MỤC LỤC

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 Thanh Cường

CHƯƠNG 1: BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM

1. 1 Nhu cầu về thông tin quan điểm và nhận xét

"Những gì người khác nghĩ" 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 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)

• Phân lớp quan điểm: Là bài toán khai thác các kỹ thuật để phân lớp các văn bản theo định hướng quan điểm (tích cực, tiêu cực hay trunglập).

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

1. Trích các đặc trưng đối tượng có nhận xét trong mỗi quan điểm.

2. 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)

3. Nhóm các cụm từ cùng nghĩa đặc trưng.

1. 4 Bài toán phân lớp quan điểm

Bài toán 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 Naive 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) theo định hướng ngữ nghĩa (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.

- 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 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: BÀI TOÁN TỰ ĐỘNG XÁC ĐỊNH CÁC ƯU, NHƯỢC ĐIỂM TRONG CÁC NHẬN XÉT ONLINE

2. 1 Tổng quan

Các tác giả giới thiệu hệ thống tự động trích các ưu nhược điểm từ các đánh giá trực tuyến. Mặc dù đã có nhiều phương pháp được phát triển để trích xuất các nhận xét từ văn bản, trong bài báo này các tác giả tập trung vào trích những lý do để đưa ra các nhận xét, mà chúng có thể là một trong hai hình thức thể hiện là thông tin thực tế hoặc quan điểm. Tận dụng các trang web xem trực tuyến với các ưu và nhược điểm đã được đưa ra trong đó, họ đề xuất một hệ thống cho việc gióng các ưu và nhược điểm với các câu trong các văn bản nhận xét. Họ sử dụng mô hình Maximum Entropy để huấn luyện tập kết quả đã gán nhãn cho các ưu, nhược điểm trích tuần tự từ các trang web nhận xét.

Kết quả thực nghiệm của họ cho thấy rằng hệ thống này xác định các ưu và nhược điểm với độ chính xác là 66% và thu hồi 76%.

2. 2. Giới Thiệu Bài Toán

Nhiều nhận xét được thể hiện trên các trang web trong các dạng như đánh giá sản phẩm, các blog cá nhân và các nhóm nhận tin phản hồi. Mọi người ngày càng gia tăng tham gia bày tỏ nhận xét của họ trên các phương tiện trực tuyến. Xu hướng này đã đưa ra nhiều đề tài nghiên cứu thú vị và đầy thử thách như phát hiện chủ quan, phân loại định hướng ngữ nghĩa, và phân loại đánh giá.

Phát hiện chủ quan là nhiệm vụ xác định các từ chủ quan, các giải thích, và câu. (Wiebe et al, 1999; Hatzivassiloglou và Wiebe, 2000; et al Riloff, 2003). Xác định chủ quan giúp phân tách các nhận xét từ các thông tin thực tế, nó có thể hữu ích trong trả lời câu hỏi, tóm tắt,..

Phân loại định hướng ngữ nghĩa là một nhiệm vụ của việc nhận xét là tích cực hay tiêu cực của các từ (Hatzivassiloglou và McKeown, 1997; Turney, 2002; Esuli và Sebastiani, 2005). Nhận xét của các cụm từ và các câu cũng đã được nghiên cứu (Kim và Hovy, 2004; Wilson et al. 2005). Phân loại nhận xét mức độ

tài liệu được thực hiện chủ yếu cho các đánh giá, nơi mà các hệ thống chỉ xác định một nhận xét là tích cực hay tiêu cực cho toàn bộ một tài liệu (Pang et al, 2002; Turney2002).

Trong công việc này, có rất nhiều vấn đề phức tạp hơn trong lĩnh vực quan điểm đã được nghiên cứu. (Bethard et al, 2004; Choi et al, 2005;. Kim và Hovy, 2006) đã xác định người đưa ra nhận xét của quan điểm được thể hiện trong các câu bằng cách sử dụng các kỹ thuật khác nhau Wilson và các cộng sự, 2004 tập trung vào sự nhấn mạnh của các mệnh đề quan điểm, tìm kiếm những quan điểm mạnh và yếu. Chklovski, 2006 giới thiệu một hệ thống tập hợp và định lượng mức độ đánh giá các nhận xét rải rác khắp các trang web.

Ngoài phân loại quan điểm ở mức độ tài liệu trong phần đánh giá sản phẩm trực tuyến, Hu và Liu, 2004. Popescu và Etzioni, 2005 tập trung vào khai thác và tóm tắt các nhận xét bằng cách trích xuất câu quan điểm liên quan đến đặc trưng sản phẩm.

Trong bài báo này, các tác giả tập trung vào một vấn đề đầy thách thức nhưng quan trọng của phân tích quan điểm, xác định lý do cho nhận xét, đặc biệt là đối với các nhận xét trong phần đánh giá sản phẩm trực tuyến. Bài toán xác định lý do nhận xét trong các đánh giá trực tuyến là tìm cách trả lời câu hỏi *"Các nguyên nhân nào mà tác giả của nhận xét là thích hay không thích sản phẩm?"*

Ví dụ, trong đánh giá của khách sạn, thông tin như "tìm thấy 189 nhận xét tích cực và 65 đánh giá tiêu cực" có thể không thỏa mãn đầy đủ các nhu cầu thông tin cho các người dùng khác nhau. Thông tin hữu ích hơn có thể là "khách sạn này là rất tốt cho các gia đình có trẻ sơ sinh".

Công việc này khác một cách quan trọng với các nghiên cứu (Hu và Liu, 2004) và (Popescu và Etzioni, 2005). Các phương pháp tiếp cận này trích xuất các đặc trưng của các sản phẩm và xác định các câu có nhận xét về những đặc trưng này bằng cách sử dụng các từ và cụm từ nhận xét. Ở đây, họ tập trung vào trích các ưu và nhược điểm trong đó bao gồm không phải là câu chỉ có chứa thể hiện nhận xét về các sản phẩm và các đặc trưng mà còn gồm các câu với những lý do tại sao một tác giả của một bài đánh giá viết nhận xét. Một số ví dụ xác định bởi hệ thống của họ:

It creates duplicate files.

Video drains battery.

It won'tplay music from all music stores

Mặc dù việc tìm kiếm lý do trong văn bản mang nhận xét là một phần quan trọng của việc đánh giá, không có một nghiên cứu nào đã được thực hiện cụ thể một phần vì không có dữ liệu được gán nhãn. Gán nhãn mỗi câu là một nhiệm vụ tốn thời gian và tốn kém. Trong bài báo này, họ đặt ra một framework để tự động xác định lý do trong các đánh giá trực tuyến và giới thiệu một kỹ thuật mới để tự động gắn nhãn dữ liệu huấn luyện cho nhiệm vụ này. Họ giả thuyết rằng trong một tài liệu đánh giá trực tuyến liên quan chặt chẽ đến các ưu và nhược điểm thể hiện trong văn bản. Họ tận dụng thực tế rằng trong các đánh giá trên một số trang web như epinions.com đã chứa các ưu và khuyết điểm bằng văn bản của cùng một tác giả như là các nhận xét. Họ sử dụng những ưu và khuyết điểm tự động gắn nhãn câu trong các nhận xét sau đó họ huấn luyện hệ thống phân loại. Sau đó áp dụng các hệ thống kết quả để trích xuất các ưu và nhược điểm từ các nhận xét trong các trang web khác mà không có các ưu và khuyết điểm xác định.

2. 3. Các ưu điểm và nhược điểm trong các nhận xét trực tuyến

Xem xét việc xác định một quan điểm trong các nghiên cứu của các tác giả về các tính toán ngôn ngữ, đây là một việc rất khó để định nghĩa thế nào là một quan điểm trong một mô hình tính toán vì khó có thể xác định đơn vị của một quan điểm. Nhìn chung, các nhà nghiên cứu phân tích quan điểm ở ba mức độ khác nhau là: mức từ, mức câu và mức tài liệu.

Phân tích quan điểm mức từ bao gồm phân loại nhận xét từ, là xem mục từ vựng duy nhất (ví dụ như tốt hay xấu) như là chứa nhận xét, cho phép phân loại các từ vào loại ngữ nghĩa tích cực và tiêu cực. Các nghiên cứu theo nhận xét mức câu coi câu như là một đơn vị nhỏ nhất của một nhận xét. Các nhà nghiên cứu cố gắng xác định câu mang nhận xét, phân loại nhận xét của chúng, và xác định các người đưa ra nhận xét và các chủ đề của các câu nhận xét. Phân tích nhận xét mức tài liệu được áp dụng chủ yếu để phân loại nhận xét, trong đó toàn bộ tài

liệu được viết cho một nhận xét được được đánh giá là chứa quan điểm là tích cực hay tiêu cực. Nhiều nhà nghiên cứu cho rằng xem xét toàn bộ tài liệu chứa quan điểm là quá thô.

Trong nghiên cứu của Kim và các cộng sự, họ đưa ra tiếp cách là một nhận xét có quan điểm chính (nhận xét hoặc không) về một sản phẩm nhất định, nhưng cũng bao gồm các lý do khác nhau cho các nhận xét hoặc không nhận xét, mà nó có giá trị để xác định. Vì vậy, họ tập trung vào việc phát hiện những lý do trong nhận xét sản phẩm trực tuyến. Họ cũng giả thuyết rằng các lý do trong bài đánh giá liên quan chặt chẽ đến ưu và khuyết điểm thể hiện trong nhận xét. Các ưu điểm trong một đánh giá sản phẩm là những lý do mô tả tại sao một tác giả của nhận xét thích sản phẩm. Các nhược điểm là lý do tại sao tác giả không thích sản phẩm. Dựa trên quan sát của họ trong các đánh giá trực tuyến, hầu hết các đánh giá có cả ưu và khuyết điểm ngay cả khi đôi khi một trong số chúng chiếm ưu thế.

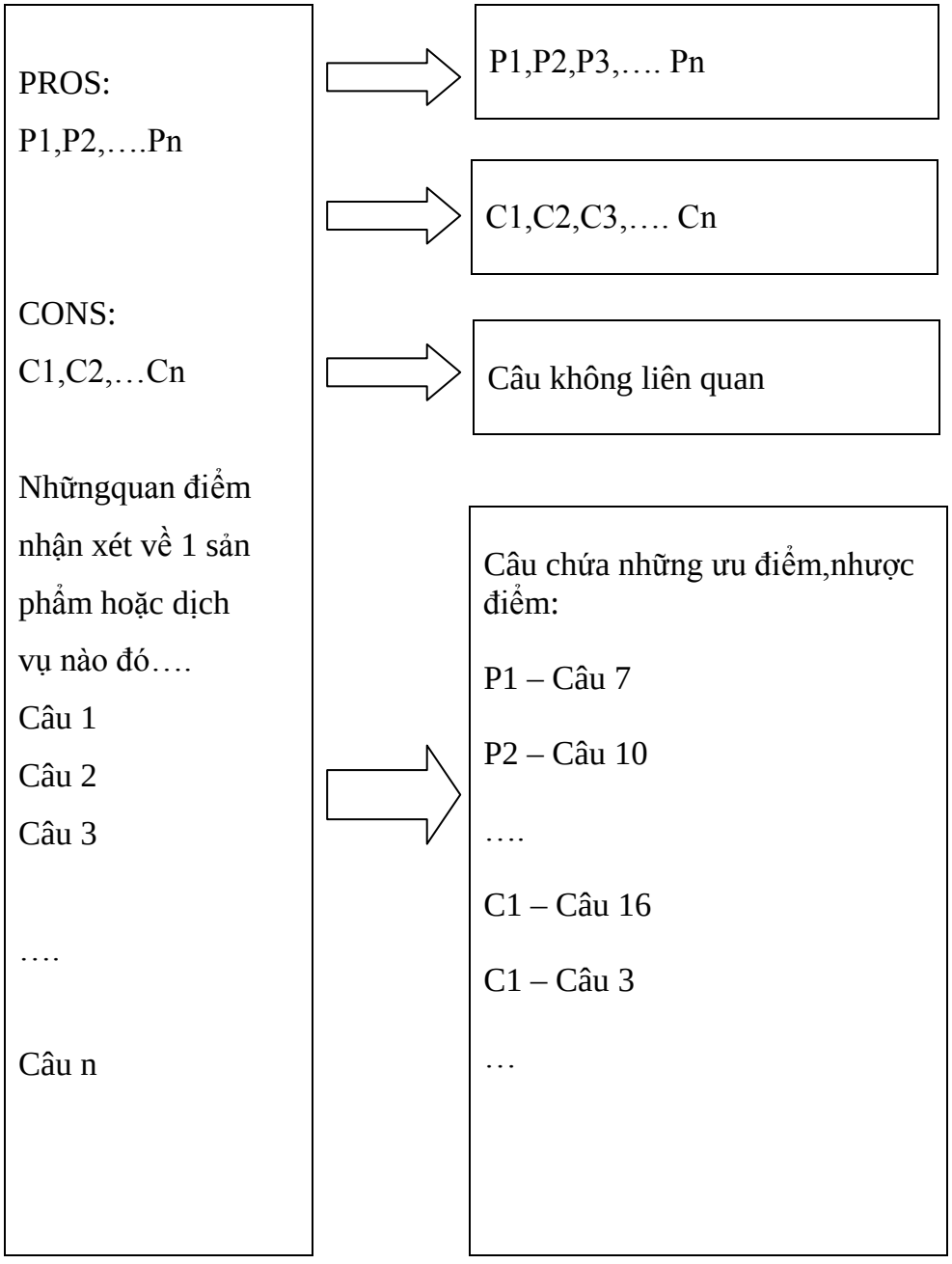
2. 4. Tìm kiếm ưu và nhược điểm

Phần này mô tả cách tiếp cận của của Kim và Hovy cho việc tìm kiếm các câu ưu và nhược điểm đã cho trong một văn bản xem xét. Trước tiên họ thu thập dữ liệu từ epinions.com và tự động gắn nhãn mỗi câu trong tập dữ liệu. Sau đó họ mô hình hệ thống sử dụng một trong những kỹ thuật học máy đã được áp dụng thành công cho các vấn đề khác nhau trong xử lý ngôn ngữ tự nhiên.

2. 4. 1 Tự động gán nhãn câu ưu điểm và nhược điểm

Trong một số trang web có đánh giá sản phẩm như amazon. com và epinions.com, một số trong đó (ví dụ như epinions.com) đưa các cụm từ thể hiện trực tiếp ưu và nhược điểm trong các tương ứng của nhận xét của mỗi tác giả cùng với các văn bản nhận xét. Đầu tiên, họ thu thập một tập lớn các bộ ba: <văn bản nhận xét, các ưu điểm, các nhược điểm> từ epinions.com. Một tài liệu đánh giá trong epinions.com bao gồm một chủ đề (một mô hình sản phẩm, tên nhà hàng, điểm đến du lịch, vv...), ưu và nhược điểm (phần lớn là một vài từ khóa nhưng đôi

khi là câu hoàn chỉnh), và văn bản nhận xét. Hệ thống gán nhãn tự động của họ trước tiên là thu thập các cụm từ trong trường ưu điểm và nhược điểm và sau đó tìm kiếm các văn bản đánh giá chính để thu thập các câu tương ứng với những cụm từ.



Một tài liệu nhận xét quá trình gán nhãn

Quá trình gán nhãn tự động các câu ưu và nhược điểm trong một nhận xét.

Hệ thống trích xuất đầu tiên các cụm từ phân cách bởi dấu phẩy từ mỗi trường ưu điểm và nhược điểm, tạo ra hai tập các cụm từ: {P1, P2,..., Pn} cho ưu điểm và {C1, C2,..., Cm} cho khuyết điểm.

Ví dụ trong hình 1: “beautiful display "có thể là P_i và" not something you want to drop"có thể là C_j.

Sau đó, hệ thống so sánh những cụm từ với các câu trong văn bản trong toàn bộ nhận xét. Đối với mỗi cụm từ trong {P_1, P_2,..., P_n} và {C_1, C_2,..., C_m}, hệ thống kiểm tra từng câu để tìm một câu mà bao trùm hầu hết các từ trong cụm từ. Sau đó, hệ thống gán nhãn câu này với nhãn "pro" hoặc "con" thích hợp. Tất cả các câu còn lại được gán nhãn không được đánh dấu là "neither". Sau khi gán nhãn tất cả các dữ liệu epinion, họ sử dụng nó để huấn luyện hệ thống nhận dạng câu.

2. 4. 2 Mô hình hóa với phân loại Maximum Entropy

Kim và Hovy sử dụng phân loại Maximum Entropy cho nhiệm vụ tìm kiếm câu ưu điểm và nhược điểm trong một đánh giá nào đó. Phân loại Maximum Entropy đã được áp dụng thành công trong nhiều nhiệm vụ trong xử lý ngôn ngữ tự nhiên, chẳng hạn như vai trò gán nhãn ngữ nghĩa, trả lời câu hỏi, và trích rút thông tin.

Class symbol	Description
PR	Sentences related to pros in a review
CR	Sentences related to cons in a review
NR	Sentences related to neither PR nor CR

Bảng 1: Các lớp được định nghĩa cho các nhiệm vụ phân lớp

Các mô hình Maximum Entropy là mô hình tốt nhất là phù hợp với tập các ràng buộc bắt buộc bởi bằng chứng nhưng dạng không thống nhất có thể (Berger et al, 1996). Họ mô hình xác suất có điều kiện của một lớp c cho một vector đặc trưng x như sau:

$$P(c|x) = \frac{1}{Z_x \exp\left(\sum_i f_i(c,x)\right)}$$

Trong đó: Z_x là nhân tố chuẩn hóa được tính như sau:

$$Z_x = \sum_c \exp\left(\sum_i \alpha_i f_i\,(c, x)\right)$$

Trong công thức đầu tiên, $f_i(c, x)$ là một hàm đặc trưng trong đó có một giá trị nhị phân 0 hoặc 1. α_i là một tham số trọng số trong hàm chức năng $f_i(c, x)$ và giá trị lớn hơn của trọng số cho thấy $f_i(c, x)$ là một đặc trưng quan trọng trong lớp c. Đối với phát triển hệ thống, họ sử dụng MegaM toolkit để thực hiện phân lớp trên.

Để xây dựng một mô hình hiệu quả, họ chia nhiệm vụ của việc tìm kiếm câu ưu và nhược điểm thành hai giai đoạn, mỗi giai đoạn là một phân lớp nhị phân. Đầu tiên là giai đoạn xác định cụm từ và thứ hai là giai đoạn phân loại. Đối với mô hình 2 giai đoạn này, họ định nghĩa một lớp bộ ba của c được liệt kê trong Bảng 1. Nhiệm vụ xác định tách các câu ứng cử là ưu điểm và nhược điểm (CR và PR trong Bảng 1) từ các câu không liên quan đến một trong số chúng (NR). Nhiệm vụ phân loại sau đó phân loại các ứng cử viên vào ưu (PR) và nhược điểm (CR).

2. 4. 3 Các đặc trưng

Phân lớp sử dụng ba loại của các đặc trưng: các đặc trưng từ vựng, các đặc trưng vị trí, và các đặc trưng từ chứa quan điểm.

Đối với các đặc trưng từ vựng, họ sử dụng unigrams, bigrams, và trigrams thu thập được từ tập huấn luyện. Họ phân tích trực giác rằng có một số từ thường được sử dụng trong các câu ưu điểm và nhược điểm, chúng có khả năng đại diện cho lý do tại sao người dùng viết trong nhận xét. Ví dụ về các từ và cụm từ đó là: *"because"* và *"that's why"*.

Đối với các đặc trưng vị trí, đầu tiên họ tìm thấy ranh giới đoạn trong văn bản nhận xét bằng cách sử dụng các thẻ html như
 và <p>. Sau khi tìm thấy ranh giới đoạn, họ thêm các đặc trưng cho biết câu đầu tiên, thứ hai, cuối cùng, và câu cuối cùng thứ hai trong một đoạn. Những đặc trưng kiểm tra trực giác được sử dụng trong tóm tắt tài liệu mà các câu quan trọng có chứa các chủ đề trong một văn bản có mẫu vị trí nhất định trong một đoạn (Lin và Hovy, 1997), nó có thể áp dụng vì các lý do như ưu và nhược điểm trong một tài liệu nhận xét là các câu quan trọng nhất để tóm tắt toàn bộ các điểm của một nhận xét.

Đối với các đặc trưng từ chứa nhận xét, họ đã sử dụng các từ chứa quan điểm đã được chọn trước được tạo ra bởi sự kết hợp của hai phương pháp. Phương pháp đầu tiên thu được một danh sách các từ chứa nhận xét từ một ngữ liệu tin tức lớn bằng cách tách các bài viết nhận xét như thư từ, bài xã luận từ các bài báo mà là các tin tức đơn giản hoặc các sự kiện. Phương pháp thứ hai được tính định hướng ngữ nghĩa của từ dựa trên tập các từ đồng nghĩa WordNet2. Trong nghiên cứu của Kim và Hovy, 2005, họ đã chứng minh rằng danh sách các từ được tạo ra bởi sự kết hợp của hai phương pháp thực hiện rất tốt trong việc phát hiện các câu chứa nhận xét.

Động cơ để tạo danh sách các từ chứa nhận xét là một trong các đặc trưng của họ là các câu ưu và nhược điểm hoàn toàn có thể chứa các biểu thức chứa nhận xét(thậm chí một số trong đó là câu thực tế), chẳng hạn như *"The waiting time was horrible"* và *"Their portion size of food was extremely generous!"* trong nhận xét nhà hàng. Họ coi các câu ưu điểm và nhược điểm chỉ chứa các sự kiện, chẳng hạn như *"The battery lasted 3 hours, not 5 hours like they advertised"*, sẽ được bắt bởi các đặc trưng từ vựng hoặc vị trí.

Loại đặc trưng (Feature category)	**Mô tả (Description)**	**Ký hiệu (Symbol)**
Các đặc trưng từ vựng (Lexical Features)	Unigram Bigram Trigram	**Lex**
Các đặc trưng vị trí (Positional Features)	Thefirst, the second,the last,the second to last sentence in a paragraph	**Pos**
Các đặc trưng từ chứa quan điểm (Opinion-bearingword features)	Pre-selected opinion-bearing words	**Op**

Bảng 2: Bảng tóm tắt các đặc trưng

Bảng trên tóm tắt các đặc trưng mà họ sử dụng cho các mô hình và các biểu tượng của chúng ta sẽ sử dụng trong phần còn lại của bài báo này.

2.5. Dữ Liệu

Họ thu thập dữ liệu từ hai nguồn khác nhau:

+ www.epinions.com

+ www.complaints.com

Dữ liệu từ epinions.com chủ yếu được sử dụng để huấn luyện hệ thống trong khi dữ liệu từ complaints.com là để kiểm tra các mô hình đã huấn luyện thực hiện trên dữ liệu mới.

Complaints.com bao gồm một cơ sở dữ liệu lớn các nhận xét của người tiêu dùng về đa dạng các sản phẩm, dịch vụ, và các công ty được thu thập hơn 6 năm. Đánh giá trong complaints.com là hơi khác so với các trang web khác mà chúng được liên kết trực tiếp hoặc gián tiếp với các trung tâm mua sắm trên mạng như amazon.com và epinions.com. Mục đích của việc đánh giá trong complaints.com là để chia sẻ kinh nghiệm của người tiêu dùng chủ yếu là tiêu cực và cảnh báo các doanh nghiệp từ phản hồi của các khách hàng. Tuy nhiên, một số nhận xét trong trung tâm mua sắm trực tuyến nhận xét liên quan là tích cực và đôi khi khuyến khích mọi người mua các sản phẩm hoặc sử dụng các dịch vụ nhiều hơn nữa. Mặc dù nó ý nghĩa nhưng tuy nhiên không có dữ liệu được gán nhãn bằng tay để chúng ta có thể sử dụng để xây dựng một hệ thống để xác định nguyên nhân của complaints.com. Để giải quyết vấn đề này, họgiả thuyết rằng lý do trong các nhận xét đánh giá tương tự như các khuyết điểm trong các nhận xét đánh giá khác và do đó nếu chúng ta, bằng cách nào đó có thể để xây dựng một hệ thống có thể xác định các khuyết điểm từ các nhận xét, chúng ta có thể áp dụng nó để xác định lý do trong các đánh giá nhận xét. Dựa trên giả thuyết này, họ học một hệ thống sử dụng các dữ liệu từ epinions.com, mà để có thể áp dụng kỹ thuật gán nhãn dữ liệu tự động của họ, và sử dụng hệ thống kết quả để xác định lý do từ các đánh giá trong complaint.com.

2. 5.1 Tập dữ liệu 1: Tự động gắn nhãn dữ liệu

Họ thu thập hai lĩnh vực nhận xét khác nhau từ epinions.com: các đánh giá sản phẩm và các đánh giá nhà hàng. Đối với các đánh giá sản phẩm, họ thu thập 3241 nhận xét (115. 029 câu) về máy nghe nhạc mp3 được thực hiện bởi các nhà sản xuất khác nhau, chẳng hạn như Apple, iRiver, Creative Lab, và Samsung. Họ cũng thu thập 7524 nhận xét (194. 393 câu) về các loại khác nhau của nhà hàng như. Family restaurants, Mexican restaurants, fast food chains, steak houses, and Asian restaurants. Các con số trung bình của các câu trong một tài liệu nhận xét là 35, 49 và 25, 89.

Mục đích của việc lựa chọn một trong các sản phẩm điện tử và các nhà hàng là các chủ đề nhận xét cho nghiên cứu của họ là để thử nghiệm phương pháp tiếp cận của họ trong hai tình huống rất khác nhau. Lý do tại sao người tiêu dùng thích hay không thích một sản phẩm trong đánh giá các thiết bị điện tử chủ yếu là về các đặc trưng cụ thể và hữu hình. Ngoài ra, có phần nào một tập cố định các đặc trưng của một dạng cụ thể của sản phẩm, ví dụ: dễ sử dụng, độ bền, pin, chất lượng hình ảnh và độ chớp cho các máy ảnh kỹ thuật số. Do đó, chúng ta có thể mong đợi nguyên nhân trong đánh giá của thiết bị điện tử có thể chia sẻ những từ đặc trưng sản phẩm và các từ đó mô tả các khía cạnh của các đặc trưng như: *"short"* hoặc *"long"* cho *"battery life"*. Thực tế này có thể làm cho các nhiệm vụ xác định lý do một cách dễ dàng.

Mặt khác, các nhận xét nhà hàng nói về khía cạnh rất đa dạng và các đặc trưng trừu tượng như là các lý do. Ví dụ, lý do như: *"You feel like you are in a train station or a busy amusement park that is ill-staffed to meet demand!"*, *"preferential treatment given to large groups"*, và *"they don't offer salads of any kind"* là khó để dự đoán.

Họ lần đầu tiên tự động gán nhãn từng câu trong các đánh giá thu thập được từ mỗi miền với các đặc trưng được mô tả như mục tự động gán nhãn ưu điểm và nhược điểm cho các câu. Họ chia các dữ liệu thành dữ liệu huấn luyện và thử nghiệm.

Sau đó, Họ huấn luyện mô hình của họ bằng cách sử dụng tập huấn luyện và thử nghiệm nó để xem nếu hệ thống có thể gán nhãn thành công câu trong tập thử nghiệm.

2.5. 2. Tập dữ liệu 2: Dữ liệu Complaints.com

Từ cơ sở dữ liệu trong complaints.com, họ tìm kiếm các chủ đề cùng nhận xét như tập dữ liệu 1: 59 khiếu nại các nhận xét phản hồi về các máy nghe nhạc mp3 và 322 đánh giá về nhà hàng. Họ đã thử nghiệm hệ thống trên tập dữ liệu này và so sánh kết quả với kết quả xác định chú thích của con người.

2.6. Kết quả và thực nghiệm

Họ mô tả hai mục đích của thực nghiệm. Trước tiên là phân tích làm thế nào để mô hình phát hiện tốt các ưu, nhược điểm với việc kết hợp các đặc trưng khác nhau thực hiện trên dữ liệu thu thập được từ epinions.com. Thứ 2 là xem xét làm thế nào mô hình huấn luyện tốt nhất thực hiện trên dữ liệu mới từ một nguồn khác, complaint.com.

Cho cả hai tập dữ liệu, họ thực hiện trên cả hai phần của thực nghiệm, cho cả hai miền các nhận xét cho máy nghe nhạc mp3 và nhà hàng. Họ phân chia 80 % dữ liệu cho huấn luyện, 10 % cho phát triển và 10% cho đánh giá kết quả.

2.7. Nghiên cứu của Kim và Hovy để tự động phát hiện các câu và các từ chứa quan điểm

Xác định chủ quan mức câu. Wilson và Wiebe (2003) phát triển một lược đồ gán nhãn chú thích cho các câu chủ quan. Họ tạo ra ngữ liệu MQPA, bao gồm các bài báo được gán nhãn bằng tay. Một số tiếp cận khác được áp dụng để học các từ và các cụm từ có dấu hiệu chủ quan. Turney (2002) và Wiebe (2000) tập trung vào học các cụm tính từ và trạng từ, Wiebe và các cộng sự (2001) tập trung vào các danh từ. Riloff và các cộng sự (2003) đã trích các danh từ và Riloff và Wiebe (2003) đã trích các mẫu cho các thể hiện chủ quan sử dụng quá trình học tăng cường.

2.7.1 Thu thập các nguồn dữ liệu

Họ phát triển một số thu thập của các từ học không chứa quan điểm và chứa quan điểm. Họ kết hợp chúng để đạt được danh sách các từ tin cậy hơn. Họ đạt đươc một danh sách các từ thêm vào từ đại học Columbia.

2.7.1.1 Thu thập 1: sử dụng WordNet.

Trước tiên, họ thu thập bằng tay một tập các từ chứa quan điểm (34 tính từ và 34 động từ). Phân lớp đầu tiên được đưa ra mà độ chính xác rất cao (hệ thống chỉ tìm các câu chứa quan điểm), nhưng khi danh sách các từ quá nhỏ, độ hồi tưởng lại rất nhỏ (nó bị mất một số). Do đó, họ sử dụng một danh sách các từ được mở rộng bằng cách sử dụng WordNet. Giả thuyết của họ là các từ đồng nghĩa và trái nghĩa của một từ chứa quan điểm có thể là từ quan điểm tốt, ví du như: "nice, virtuous, pleasing, well-behaved, gracious, honorable, righteous" là các từ đồng nghĩa cho “good” hoặc "bad, evil, disreputable, unrighteous" là các từ trái nghĩa. Tuy nhiên, không phái tất cả các từ đồng nghĩa và trái nghĩa có thể sử dụng được: một số từ này dường như có mặt trong cả ngữ cảnh chứa quan điểm và không chứa quan điểm, như "solid, hot, full, ample" cho "good". Điều này cho thấy cần thiết cho một mức độ của độ lớn giá trị. Nếu chúng ta có thể đo độ :”gần nhất chứa quan điểm” của các từ đồng nghĩa và trái nghĩa để nhận biết quan điểm, sau đó có thể xác định có hay không trong tập được mở rộng. Để tính toán tỷ lệ, họ tạo ra một danh sách từ chứa quan điểm bằng tay và tạo ra các từ liên quan cho chúng sử dụng WordNet. Để tránh thu thập các từ không phổ biến, họ bắt đầu với danh sách từ cơ bản và phổ biến cho các sinh viên nước ngoài chuẩn bị cho thi TOEFL. Từ đó, họ lấy ngẫu nhiên 462 tính từ và 502 động từ cho gán nhãn bằng tay. Human1 và human2 được gán nhãn 462 tính từ và human3, human4 được gán nhãn 502 động từ, gán nhãn mỗi từ là từ chứa quan điểm hoặc không chứa quan điểm của từ khác.

Để đạt được độ đo của độ nhấn mạnh quan điểm/không quan điểm, họ đo khảng cách WordNet của từ đích (đồng nghĩa hoặc trái nghĩa) với hai tập các từ giống được chọn bằng tay cộng với các từ mở rộng hiện tại (hình 1). Họ xác định một từ mới vào phân loại gần hơn.

Công thức cho tiếp cận này như sau:

$$arg_c \, max \, P(c|w) \cong arg_c \max P(c|syn_1, syn_2 \ldots syn_n) \quad (1)$$

Trong đó: *c* là một bình luận (chứa quan điểm hoặc không chứa quan điểm)

w là từ đích.

Syn_n là các từ đồng nghĩa hoặc trái nghĩa của từ đã cho bởi WordNet.

Để tính toán công thức (1), họ xây dựng mô hình phân loại, công thức (2):

$$arg_c \max P(c|w) = arg_c \max P(c)P(w|c)$$
$$= arg_c \max P(c)P(syn_1, syn_2 \ldots syn_n|c)$$
$$= arg_c max P(c) \prod_{k=1}^{m} P(f_k|c)^{count(f_k synset(w))}$$

Trong đó f_k là đặc trưng của c, nó cũng là thành viên của tập các từ mục tiêu *w*.

Count(f_k, synset(w)) là tổng tất cả sự xuất hiện của *fk* trong tập các từ đồng nghĩa của w.

Mục đích của mô hình này là phân loại tài liệu. (Mặc dù họ sử dụng tập từ đồng nghĩa của các từ giống thu được từ WordNet, họ có thể thay thế từ các đặc trưng thu được từ một ngữ liệu.) Sau khi mở rộng, họ đạt được 2682 tính từ chứa quan điểm và 2548 tính từ không chứa quan điểm, 1329 động từ chứa quan điểm và 1760 động từ không chứa quan điểm, với các giá trị nhấn mạnh. Bằng cách sử dụng các từ như là các đặc trưng, họ đã xây dựng phân lớp Naive bayesian và phân lớp được 32373 từ.

2.7.1.2 Thu thập 2: Dữ liệu WSJ

Các Thực nghiệm với tập ở trên đã không cho các kết quả khả quan trên một văn bản tùy ý. Vì một lý do là các kết nối từ đồng nghĩa của từ điển WordNet đơn giản là không đủ mở rộng. Tuy nhiên, nếu chúng ta biết tần suất tương đối của một từ trong các văn bản chứa quan điểm so với văn bản không chứa quan điểm, chúng ta có thể sử dụng thông tin thống kê thay vì thông tin từ vựng. Đối với điều này, họ đã thu thập được một số lượng lớn dữ liệu để bù cho những hạn chế của bộ sưu tập 1.

Theo quan điểm của Yu và Hatzivassi-loglou (2003), họ thiết lập giả định cơ bản và thô là các từ mà xuất hiện thường xuyên hơn trong các bài xã luận báo chí và thư cho người biên tập hơn trong các bài báo không biên tập có thể là các từ tiềm năng chứa quan điểm (mặc dù bài xã luận bao gồm các câu sự kiện thực tế). Họ sử dụng bộ sưu tập TREC để thu thập dữ liệu, trích rút và phân loại tất cả các tài liệu Wall Street Journal từ nó hoặc là Editorial hoặc nin-Editorial dựa trên sự xuất hiện của các từ khóa "Letters to Editor" "Letter to Editor", hoặc "Editor" hiện diện trong tiêu đề của nó. Việc này tạo ra tổng số 7053 tài liệu biện tập và 166. 025 tài liệu không biên tập.

Họ tách ra các từ quan điểm từ những từ không có quan điểm bằng cách xem xét tần suất liên quan của chúng trong hai bộ sưu tập, dưới dạng xác suất, bằng cách sử dụng SRILM, bộ công cụ ngôn ngữ mô hình SRI. Với mỗi từ *W* xuất hiện ở một trong những bộ tài liệu, họ tính toán như sau:

$$EditorialProb(W) = \frac{\#WinEditorialdocuments}{totalwordsinEditorialdocuments}$$

$$nonEditorialPro(W) = \frac{\#WinnonEditorialdocs}{totalwordsinnonEditorialdocs}$$

Họ sử dụng Kneser-Ney làm mịn (Kneser và Ney, 1995) để xử lý các từ chưa biết / hiếm. Để có xác suất trên họ tính toán số điểm của *W* như tỷ lệ sau đây:

$$Score(W) = \frac{EditorialProb(W)}{nonEditorialProb(W)}$$

Score(W) dấu hiệu xu hướng của mỗi từ đối với văn bản biên tập hoặc không biên tập. Họ tính toán các điểm cho 86. 674. 738 từ tố. Đương nhiên, các từ với số điểm gần 1 là không đáng tin cậy. Để loại bỏ các từ này, họ áp dụng một bộ lọc đơn giản như sau: họ chia mỗi bộ sưu tập Editorial và non-Editorial thành 3 tập con. Với mỗi từ trong một cặp con {Editorial, non-Editorial} họ tính toán điểm (W). Họ chỉ giữ lại những từ mà các điểm trong tất cả 3 cặp tập hợp con đều lớn hơn 1 hoặc nhỏ hơn 1. Nói cách khác, họ chỉ giữ các từ lặp đi lặp lại với xu hướng lặp lại theo Editorial hoặc non-Editorial. Thủ tục này đã giúp loại bỏ một số các từ không cần thiết, trả về là 15. 568 từ.

2.7.1.3 Thu thập 3: với Columbia Wordlist

Phân đoạn đơn giản các bài báo WSJ vào viết vào Editorial/non-Editorial là một sự khác biệt rất rõ ràng. Để so sánh hiệu quả của việc thực hiện của họ về ý tưởng này với việc thực hiện của Yu và Hatzivassiloglou của Đại học Columbia, họ truy vấn danh sách từ của họ. Danh sách này chứa 167. 020 tính từ, 72. 352 động từ, 168. 614 danh từ, và 9884 trạng từ. Tuy nhiên, con số này đang tăng cao đáng kể do đếm dự phòng của các từ với các biến thể trong vốn từ và một dấu chấm câu. Họ trộn danh sách này và có được bộ sưu tập 4. Trong số những từ này, họ chỉ lấy 2000 từ chứa quan điểm và 2000 từ không chứa quan điểm cho danh sách từ cuối cùng.

2.7.1.4 Thu thập 4: Trộn dữ liệu cuối cùng

Cho đến nay, họ đã phân loại các từ hoặc là chứa quan điểm hoặc không chứa quan điểm bằng hai phương pháp khác nhau. Phương pháp đầu tiên tính toán mức độ gần với các tập từ chứa quan điểm và không chứa quan điểm được chọn bằng tay trong từ điển WordNet và quyết định lớp và độ chắc chắn của chúng. Khi từ có mức độ gần bằng nhau cho cả hai lớp thì sẽ khó khăn để quyết tính định chủ

quan của nó, và khi WordNet không chứa một từ hoặc các từ đồng nghĩa của nó, chẳng hạn như từ "*antihomosexsual*", họ không phân loại nó.

Phương pháp thứ hai, phân loại các từ sử dụng các văn bản WSJ, là ít đáng tin cậy hơn so với các phương pháp từ vựng. Tuy nhiên, nó xử lý thành công cho ví dụ "antihomosexual". Vì vậy, họ kết hợp các kết quả của hai phương pháp (collections 1 và 2), khi các đặc tính khác nhau của chúng bù đắp cho nhau. Sau đó họ cũng kết hợp 4000 từ từ danh sách từ Columbia để cho danh sách từ cuối cùng là 43700. Khi tất cả ba danh sách bao chứa mức độ giữa 0 và 1, họ lấy trung bình cộng chúng, và bình thường hóa mức độ khoảng từ -1 đến +1, với giá trị quan điểm lớn hơn gần với 1. Các từ có một giá trị hấp dẫn cao trong tất cả ba bộ sưu tập có một mức độ tích cực toàn bộ cao nhất. Khi có một cuộc bỏ phiếu xung đột trong số ba cho một từ, nó sẽ tự động suy yếu.

CHƯƠNG 3: THỰC NGHIỆM

3.1 Công cụ và ngôn ngữ lập trình

3.1.1 Ngôn ngữ JAVA

Có thể nói rằng Java là một ngôn ngữ lập trình mạnh đang được sử dụng rất rộng rãi hiện nay trên toàn thế giới. Trên thực tế, Java được biết đến không chỉ là một ngôn ngữ lập trình mà còn là một platform một môi trường và công nghệ phát triển riêng biệt. Khi làm việc với Java, người lập trình được sở hữu một thư viện lớn, có tính mở với một lượng mã nguồn tái sử dụng khổng lồ luôn có trên internet. Ngoài ra, các chương trình viết bằng Java có môi trường thực thi riêng với các tính năng bảo mật, khả năng triển khai trên nhiều hệ điều hành khác nhau.

Java là một ngôn ngữ lập trình hướng đối tượng (OOP). Khác với phần lớn ngôn ngữ lập trình thông thường, thay vì biên dịch mã nguồn thành mã máy hoặc thông dịch mã nguồn khi chạy, Java được thiết kế để biên dịch mã nguồn thành bytecode, bytecode sau đó sẽ được môi trường thực thi (runtime environment) chạy. Bằng cách này, Java thường chạy nhanh hơn những ngôn ngữ lập trình thông dịch khác như Python, Perl, PHP,… Java được phát triển từ C ++ và C++ là hậu duệ trực tiếp của C, đo đó Java kế thừa cú pháp của C và tính hướng đối tượng của C++ nhưng có cú pháp hướng đối tượng đơn giản hơn và ít tính năng xử lý cấp thấp hơn.

3.1.2 Bộ công cụ NetBeans IDE 7

NetBeans IDE là một "môi trường phát triển tích hợp" (Integrated Development Environment) kiểu như Visual Studio của Microsoft và được xem là một bộ ứng dụng cần phải có dành cho các nhà phát triển phần mềm.

NetBeans IDE hỗ trợ nhiều hệ điều hành khác nhau như Windows, Mac, Linux, và Solaris. NetBean bao gồm một IED mã nguồn mở và một nền tảng ứng dụng cho phép nhà phát triển nhanh chóng tạo nên các ứng dụng dành cho web, doanh nghiệp,desktop và thiết bị di động bằng các ngôn ngữ lập trình Java, C/C++, JavaScript, Ruby,Groovy,và PHP.

3. 2 Chương trình thực nghiệm

3. 2. 1 Bài toán

Bài toán tự động xác định ưu điểm và nhược điểm của các nhận xét online được thực hiện gồm có hai pha làm việc như sau:

1. **Khai thác dữ liệu từ các trang Web chứa các bình luận có dạng: Pros, Cons, và quan điểm chi tiết: (Epinion.com)**

   Ví dụ:

   Pros: Great photos easy to use, very small

   Cons: Battery usage: included memory is stingy

I had never used a digital camera prior to purchasing have always used a SLR…

Thực hiện quá trình gán nhãn một cách tự động cho các câu là ưu điểm, nhược điểm dựa vào các tóm tắt ưu, nhược điểm ở mỗi bình luận.

Các dữ liệu được gán nhãn này sẽ được sử dụng làm dữ liệu huấn luyện để áp dụng các phương pháp phân lớp quan điểm cho các bình luận không có dạng ưu và nhược điểm như trên.

Phần lớn các phương pháp được sử dụng để phân lớp quan điểm đều sử dụng trên dữ liệu đã gán nhãn và được thực hiện bằng tay. Do đó, chi phí để thực hiện cho xây dựng các ngữ liệu huấn luyện này là đắt đỏ và tốn kém.

Việc xây dựng các phương pháp gán nhãn tự động để tạo ra bộ ngữ liệu có ý nghĩa quan trọng mang lại lợi ích cho trong cả nghiên cứu lý thuyết lẫn lợi ích kinh tế. Dữ liệu được khai thác tự động sẽ giúp ích cho việc khai thác thác các thông tin quan điểm đa dạng và phong phú từ đó phát triển các ứng dụng trong thực tế đặt ra.

2. Sử dụng các phương pháp phân lớp để xác định các ưu và nhược điểm trên các nhận xét online dựa trên dữ liệu huấn luyện đã thu thập được.

Trong khuôn khổ của đồ án này, chúng tôi thực hiện việc gán nhãn tự động các câu trong một bình luận tương ứng với các ưu điểm và nhược điểm về một sản phẩm hoặc dịch vụ nào đó. Dữ liệu này sau đó sẽ được sử dụng để huấn luyện phân lớp các câu chứa ưu, nhược điểm phục vụ cho việc xác định thông tin về một sản phẩm hoặc dịch vụ nào đó có phải là xu hướng của người dùng hay không.

Input: Các bình luận được thu thập từ các trang Web.

Output : Các câu trong bình luận được gán nhãn tương ứng với các ưu điểm,nhược điểm được tóm tắt ở đầu mỗi bình luận.

3. 2. 2. Bộ dữ liệu

Chúng tôi thu thập 50 bình luận về thế hệ điện thoại thông minh từ trang http: //www. epinions.com để làm dữ liệu đầu vào cho bài toán.

Một bình luận có dạng:

= No1 =
Pros : good lookc slim, usable, fast Internet services, good picture quality.
Cons : slippery, typing problem

My papa brought me Samsung Galaxy S II Smart phone on my birthday.
It is really useful and fantastic phone to use.
I can browse the Internet through this and can take pictures, videos.I like to watch videos in my phone.
I can easily find my nearest restaurants and other places through it.
it has high picture quality and videos.
I like to play games on it.
The games are interesting and easily down loadable.
I have taken my baby pictures with this product.
The photos are great.
The smart phone is usable for my husband for checking mails for job purpose.
Thin and light smartphones are easy to carry around and easier on the hand with extended use.
Since it is based on Android 4.0 we already know that it has a solid base.
Samsung flexed its software chops with the S III in a way that sets the phone apart from its competitors.
Samsung did some really interesting things with the camera.
The curved bump on the phone's bottom can make it hard to hold, and the battery life needs improvement.
The faster the processor, the smoother a phone's user interface, and the quicker you can browse the web, run apps and stream media.
s front-facing camera is great for video chatting, and the 8-megapixel primary camera on the back snaps high-quality photos – an LED flash helps.
The device's total storage capacity is an impressive 48GB, with 16GB installed and up to an additional 32GB through an external microSD card.
The larger screen allowed me to adjust from a hard qwerty keyboard to the soft one on screen with ease.

The display provides great viewing and the ability to join emails and other social accounts is the best.
Once I figure out how to make the Voice Actions work I'm sure it will be a total plus for me.

Chúng tôi tiến hành gán nhãn các câu trong 50 bình luận tương ứng với các ưu điểm, nhược điểm đã được liệt kê bằng tay để dùng cho đánh giá dữ liệu.

Một bình luận được gán nhãn có dạng sau:

Pros:P1=good look,P2= slim,P3= usable, P4=fast Internet services,P5= good picture quality.
Cons:C1=slippery,C2=typing problem

My papa brought me Samsung Galaxy S II Smart phone on my birthday.

P3:It is really useful and fantastic phone to use.I can browse the Internet through this and can take pictures, videos.

I like to watch videos in my phone.

P1:I can easily find my nearest restaurants and other places through it.it has high picture quality and videos.

I like to play games on it. The games are interesting and easily down loadable.

P5:I have taken my baby pictures with this product. The photos are great.

The smart phone is usable for my husband for checking mails for job purpose.

P2=Thin and light smartphones are easy to carry around and easier on the hand with extended use.

Since it is based on Android 4.0 we already know that it has a solid base.

Samsung flexed its software chops with the S III in a way that sets the phone apart from its competitors.

Samsung did some really interesting things with the camera.

C1:The curved bump on the phone's bottom can make it hard to hold, and the battery life needs improvement.

P4:The faster the processor, the smoother a phone's user interface, and the quicker you can browse the web, run apps and stream media.

front-facing camera is great for video chatting, and the 8-megapixel primary camera on the back snaps high-quality photos – an LED flash helps..

The device's total storage capacity is an impressive 48GB, with 16GB installed and up to an additional 32GB through an external microSD card.

The larger screen allowed me to adjust from a hard qwerty keyboard to the soft one on screen with ease.

P9=The display provides great viewing and the ability to join emails and other social accounts is the best.

Once I figure out how to make the Voice Actions work I'm sure it will be a total plus for me.

3.2.3 Phương pháp

Dựa vào việc phân tích các dữ liệu thực tế và các nghiên cứu trước đó cho thấy, người dùng thường sử dụng các đặc trưng tính từ, trạng từ và một số dạng mở rộng của động từ để thể hiện các nhận xét của họ về các đặc trưng hay trực tiếp về các sản phẩm hay dịch vụ. Các đặc trưng hay chính các sản phẩm, dịch vụ được thể hiện bằng các danh từ.

Ví dụ: trong một đánh giá ưu điểm: "*good picture quality"*

Thì "*pictrure quality"* là cụm danh từ thể hiện đặc trưng của sản phẩm, còn "*good"* là tính từ thể hiện nhận xét về đặc trưng đó.

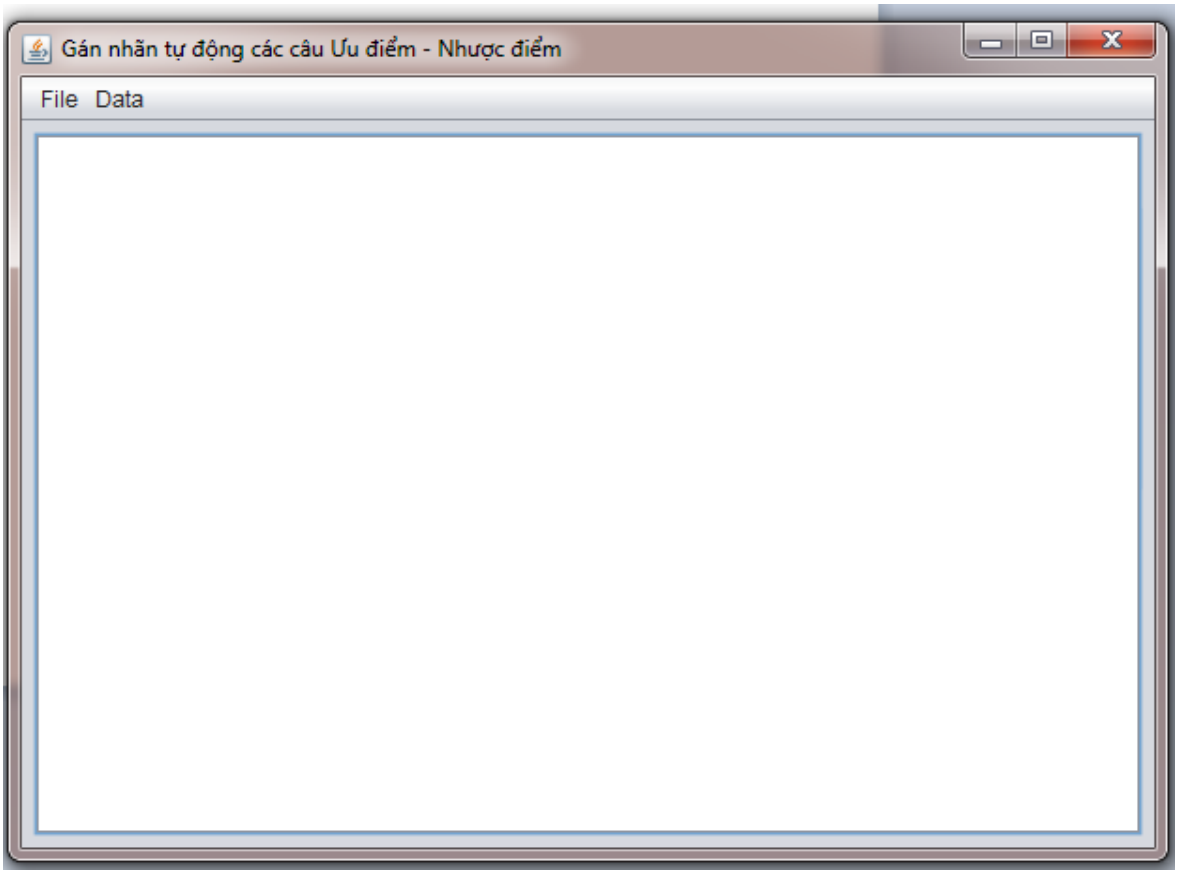
Do đó, chúng tôi sử dụng một phương pháp đơn giản là tìm kiếm các danh từ, sự xuất hiện tương ứng của các danh từ, tính từ, trạng từ và động từ được nhắc tới trong các ưu, nhược điểm đã được tóm tắt ở đầu bình luận trong các câu nhận xét để làm dấu hiệu gán nhãn ưu, nhược điểm cho các câu.

3. 3 Kết Quả

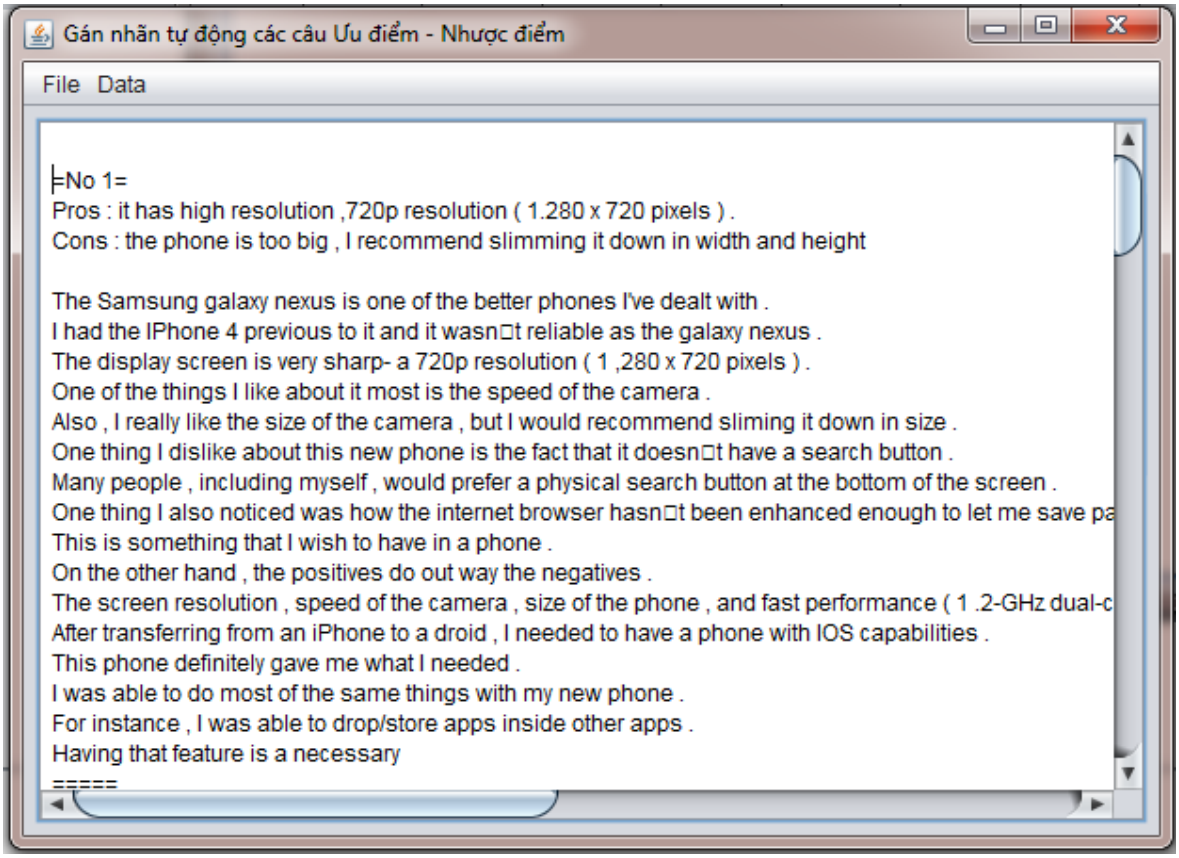
Chương trình thử nghiệm gán nhãn cho 50 bình luận thu thập từ trang http: //www. epinions.com. Chúng tôi so sánh kết quả với bộ dữ liệu đã gán nhãn bằng tay.

3.3.1 Một số giao diện chương trình:

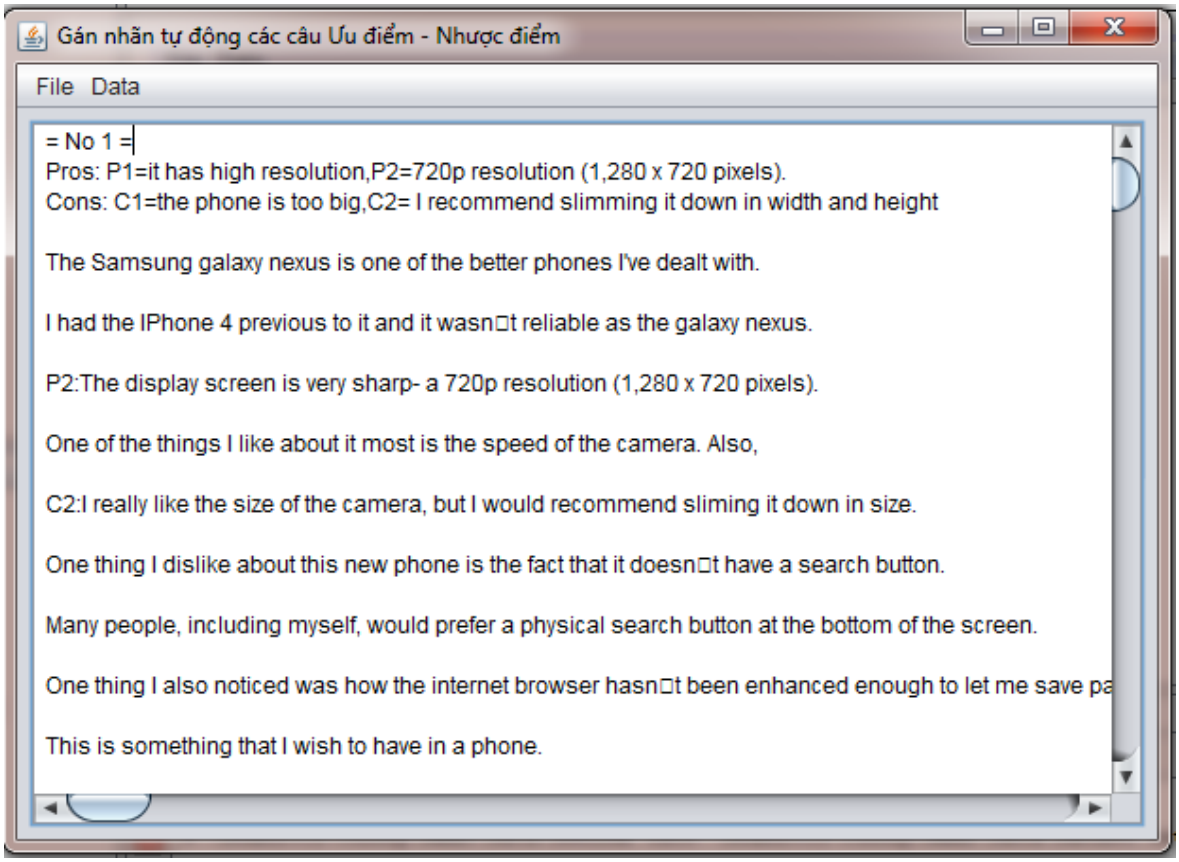
3.3.2 Giao diện chính



Mở file dữ liệu:



Kết quả chạy:


Gán nhãn tự động các câu Ưu điểm - Nhược điểm
File Data
= No 1 =
Pros: P1=it has high resolution,P2=720p resolution (1,280 x 720 pixels).
Cons: C1=the phone is too big,C2= I recommend slimming it down in width and height
The Samsung galaxy nexus is one of the better phones I've dealt with.
I had the IPhone 4 previous to it and it wasn□t reliable as the galaxy nexus.
P2:The display screen is very sharp- a 720p resolution (1,280 x 720 pixels).
One of the things I like about it most is the speed of the camera. Also,
C2:I really like the size of the camera, but I would recommend sliming it down in size.
One thing I dislike about this new phone is the fact that it doesn□t have a search button.
Many people, including myself, would prefer a physical search button at the bottom of the screen.
One thing I also noticed was how the internet browser hasn□t been enhanced enough to let me save pa
This is something that I wish to have in a phone.

KẾT LUẬN

Luận văn hướng tới mục tiêu:

Tự động trích những ưu điểm, nhược điểm của các nhận xét online.

Đưa ra kết luận những sản phẩm hoặc dịch vụ có là xu hướng của người dùng hay không.

Tuy đã xem xét được tất cả các mục tiêu như trong phần giới thiệu nhưng do thời gian có hạn, nên chúng tôi chưa thể thực hiện thành công tất cả những mục tiêu đưa ra. Tuy nhiên, luận văn cũng đạt được một số kết quả:

- Nghiên cứu và trình bày về bài toán phân tích quan điểm.
- Nghiên cứu và trình bày bài toán tự động xác định ưu,nhược điêm của các nhận xét online.
- Cài đặt thành công chương trình tự động gán nhãn các câu là ưu điểm và nhược điểm trong một bình luận trên nền JAVA.

Do thời gian có hạn, nên chúng tôi mới chỉ nghiên cứu, thu thập gán nhãn một bộ dữ liệu nhỏ và tiến hành thực nghiệm để gán nhãn tự động các câu là ưu điểm và nhược điểm trong một bình luận.

Trong thời gian tới, chúng tôi tiếp tục phát triển việc gán nhãn tự động các câu ưu, nhược điểm sử dụng thêm thông tin từ từ điển WordNet và tiến hành phân loại các câu ưu và nhược điểm.Tự động đưa ra được kết luận là sản phảm hay dịch vụ nào đó có là xu hướng của người dùng hay không? Dựa vào đây những nhà đầu tư,các nhà sản xuất sẽ có căn cứ để điều chỉnh sản phẩm,dịch vụ theo xu hướng của đa số người dùng.

Chúng tôi rất mong nhận được những ý kiến đóng góp từ các Thầy, Cô và các bạn.

Trân trọng cảm ơn.

TÀI LIỆU THAM KHẢO

Tiếng Việt :

1. Ths. Nguyễn Thị Xuân Hương và Ths. Lê Thụy về "phân tích quan điểm và một số hướng tiếp cận". Hội nghị khoa học lần thứ nhất, 2012, trường ĐHDL Hải Phòng
2. Đặng Thị Ngọc Thanh,Trích và sắp xếp các đặc trưng của bài phân tích quan điểm khoá luận tốt nghiệp hệ đại học ngành Công nghệ thông tin, Đại học Dân lập Hải Phòng, 2011.

Tiếng Anh :

1. Kim, Soo-Min and Eduard Hovy. 2005. Automatic Detection of Opinion Bearing Words and Sentences. In the Companion Volume of the Proceedings of IJCNLP-05, Jeju Island, Republic of Korea.
2. Kim, Soo-Min & Eduard Hovy (2006a). Automatic identification of pro and con reasons in online reviews. In Proceedings of the Poster Session at the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006, pp. 483–490.
3. http: //www. epinions.com
4. http: //www.complaints.com