng O là: sản phẩm, người, sự kiện, tổ chức hoặc chủ đề.
- Biểu diễn O: Hệ thông phân cấp, O: là nút gốc, mỗi nút là một thành phần (component) và được kết hợp với tập các thuộc tính (attributes) của nó
- Một quan điểm có thể được thể hiện trong một nút hoặc thuộc tính của nút.
- Sử dụng các đặc trưng (features) thay cho các thành phần và thuộc tính.

Ví dụ: biểu diễn cho một thực thể là máy ảnh Cannon S500:



1.3.3. Mô hình của một bình luận cho đối tượng:

Một nhận xét, đánh giá của người dùng cho đối tượng O có thể được thể hiện qua mô hình sau:

- **Một đối tượng O được biểu diễn bằng một tập hữu hạn các đặc trưng: $F = \{f1, f2, \ldots, fn\}$.**

  - Mỗi đặc trưng f_i trong F là một tập hữu hạn các từ hoặc cụm từ W_i (các từ đồng nghĩa – Synonyms)

  - Có tập các từ đồng nghĩa tương ứng: $W = \{W1, W2, \ldots, Wn\}$

- **Mô hình của một quan điểm: Môt opinion holder j nhận xét một tập các đặc trưng $S_j \subseteq F$ của đối tượng O**

  - Mỗi đặc trưng $f_k \in S_j$ là nhận xét của j

    + Chọn một từ hoặc cụm từ từ W_k để mô tả đặc trưng

    + Thể hiện quan điểm là tích cực, tiêu cực, hoặc trung lập trong f_k.

Một quan điểm là bộ 5 thành phần(quintuple)

$(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$,

- o_j là một đối tượng đích
- f_{jk} là một đặc trưng của đối tượng o_j.
- so_{ijkl} là giá trị quan điểm của người nhận xét h_i trong đặc trưng f_{jk} của đối tượng o_j ở thời gian t_l. so_{ijkl} là +ve, -ve, or neu, hoặc các sắp xếp khác.
- h_i là một opinion holder.
- t_l là thời gian quan điểm được đưa ra.

1.4. 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.*
- ✓ *Trích các thông tin chứa quan điểm*
- ✓ *Tóm tắt quan điểm*

1.4.1. 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

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

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

1.4.2. 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

Một số đặc trưng trong dữ liệu văn bản thường được sử dụng trong khai thác quan điểm:

- **Tần suất xuất hiện (Term Presence vs. Frequency)**

  Trong phân mức độ thể hiện quan điểm (polarity classification) việc sử dụng các vector đặc trưng nhị phân là hiệu quả hơn sử dụng tần xuất của các từ thể hiện quan điểm (Pang et al., 2002). Trong khi đó, phân loại văn bản dựa trên chủ đề (topic) lại sử dụng tần xuất xuất hiện của các từ khoá chắc chắn.

Nhưng trên thực tế, các từ xuất hiện chỉ một lần trong văn bản lại có thể là từ chủ quan với độ chính xác cao (Wiebe et al., 2004); Yang et al.,

2006 xem các từ không được liệt kê trong từ điển có trước có thể là từ mới chủ quan dùng để nhấn mạnh trong các bình luận.

- **Mô hình ngôn ngữ: sử dụng các n-grams**

Vị trí của từ có khả năng tác động quan trọng đến cảm xúc hoặc trạng thái chủ quan trong văn bản. Trong Kim and E. Hovy, 2006; Pang et al., 2002, vị trí của từ được mã hoá thành vector đặc trưng và sử dụng cho bài toán phân tích quan điểm.

Thảo luận về việc sử dụng n-grams mức cao là hữu ích, Pang et al., 2002 cho thấy uni-grams thực hiện tốt hơn bigrams trong phân lớp các quan điểm theo các mức cảm xúc cho dữ liệu phim ảnh. Nhưng theo Dave et al., 2003 thì bigrams, trigrams thực hiện tốt hơn trong phân loại phân cực đánh giá sản phẩm.

Riloff et al., 2006 sử dụng một phân cấp tiền đề con để chính thức xác định các loại khác nhau của các đặc trưng từ vựng và các mối quan hệ giữa chúng để xác định các đặc trưng phức tạp hữu ích cho phân tích ý kiến.

- **Thông tin từ loại (Parts of Speech)**

Một số nhà nghiên cứu Mullen và Collier, 2004, Whitelaw et el., 2005, sử dụng các tính từ như các đặc trưng. Hatzivassiloglou và McKeown, 1997 dự đoán data-driven của tiếp cận ngữ nghĩa với từ được phát triển cho các tính từ.

Turney, 2002 đề xuất để phát hiện cảm xúc dựa trên cụm từ được lựa chọn thông qua số lượng xác định trước câu mẫu gán nhãn từ loại có trước, phần lớn bao gồm một tính từ hoặc một trạng từ.

Các nhà nghiên cứu chỉ ra rằng sử dụng các danh từ, động từ có thể là chỉ dẫn mạnh mẽ cho cảm xúc, Riloff et al., 2003.

Một số nghiên cứu Benamara et al., 2007; Nasukawa và Yi, 2003; Wiebe et al., 2004 so sánh hiệu quả của các tính từ, động từ, trạng từ khi phân loại.

- **Phân tích cú pháp (Syntax)**

Những phân tích ngôn ngữ sâu hơn xem như liên quan đặc biệt đến một đoạn của văn bản. Kudo và Matsumoto, 2004 cho rằng hai phân loại mức câu, phân loại cảm xúc và xác định phương thức ("ý kiến", "khẳng định," hoặc "mô tả"), sử dụng học tăng cường dựa trên cây con với các đặc trưng dựa trên cây phụ thuộc thực hiện tốt hơn phương pháp cơ bản thực hiện trên nhóm các từ.

Phân tích cú pháp văn bản có thể là cơ sở cho mô hình hóa valence shifters như phủ định (negative), tăng cường (intensifiers) , và giảm bớt (diminishers) Kennedy và Inkpen, 2006.

Các sắp đặt thứ tự và các mẫu cú pháp phức tạp hơn cũng được sử dụng hữu ích cho phát hiện chủ quan Rilo và Wiebe, 2003; Wiebe et al., 2004.

- **Xử lý phủ định (Negation):là một mối quan tâm quan trọng**

Mô hình hoá phủ định trực tiếp có thể được mã hoá trực tiếp trong định nghĩa các đặc trưng. Das và Chen 2001 thêm NOT vào các từ xuất hiện gần với thuật nhữ như “no” hoặc “don’t”.

Na et al., 2004 mô hình phủ định chính xác hơn bằng cách tìm kiếm các mẫu gán nhãn từ loại đặc biệt để gán nhãn các cụm từ phủ định.

Phủ định có thể được diễn đạt một cách tinh tế khó phát hiện, VD: “[it] avoids all clich´es and predictability found in Hollywood movies”, từ avoid thể hiện ý nghĩa đảo ngược.

Wilson et al., 2005 thảo luận về các tác động phủ định phức tạp khác.

- **Các đặc trưng hướng chủ đề (Topic-Oriented Features)**

Tương tác giữa chủ đề và cảm xúc đóng vai trò quan trọng trong opinion mining. Hagedorn, 2007, về quy mô, thông tin chủ đề có thể kết hợp vào trong các đặc trưng.

Mullen và Collier, 2004 kiểm tra hiệu quả của các đặc trưng khác nhau dựa trên chủ đề (VD, họ đưa vào tính toán khi một cụm từ theo sau một suy dẫn đến chủ đề đang được thảo luận) điều kiện trong thực nghiệm là các suy luận chủ đề được gán nhãn bằng tay.

Kim và Hovy, 2007 đề xuất sử dụng đặc trưng tổng quát để phân tích các quan điểm dự đoán và sau đó tìm trích chọn như là các đặc trưng n-gram. Lược đồ sư dụng đặc trưng n-gram thực hiện tốt hơn 10% độ chính xác trong thực nghiệm của họ.

Sự tương tác topic-sentiment được mô hình hoá thông qua phân tích cây các đặc trưng. Popescu và Etzioni, 2005 sử dụng cây phụ thuộc thể hiện mối quan hệ giữa các cụm quan điểm ứng cử và chủ đề

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

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%.

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: PHƯƠNG PHÁP XẾP HẠNG CÁC ĐẶC TRƯNG SẢN PHẨM CHO XẾP HẠNG CÁC SẢN PHẨM

2.1. Giới thiệu

Một nhiệm vụ khác của khai thác quan điểm nhằm mục đích tóm tắt nội dung các ý kiến cho một thương hiệu, một sản phẩm hoặc một nhà sản xuất cụ thể nào đó. Tuy nhiên, mong muốn thực tế của người dùng thường là được thực hiện theo từng cấp độ, được hỗ trợ tạo ra các xếp hạng hạng tương ứng với nhu cầu cụ thể. Ví dụ như theo một số tiêu chí là đặc trưng của sản phẩm được quan tâm.

Mặt khác, câu hỏi làm thế nào để biết được sản phẩm nào được đánh giá tốt, các tính năng (đặc trưng) của sản phẩm nào đang được người dùng quan tâm nhiều hơn và mang yếu tố sống còn cho sản phẩm cũng thường được đặt ra.

Wiltrud Kessler và các cộng sự đã giới thiệu phương pháp để xếp hạng các sản phẩm dựa trên các thông tin cảm xúc và các bước để thực hiện nhiệm vụ này. Họ xây dựng phương pháp để đưa ra một danh sách xếp hạng các sản phẩm và đưa ra giả thuyết rằng một thứ hạng như vậy sẽ có ích hơn cho người dùng khi họ cần lựa chọn một sản phẩm dựa trên nhu cầu cụ thể hơn so với giá trị cố định.

Có hai điều kiện tiên quyết chính để có thể đạt được mục tiêu đó:

Thứ nhất là cần có chuẩn vàng thông tin xếp hạng, dựa vào đó như là nền tảng để đánh giá. Các xếp hạng này có thể bổ sung để sử dụng tối ưu hóa định hướng dữ liệu của phương pháp để tự động tạo ra các xếp hạng này dựa trên cấu trúc hoặc thông tin nhận xét dạng văn bản.

Trong tiếp cận này, họ sử dụng hai tiêu chuẩn vàng bên đó là xếp hạng bán hàng của Amazon.com và xếp hạng đánh giá cho các đặc trưng sản phẩm của Snapsort.com.

Thứ hai là các tiếp cận khác nhau để sử dụng các phương pháp khai phá quan điểm để tạo ra các thứ hạng cho các sản phẩm. Họ tập trung vào các phương pháp làm mịn dần với sự kết hợp thể hiện quan điểm của từng đặc trưng khác nhau. Họ tạo ra bảng xếp hạng với từng đặc trưng cụ thể với những đánh giá cho đặc trưng đó của sản phẩm. Việc kết hợp các xếp hạng cho từng đặc trưng sẽ cho chúng ta xếp hạng của sản phẩm đó. Các xếp hạng đặc trưng có thể được sử dụng để xác định ảnh hưởng của một đặc trưng trên bảng xếp hạng tổng thể.

Công trình đã mang lại các đóng góp sau:

Thảo luận về nhiệm vụ của dự đoán xếp hạng đầy đủ cảu các sản phẩm bên cạnh dự đoán riêng biệt của các bình chọn.

Chứng minh làm thế nào phương pháp khai phá quan điểm dựa trên so sánh và hướng mục tiêu có thể được sử dụng cho dự đoán các thứ hạng sản phẩm. Họ sử dụng dữ liệu thực tế cho các xếp hạng, sử dụng thông tin xếp hạng bán hàng từ Amazon.com và xếp hạng chất lượng từ Snapsort.com.

Chỉ ra rằng phương pháp khai thác quan điểm bằng cách làm mịn dần (xếp hạng các đặc trưng trước) đạt được hiệu suất đáng kể trong việc dự đoán các thứ hạng từ thông tin văn bản.

Giới thiệu các xếp hạng đặc trưng cho phép hiểu được tác động của từng khía cạnh cho các xếp hạng chung của sản phẩm.

2.2. Định hướng xếp hạng dựa trên đặc trưng của các sản phẩm

Phần lớn các cách tiếp cận khai thác quan điểm thực hiện trích các đánh giá của các sản phẩm và các đặc trưng để làm kết quả của quá trình phân tích. Đây chính là quá trình giải thích cho người dùng cuối các thứ hạng cho các đặc trưng khác nhau. Tuy nhiên, các giả định cơ bản là người dùng cuối này có thể kết hợp thông tin này theo một cách nào đó để đưa ra các quyết định riêng. Tính tiện ích của thông tin từ các hệ thống khai thác quan điểm rõ ràng là tùy thuộc vào các trường hợp sử dụng cụ thể và nhu cầu chủ quan. Do đó, các đặc trưng quan trọng của một thứ hạng của các sản phẩm chính là:

Việc xếp hạng hỗ trợ các nhu cầu cụ thể của một cá nhân hay của một nhiệm vụ đầu/cuối.

• Việc xếp hạng có thể hoàn toàn chủ quan hoặc nửa chủ quan.

• Một người sử dụng có thể bị ảnh hưởng bởi những yếu tố tác động đến sở thích dù có thứ hạng hay không.

Một ví dụ của một thứ hạng là nó đã có sẵn từ cấu trúc siêu dữ liệu chính là bảng xếp hạng của một chủng loại sản phẩm từ một cửa hàng bán hàng trực tuyến (trong công việc này, là các thứ hạng doanh số bán hàng của Amazon.com).

Thứ hạng này xác định cho trường hợp người quản lý có nhu cầu tối đa hóa sự phổ biến của một sản phẩm. thứ hạng này là nửa chủ quan và người sử dụng thường không nhận thức đầy đủ của tất cả các yếu tố ảnh hưởng đến thứ hạng. Các yếu tố đó là giá của sản phẩm, chất lượng, tỷ lệ hiệu năng của giá cả, quảng cáo, vv. Do đó, thực hiện tính toán thông tin được sinh ra bằng các phương pháp khai thác quan điểm theo cách làm mịn dần có thể làm sáng tỏ đến tác động của từng khía cạnh trên các xếp hạng này. Nếu các đánh giá và xếp thứ hạng bán hàng xuất phát từ cùng một nguồn, số các ý kiến đánh giá đang được sẵn sàng cho một sản phẩm có thể được coi là tương quan (hoặc ít nhất là tương tác) với số lượng bán ra.

Các nhận xét đóng một vai trò quan trọng đối với một quyết định mua hàng, vì vậy sự tương tác cũng sẽ làm việc theo một hướng khác, khi một sản phẩm có nhiều đánh giá và hầu hết trong số đó là tích cực, cơ hội sẽ tăng lên và mọi người sẽ mua nó.

Một trường hợp khác của nguồn Một thể hiện của một nguồn thông tin đã có là xếp hạng chuyên gia, trong đó một chuyên gia miền so sánh các sản phẩm khác nhau và các đặc trưng khác nhau của chúng và đặt chúng theo một thứ tự.

Một nguồn tin phổ biến cho xếp hạng này là các trang báo hoặc các trang web cụ thể của miền với mục đích cung cấp cho người dùng với một nguồn đầy đủ thông tin hỗ trợ ra quyết định mua hàng của họ. Xếp hạng này thường hoàn toàn chủ quan, tuy nhiên, các yếu tố khác nhau được đưa

vào tính toán, nó có thể được tiết lộ hay không. Ở đây, họ sử dụng các thông tin sẵn có từ Snapsort.com

Đây là một dịch vụ thu thập thông tin chi tiết về máy ảnh và cung cấp sự so sánh giữa chúng. Điểm số của chúng kết hợp các đặc trưng từ thông số kỹ thuật như màn trập, kích thước ngắm, có hay không sự ổn định của việc định hình ảnh, cũng như tính phổ biến (các máy ảnh đã được xem bao nhiêu lần trên các trang web) hoặc số ống kính có sẵn. Thứ hạng như vậy đã được sử dụng trong công việc trước đây công bố gần đây của Tkachenko và Lauw (2014), người sử dụng một phần của đánh giá chuyên gia tiêu chuẩn vàng khi họ xác định các đặc điểm được xác định trước cho sản phẩm của họ (ví dụ: máy ảnh nhỏ hơn được đánh giá tốt) và đánh giá lần nữa đối với các xếp hạng đặc trưng cụ thể.

Cả xếp hạng doanh thu và xếp hạng chuyên gia đều đang cố gắng để kết hợp ý kiến từ hoặc một tập hợp các người dùng. Tuy nhiên, các xếp hạng các sản phẩm có thể là rất chủ quan. Vì vậy, việc giới thiệu một xếp hạng thực tế phải dựa trên cộng đồng mà không làm mịn trước những đặc trưng được đưa vào tính toán để đưa ra quyết định.

Thông thường trong việc gán nhãn xếp hạng, yêu cầu một xếp hạng đầy đủ của một danh sách các sản phẩm từ những người gán nhãn là một thách thức rườm. Vì vậy, đề xuất nhiệm vụ cộng đồng như vậy cần được thiết lập trong học xếp hạng, khi đó những người gán nhãn được yêu cầu xác định ưu tiên cho một cặp sản phẩm. Các nhãn như vậy có được sử dụng sau đó để tạo ra một thứ hạng nửa chủ quan cũng như thứ hạng cá nhân. Cách tiếp cận này không được thực hiện trong bài báo này nhưng có thể mang lại những đóng góp cho các nghiên cứu trong tương lai.

Từ các thứ hạng như vậy, một chức năng sở thích cá nhân có thể được học với trọng số khác nhau của mỗi đặc trưng khác nhau với nhau, thậm chí cả khi người dùng không nhận thức được các nhân tố này.

2.3. Các phương pháp

Nhiệm vụ của bài báo này là tạo ra một danh sách thứ hạng của các sản phẩm dựa trên thông tin cảm xúc. Để xếp thứ hạng các sản phẩm, các

tác giả thực hiện 3 phương pháp cho phân tích văn bản và 2 phương pháp cơ bản (baselines).

Có hai cách tiếp cận dựa trên tính các từ hoặc các cụm từ có thể hiện tích cực và tiêu cực.

Đầu tiên là xác định các mức độ quan điểm dựa trên từ điển với lớp tương ứng được quy định rõ ràng.

Điểm thể thiện cảm xúc score(p) cho mỗi sản phẩm p được tính bằng số các từ tích cực (pos) trên toàn bộ các nhận xét cho sản phẩm này trừ đi số các từ tiêu cực (neg).

$$score_{dict}(p) = pos(p) - neg(p) (1)$$

Để tính sự tác động cho các nhận xét dài hơn, họ chuẩn hóa số các từ trong toàn bộ các nhận xét cho các sản phẩm đặc biệt all_p:

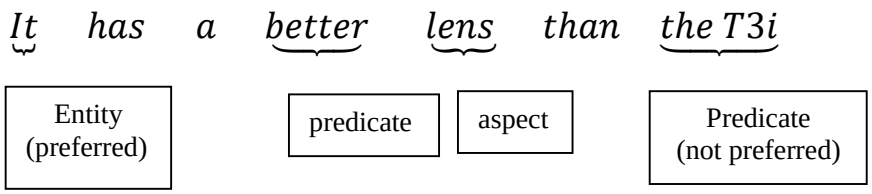
$$\overline{score_{dict}(p)} = \frac{score(p)}{all_p} (2)$$

Danh sách được xếp hạng của các sản phẩm được tạo bởi việc sắp xếp theo các điểm này. Quan tâm đến hai biến thể của phương pháp này là DICT và DICTNorm.

Đây là phương pháp đầu tiên dựa trên từ điển dễ dang thực hiện và sử dụng. Tuy nhiên, nó không thể đưa vào công thức này các thể hiện chứa mức độ quan điểm. Do vậy, phương pháp thứ hai được lựa chọn là phát hiện dựa trên học máy cho các cụm khách quan với các mức độ thể hiện quan điểm của chúng trong ngữ cảnh, sử dụng JPFA (Joint Fine-Grained Sentiment Analysis Tool, Kingler và Cimiano, 2013).

Tính toán điểm cho sản phẩm và xếp hạng được thực hiện tương tự như cách tiếp cận dựa trên từ điển. Họ đề cập đến hai biến thể của phương pháp này là JFSA và JFSA-NORM.

Để tạo ra một danh sách được xếp hạng các sản phẩm, họ hướng đến việc thực hiện khai thác các thể hiện so sánh văn bản, như trong ví dụ sau:



Để trích các so sánh này, sử dụng công cụ được giới thiệu cho CSRL (Comparision Semantic Role-Labeler, Kessler và Kuhn, 2013). Hệ thống này phát hiện và trích các vị từ so sánh ("better"), hai thực thể liên quan là "It" và "the T3i", trong đó "It" được quan tâm hơn và đặc trưng được so sánh là "lens".

Để xác định các sản phẩm nào được yêu thích hơn, họ kết hợp với thực thể được đề cập cho tên sản phẩm (hoặc các tên đại diện) với độ tương tự cosin tối thiểu trong mức từ.

Ở ví dụ trên, "T3i" được kết hợp với "Canon EOS Rebel T3i"; đại từ "It" được xác định với sản phẩm đang được đánh giá.

Điểm cho một sản phẩm được tính dựa trên số lần nó xuất hiện là sản phẩm được thích hơn (pref) trừ đi số lần nó không được thích hơn (npref):

$$score_{CSRL}(p) = pref(p) - npref(p) (3)$$

Điểm trả về cho từng sản phẩm được sử dụng để sắp xếp tương tự như đề cập ở trên. Phương pháp này được gọi là C_{SRL}.

Sử dụng hai phương pháp cơ bản để xác định thông tin văn bản của một bình luận:

Phương pháp đầu tiên là phân loại sản phẩm theo xếp hạng sao trung bình (từ một đến năm sao, được xác định bởi tác giả của một bài đánh giá) của tất cả các đánh giá các sản phẩm tương ứng (STAR).

Phương pháp thứ hai sắp xếp các sản phẩm bằng của số ý kiến đã nhận được (NUMREVIEWS). Băng trực giác có thể thấy sản phẩm được bán ra thường xuyên sẽ có nhiều đánh giá hơn.

Hai phương pháp đề xuất là J_{FSA} và C_{SRL} nhận dạng các đặc trưng của sản phẩm cùng với các cụm từ đánh giá khách quan hoặc so sánh tương ứng.

Bên cạnh việc tạo một các thứ hạng được xếp, phương pháp còn kết hợp độ đo cảu tất cả các đặc trưng của sản phẩm, sử dụng các tùy chọn để chỉ sử dụng để đánh giá về các đặc trưng cụ thể từ đó trả về kết quả là danh sách các đặc trưng được xếp hạng. Khi một đặc trưng được đề cập đến với nhiều thể hiện, họ sử dụng hàm chuẩn hóa để lọc thông tin cần thiết.

Khi tiến hành thực nghiệm, họ sử dụng một danh sách được thực hiện thủ công các đánh giá văn bản cho các đặc trưng xuất hiện thường xuyên nhất trong tập dữ liệu. Trong phiên bản tiếp theo của phương pháp, các cụm từ chủ quan hoặc các thực thể xem xét chỉ tính giá trị của sản phẩm nếu có một từ trùng giữa đặc trưng được nhận dạng và một văn bản biến thể của đặc trưng mục tiêu.

Method	Amazon	Snapsort
STARS	-0.027	0.436*
NUMREVIEWS	0.331*	0.095
DICT-NORM (GI)	0.125*	-0.148
DICT-NORM (MPQA)	0.142*	-0.145
DICT (GI)	0.219*	0.426*
DICT (MPQA)	0.222*	0.441*
JFSA-NORM	0.151*	-0.230
JFSA	0.234*	0.404*
CSRL	0.183*	0.511*

Bảng 1: Kết quả của các phương pháp target-agnostic cho sự đoán xếp hạng bán hàng của amazon và xếp hạng chất lượng của Snapsort. Sự cải thiện vượt quá ngẫu nhiên được đánh dấu * ($p<0.05$). Phương pháp cơ bản tốt nhất được in đậm.

2.2.1 Các thực nghiệm

Các thiết lập cho thực nghiệm

Để đánh giá phương pháp, sử dụng các nhận xét được lấy từ trang Amazon với các sản phẩm: "camera" và "camera" trong kết nối với "fuji", "fuji-hTm", "canon", "panasonic", "olympus", "nikon", "sigma", "hasselblad", "leica", "pentax", "rollei", "Samsung", "sony", "olympus"

Sử dụng cho chuẩn vàng thứ nhất, dữ liệu được lấy từ trang xếp hạng bán hàng Amazon cho các mô tả sản phẩm. (Xếp hạn bán hàng tốt nhất trên Amazon cho loại Máy ảnh và Photo) trong khoảng thời gian từ 14-18/04/2015, và bao gồm chỉ các sản phẩm được cung cấp xếp hạng. Kết quả trả về danh sách 920 sản phẩm với tổng số 71.409 nhận xét. Các tên của sản phẩm được trích từ tiêu đề của trang và sử dụng 6 ký tự đầu tiên.

Đối với chuẩn vàng thứ hai, sử dụng thứ hạng cho chất lượng sản phẩm được cung cấp bởi Snapsort, trong số 150 sản phẩm hàng đầu trong bảng xếp hạng doanh số bán hàng của Amazon thì có 56 sản phẩm cuất hiện tên Snapsort. Sử dụng các thứ hạng trong loại "Best overall" (tổng thể tốt nhất) của "tất cả các máy ảnh kỹ thuật số công bố trong 48 tháng cuối cùng" được truy hồi vào ngày 12 Tháng Sáu 2015.

J_{FSA} được huấn luyện trên dữ liệu về máy ảnh được thiết lập bởi Kessler et al. (2010). C_{SRL} được huấn luyện về dữ liệu máy ảnh của Kessler và Kuhn (2014). Đối với các phương pháp dict và dict-NORM, các tác giả thử trên hai nguồn từ quan điểm khác nhau, từ điển người điều tra chung (Stone et al., 1996) và các đầu mối chủ quan từ hệ hỏi đáp MPQA (Wilson et al., 2005).

Để đo lường sự tương quan của xếp hạng được tạo ra bằng các phương pháp khác nhau các tác giả sử dụng thứ hạng vàng, tính toán hệ số điều chỉnh tương quan thứ hạng của Spearman là p (Spearman, 1904). Kiểm tra tính khả quan với các thử nghiệm Steiger (Steiger, 1980).

2.2.2. Các kết quả

Xem xét hai xếp hạng khác nhau cho đánh giá: xếp hạng bán hàng bao gồm 920 sản phẩm, đây là một ví dụ cho một xếp hạng có thể hữu ích cho các nhà quản lý bán hàng và các nhà sản xuất sản phẩm.

Thứ hai là xếp hạng chuyên gia bởi Snapsort.com bao gồm 56 sản phẩm. Đây là hai thứ hạng cho hai khái niệm khác nhau và không có độ tương quan giữa hai xếp hạng (p= -0.04).

Theo các tác giả, bảng 1 là sự so sánh kết quả của các phương pháp cơ sở và các phương pháp đề xuất.

Kết quả tốt nhất trên Amazon bằng các đếm số nhận xét (p=0.33, NUMREVIEWS)

Với Snapsort, NUMREVIEWS chỉ cho p = 0.1. Nhân tố tạo ra sự khác biệt trong trường hợp của Amazon là đánh giá và xếp hạng đến từ cùng một nguồn và nó không rõ ràng khi mà có hay không sự phổ biến của một sản phẩm dẫn đến có nhiều nhận xét đánh giá hay sản phẩm dẫn đến nhiều nhận xét hay số đánh giá nhiều dẫn đến danh số bán hàng cao hơn. Và mặc dù "phổ biến" là một trong những khía cạnh ảnh hưởng đến đánh giá trên Snapsort, nhưng nó không đáng chú ý.

Hiệu suất của phương pháp cơ bản STARS không khác biệt đáng kể khi lấy ngẫu nhiên từ Amazon. Điều này giải thích một phần bởi thực tế là trong số các sản phẩm với đánh giá 5* chỉ có rất ít nhận xét (dưới 10). Đây là một vấn đề yếu trong xếp hạng của Snapsort. Bên cạnh đó, mong muốn nội dung của các đánh giá là các quyết định chất lượng và gần với những gì người dùng Snapsort sử dụng để đánh giá hơn là những ảnh hưởng của doanh số bán hàng.

Xếp hạng dựa trên xác định mức độ quan điểm theo từ điển (DlCT) xấp xỉ xếp hạng doanh thu bán hàng với p = 0,22, cho cả MPQA và GI. Chuẩn hóa các điểm mức độ quan điểm làm giảm sự tương quan. Sự tương tự của các kết quả thu được của hai bộ từ điển khác nhau được phản ánh trong các mối tương quan rất cao của các xếp hạng trả về (không chuẩn hóa: p = 0,99; chuẩn hóa: p = 0,8). Tuy nhiên, các xếp hạng với không

chuẩn hóa là không tương quan với các xếp hạng chuẩn hóa của cùng từ điển. (GI p = -0.16, MPQA p = -0,14).

Việc xếp hạng dựa trên từ điển tốt hơn một chút với JFSA, p = 0,23. Chuẩn hóa số từ tố (do đó tác động đến số nhận xét) làm giảm hiệu suất p = 0,15. Sự khác biệt của JFSA với dict-NORM (GI) và DlCT (MPQA và GI) là khả quan (p <0,05). Đối với Snapsort, chuẩn hóa có tác động rất không tốt.

Trên Amazon, xếp hạng đạt được với CSRL là bình thường so với các phương pháp khác. CSRL chịu sự ảnh hưởng của dữ liệu thưa (số lượng cao nhất của các cụm từ quan điểm cho một sản phẩm được tìm thấy trong JFSA là hơn 9000, trong khi số lượng cao nhất của sự so sánh đó đề cập đến một sản phẩm đã cho là 662 cho CSRL). Tuy nhiên trong xếp hạng ở Snapsort, CSRL cho kết quả tốt nhất của tất cả các thực nghiệm với p = 0,51.

So sánh việc sử dụng tất cả các thông tin từ các ý kiến để tạo ra các xếp hạng, các kết quả đặc trưng cụ thể cho thấy sự hiểu biết về tác động của từng đặc trưng trên xếp hạng vàng. Các xếp hạng đặc trưng cụ thể đối với các đặc trưng quan trọng liên quan chặt chẽ với xếp hạng vàng, trong khi những đặc trưng hoàn toàn không liên quan có một tương quan gần ngẫu nhiên.

Aspect	#	p	σ
performance	637	0.301	0.009
Video	600	0.278	0.013
Size	513	0.218	0.017
pictures	790	0.213	0.003
battery	541	0.208	0.012
Price	625	0.198	0.008

Zoom	514	0.196	0.013
shutter	410	0.191	0.016
features	629	0.190	0.009
autofocus	403	0.175	0.013
screen	501	0.136	0.012
Lens	457	0.099	0.012
Flash	591	0.093	0.011

Bảng 2: Các kết quả của phương pháp JFSA cho dự đoán thứ hạng doanh số bán hàng khi chỉ sử dụng các cụm từ được xem xét cho đặc trưng mục tiêu đã xác định.

Các kết quả cho xếp hạng doanh số bán hàng Amazon và JFSA được thể hiện trong Bảng 2. Do sự thưa thớt dữ liệu, một số lượng lớn các sản phẩm nhận được một số điểm là 0. Để loại bỏ những kết phản ánh phát giả của p trong khi cho phép so sánh giữa các phương pháp với nhau về số lượng sản phẩm được lưu, họ thêm các sản phẩm điểm 0 theo thứ tự ngẫu nhiên và có hơn 100 danh sách xếp hạng khác nhau ngẫu nhiên. Bỏ qua các kết quả cho CSRL và các kết quả trên Snapsort mà tất cả đều gần ngẫu nhiên.

Đối với bảng xếp hạng được tạo bởi JFSA, hiệu suất của đặc trưng đóng góp nhiều nhất gần với xấp xỉ xếp hạng của doanh thu (p = 0,30) tiếp theo là video (p = 0,28). Cả hai kết quả tốt hơn xếp hạng target-agnostic của JFSA (p = 0,23) (đáng kể về mặt hiệu suất).

2.3. Tổng kết

Các tác giả giới thiệu công việc dự đoán thứ hạng của các sản phẩm và giới thiệu ba nguồn tiềm năng cho các thứ hạng vàng: xếp hạng doanh thu bán hàng và xếp hạng dựa trên ý kiến đánh giá của chuyên gia đã được

sử dụng trong các thực nghiệm. Thêm nữa là các thảo luận làm thế nào để gán nhãn dữ liệu xếp hạng đựa trên cộng đồng. Chứng minh các các kết quả ban đầu làm thế nào để sử dụng các phương pháp khai thác quan điểm khác nhau (dựa trên từ điển, máy học, dựa vào so sánh) để dự đoán xếp hạng. Và thực nghiệm về cách xếp hạng các đặc trưng cụ thể có thể được sử dụng cho đo lường tác động của các thông tin quan trọng trong xếp hạng.

Các phương pháp thảo luận cho thấy một hiệu suất còn hạn chế, tuy nhiên, những kết quả xấp xỉ một thứ hạng ở thế giới thực là có triển vọng và khuyến khích nghiên cứu thêm. Mặc dù điểm số tương quan là tương đối thấp, nhưng nó cho phép cho một phân tích về ảnh hưởng của một đặc trưng cụ thể trong xếp hạng như cho xếp hạng doanh thu trên Amazon.

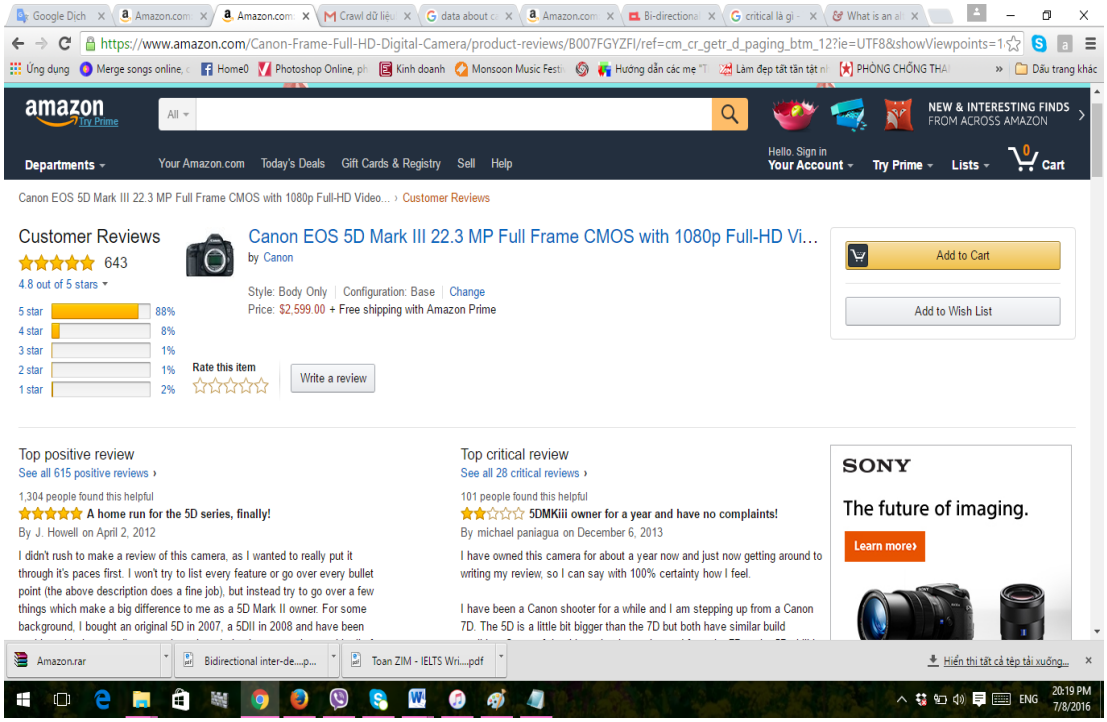
Kết quả tốt nhất cho việc xếp hạng doanh số bán hàng của Amazon đạt được dựa trên số đánh giá (NUMREVIEWS). Điều này có thể được xem như là một trường hợp của con gà và quả trứng, và nó có thể là trường hợp mà có rất nhiều đánh giá bởi vì sản phẩm đã được bán nhiều lần. Hiệu ứng tương tự không xuất hiện trên Snapsort. Xếp hạng sao trung bình (STARS) không phải là thông tin hướng tới cho xếp hạng bán hàng trên Amazon, nhưng cho kết quả tốt trên Snapsort.

Các phương pháp này xem xét đến mức độ quan điểm của các cụm từ mang lại kết quả tốt thứ hai (JFSA và DICT) trên Amazon. Với Snapsort, phương pháp dựa trên sự so sánh CSRL thực hiện tốt nhất trên tất cả các phương pháp khác và cho hiệu suất cao nhất trên mọi thực nghiệm ở đây ($p = 0.51$).

CHƯƠNG 3: THỬ NGHIỆM TRÊN DỮ LIỆU

3.1. Dữ liệu thử nghiệm cho đồ án

Dữ liệu được sử dụng: Trong phần thực hiện thử nghiệm cho phương pháp sắp xếp thứ hạng các đặc trưng phục vụ cho xếp thứ hạng các sản phẩm, em tìm hiểu và phân tích dữ liệu trên trang Amazon.com với các đánh giá của khách hàng cho sản phẩm cung máy ảnh.



Dữ liệu bình luận của khách hàng được crawl về cho danh mục sản phẩm Camera and Photo có dạng sau:

```
<title>Canon EOS 5D Mark III</title>

<link>  https://www.amazon.com/Canon-Frame-Full-HD-Digital-Camera/dp/B007FGYZFI/ref=sr_1_1?s=electronics&ie=UTF8&qid=1467980819&sr=1-1&refinements=p_89%3ACanon#customerReviews</link>

<NumberOfPosts>643</NumberOfPosts>

<lastBuildDate>07 june 2016 04:26:48 AM</lastBuildDate>

<!-- Comments -->

<items>

<tag  value="0">
```

<properties CustomerReviews="3" TopReviewerRanking="134" HelpfulVotes="3" author=" Nelly “date="02/18/2016">

</properties>

<infomation star="5" title="All I have is one word to describe this camera... HOLY CRAP!! Yes I know that's more than one word!">

</infomation>

<comment>OK so I didn't get my Canon 5d III from amazon because of financing options elsewhere but I just had to leave a review here. Ok so I jumped from a canon t3i with the 18-55mm kit lens straight into this monster 5d III with the canon 24-70mm f/2.8 mkII zoom lens.

The Canon 5d III is better than the Canon t3i in just about every department. I bought it about a week before Christmas 2015 and I'm writing this review now about 4 weeks later after I've had time to actually play with it and take a few 100 shots during Christmas and New Years parties and a small portrait session. I am blown away at the image quality this camera and lens produces. I'm extremely thrilled to be producing those same sharp clean images that I would see online from night club, sports, and portrait photographers.

Comparisons between the 5dIII and t3i

1. The image quality is MUCH BETTER, SHARPER, AND CLEANER.
2. Better capability at low light shooting with higher ISO's.
3. Incredible autofocusing system with 61 AF points (41 of them are cross type) that I am still learning as I go. No more focus and recompose. Use the (orientation linked AF point) option in the auto focus menu and you'll see what I mean. (That's just one of many many useful features of the AF system)
4. The extra buttons at the top of the camera give you more flexibility at changing almost any setting at just a push of a button and a turn of either the top wheel with your index finger or the bottom wheel with your thumb. Changing some of those same settings on the t3i requires going a little further into the menu which takes a couple extra seconds to push a couple extra buttons decreasing your chances of getting that candid shot that you want to get in a hurry.

5. In-camera HDR. Helpful in properly exposing shadow areas when shooting towards the sun or other bright areas without over exposing the brighter areas.

6. You can rearrange and customize a number of buttons to suit your shooting style and needs.

7. You can calibrate (micro-adjust) any lens if needed.

8. Better selection of higher quality lenses.

9. Weather sealed.

I'm sure I'm missing a few more points that I can't think of right now. There are only three things that the t3i is better at than the 5diii... Smaller, lighter, and cheaper. Other than that, the 5diii ate the t3i for lunch and pooped it out by dinner time. Don't get me wrong I must emphasize that the Canon t3i was a great little starter camera and I have produced plenty of great images especially when paired with the Canon 50mm 1.8 but it was time to step up my game and start making some money on the side with this monster camera and lens.

Even though the 5d III price dropped about $800 around the beginning of 2015 it's still pretty expensive. I know it was crazy to spend about $1,100 more on the 5dIII vs the 6d (which has the same great image quality) just to get an incredible AF system and an extra storage slot (which I don't care for too much right now) and better ease of use of the custom functions, settings buttons layout that more than likely you will be changing frequently throughout... but I wanted to be prepared for any kind of photography event that comes my way. So there ya have it, My review.</comment>

</tag>

Nhận xét:

Dữ liệu được truy hồi từ trang Amazon.com phục vụ cho thực nghiệm chứa các nội dung sau:

1. Thông tin đánh giá sao: <infomation star>: được sử dụng cho đánh giá chuẩn vàng xếp hạng
2. Số người xem xếp hạng: <TopReview of ranking> được sử dụng cho đánh giá chuẩn vàng xếp hạng

3. Bình luận của khách hàng: <comment> được sử dụng để trích thông tin xếp hạng cho các đặc trưng để xếp hạng cho sản phẩm

3.2. Phương pháp

Thuật toán được thực hiện như sau:

1. Thu thập dữ liệu đánh giá của khách hàng theo định dạng như phần 3.1
2. Thực hiện tiền xử lý dữ liệu: tách từ, xóa bỏ các khoảng trống không cần thiết.
3. Sử dụng công cụ JFSA và CSRL để trích các cụm từ chứa quan điểm hoặc các so sánh quan điểm cho từng đặc trưng của sản phẩm.
4. Tính điểm và xếp thứ hạng cho các đặc trưng theo công thức (1),(2) cho JFSA và (3) cho CSRL.
5. Sử dụng công đánh giá của Speaman, 1980 để đo độ tương tự giữa các kết quả xếp hạng của các phương pháp.

3.3. Giới thiệu công cụ JFSA

JFSA là một phần mềm mã nguồn mở được phát triển bởi Roman Klinger, 2015 sử dụng để thực hiện các thực nghiệm với mô hình xác suất cho việc trích các đặc trưng và cụm từ chủ quan thể hiện các đánh giá tương ứng.

- **Cấu trúc của thư mục như sau:**

src/ bao gồm tất cả các file nguồn

bin/ bao gồm các kịch bản trợ giúp để biên dịch chương trình

3rdparty/ bao gồm ark-tweet-nlp-0.3.2.jar

data/ gồm các dữ liệu ví dụ, các ngữ liệu sử dụng và các từ điển được sử dụng trong mô hình

ini/ gồm các file khởi tạo

models/ gồm các mô hình đã được huấn luyện trước.

- Phần mềm được cài đặt trên hệ điều hành linux với Java 1.7 và Maven 2.0
- **Để cài đặt, chúng ta thực hiện các thao tác sau:**
  - Cài đặt ark-tweet trên thư mục Maven
    source bin/install-ark-tweet-nlp.sh
  - Biên dịch maven và tạo một file jar
    Kết quả : tạo ra một file jar:
    jfsa-0.1-jar-with-dependencies.jar
- **Dữ liệu:** Phần mềm này thực hiện trích đặc trưng và các cụm từ chứa quan điểm trên dữ liệu không gán nhãn.
  - **Dữ liệu vào**: là tệp .txt chứa dữ liệu đánh giá có cấu trúc như sau:

Cột đầu tiên: là số thứ tự (các bình luận)

Cột thứ 2: chưa sử dụng: dành cho các phát triển sau

Cột thứ 3: văn bản chứa dữ liệu đánh giá

- **Dữ liệu ra:**

Các đặc trưng và cụm từ chứa nhận xét tương ứng được chứa trong file .csv

Các mối quan hệ so sánh được chứa trong file .rel

- **Chạy hệ thống trên mô hình đã được huấn luyện trước:**

`java-Xmx2g-cptarget/jfsa-0.1.jar:target/jfsa-0.1-jar-with dependencies.jarsc.rk.targsubj.TargSubjSpanNERmodelfile.jfsainputdata.txt outputdata.txt

Hoặc: ./bin/run.sh modelfile.jfsa inputdata.txt outputdata.txt

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 dữ liệu người dùng đánh giá sản phẩm cho máy ảnh trên trang Amazon.com, mẫu dữ liệu quan điểm được crawl về từ trang này để phân tích thuật toán áp dụng trên dữ liệu đó.
- Chuẩn bị dữ liệu cho thực nghiệm
- Tìm hiểu sử dụng công cụ trích các đặc trưng và từ quan điểm tương ứng trong văn bản chứa nhận xét.

Chủ đề nghiên cứu của đồ án này là một lĩnh vực kiến thức mới hoàn toàn mới mà chúng em chưa được học. Do đó việc đọc tài liệu để tìm hiểu và phân tích đã giúp em hiểu biết thêm rất nhiều cho những bài toán có ý nghĩa trên thực tế. Do thời gian có hạn nên đề tài mới chỉ bước đầu phân tích dữ liệu và xác định thuật toán cho chương trình thực nghiệm. Trong thời gian tới, em sẽ tiếp tục phát triển đề tài, đánh giá kết quả thực nghiệm của phương pháp.

Trong quá trình thực hiện đề tài và trình bày nội dung đã tìm hiểu được chắc em không tránh khỏi có 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 !

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]. Đặng Thị Ngọc Thanh, Tìm hiểu về phương pháp trích và sắp xếp các đặc trưng sản phẩm trong tài liệu chứa quan điểm. Đồ án tốt nghiệp ngành **C**ông nghệ Thông tin, trường ĐHDL Hải Phòng, 2012.

[3]. Bing Liu, Sentiment Analysis Tutorial 2011.

[4]. Wiltrud Kessler and Jonas Kuhn. 2013. Detection of product comparisons - How far does an out-of-thebox semantic role labeling system take you? In EMNLP, pages 1892–1897. ACL

[5] Wiltrud Kessler, Roman Klinger, and Jonas Kuhn. 2015. Towards Opinion Mining from Reviews for the Prediction of Product Rankings. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. @ Association for Computational Linguistics 2015.

[6]. James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2):245–251.

[7]. https://java.com/en/download/chrome.jsp

[8]. http://maven.apache.org/download.cgi

[9]. https://bitbucket.org/rklinger/jfsa/downloads