

**ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Vũ Xuân Sơn**

**TỔNG HỢP QUAN ĐIỂM DỰA TRÊN MÔ HÌNH  
THỐNG KÊ VÀ ỨNG DỤNG VÀO KHAI PHÁ QUAN  
ĐIỂM TRONG VĂN BẢN  
TIN TỨC TIẾNG VIỆT**

**KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC CHÍNH QUY**

**Ngành: Công nghệ thông tin**

**HÀ NỘI - 2011**

**ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Vũ Xuân Sơn**

**TỔNG HỢP QUAN ĐIỂM DỰA TRÊN MÔ HÌNH  
THỐNG KÊ VÀ ỨNG DỤNG VÀO KHAI PHÁ QUAN  
DIỂM TRONG VĂN BẢN  
TIN TỨC TIẾNG VIỆT**

**KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC CHÍNH QUY**

**Ngành: Công nghệ thông tin**

**Cán bộ hướng dẫn: Th.S Nguyễn Thu Trang**

**Cán bộ đồng hướng dẫn: CN. Nguyễn Tiến Thanh**

**HÀ NỘI - 2011**

## **Lời cảm ơn**

Lời đầu tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc tới PGS.TS Hà Quang Thụy, ThS. Nguyễn Thu Trang và CN. Nguyễn Tiến Thanh đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi cũng xin gửi lời cảm ơn tới CN. Vũ Tiến Thành, CN. Trần Bình Giang và các anh chị, các bạn sinh viên tại phòng thí nghiệm KT-Sislab đã hỗ trợ tôi rất nhiều trong quá trình thực hiện khóa luận. Tôi xin gửi lời cảm ơn tới các bạn trong lớp K52CB và K52CHTTT đã ủng hộ và khích lệ tôi trong suốt thời gian học tập tại trường.

Tôi chân thành cảm ơn các thầy, cô đã tạo cho tôi những điều kiện thuận lợi giúp tôi học tập và nghiên cứu tại trường Đại học Công Nghệ. Xin cảm ơn sự hỗ trợ từ đề tài QG.10.38 trong thời gian tôi thực hiện khóa luận.

Cuối cùng, tôi muốn gửi lời cảm ơn vô hạn tới gia đình, bạn bè, những người thân yêu luôn bên cạnh và động viên tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi xin chân thành cảm ơn!

Sinh Viên

Vũ Xuân Sơn

## Tóm tắt nội dung

Khai phá quan điểm trên miền tin tức là một lĩnh vực mới, nhận được nhiều sự quan tâm trong những năm gần đây, và đánh dấu một bước phát triển trong khai phá văn bản (text mining). Khai phá văn bản hướng tới việc phân tích ngữ nghĩa, giúp máy móc thực sự “hiểu” nội dung văn bản nói và quan điểm của người viết như thế nào (ví dụ: khen/chê) trong văn bản đó.

Nhu cầu một máy tìm kiếm quan điểm được đặt ra đáp ứng nhu cầu tìm kiếm quan điểm người dùng. Máy tìm kiếm quan điểm nhận đầu vào là một truy vấn từ người dùng và kết quả trả về là những quan điểm về vấn đề mà người dùng quan tâm, thay vì trả về tập các văn bản liên quan tới truy vấn của người dùng như các máy tìm kiếm thông thường.

Khóa luận tập trung nghiên cứu phương pháp và xây dựng mô hình thống kê cho tổng hợp quan điểm trên miền ứng dụng tin tức tiếng Việt nhằm ứng dụng vào máy tìm kiếm quan điểm trên miền dữ liệu tin tức tiếng Việt. Với đầu vào là một danh từ chỉ tên thực thể người dùng quan tâm, hệ thống tiến hành gửi truy vấn lên các máy tìm kiếm (Google, Yahoo..) và lấy về các trang tin có chứa bình luận của người dùng. Với tập các trang tin thu thập được, hệ thống tiến hành tổng hợp quan điểm và trả về kết quả tổng hợp cho người dùng.

Với mô hình đề xuất, khóa luận tiến hành xây dựng thử nghiệm áp dụng mô hình trên miền dữ liệu là các bình luận từ trang tin VnExpress. Trong [DK08], Hoa và cộng sự đã đưa ra phương pháp đánh giá kết quả cho máy tìm kiếm dựa vào chuyên gia. Thử nghiệm cho kết quả trên mức điểm là 5, giá trị đáp ứng trung bình và chất lượng tổng hợp đạt mức điểm khả quan trên 3. Kết quả này cho thấy mô hình đề xuất là đúng đắn và có thể triển khai thực tế.

## **Lời cam đoan**

Tôi xin cam đoan khóa luận với đề tài “Tổng hợp quan điểm dựa trên mô hình thống kê và ứng dụng vào khai phá quan điểm trong văn bản tin tức tiếng Việt” là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả được trình bày trong khóa luận là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác.

Tôi đã trích dẫn đầy đủ các tài liệu tham khảo, công trình nghiên cứu liên quan ở trong nước và quốc tế.

Trong các công trình khoa học được công bố trong khóa luận, tôi đã thể hiện rõ ràng và chính xác những gì do tôi đã đóng góp.

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

Tác giả

Vũ Xuân Sơn

## Mục lục

Tóm tắt nội dung.....	4
Bảng các ký hiệu và chữ viết tắt.....	8
Danh sách bảng biểu.....	9
Danh sách hình ảnh.....	10
Mở đầu.....	11
Chương 1: Giới thiệu chung .....	13
1.1 Khai phá quan điểm .....	13
1.1.1 Khái quát khai phá quan điểm.....	14
1.1.2 Ý nghĩa và ứng dụng bài toán .....	20
1.2 Khai phá quan điểm trên miền tin tức.....	22
1.2.1 Giới thiệu bài toán.....	22
1.2.2 Máy tìm kiếm quan điểm .....	23
1.2.3 Tổng hợp quan điểm dựa trên truy vấn .....	25
Tóm tắt chương 1 .....	25
Chương 2: Các phương pháp tiếp cận giải quyết bài toán khai phá quan điểm trên văn bản tin tức .....	26
2.1 Một số nghiên cứu liên quan.....	26
2.2 Phương pháp tóm tắt quan điểm dựa trên mô hình thống kê.....	26
Bước 1. Thu thập từ nhận định .....	29
Bước 2: Trích xuất quan điểm.....	29
Bước 3. Tổng hợp quan điểm trích xuất được. ....	30
2.3 Phương pháp tóm tắt quan điểm dựa trên mô hình học máy .....	32
2.3.1 Mô tả hệ thống .....	32
2.3.2 Dữ liệu.....	37

2.3.3 Phương pháp thực hiện .....	38
2.4 Nhận xét.....	40
Tóm tắt chương 2.....	40
Chương 3: Tổng hợp quan điểm dựa trên mô hình thống kê .....	41
3.1 Cơ sở lý thuyết.....	41
3.1.1 Kho ngữ liệu khai phá quan điểm .....	41
3.1.2 Phương pháp trích rút đặc trưng văn bản.....	43
3.1.3 Phương pháp tổng hợp quan điểm dựa vào từ điển .....	45
3.2 Mô hình thống kê áp dụng tổng hợp quan điểm cho văn bản tin tức tiếng Việt .....	46
3.2.1 Phân tích mô hình và đề xuất .....	46
3.2.2 Phân tích phương pháp và đề xuất .....	49
Tóm tắt chương 3 .....	53
Chương 4: Thực nghiệm và đánh giá .....	54
4.1. Môi trường và các công cụ sử dụng thực nghiệm .....	54
4.2 Dữ liệu thử nghiệm .....	55
4.2.1 Đặc trưng trang tin tức VnExpress.....	55
4.2.2 Thu thập dữ liệu .....	57
4.3 Thực nghiệm .....	58
4.3.1 Mô tả cài đặt chương trình .....	58
4.3.2 Thực nghiệm hệ thống .....	58
4.3.3 Đánh giá kết quả thực nghiệm .....	61
Tóm tắt chương 4.....	63
Kết luận và định hướng phát triển .....	64
Phụ lục .....	66
Tài liệu tham khảo .....	66

### **Bảng các ký hiệu và chữ viết tắt**

Ký hiệu viết tắt	Viết đầy đủ
POS	Part Of Speech
TF-IDF	Term Frequency-Inverse Document Frequency
Pos(s)	Positive Score
Neg(s)	Negative Score



## Danh sách bảng biểu

Bảng 1: Mẫu các nhãn POS trích chọn quan điểm.....	17
Bảng 2: Kết quả trích xuất từ quan điểm từ tập dữ liệu .....	29
Bảng 3: Sự khác nhau giữa TAC 2008 và nghiên cứu của các tác giả.....	35
Bảng 4: Danh sách máy tìm kiếm blog và thuộc tính .....	38
Bảng 5: Hướng dẫn đánh giá khả năng trả lời câu hỏi .....	39
Bảng 6: Hướng dẫn đánh giá chất lượng ngôn ngữ học.....	39
Bảng 7. Ví dụ một synset trong từ điển VietSentiWordNet.....	42
Bảng 8. Một số từ trong tập từ điển phủ định .....	43
Bảng 9. Một số từ trong từ điển thể hiện sắc thái.....	43
Bảng 10. Cấu hình hệ thống thử nghiệm.....	54
Bảng 11. Công cụ phần mềm sử dụng.....	54
Bảng 12: Thành phần trong bài tin và định dạng HTML.....	57
Bảng 13: Các gói cài đặt trong thực nghiệm .....	58
Bảng 14: Một số đoạn bình luận liên quan tới từ khóa “Rùa Hồ Gươm” .....	60
Bảng 15: Kết quả tổng hợp quan điểm với từ khóa truy vấn “Rùa Hồ Gươm” .....	61
Bảng 16: Thang điểm đánh giá khả năng trả lời câu hỏi của hệ thống đề xuất.....	62
Bảng 17: Thang điểm đánh giá chất lượng ngôn ngữ học.....	62
Bảng 18: Kết quả đánh giá thực nghiệm với 5 truy vấn.....	63

## **Danh sách hình ảnh**

Hình 1. Trang web Twitter Sentiment với từ khóa search là Obama.....	21
Hình 2. Trang web tweetfeel với từ khóa search Steve Jobs.....	22
Hình 3. Mô hình thống kê tổng hợp quan điểm .....	28
Hình 4. Kiến trúc FastSum cho tổng hợp quan điểm Blog .....	34
Hình 5. Mô hình tổng hợp quan điểm dựa trên phương pháp thống kê .....	48
Hình 6. Truy vấn máy tìm kiếm lấy các trang liên quan .....	50
Hình 7: Bảng xếp hạng của VnExpress.Net trên Alexa .....	55
Hình 8: Một bài tin trên trang VnExpress.Net .....	56
Hình 9: Thực nghiệm pha thu thập tài liệu liên quan .....	59
Hình 10: Ví dụ một tài liệu sau bước tiền xử lý .....	59
Hình 11: Thực nghiệm pha trích xuất quan điểm với từ khóa “Rùa Hồ Gươm” .....	60
Hình 12. Định dạng lại dữ liệu lấy về từ VnExpress.Net sau khi trích xuất thông tin.....	66

## Mở đầu

Khi sự phát triển mạnh mẽ của các mạng xã hội và blog cá nhân, các thông tin cá nhân và quan điểm người dùng được đưa lên Internet ngày càng tăng. Bài toán đặt ra là làm thế nào để tìm kiếm các quan điểm của người khác về các thực thể mà người dùng quan tâm? Giải quyết được bài toán chính là đưa ra được câu trả lời cho câu hỏi “những người khác nghĩ gì về vấn đề mà người dùng đang quan tâm?”. Từ đó giúp người dùng có một cái nhìn khái quát quan điểm của mọi người về đối tượng đang được quan tâm.

Trong những năm gần đây, có nhiều nghiên cứu như [JJLF08, AMT08, KCL06] được đưa ra nhằm giải quyết vấn đề tổng hợp quan điểm tin tức và blog. Tuy nhiên, đối với miền dữ liệu tiếng Việt, chưa có một nghiên cứu nào được công bố. Với các máy tìm kiếm hiện tại, để tìm kiếm quan điểm người dùng cần duyệt từng kết quả trả về từ máy tìm kiếm để lấy ra được các quan điểm về vấn đề mình đang quan tâm.

Khóa luận giới thiệu phương pháp tổng hợp dựa trên mô hình thống kê của Sushant Kumar và Diptesh Chatterjee [SD08], cùng phương pháp tổng hợp dựa trên hệ thống FastSum sử dụng mô hình học máy SVM của Jack G. Conrad và cộng sự [JJLF08]. Từ đó khóa luận đề xuất phương pháp tổng hợp quan điểm dựa trên mô hình thống kê áp dụng vào bài toán khai phá quan điểm trong văn bản tin tức tiếng Việt. Phương pháp được đưa ra với các pha xử lý được điều chỉnh phù hợp với miền dữ liệu tiếng Việt. Và cải tiến bằng việc kết hợp với phương pháp tổng hợp quan điểm sử dụng từ điển của Ku và Liang đề xuất [KCL06]. Kết quả thực nghiệm đánh giá hệ thống cho thấy mô hình đề xuất là đúng đắn và khả quan để đưa vào áp dụng thực tế.

Nội dung khóa luận gồm có 5 chương:

**Chương 1:** Giới thiệu khái quát về khai phá quan điểm và bài toán tổng hợp quan điểm trên miền tin tức.

**Chương 2:** Giới thiệu về các phương pháp giải quyết bài toán tổng hợp quan điểm trên miền tin tức trên thế giới. Khóa luận giới thiệu hai phương pháp tiêu biểu cho tổng hợp quan điểm dựa trên truy vấn là phương pháp tổng hợp dựa trên mô hình thống kê và phương pháp tổng hợp dựa trên mô hình học máy. Đây là cơ sở phương pháp luận để khóa luận đưa ra mô hình áp dụng với bài toán tổng hợp quan điểm dựa trên mô hình thống kê ứng dụng cho khai phá quan điểm tin tức tiếng Việt.

**Chương 3:** Trên cơ sở phân tích ưu và nhược điểm của các phương pháp trình bày trong chương 2, phương pháp tổng hợp quan điểm dựa trên mô hình thống kê được đề xuất và các pha xử lý được cụ thể hóa. Với truy vấn đầu vào của người dùng là tên thực thể: danh từ chỉ tên người, địa điểm..., hệ thống gửi truy vấn lên máy tìm kiếm để lấy về những trang web có nhiều thông tin bình luận từ người dùng. Tiếp đó dữ liệu được đưa qua các pha để tiến hành tổng hợp quan điểm dựa dựa ra kết quả cho người dùng.

**Chương 4:** Thử nghiệm, và đánh giá kết quả tổng hợp quan điểm. Chương này trình bày về các bước cài đặt và thử nghiệm hệ thống cài đặt theo mô hình đề xuất. Đồng thời tiến hành đánh giá kết quả hệ thống thử nghiệm. Kết quả thực nghiệm cho thấy tính đúng đắn và khả năng áp dụng vào thực tế của mô hình đề xuất là khả quan.

**Phần kết luận và định hướng phát triển khóa luận:** Tóm lược những nội dung chính đạt được của khóa luận, đồng thời cũng chỉ ra những hướng cần khắc phục và đưa ra định hướng nghiên cứu tiếp theo.

## Chương 1: Giới thiệu chung

Nội dung chính của khóa luận là đề xuất mô hình thống kê cho khai phá quan điểm trong văn bản tin tức tiếng Việt. Chương này sẽ giới thiệu các khái niệm trong khai phá quan điểm cũng như bài toán khai phá quan điểm trên miền ứng dụng tin tức.

### 1.1 Khai phá quan điểm

Thông tin văn bản (text) có thể được phân làm hai loại chính là: sự kiện (facts) và quan điểm (opinions). Sự kiện là các đối tượng thực thể và các sự việc (events) trong thế giới thực. Quan điểm là các ý kiến chủ quan mà con người nói về thực thể và sự việc.

Khai phá quan điểm, là một lĩnh vực mới, dành được nhiều quan tâm trong thời gian gần đây và chỉ mới đạt được một số kết quả bước đầu, do đó còn rất nhiều vấn đề trong khai phá quan điểm chưa được giải quyết trên thế giới cũng như ở Việt Nam.

Quan điểm có vai trò rất quan trọng, bởi khi chúng ta cần quyết định một vấn đề gì chúng ta thường đặt ra câu hỏi “*Người khác nghĩ về vấn đề đó như thế nào?*”. Chẳng hạn khi bạn muốn mua một chiếc laptop HP Pavilion DV6 bạn sẽ muốn hỏi bạn bè và người thân “*Máy HP có tốt không? Dòng Pavilion của HP thế nào? Pin dùng có lâu không?...v.v*”. Như vậy quan điểm của người khác giúp các cá nhân có thêm thông tin trước khi quyết định một vấn đề. Ngoài ra khai phá quan điểm giúp các công ty, tổ chức biết được ý kiến, quan điểm của một bộ phận người quan tâm về vấn đề của công ty, tổ chức.

Trong [BoLee08], Bo Pang và Lillian Lee đã chứng minh vai trò rất quan trọng của khai phá quan điểm. Các tác giả nêu ra cuộc điều tra vào năm 2006 với 2500 thanh niên Mỹ về hoạt động khi sử dụng internet. Kết quả cho thấy 27% để tìm kiếm online, 28% hoạt động trực tuyến để tham gia các cộng đồng mạng, 28% sử dụng để chia sẻ quan điểm của họ và 8% để bình luận chính trị. Như vậy ta thấy tỷ lệ người sử dụng Internet để chia sẻ quan điểm và bình luận là rất lớn, là kho dữ liệu giàu thông tin cho khai phá quan điểm. Lerman và cộng sự cũng đã thực hiện đánh giá trong [KSR09], cho thấy người dùng rất quan tâm tới mô hình tổng hợp quan điểm.

Ở Việt Nam, con số những người sử dụng Internet ngày càng lớn, theo thống kê của VNNIC<sup>1</sup> tính đến tháng 10/2010, số người sử dụng Internet ở Việt Nam đã đạt con số 26 triệu, chiếm hơn 30% tổng số gần 90 triệu dân của cả nước. Cùng với sự phát triển của các mạng xã hội, blog thì ngày càng nhiều các thông tin cá nhân, quan điểm cá nhân được đưa lên internet, tạo kho dữ liệu lớn cho khai phá và tổng hợp quan điểm. Đây là một lợi thế nhưng cũng là một thách thức cho bài toán khai phá quan điểm.

### ***1.1.1 Khái quát khai phá quan điểm***

Trong [BL07], Bing Liu đã đưa ra khái quát về khai phá quan điểm như các khái niệm được dùng trong khai phá quan điểm, các loại bài toán trong khai phá quan điểm:

#### **a. Các khái niệm dùng trong khai phá quan điểm:**

- ***Đối tượng (object):*** Dùng để chỉ thực thể (người, sản phẩm, sự kiện, chủ đề...) được đánh giá. Mỗi đối tượng có một tập các thành phần (components) hay thuộc tính (attributes): gọi chung là các đặc trưng (features). Mỗi thành phần hay thuộc tính lại có một tập các thành phần con hay thuộc tính con. Như vậy, một đối tượng  $O$  được biểu diễn bởi một cặp  $\{T, A\}$ :
  - $T$ : là cấu trúc phân cấp thành phần cha – thành phần con
  - $A$ : tập các thuộc tính của đối tượng  $O$

Ví dụ:

Máy quay phim có một tập các thành phần: ống kính, pin... và các thuộc tính: kích cỡ, khối lượng, chất lượng ảnh. Thành phần pin có thuộc tính con: kích cỡ, thời gian...

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

Ví dụ:

“Thời lượng pin của máy ảnh này rất tốt”: đặc trưng “thời lượng pin” là đặc trưng hiện. “Máy ảnh này quá to”: đặc trưng “kích cỡ” là đặc trưng ẩn

---

<sup>1</sup> Trung tâm Internet Việt Nam <http://vnnic.vn>

- **Đoạn đánh giá (opinion passage) về một đặc trưng:** Đoạn đánh giá về một đặc trưng  $f$  của đối tượng  $O$  trong  $r$  là một tập các câu liên tiếp trong  $r$  diễn tả quan điểm tích cực hay tiêu cực về đặc trưng  $f$ . Đoạn đánh giá bao gồm tối thiểu ít nhất một câu. Hầu hết các nghiên cứu hiện tại tập trung vào mức câu: mỗi một đoạn bao gồm một câu. Khái niệm đoạn và câu được dùng tương đương về ngữ nghĩa trong ngữ cảnh này.
- **Quan điểm hiện, ẩn:** Quan điểm hiện (explicit opinion) về một đặc trưng  $f$  là một câu thể hiện quan điểm mang tính chủ quan, diễn tả trực tiếp quan điểm tích cực hay tiêu cực của tác giả. Quan điểm ẩn (implicit opinion) về một đặc trưng  $f$  là câu thể hiện quan điểm tích cực hay tiêu cực một cách không tường minh (ngụ ý, ẩn ý)

Ví dụ:

*“Cái laptop này rất bền” “Tai nghe mới mua mà đã hỏng”*

- **Người đánh giá (opinion holder):** Là người hay tổ chức cụ thể đưa ra lời đánh giá. Với các đánh giá trên forum, blogs, người đánh giá chính là các tác giả của đánh giá hay bài viết đó.

Ví dụ:

*“Ông A rất hài lòng với kết quả của bản hợp đồng”*

#### **b. Bài toán trong khai phá quan điểm:**

Khai phá quan điểm hay còn gọi là phân lớp nhận định có 3 bài toán điển hình nhất đó là:

- Bài toán phân lớp quan điểm
- Bài toán khai phá và tổng hợp quan điểm dựa trên đặc trưng
- Bài toán khai phá quan hệ (so sánh).

**Bài toán phân lớp quan điểm:** Cũng giống bài toán phân lớp văn bản, theo đó mỗi văn bản sau khi phân lớp sẽ thuộc về một trong các lớp được xác định trước, trong phân lớp quan điểm xác định hai lớp tích cực (Positive) hoặc tiêu cực (Negative). Ví dụ, cho một tập các đánh giá sản phẩm, hệ thống sẽ quyết định đánh giá nào là tích cực, tiêu cực. Và việc phân loại thường ở mức tài liệu và không quan tâm tới vấn đề chi tiết hơn như người đánh giá sản phẩm thích hay không thích đặc trưng nào của sản phẩm.

Mô hình bài toán:

- Tập đánh giá  $D = \{d_i\}$
- Hai lớp đánh giá Pos (tích cực) và Neg (tiêu cực)
- Bộ phân lớp sẽ phân  $d_i$  vào một trong 2 lớp Pos/Neg

Ví dụ:

Với một đánh giá về bộ phim A, hệ thống sẽ xác định quan điểm chủ đạo của đánh giá này là *hay* (nên xem) hay là *không hay* (không nên xem).

Bài toán phân lớp quan điểm và phân lớp văn bản về cơ bản là tương tự nhau, tuy nhiên có một số khác biệt như sau:

- Phân lớp văn bản:
  - Phân lớp văn bản dựa vào các chủ đề được xác định trước: chính trị, thể thao, ca nhạc, hội họa,...
  - Các từ khóa liên quan tới chủ đề là quan trọng
- Phân lớp quan điểm:
  - Phân lớp các quan điểm vào hai nhóm: Pos và Neg
  - Từ khóa diễn tả quan điểm, tình cảm đóng vai trò quan trọng.

Về phương pháp phân lớp quan điểm, có một số phương pháp điển hình như:

- Phân lớp dựa vào cụm từ thể hiện quan điểm: phương pháp thực hiện gồm ba bước:

**Bước 1:** Trích chọn các từ, cụm từ chứa tính từ hay các trạng từ. Bởi trong câu quan điểm thì những tính từ và trạng từ là những thành phần tốt để biểu diễn quan điểm. Tuy nhiên, có thể sẽ không có thông tin ngữ cảnh để xác định xu hướng quan điểm của chúng là tích cực hay tiêu cực:

Ví dụ:

*“Cây cầu này “dài” quá” và “Bài diễn văn “dài” quá”*

Ở đây tính từ thể hiện quan điểm “dài” mang nghĩa tích cực ở câu thứ nhất và mang nghĩa tiêu cực ở câu thứ hai.



Các cặp từ sẽ được trích chọn nếu các nhãn POS của chúng khớp với các mẫu:

*Bảng 1: Mẫu các nhãn POS trích chọn quan điểm*

First word	Second word	Third word (Not extracted)
JJ	NN or NNS	Anything
RB, RBR, or RBS	JJ	Not NN or NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, or VBG	Anything

Các nhãn sử dụng theo nhãn của Penn Treebank được tác giả liệt kê chi tiết trong [BL07].

Ví dụ, câu “*This camera produces beautiful pictures*” thì cụm từ “*beautifulpictures*” được trích chọn do khớp với mẫu thứ nhất.

**Bước 2:** Xác định xu hướng quan điểm của cụm từ thu được theo độ đo *PMI*:

- Độ đo *PMI* là độ đo sự tương đồng ngữ nghĩa giữa hai cụm từ tính theo công thức:

$$PMI(term_1, term_2) = \log_2 \left\{ \frac{\Pr(term_1 \cap term_2)}{\Pr(term_1) \Pr(term_2)} \right\}$$

Trong đó:

- $\Pr(term_1 \cap term_2)$  là xác suất đồng xuất hiện của  $term_1$  và  $term_2$ .
- $\Pr(term_1), \Pr(term_2)$  là xác suất mà  $term_1, term_2$  xuất hiện khi thống kê chúng riêng rẽ.
- Log của tỉ lệ trên là lượng thông tin mà ta có được về sự hiện diện của một *term* khi ta quan sát *term* kia.
- Xu hướng ngữ nghĩa, hay quan điểm của một từ/cụm từ được tính dựa trên việc tính toán độ đo *PMI* của từ/cụm từ đó với hai từ “excellent” và “poor” theo công thức:

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

Hoặc sử dụng máy tìm kiếm để tính toán *PMI*, theo đó:

$Pr(term)$ : số kết quả trả về (hits) của máy tìm kiếm khi truy vấn là  $term$ . Thêm 0.01 vào hits để tránh trường hợp chia cho 0.  $Pr(term_1), Pr(term_2)$  là số kết quả trả về khi sử dụng máy tìm kiếm Alta Vista sử dụng thêm toán tử NEAR:

$$SO(phrase) = \log_2 \left( \frac{hits(phrase \text{ NEAR excellent})hits("poor")}{hits(phrase \text{ NEAR poor})hits(excellent)} \right)$$

**Bước 3:** Với mỗi đánh giá, hệ thống sẽ tính trung bình các chỉ số SO của tất cả các cụm từ trích chọn được. Nếu chỉ số dương thì sẽ xếp vào lớp Pos, chỉ số âm xếp vào lớp Neg.

- Phân lớp dựa vào các phương pháp phân lớp văn bản: phương pháp này sử dụng các thuật toán phân lớp văn bản như Naïve Bayesian, SVM, kNN... để tiến hành phân lớp các đánh giá về lớp tích cực/tiêu cực.
  - Phân lớp dựa vào hàm tính điểm số: Bước 1: Tính điểm các từ trong văn bản của tập dữ liệu học theo công thức:

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

Trong đó  $t_i$  là từ cần được tính điểm,  $C$  là một lớp quan điểm,  $C'$  là lớp phản bù của  $C$  (not  $C$ ).  $Pr(t_i|C)$ : xác suất xuất hiện ở lớp  $C$  của  $t$ . Điểm số được chuẩn hóa trong đoạn  $[-1,1]$ .

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

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

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

Phương pháp phân lớp quan điểm trên có ưu điểm: cung cấp một cái nhìn tổng thể của một ý kiến, quan điểm, đánh giá về một đối tượng. Tuy nhiên, nó có rất nhiều các nhược điểm như: không đưa ra chi tiết người đánh giá thích/không thích cái gì. Và không thích hợp áp dụng phân lớp cho các văn bản không phải là đánh giá như các bình luận ở blog, diễn đàn. Để giải quyết được các nhược điểm này, bài toán cần đi vào mức sâu hơn là mức câu, đặc trưng.

**Bài toán khai phá và tổng hợp quan điểm dựa trên đặc trưng:** Ở bài toán này, đi chi tiết vào mức câu để làm rõ đối tượng mà người đưa ra quan điểm thích hay không

thích. Đối tượng ở đây có thể là sản phẩm, dịch vụ, một chủ đề, một cá nhân hay tổ chức. Ví dụ, trong đánh giá sản phẩm, người đánh giá đưa ra các bình luận tích cực/tiêu cực về một đặc trưng của sản phẩm. Như trong câu “*tuổi thọ pin của chiếc camera hay hơi ngắn*” thì đối tượng được đưa ra bình luận ở đây là “*tuổi thọ pin*” và quan điểm này là quan điểm tiêu cực. Có hai bài toán đặt ra:

- **Bài toán 1:** Xác định và trích chọn các đặc trưng của đối tượng mà người dùng đánh giá. Ví dụ: “*hiệu năng xử lý của chiếc laptop này rất cao*” thì đặc trưng của đối tượng “*laptop*” ở đây là “*hiệu năng xử lý*”.
- **Bài toán 2:** Xác định và xem quan điểm của người đánh giá về đặc trưng của đối tượng đó là tích cực, tiêu cực, hay trung lập. Ví dụ: trong đánh giá của người dùng về *hiệu năng xử lý* của *laptop* thì quan điểm đưa ra là tích cực.

**Bài toán khai phá quan hệ so sánh:** Ngoài cách biểu diễn các quan điểm bằng cách trực tiếp nhận xét về đối tượng còn có một cách đánh giá là bằng cách so sánh đối tượng muốn nhận xét với một đối tượng khác. Ví dụ, khi một người nói một cái gì đó là tốt hay xấu, người ta thường yêu cầu “*so với cái gì ?*”. Vì vậy, một trong những cách quan trọng nhất của đánh giá đối tượng là so sánh trực tiếp nó với một đối tượng tương tự khác.

Ví dụ:

“*Laptop HP Pavilion DV6 thì nhanh hết pin hơn so với dòng Pavilion DV4*” ở đây đặc trưng “*thời lượng pin*” của Pavilion DV6 là đối tượng được nhận xét.

Trong nội dung khóa luận, chúng tôi đề cập liên quan tới bài toán thứ nhất là bài toán phân lớp quan điểm: coi khai phá quan điểm như là phân lớp văn bản. Coi mỗi quan điểm là một văn bản và quá trình phân lớp quan điểm chính là phân lớp văn bản. Các quan điểm sẽ được phân vào hai lớp tích cực (tốt) và tiêu cực (xấu), không quan tâm tới lớp trung lập (neutral) bởi những nhận định mang tính trung lập không ảnh hưởng tới kết quả tổng hợp quan điểm. Ở đây, thay vì phân lớp văn bản, chúng tôi tiến hành phân lớp các câu quan điểm liên quan tới truy vấn của người dùng về một thực thể mà người dùng quan tâm và không quan tâm tới mức đặc trưng, tức coi quan điểm được đưa ra là cho đối tượng. Mục tiêu chủ đạo là nhanh chóng xác định quan điểm đánh giá về một thực thể liên quan tới truy vấn là tốt hay xấu và tỷ lệ phần trăm tốt xấu.

### ***1.1.2 Ý nghĩa và ứng dụng bài toán***

Nghiên cứu khai phá quan điểm bắt đầu bằng việc xác định những từ thể hiện quan điểm (nhận định) như: tuyệt vời (great) , tuyệt diệu (wonderful), tốt (good), xấu (bad). Đã có nhiều nghiên cứu về việc xác định xu hướng quan điểm (tốt/xấu) của một từ. Trong phạm vi khóa luận, chúng tôi tập trung vào nhiệm vụ tổng hợp quan điểm dựa vào truy vấn của người dùng trên miền dữ liệu là các bình luận của độc giả trên trang tin tức VnExpress.Net.

Trên thế giới đã có nhiều các nghiên cứu tới khai phá quan điểm tin tức, trong đó cần phải kể đến hai trang web Twitter Sentiment<sup>2</sup> và TweetFeel<sup>3</sup>. Với đầu vào là tên thực thể người dùng cần nắm quan điểm, hệ thống đưa ra tổng hợp các bình luận của người dùng Twitter về thực thể, đồng thời đưa ra tỷ lệ tích cực/tiêu cực các quan điểm về thực thể đó.

---

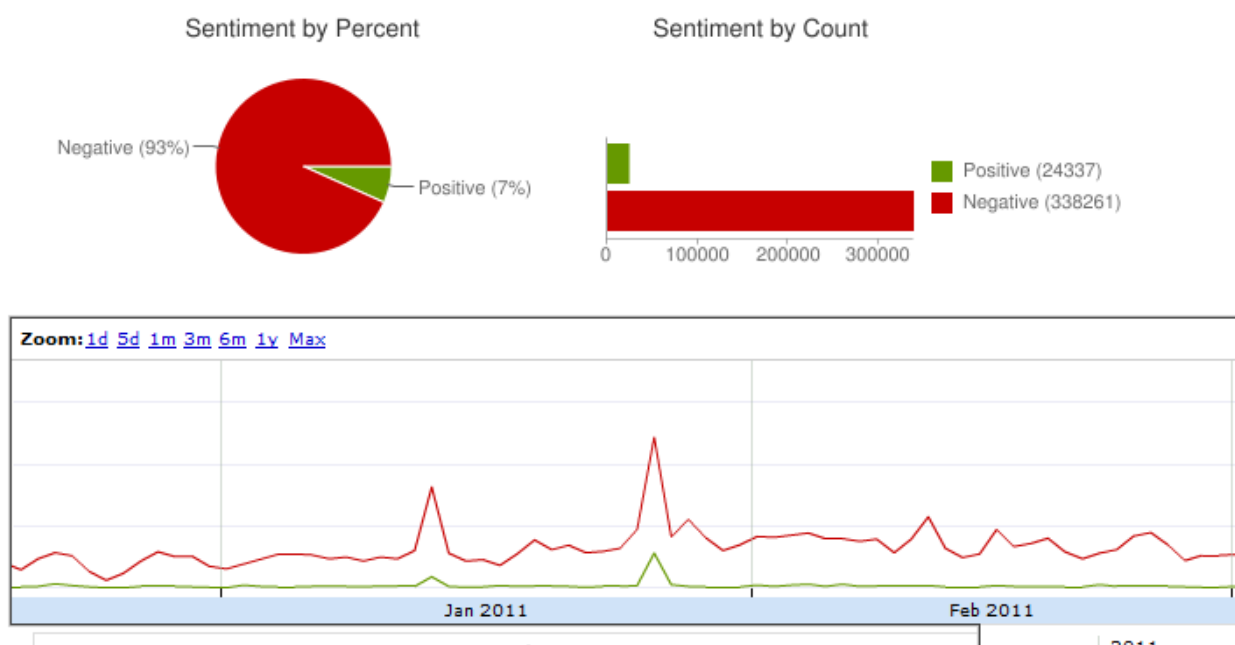
<sup>2</sup><http://twittersentiment.appspot.com/>

<sup>3</sup><http://www.tweetfeel.com>

## 1. Trang web Twitter Sentiment

Twitter Sentiment là trang web tổng hợp và theo dõi quan điểm theo thời gian thực về một thực thể, với tập dữ liệu là các tin nhắn (blash) của người dùng trên mạng xã hội Twitter. Kết quả tổng hợp được đưa ra kèm theo thời điểm cùng với thống kê phần trăm Pos/Neg của các quan điểm. Thêm vào đó, hệ thống còn cho phép người dùng đánh giá lại hướng quan điểm của hệ thống đưa ra, điều này giúp hệ thống cải thiện được kết quả đánh giá quan điểm.

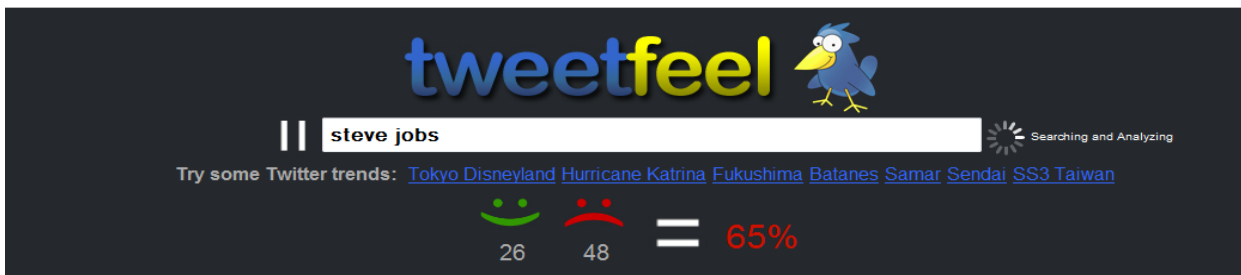
### Sentiment analysis for obama



Hình 1. Trang web Twitter Sentiment với từ khóa search là Obama

## 2. Trang web TweetFeel

Là trang web tổng hợp quan điểm về một thực thể (entity) với tập dữ liệu là các blash của người dùng trên mạng xã hội <http://twitter.com/>.



Hình 2. Trang web tweetfeel với từ khóa search Steve Jobs

### 1.2 Khai phá quan điểm trên miền tin tức

#### 1.2.1 Giới thiệu bài toán

Quan điểm có ở khắp mọi nơi trên Internet từ các trang web tin tức, các trang web đánh giá sản phẩm, các trang blog cá nhân,... Tuy nhiên, trên mỗi miền dữ liệu, thông tin chứa quan điểm có các hình thức thể hiện riêng. Qua quá trình phân tích các miền dữ liệu, chúng tôi nhận thấy sự khác biệt giữa các miền như sau:

- Đối với miền dữ liệu là các trang web đánh giá sản phẩm, cấu trúc dữ liệu thường phức tạp, bài viết có những ngôn ngữ đặc thù, khó nhận biết quan điểm. Hơn nữa, số lượng các trang web đánh giá sản phẩm ở Việt Nam chưa nhiều, cùng với sự quản lý bài viết từ người dùng trên các trang này còn hạn chế, do đó các bài đánh giá chứa ít thông tin và có nhiều dữ liệu nhiễu. Điều này gây khó khăn cho việc xác định quan điểm của người viết.

- Các trang tin tức và các trang blog là hai miền tin tức giàu các thông tin quan điểm với cấu trúc bài viết tương đối giống nhau, văn phong giản dị và ít bị nhập nhằng ngữ nghĩa, đặc biệt là ngôn ngữ sử dụng chuẩn tiếng Việt. Đây là thuận lợi lớn cho thực hiện khai phá quan điểm trên miền này.

Hiện bài toán khai phá quan điểm trên miền tin tức vẫn là một bài toán mới mẻ trên thế giới cũng như ở Việt Nam. Ở Việt Nam chưa có ứng dụng nào được công bố chính thức về khai phá quan điểm tin tức. Như một bước đệm cho khóa luận này, trong công trình sinh viên NCKH [SHH11] tôi và các đồng tác giả đã xây dựng bộ từ điển

VietSentiWordNet cho miền tin tức tiếng Việt và áp dụng vào trích xuất và tổng hợp quan điểm tin tức ở mức câu, mức đoạn và mức tài liệu. Hệ thống của chúng tôi cho kết quả với với độ chính xác F1 cao nhất là 70%. Dựa vào bộ từ điểm này, tôi đề xuất mô hình khai pháp quan điểm trên miền tin tức dựa trên mô hình thống kê để phát hiện và tổng hợp những quan điểm, bình luận của người đọc liên quan tới từ khóa truy vấn.

Khai phá quan điểm trên miền tin tức bao gồm ba bài toán con: tìm kiếm, trích chọn và tổng hợp quan điểm.

Để giúp người dùng tìm kiếm quan điểm, cần có hệ thống có tìm kiếm phát hiện quan điểm của người trên các bài báo tin tức. Khác với các máy tìm kiếm truyền thống là trả về các tài liệu chứa từ khóa truy vấn, và người dùng phải duyệt qua các tài liệu để lấy thông tin mình cần, máy tìm kiếm quan điểm sẽ tổng hợp và trả về các quan điểm liên quan tới truy vấn người dùng.

Một trong những khâu quan trọng của máy tìm kiếm quan điểm là tổng hợp quan điểm dựa vào truy vấn người dùng. Do đó khóa luận tập trung giải quyết khâu này. Để tiến hành được bước tổng hợp quan điểm dựa vào truy vấn, khóa luận cũng tiến hành các bước tìm kiếm và trích chọn những quan điểm liên quan tới truy vấn người dùng.

Trong khóa luận này, chúng tôi sử dụng mô hình thống kê để phát hiện và tổng hợp những quan điểm, bình luận của người đọc liên quan tới từ khóa truy vấn. Khai phá quan điểm ứng dụng cho máy tìm kiếm quan điểm là trích xuất và tổng hợp các quan điểm về thực thể mà người dùng đang quan tâm. Theo đó công việc chính của khai phá quan điểm trên miền tin tức là khai phá quan điểm từ bình luận của độc giả trên các bài báo.

### ***1.2.2 Máy tìm kiếm quan điểm***

Hàng ngày, một số lượng lớn các quan điểm được người dùng đưa lên các trang blog cá nhân về tất cả các chủ đề và các tin tức khác nhau, khi đó số lượng các bài viết tăng lên và trở thành một kho dữ liệu lớn. Một bài toán được đặt ra là làm sao để giải quyết vấn đề giàu về dữ liệu mà nghèo về tri thức. Để giải quyết bài toán này, cần thiết phải có một hệ thống tìm kiếm quan điểm. Hệ thống giúp cho người dùng biết được những người khác nghĩ thế nào về vấn đề mà họ đang quan tâm.

Giống như các máy tìm kiếm thông thường, đầu tiên hệ thống cần lấy các nội dung từ người dùng trên web và cung cấp một dịch vụ tìm kiếm quan điểm. Máy tìm kiếm cho

phép tìm kiếm các quan điểm về bất kỳ một đối tượng nào. Trong [BL07], Bing Liu đã đưa ra các truy vấn thông thường về tìm kiếm quan điểm như:

1. Tìm kiếm quan điểm về một đối tượng hoặc một đặc trưng của đối tượng riêng biệt. Ví dụ: quan điểm khách hàng về một máy camera hoặc về chất lượng ảnh của máy camera hoặc quan điểm của người dân về các chủ đề chính trị. Các đối tượng của tìm kiếm quan điểm có thể là một sản phẩm, một tổ chức, hoặc một chủ đề nào đó.
2. Tìm kiếm quan điểm của một người, hoặc một tổ chức về một chủ đề riêng biệt. Ví dụ: Người ta có thể tìm kiếm quan điểm của Bill Clinton về nạn người nhập cư bất hợp pháp hoặc về một khía cạnh đặc biệt của nó. Những kiểu tìm kiếm thường liên quan tới các tài liệu về tin tức, nơi các cá nhân, hoặc tổ chức đưa ra quan điểm của mình. Đối với các trang web do người dùng tự biên soạn nội dung, người viết bài chính là người đưa ra quan điểm.

Đối với kiểu truy vấn thứ nhất, người dùng có thể đơn giản đưa truy vấn vào là một đối tượng hoặc đặc trưng của đối tượng. Với truy vấn thứ hai, người dùng có thể đưa truy vấn là tên người đưa ra quan điểm và tên đối tượng. Rõ ràng, khó có thể áp dụng kết hợp từ khóa cho các loại truy vấn khác nhau bởi vì một tài liệu có thể chứa từ khóa nhưng lại không chứa quan điểm. Ví dụ: nhiều cuộc thảo luận trên các diễn đàn và blog không chứa quan điểm, nhưng chỉ chứa các câu hỏi và trả lời về một vài đối tượng. Những câu hoặc tài liệu chứa quan điểm cần được xác định trước khi cho phép tìm kiếm. Như vậy, hình thức đơn giản nhất của tìm kiếm quan điểm là áp dụng tìm kiếm dựa trên từ khóa để xác định những câu/tài liệu liên quan.

Cho việc xếp hạng, các công cụ tìm kiếm web truyền thống xếp hạng trang web dựa vào độ tin cậy và các trọng số liên quan. Với kiểu truy vấn thứ hai thì việc xếp hạng các trang web có chứa thông tin người dùng tìm kiếm là cần thiết, do những người đưa ra quan điểm thường chỉ đưa ra một quan điểm về đối tượng tìm kiếm và quan điểm thường chứa trong một tài liệu hoặc một trang tin. Tuy nhiên, với kiểu truy vấn quan điểm đầu tiên, tập các tài liệu thứ hạng đầu tiên chỉ chứa quan điểm của một vài người. Do đó, cần tổng hợp và đưa ra phần trăm tích cực/tiêu cực của toàn bộ tài liệu liên quan tới thực thể được truy vấn thay vì chỉ một vài tài liệu có thứ hạng cao ở đầu tiên. Một vài trường hợp, những tài liệu chứa quan điểm rất dài (chẳng hạn như các đánh giá), điều này gây khó khăn cho người dùng khi phải đọc toàn bộ tài liệu để hiểu được quan điểm của người viết.



Do đó, nhu cầu cần tóm tắt quan điểm, có thể là một đánh giá trung bình về tỷ lệ tích cực/tiêu cực về các tài liệu thể hiện quan điểm, hoặc phức tạp hơn là tổng kết quan điểm ở mức đặc trưng.

### ***1.2.3 Tổng hợp quan điểm dựa trên truy vấn***

Với nhiệm vụ của tổng hợp quan điểm ứng dụng cho máy tìm kiếm quan điểm là tạo ra các tổng hợp dựa trên truy vấn. Miền dữ liệu có thể là các quan điểm người dùng về chính trị, phim ảnh, âm nhạc, hoặc về các sản phẩm mới ra trên thị trường. Việc tổng hợp quan điểm dựa vào câu truy vấn người dùng nhằm đưa ra những câu trả lời chính xác là *những quan điểm liên quan tới từ khóa truy vấn, thay vì đưa ra một tập tài liệu cho người dùng.*

Trong nội dung khóa luận, chúng tôi tiến hành tổng hợp quan điểm dựa trên truy vấn của người dùng là tên các sự kiện, thực thể, từ đó tìm ra các quan điểm của độc giả bình luận trên trang tin VnExpress.Net về sự kiện, thực thể người dùng quan tâm.

Ví dụ:

Khi người dùng đưa vào truy vấn là “*Rùa Hồ Gươm*” hệ thống sẽ tìm các quan điểm người dùng liên quan tới từ khóa truy vấn và tiến hành tổng hợp quan điểm.

### **Tóm tắt chương 1**

Trong chương này, chúng tôi đã giới thiệu khái quát các khái niệm liên quan tới khai phá quan điểm, các bài toán trong khai phá quan điểm. Khóa luận cũng giới thiệu bài toán khai phá quan điểm trên miền ứng dụng tin tức và ứng dụng vào tìm kiếm quan điểm tin tức.

Trong chương tiếp theo, khóa luận mô tả một số phương pháp giải quyết bài toán khai phá quan điểm miền ứng dụng tin tức trên thế giới.

## **Chương 2: Các phương pháp tiếp cận giải quyết bài toán khai phá quan điểm trên văn bản tin tức**

Có nhiều kỹ thuật và phương pháp được sử dụng để giải quyết bài toán tổng hợp quan điểm dựa trên truy vấn. Chương này giới thiệu các nghiên cứu liên quan tới bài toán tổng hợp quan điểm và tập trung giới thiệu mô hình thống kê cho tổng hợp quan điểm dựa trên truy vấn người dùng trong [JJLF08] và phương pháp dựa trên mô hình học máy SVM trong [SD08].

### **2.1 Một số nghiên cứu liên quan**

Các phương pháp tổng hợp quan điểm đã nhận được nhiều sự quan tâm của các nhà nghiên cứu trên thế giới. Trong [AMT08], Aurélien Bossard và cộng sự đã tiếp cận hướng tổng hợp quan điểm trên nhiều tài liệu sử dụng học máy SVM, với dữ liệu là các bài viết trên blog. Trong [HL04], Hu và Liu đề xuất phương pháp tổng hợp đánh giá người dùng về sản phẩm bằng cách biểu diễn những quan điểm tích cực/tiêu cực của người dùng về những đặc trưng của sản phẩm. Trong [ADSB10], Amitava Das và cộng sự đã đưa ra phương pháp tổng hợp quan điểm dựa trên chủ đề sử dụng từ điển Bengali SentiWordNet.

Nhiều kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được áp dụng, các phương pháp học máy như phân lớp Naïve Bayes, cực đại hóa entropy và SVM được nghiên cứu và áp dụng thử nghiệm. Tuy nhiên, trong [PLV02], Lang, Pee và Vaithyanathan đã chứng minh rằng các kỹ thuật NLP không thực hiện tốt như phân lớp tình cảm dựa trên chủ đề truyền thống. Trong [SD08, JJLF08], các tác giả đã đưa ra mô hình và phương pháp tổng hợp quan điểm dựa trên truy vấn dưới hai góc độ tiếp cận và học máy SVM. Kết quả nghiên cứu của hai nhóm này là cơ sở quan trọng để chúng tôi phân tích và đưa ra được mô hình áp dụng phù hợp cho bài toán.

### **2.2 Phương pháp tóm tắt quan điểm dựa trên mô hình thống kê**

Trong [SD08], Sushant Kumar và cộng sự đã đưa ra mô hình thống kê cho bài toán tổng hợp quan điểm dựa vào truy vấn. Mô hình của Sushant Kumar và cộng sự được đăng tại hội nghị TAC 2008 thể hiện được nhiều ưu điểm vượt trội. Hệ thống có bapcha chính là các pha:

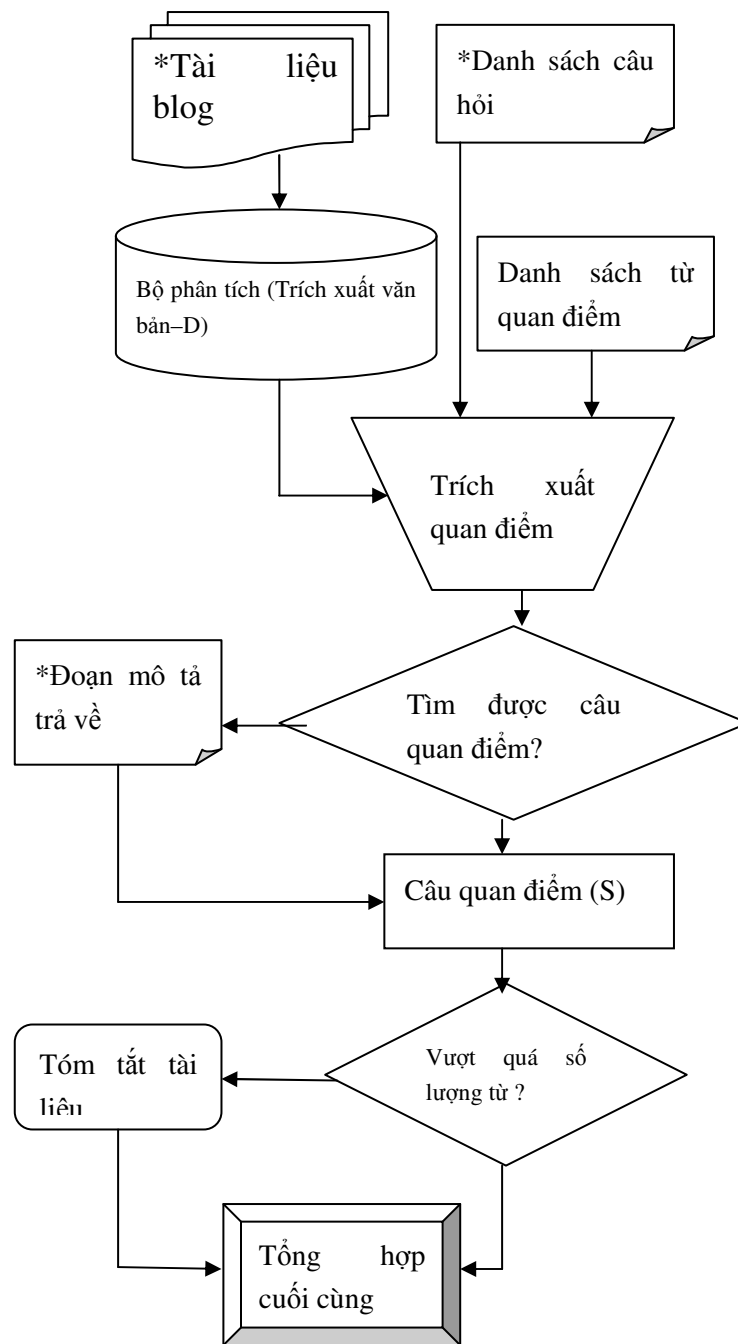
1. Pha trích xuất văn bản
2. Pha trích xuất quan điểm

### 3. Pha tổng hợp quan điểm

Với phương pháp tiếp cận các tác giả đưa ra theo các bước chính:

1. Đầu tiên là phân tích cú pháp những văn bản có được. Pha này cho ta tập  $D$  là tất cả câu từ các bài viết trên blog.
2. Từ tập các câu  $D$ , trích ra các câu nhận định  $S$  ( $S \subseteq D$ ) bằng cách: Đầu tiên, lọc các câu liên quan tới truy vấn người dùng từ tập  $D$ . Sau đó trích ra các câu quan điểm từ những câu liên quan tới truy vấn người dùng. Trong trường hợp không có câu nào liên quan tới truy vấn thì sử dụng phần tóm tắt trả về từ máy tìm kiếm (snippet) để lấy các quan điểm về tài liệu đó.
3. Nếu các câu quan điểm (tập  $S$ ) được trích ra nằm trong các từ giới hạn tóm tắt, thì những câu này được sắp xếp lại theo chỉ mục sao hai câu cùng ở trong một tài liệu thì sẽ đi liền với nhau.
4. Nếu những câu quan điểm được trích không nằm trong giới hạn từ tóm tắt, sử dụng pha trích tổng hợp để đưa ra tổng hợp cuối cùng. Người tổng hợp chỉ sử dụng câu quan điểm để tạo tóm tắt và bỏ qua tất cả các văn bản khác trên tài liệu.

Mô hình hệ thống được đưa ra bởi TAC:



Hình 3. Mô hình thống kê tổng hợp quan điểm

Các thuật toán để trích xuất từ nhận định sử dụng danh sách từ nhận định để quyết định độ phân cực của văn bản và dựa trên công thức xếp hạng để quyết định nội dung quan điểm của bản. Các tác giả sử dụng hệ thống trích xuất tổng hợp dựa trên thuật toán Lexrank cho bước tổng hợp với công thức xếp hạng ở mức câu đã được cải tiến.

### ***Bước 1. Thu thập từ nhận định***

Với tập từ vựng ban đầu là 100 từ quan điểm tích cực và 100 từ quan điểm tiêu cực.

*Bảng 2: Kết quả trích xuất từ quan điểm từ tập dữ liệu*

Số lượng	Số lượng từ nhận được	Số lượng từ phân lớp đúng	Phần trăm chính xác
Đầu tiên	1634	1458	89.22%
Thứ hai	3872	3109	80.29%
Thứ ba	6982	6108	87.48%

Sau khi có tập các từ thể hiện quan điểm tích cực/tiêu cực, các tác giả chia nhóm từ theo các loại sau:

- Rất tích cực (Score = 2)
- Tích cực (Score = 1)
- Trung tính (Score = 0)
- Tiêu cực (Score = -1)
- Rất tiêu cực (Score = -2)

### ***Bước 2: Trích xuất quan điểm***

**Thuật toán:** thuật toán thực hiện trên một tài liệu thuộc về một chủ đề riêng biệt. Không thực hiện cho trích trọn quan điểm trên nhiều tài liệu. Hệ thống tập trung vào truy vấn, với các bước thuật toán thực hiện như sau:

- **Bước 1:** Các câu truy vấn được phân tích để trích xuất ra tên trong câu truy vấn. Những tên này sẽ hình thành từ khóa tìm kiếm. Ngoài ra, các từ trong câu truy vấn sẽ được kết hợp với danh sách từ nhận định để tìm ra loại quan điểm đang được người dùng truy vấn. Trong trường hợp không có từ nào trong danh sách từ quan điểm phù hợp thì các từ sẽ được chuyển qua WordNet và lấy ra các từ đồng nghĩa và lại tiến hành tìm từ phù hợp trong danh sách từ.
- **Bước 2:** tách câu trong tài liệu

- **Bước 3:** với mỗi câu, tìm sự xuất hiện của từ khóa truy vấn  $w$ . Giả sử vị trí của  $w$  là vị trí từ  $i$
- **Bước 4:** Kiểm tra tất cả các từ ở vị trí từ  $i-6$  đến  $i+6$ . Nếu có một từ nhận định trong khoảng đó, đánh dấu đó là một câu quan điểm về từ  $w$ . Nếu không thì quay lại bước 2.
- **Bước 5:** nếu từ quan điểm xuất hiện ở vị trí thứ  $k$ , kiểm tra các từ ở vị trí  $(k-2)$  đến  $(k+2)$ . Nếu có từ ở trong khoảng này nằm trong danh sách từ phủ định thì tiến hành đảo ngược quan điểm của câu.
- **Bước 6:** Từ danh sách trọng số, tính toán trọng số của từ nhận định. Việc tính toán này được lưu lại để đảo chiều quan điểm khi cần. Ví dụ một từ có trọng số là  $+1$ , nếu đảo lại thì sẽ có trọng số là  $-1$ . Tiếp theo tính độ phân cực trung bình của câu bằng cách chia cho tổng số từ nhận định tìm thấy.
- **Bước 7:** Phân cực quan điểm của câu được chia như sau: (gọi  $S$  là phân cực trung bình)
  - $S > 1 \Rightarrow$  quan điểm rất tích cực
  - $0,3 < S < 1 \Rightarrow$  quan điểm tích cực
  - $-0,3 < S < 0,3 \Rightarrow$  quan điểm trung tính
  - $-1 < S < -0,3 \Rightarrow$  quan điểm tiêu cực
  - $S < -1 \Rightarrow$  quan điểm rất tiêu cực
- **Bước 8:** Trở về bước 2.

Các tác giả đã tiến hành thực nghiệm và cho thấy vị trí các từ quan điểm thường xuất hiện ở vị trí trong khoảng  $i-6$  đến  $i+6$  với  $i$  là vị trí của từ khóa tìm kiếm. Kết quả đánh giá thực nghiệm cũng cho thấy độ hồi tưởng và độ chính xác là đạt kết quả cao nhất.

### ***Bước 3. Tổng hợp quan điểm trích xuất được.***

Với danh sách các câu quan điểm đã trích xuất được từ tài liệu được tiến hành tổng hợp. Thuật toán tổng hợp dựa trên danh sách tần suất thuật ngữ và xác định trọng số lớn nhất của câu (với công thức được sử dụng) và nhóm những câu có trọng số lớn nhất và khác nhau để đưa vào tổng hợp. Sử dụng unigram đơn giản để khớp những từ liên quan (danh từ, động từ) để tìm độ tương tự giữa hai câu bất kỳ. Do đó ở bản tổng hợp cuối cùng sẽ không có thông tin nào bị trùng nhau.

### Thuật toán tổng hợp:

1. Chuẩn bị danh sách tần xuất các từ trong tài liệu mà không xem xét tới từ dừng (stopwords)
2. Theo các quy tắc dưới đây để lấy một tài liệu mới từ tài liệu đã có:
  - a. Đầu tiên, tìm tất cả các danh từ, đại từ và liên kết của chúng nếu tương thích. Nếu không tìm về các từ trước đó trong tài liệu để lấy những danh từ và đại từ tương thích
  - b. Với các từ trong ngoặc kép, như các động từ “say”, “told”, “said” thì thường liên quan tới các danh từ chỉ người ở trước như đại từ “I” v.v
3. Với mỗi câu đã được tính trọng số cơ bản dựa vào danh sách tần xuất từ. Xác định một giá trị ngưỡng bằng thực nghiệm và tất cả các từ có tần số cao hơn ngưỡng được lấy làm trọng số của câu. Với mỗi câu  $S$  thì  $Weight=W(S) = \frac{1}{n} \sum_{i=1}^n w_i$  trong đó  $w_i$  là tần xuất của từ lớn hơn ngưỡng.
4. Tìm độ tương đồng giữa 2 câu sử dụng unigram đơn giản. Định nghĩa giá trị quan hệ hệ số (RC – Relation Coefficient) để thể hiện mối quan hệ tương tự giữa 2 câu bất kỳ. Công thức  $RC = \text{số unigram phù hợp} / \max(\text{unigram của một trong 2 câu phù hợp})$

Ví dụ:

$S_1$ = My name **is** **Tom** Sawyer

$S_2$ =**Tom** **is** friend with Huck Finn.

5. Unigram bắt được là “Tom” và “is”. Độ dài  $S_1 = 5$ ,  $S_2 = 6$  do đó  $S_{12} = (2/6) = 0.3333$
6. Lấy những câu có trọng số cao nhất. Gọi các trọng số là  $S_i$ ,  $S_j$  của các câu được đưa vào danh sách tổng hợp và loại bỏ khỏi danh sách câu. Để giảm thiểu sự dư thừa, tất cả các câu có giá trị  $RC \geq 0.5$  đều bị loại bỏ. Do theo thống kê các tác giả cho thấy những câu có giá trị  $RC \geq 0.5$  là những câu có độ trùng lặp thông tin cao. Nếu danh sách câu còn các câu chưa xét thì lặp lại bước 5.
7. Lặp lại từ bước 1 đến bước 6 cho tới khi đạt ngưỡng tổng hợp.
8. Cuối cùng đưa ra tất cả các câu tổng hợp trong danh sách tổng hợp sắp xếp theo thứ tự ưu tiên theo chỉ số sao cho các câu cùng ở một tài liệu thì sẽ ở gần nhau.

## 2.3 Phương pháp tóm tắt quan điểm dựa trên mô hình học máy

Trong [JJLF08], Jack G. Conrad và cộng sự đã áp dụng hệ thống học máy FastSum vào hệ thống tóm tắt quan điểm dựa vào truy vấn người dùng với mục đích, đưa ra được một tóm tắt về các quan điểm mà người dùng quan tâm. Ví dụ với từ khóa truy vấn là một câu hỏi về tin tức “*Có phải hầu hết mọi người phản đối chính phủ liên bang cứu trợ tài chính cho các ngân hàng và tổ chức tài chính ?*”. Đầu tiên hệ thống sẽ tiến hành tìm mục tiêu truy vấn mà ở đây mục tiêu truy vấn chính là “*chính phủ liên bang*” và “*cứu trợ tài chính*”, tiếp theo gửi mục tiêu truy vấn lên máy tìm kiếm blog để lấy ra tập kết quả đầu tiên, cho qua bộ lọc để xác định độ phù hợp của các kết quả trả về. Cuối cùng tiến hành tổng hợp và đưa ra bản tổng hợp quan điểm khoảng 250 từ phù hợp với truy vấn đầu vào:

*“...Chi phí chương trình hai năm của chính phủ Obama dự kiến sẽ tốn khoảng 800 tỷ USD. Các ngân hàng đã nhận được 200 tỷ USD vốn mới từ Bộ Tài chính kể từ mùa thu năm ngoái và đã vay hàng trăm tỷ đô lan từ Fed. Một khi thị trường ổn định, các ngân hàng sẽ mua cổ phần của họ trở lại từ chính phủ. Ví dụ rõ ràng nhất mà hệ thống ngân hàng cần được giúp đỡ hơn nữa là Citigroup. FERRE: Mặc dù các ngân hàng và tổ chức tài chính nhận được 350.000.000.000 đô la viện trợ khẩn cấp đối tượng nộp thuế tài chính, cho người tiêu dùng như Baltiera, chi phí tín dụng vẫn còn cao...”*

Nghiên cứu dựa trên tập câu hỏi với pha tìm kiếm thông tin blog và hệ thống FastSum, hệ thống tự động trích xuất và tổng hợp quan điểm trên nhiều tài liệu. Nghiên cứu của các tác giả cũng đưa ra phương pháp đánh giá kết quả tổng hợp quan điểm sử dụng đánh giá của các chuyên gia [DUK08]. Các tác giả cũng thực hiện đánh giá phương pháp tổng hợp cơ sở (baseline) cho tổng hợp quan điểm dựa truy vấn với dữ liệu là blog. Kết quả cho thấy trên mức điểm là 5, giá trị đáp ứng trung bình của hệ thống và chất lượng ngôn ngữ kết quả trả về của hệ thống là lớn hơn 2.

### 2.3.1 Mô tả hệ thống

FastSum là hệ thống tổng hợp đa văn bản đã được Jack G. Conrad và cộng sự thay đổi cho tổng hợp quan điểm. FastSum sử dụng SVM hồi quy để học phân lớp quan điểm mức câu. Phần quan trọng của FastSum là thành phần lọc để xác định và loại những câu ít có khả năng sử dụng làm tổng hợp quan điểm. Ngoài ra, có một bộ lọc khác xem xét tới nhận định của câu. Bộ lọc này được các tác giả đưa thêm vào hệ thống FastSum để thực hiện nhiệm vụ tổng hợp quan điểm.

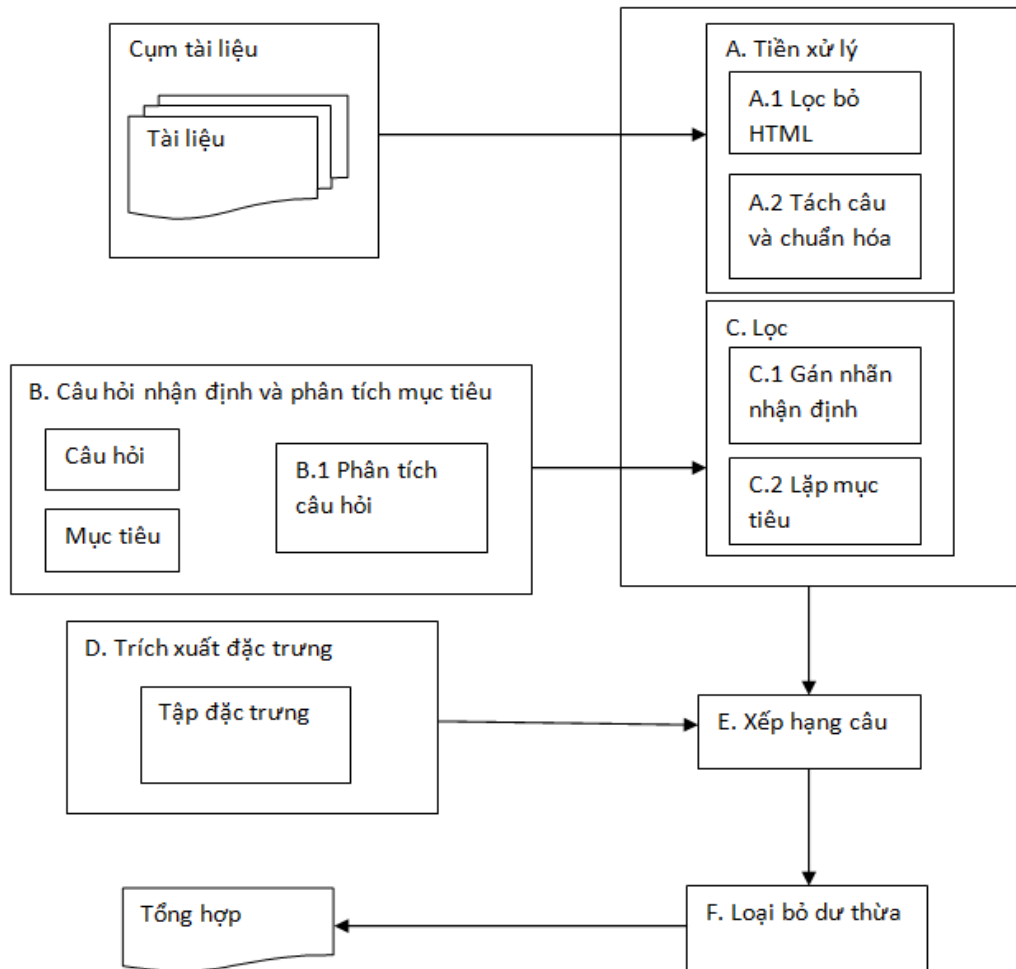


Mô hình toàn bộ hệ thống tổng hợp quan điểm blog được mô tả trong [FRJJ08], như hình dưới. Về tổng quan, hệ thống tổng hợp quan điểm FastSum được thực hiện theo trình tự:

- A. Tiền xử lý
- B. Câu hỏi quan điểm và phân tích mục đích
- C. Bộ lọc
- D. Trích xuất đặc trưng
- E. Xếp hạng câu
- F. Loại bỏ dư thừa

Đây là những thành phần cơ bản về mô hình chung của hệ thống, Jack G. Conrad và cộng sự đã có một số thay đổi trong hệ thống FastSum để áp dụng cho tổng hợp quan điểm blog dựa truy vấn như sau:

- A.1 Bộ phân tích và loại bỏ HTML
- B.1 Bộ câu hỏi quan điểm và phân tích mục tiêu
- C.1 Gán nhãn nhận định
- C.2 Bao phủ mục tiêu (*Target overlap*)



Hình 4. Kiến trúc FastSum cho tổng hợp quan điểm Blog

### 1. Tiền xử lý, phân tích truy vấn và bộ lọc

Bộ tiền xử lý thực hiện tách từ và tách câu. Thêm vào đó thành phần câu đơn giản dựa vào một vài biểu thức để loại thành phần không quan trọng trong câu. Bước xử lý này không bao gồm bước xử lý cú pháp nào. Bộ câu hỏi nhận định và phân tích mục đích quyết định độ phân cực và mục đích của câu hỏi. Với thực nghiệm hiện tại, độ phân cực và mục đích câu hỏi được xác định bằng tay.

- **Tiền xử lý:** Các tác giả thay đổi FastSum theo xử lý cho blogs bằng cách phân tích HTML để trích xuất văn bản từ các trang blog và loại bỏ tất cả ngôn ngữ trên blog. Các tác giả sử dụng bộ Jericho htmlParser<sup>4</sup> để phân tích tài liệu

<sup>4</sup><http://jerichohtml.sourceforge.net/doc/index.html>

HTML. Xóa bỏ ngôn ngữ của soạn giả, tính mật độ các từ viết hoa trong một câu bằng cách kết hợp một biểu thức ngôn ngữ được sử dụng thường xuyên trên blog.

- **Câu hỏi nhận định và phân tích mục đích:**

*Bảng 3: Sự khác nhau giữa TAC 2008 và nghiên cứu của các tác giả*

	TAC 2008	Nghiên cứu
Mục tiêu	Tên thực thể được cung cấp bởi NIST (bằng tay)	Được cung cấp bởi cụm danh từ (bằng tay)
Phân tích câu hỏi	Các mẫu và từ khóa	<không có>

- **Bộ lọc:**Như bộ lọc được thiết lập ban đầu, các tác giả loại bỏ tất cả các câu mà không chứa chính xác hai từ hoặc ít nhất ba từ mờ (three fuzzy matched) phù hợp với chủ đề mô tả. Câu được lựa chọn phụ thuộc vào nhận định và sự liên quan tới mục đích câu hỏi. Trong bộ lọc có các bước gán nhãn nhận định và xác định mục đích truy vấn.

- **Gán nhãn nhận định:** Các tác giả tiến hành gán nhãn phân cực nhận định dựa vào tìm kiếm các thuật ngữ đơn. Việc gán nhãn dựa trên tìm kiếm các cụm từ, đếm các từ tích cực/tiêu cực và gán các nhãn theo điều kiện:

$$\begin{cases} \text{Negative nu polarity} < -1 \\ \text{Neutral nu} - 1 \leq \text{polarity} \leq 1 \\ \text{Positive nu polarity} > +1 \end{cases}$$

ở đây,  $Polarity = (\#PositiveTok - \#NegativeTok) / \#AllTok$

- **Xác định mục tiêu truy vấn:**Trong hệ thống FastSum cho tổng hợp quan điểm, các tác giả sử dụng kỹ thuật xác định những câu chứa các thực thể của mục đích truy vấn. Thực nghiệm xác định mục đích truy vấn cũng được thực hiện, mặc dù mục đích truy vấn được mô tả trừu tượng hơn so với định nghĩa trong TAC.Các mục tiêu (target) không nhất thiết phải có mặt trong câu được xét, miễn là nó xuất hiện trong vùng mô tả mục tiêu. Các tác giả khớp các từ với mục tiêu bằng hàm

tương tự của Jaro Winkler. Sử dụng hàm Cosine để gán “targetness” (gần với mục tiêu nhất) sau một mục tiêu được xác định. Do đó, một câu tiếp theo vẫn có thể được xem xét đưa vào tổng hợp bởi những câu gần với mô tả mục tiêu ở câu trước. Công việc sắp tới các tác giả mong muốn là tập trung vào cách xác định các câu có liên quan và tách bỏ được các câu không liên quan tới mục đích truy vấn.

## 2. Xác định đặc trưng xếp hạng câu cho SVM

Đặc trưng phụ thuộc vào tần xuất của từ trong câu, cụm, tài liệu và chủ đề. Đặc trưng các tác giả sử dụng được chia làm hai mức: mức từ (word based) và mức câu (sentence based).

- **Đặc trưng mức từ:** Được tính toán liên quan tới tần suất của từ trong những đoạn khác nhau (cụm, tài liệu, tiêu đề và mô tả). Về thời gian chạy, tần suất liên quan giữa tất cả các từ trong câu ứng viên  $s$  được cộng lên và được chuẩn hóa bằng cách chia cho độ dài  $|s|$ .
- **Đặc trưng mức câu:** Bao gồm độ dài và vị trí của câu trong tài liệu.
- **Tần suất tiêu đề của chủ đề:** Tên chủ đề và tần suất tiêu đề  $T$  cho mỗi câu  $s$  được tính theo công thức

$$\frac{\sum_{i=1}^{|a|} f_T(t_i)}{|s|}$$

Trong đó  $f_T = \begin{cases} 1 : t_i \in T \\ 0 : \text{còn lại} \end{cases}$

- **Tần suất từ trong nội dung:** Tần suất từ trong nội dung liên quan  $p_c(t_i)$  của tất cả từ nội dung  $t_{1...|s|}$  xuất hiện trong câu  $s$ . Xác suất từ nội dung được định nghĩa:  $p_c(t_i) = \frac{n}{N}$ , trong đó  $n$  là số lần từ được xuất hiện trong cụm và  $N$  là tổng số từ trong cụm  $\frac{\sum_{i=1}^{|a|} p_c(t_i)}{|s|}$
- **Tần suất tài liệu:** Tần suất tài liệu liên quan  $p_d(t_i)$  của các từ trong nội dung  $t_{1...|s|}$  xuất hiện trong câu  $s$ . Xác suất tài liệu được định nghĩa  $p_d(t_i) = \frac{n}{D}$ , trong đó  $d$  là số tài liệu từ  $t_i$  xuất hiện trong cụm và  $D$  là tổng số tài liệu trong cụm  $\frac{\sum_{i=1}^{|s|} p_d(t_i)}{|s|}$

- **Tần suất tiêu đề:** Tần suất từ trong tiêu đề liên quan của tất cả các nội dung từ trong câu  $s$ . Xác suất tiêu đề được định nghĩa  $p_h(t_i) = \frac{h}{H}$  trong đó  $h$  là số lần từ xuất hiện trong tiêu đề và  $H$  là tổng số từ có trong tiêu đề:  $\frac{\sum_{i=1}^{|s|} p_h(t_i)}{|s|}$
- **Độ dài câu:** Đặc trưng nhị phân với giá trị bằng 1 nếu số từ nằm trong khoảng 8 đến 50. Giá trị bằng 0 nếu thuộc trường hợp còn lại.
- **Vị trí câu (nhị phân):** Chỉ ra liệu các vị trí của câu là nhỏ hơn một ngưỡng nhất định.
- **Vị trí của câu (giá trị thực):** Tỷ lệ vị trí của câu trong số các câu trong tài liệu.

### 3. Học xếp hạng câu

Để học các trọng số đặc trưng, các tác giả huấn luyện SVM hồi quy được giới thiệu tại hội nghị DUC07 với dữ liệu tin tức sử dụng cùng tập đặc trưng. Trong hồi quy, yêu cầu hàm ước lượng sự phụ thuộc của một biến vào tập các biến phụ thuộc. Trong trường hợp này, mục đích là để ước lượng “độ phù hợp tổng kết” của một câu dựa trên tập đặc trưng.

### 4. Loại bỏ thông tin dư thừa

Là bước cuối cùng, sử dụng thuật toán trong [JCD01] để xử lý loại bỏ dư thừa. Với ý tưởng cơ bản: tránh sự dư thừa bằng cách thay đổi tầm quan trọng của các câu còn lại dựa vào những câu đã được lựa chọn. Tổng hợp cuối cùng được tạo ra bằng cách xếp hạng các câu sau khi loại bỏ dư thừa.

#### 2.3.2 Dữ liệu

Dữ liệu được lấy về bằng cách tạo một vài truy vấn giống truy vấn được nêu ra trong hội nghị TAC08. Bao gồm các truy vấn đánh giá về luật và dữ liệu tạp chí luật pháp. Dữ liệu các tác giả sử dụng lấy về từ 6 công cụ tìm kiếm blog. Tập trung vào các blog luật pháp. Các công cụ được trình bày trong bảng 3.

*Bảng 4: Danh sách máy tìm kiếm blog và thuộc tính*

Lĩnh vực	Máy tìm kiếm	Thuộc tính (được lựa chọn)
Máy tìm kiếm blog chung (tập trung: blogosphere)	Technorati.com	Kết quả bao gồm độ quan trọng của trang
	Blogsearch.google.com	Xếp hạng theo ngày hoặc theo độ liên quan hoặc theo xếp hạng thích hợp
	www.blogsearchengine.com	Tập trung vào nội dung hơn
Máy tìm kiếm blog luật pháp(tập trung: blawgosphere)	www.blawg.com	Các mục kết quả trả về thường ngắn hơn
	Blawsearch.justia.com	Các mục kết quả trả về xếp hạng theo ngày hoặc theo độ liên quan
	www.blawgrepublic.com	Các mục kết quả trả về thường ngắn hơn

Các tác giả tiếp cận các hệ thống tìm kiếm blog theo hai hướng. Hướng thứ nhất, các mục được trả về theo thứ tự sắp xếp theo thời gian: Google và Justia cho phép người sử dụng lựa chọn sắp xếp theo ngày hoặc theo thứ hạng liên quan. Hướng thứ hai, các mục trả về có nội dung ngắn gọn.

### **2.3.3 Phương pháp thực hiện**

Mô hình các bước xử lý hệ thống được mô tả trong hình 4. Một vài bước tiền xử lý: chuyển các *chủ đề quan điểm về luật pháp* thành những *câu truy vấn*, sau đó xác định các *thực thể* hoặc *khái niệm* cho những *câu truy vấn*. Và *cuối cùng* được đưa vào hệ thống FastSum để tiến hành tổng hợp. Tiếp theo, đưa những câu truy vấn vào máy tìm kiếm blog, lấy ra tập kết quả trả về đầu tiên và cho kết quả chạy qua bộ lọc để kiểm tra độ phù hợp của dữ liệu với truy vấn.

Kết quả của hệ thống FastSum là một bản tổng hợp với khoảng 250 từ là những quan điểm phù hợp với truy vấn là câu hỏi của người dùng về vấn đề họ quan tâm. Thực nghiệm cho thấy, các kết quả trả về từ máy tìm kiếm blogsearchengine.com cho kết quả

tổng hợp tốt nhất. Các blog đề cập tới nhiều chủ đề khác nhau, từ các quyền dân sự như quyền riêng tư trên internet tới các hoạt động của chính phủ. Các truy vấn sau khi gửi vào máy tìm kiếm blog và đưa ra kết quả, hai kết quả đầu tiên được đưa vào FastSum để tiến hành học và chuẩn hóa. Mười kết quả tiếp theo được sử dụng để tiến hành tổng hợp, nếu như mười kết quả này không phù hợp thì mười kết quả tiếp theo sẽ được sử dụng.

Hệ thống được đánh giá bởi hai giám định viên là luật sư với nhiều năm kinh nghiệm trong chú thích và đánh giá. Để đánh giá chất lượng các tổng hợp về các truy vấn liên quan tới pháp luật, các tác giả sử dụng hai độ đo: (1) là khả năng trả lời truy vấn (mức độ và nội dung thông tin trong tổng hợp có liên quan tới truy vấn) và (2) là chất lượng ngôn từ. Cả hai độ đo này được sử dụng trong TAC08. Bảng 5 và bảng 6 mô tả mức độ trả lời truy vấn và chất lượng ngôn ngữ:

*Bảng 5: Hướng dẫn đánh giá khả năng trả lời câu hỏi*

Bậc	Ý nghĩa	Mô tả
5	Rất tốt	Liên quan tới câu hỏi, bao gồm cả phân cực quan điểm
4	Tốt	Có liên quan tới câu hỏi, bao gồm một phần phân cực quan điểm
3	Trung bình	Hơi liên quan tới câu hỏi và có sự phân cực quan điểm
2	Kém	Có sự trùng lặp với chủ đề câu hỏi và có phân cực quan điểm
1	Rất kém	Không tập trung vào câu hỏi, phân cực về một phía (chỉ có một trong các quan điểm tích cực, tiêu cực hoặc trung lập)

*Bảng 6: Hướng dẫn đánh giá chất lượng ngôn ngữ học*

Độ đo	Điểm quan tâm
Ngữ pháp	Không có ngày, hệ thống định dạng, đoạn, các thiếu sót, lỗi ..
Không có thông tin dư thừa	Không có sự lặp lại nội dung, sự kiện, cụm danh từ...
Tham chiếu rõ ràng	Dễ dàng nhận dạng đại từ và cụm danh từ...
Tính tập trung	Cần có trọng tâm rõ ràng, thông tin đầy đủ...
Cấu trúc	Nên có cùng cấu trúc và các câu có sự liên quan tới nhau..

## **2.4 Nhận xét**

Cả hai mô hình thống kê và mô hình học máy đều thể hiện được những ưu điểm riêng. Trong [BL07], Bing Liu đã đưa ra nhận xét: ứng dụng học máy trong phân lớp quan điểm là không phù hợp và thực tế nghiên cứu trong [KLC06] [DB10] đã cho thấy kết quả phân lớp quan điểm mức câu đối với phương pháp sử dụng học máy SVM và Cây quyết định cho kết quả rất thấp so với phân lớp dựa trên từ điển. Việc học máy SVM cho kết quả thấp hơn bởi để tạo được bộ phân lớp cho SVM thì yêu cầu cần phải có một bộ dữ liệu học đủ lớn và công việc này đòi hỏi rất nhiều công sức về nhân lực cũng như về thời gian. Do đó, để kết hợp được ưu điểm và hạn chế được nhược điểm của mỗi phương pháp, trong khóa luận này dựa trên hai mô hình thống kê và mô hình học máy SVM, chúng tôi đề xuất ra mô hình mới, trong đó thay vì sử dụng SVM hồi quy để phân lớp quan điểm thì chúng tôi sử dụng phương pháp thống kê có sử dụng từ điển VietSentiWordNet. Chi tiết về phương pháp và mô hình chúng tôi nêu rõ ở chương 3.

## **Tóm tắt chương 2**

Trong chương 2, khóa luận đã nêu được những nghiên cứu liên quan tới bài toán tổng hợp quan điểm dựa trên truy vấn. Khóa luận còn nêu được hai phương pháp điển hình trong tổng hợp quan điểm đa văn bản dựa vào truy vấn, đây là cơ sở lý thuyết quan trọng để chúng tôi đưa ra mô hình đề xuất trong chương 3.



## Chương 3: Tổng hợp quan điểm dựa trên mô hình thống kê

Chương này, chúng tôi giới thiệu các cơ sở lý thuyết, và phân tích mô hình hệ thống của [SD08, JJLF08], từ đó đưa ra mô hình đề xuất giải quyết bài toán.

### 3.1 Cơ sở lý thuyết

Phần này, khóa luận nêu ra những cơ sở lý thuyết và các kiến thức nền tảng để áp dụng trong mô hình giải quyết bài toán.

#### 3.1.1 Kho ngữ liệu khai phá quan điểm

Để thực hiện bài toán khai phá quan điểm, nhu cầu về một kho ngữ liệu chứa các từ quan điểm là không thể thiếu. Thực tế cho thấy, trong tiếng Anh, tiếng Ấn Độ đã được xây dựng từ điển SentiWordNet cho khai phá quan điểm. Trong [AF06], Andrea Esuli và cộng sự phát triển SentiWordNet tiếng Anh nhằm hỗ trợ cho khai phá quan điểm. Trong [DB10], A. Das và cộng sự cũng đã phát triển và ứng dụng SentiWordNet vào khai phá quan điểm cho tiếng Ấn Độ. A. Das và cộng sự phát triển SentiWordNet Ấn Độ cho 3 bộ ngôn ngữ Bengali, Hindi và Telugu. Kết quả ứng dụng từ điển SentiWordNet vào khai phá quan điểm của A. Das và cộng sự cho kết quả độ chính xác cao nhất là 75.57%. Kết quả này cho thấy việc áp dụng SentiWordNet vào khai phá quan điểm là khả quan.

Trong [KCL06], Ku và Liang cũng đưa ra phương pháp tổng hợp quan điểm sử dụng từ điển cho tin tức tiếng Trung. Trong [KD], Kerstin Denecke đã nghiên cứu khả năng sử dụng SentiWordNet vào khai phá quan điểm trên nhiều miền lĩnh vực khác nhau. Tác giả đã sử dụng hai phương pháp phân lớp quan điểm: phương pháp phân lớp dựa trên học máy và phương pháp dựa trên luật sử dụng SentiWordNet. Kết quả cho thấy SentiWordNet có khả năng ứng dụng vào để phân loại quan điểm ở nhiều lĩnh vực khác nhau. Nghiên cứu của Kerstin Denecke còn cho thấy khả năng cải thiện kết quả phân loại quan điểm khi áp dụng học máy cho xây dựng từ điển trên một lĩnh vực riêng biệt.

Các nghiên cứu của các tác giả được nêu ở trên đã cho thấy khả năng ứng dụng, và tính cần thiết của từ điển SentiWordNet vào khai phá quan điểm. Đối với miền dữ liệu tiếng Việt, tính cần thiết sử dụng SentiWordNet vào khai phá quan điểm càng quan trọng hơn khi mà sự đa hình đa nghĩa khiến việc “hiểu” nội dung trong tiếng Việt là khó khăn.

Trong nội dung khóa luận, để tiến hành phân loại quan điểm và tổng hợp quan điểm, chúng tôi có sử dụng từ điển VietSentiWordNet trong [SHH11]. Từ điển

VietSentiWordNet là kết quả từ công trình SVNCKH năm 2011 của Vũ Xuân Sơn và cộng sự. Từ điển có cấu trúc như từ điển SentiWordNet tiếng Anh 3.0. Nhóm tác giả cũng tiến hành ứng dụng từ điển vào khai phá quan điểm tin tức tiếng Việt và độ chính xác F1 cao nhất đạt 70%. Kết quả này cho thấy ứng dụng từ điển VietSentiWordNet vào khai phá quan điểm tin tức tiếng Việt là khả quan.

### Các khái niệm được sử dụng trong SentiWordNet:

✓ **Synset**: là một bản ghi trong từ điển, cấu tạo bởi 6 cột, các cột phân cách bởi dấu <tab>:

- POS: từ loại của từ
- ID: mã đại diện cho synset
- PosScore (Pos(s)): trọng số tích cực của từ
- NegScore (Neg(s)): trọng số tiêu cực của từ
- SynsetTerms: chứa những từ nhận định trong synset.

✓ **Term**: là những từ nhận định trong synset. Một synset có thể chứa nhiều term và các term này là từ đồng nghĩa với nhau. Một term có thể có nhiều ngữ cảnh khác nhau và trọng số Pos(s)/Neg(s) sẽ khác, do đó các term này sẽ được gán kèm theo số hiệu để phân biệt các term. Ví dụ: *term hope#4 có trọng số Pos(s)/Neg(s) là 0/0.375. Term hope#1 có trọng số Pos(s)/Neg(s) là 0.125/0.125*

✓ **Gloss**: là cột giải nghĩa và ngữ cảnh sử dụng của từ.

SentiWordNet được xây dựng dựa trên từ điển thuật ngữ và quan hệ WordNet tiếng Anh. Trong đó mỗi term trong WordNet đều có trọng số điểm Pos(s)/Neg(s) nằm trong đoạn từ [0,1].

Bảng 7. Ví dụ một synset trong từ điển VietSentiWordNet

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
n	07543288	0.625	0	yêu#1	Cảm xúc mạnh mẽ của tình cảm, “tình yêu cho công việc”, “trẻ em rất cần tình yêu”

Từ điển Negdic là từ điển chứa các từ phủ định trong tiếng Việt. Có cấu trúc hai cột, một cột là từ phủ định và một cột là trọng số phủ định của từ tương ứng. Tới thời điểm báo cáo từ điển có 10 từ phủ định và trọng số tương ứng.

*Bảng 8. Một số từ trong tập từ điển phủ định*

<b>Từ phủ định</b>	<b>Trọng số</b>
Không	-1.0
Không thể	-0.8

*Từ điển thể hiện sắc thái:* Có cấu trúc 2 cột. Cột thứ nhất là từ thể hiện sắc thái như (rất, lắm, cũng, chỉ, ..) và cột thứ 2 là trọng số tương ứng cho mỗi từ. Hai cột được phân cách bởi dấu tab.

*Bảng 9. Một số từ trong từ điển thể hiện sắc thái*

<b>Từ</b>	<b>Trọng số</b>
Rất	2.0
Lắm	1.5
Cũng	0.8

Tập từ điển thể hiện sắc thái được lấy ra qua việc khảo sát ngôn ngữ do độc giả bình luận trên tập dữ liệu lấy về. Tới thời điểm thực hiện báo cáo này tập từ điển sắc thái đã có 18 từ.

Việc đánh giá trọng số cho các từ thể hiện độ mạnh và từ thể hiện độ phủ định được đánh giá theo cảm nhận của người xây dựng từ điển. Trong [MKG, TWU10] Mike Thelwall và cộng sự đã đưa ra phương pháp phát hiện độ mạnh của nhận định liên quan tới đại số gia từ. Do thời gian thực hiện khóa luận tốt nghiệp có hạn nên chúng tôi chưa áp dụng phương pháp của Mike Thelwall và cộng sự để cải tiến các trọng số của từ điển. Đây sẽ là một hướng phát triển tiếp theo của khóa luận sau này.

### ***3.1.2 Phương pháp trích rút đặc trưng văn bản***

Khai phá quan điểm bao gồm nhiệm vụ tổng hợp và tìm kiếm quan điểm. Để tìm kiếm quan điểm trong lĩnh vực tin tức, một nhiệm vụ đặt ra là cần biểu diễn các văn bản tin tức bằng các từ khóa đặc trưng. Để với một truy vấn đầu vào, hệ thống cần tìm ra được các văn bản tin tức liên quan tới truy vấn để tiến hành tổng hợp quan điểm liên quan tới

truy vấn. Phần này, chúng tôi xin giới thiệu bài toán và phương pháp trích rút đặc trưng cho văn bản.

Để trích rút đặc trưng, cần thực hiện đánh trọng số cho các từ trong văn bản. Trong [THST09], có nêu phương pháp đánh trọng số dựa trên tần số từ khóa TF (Term Frequency) và phương pháp dựa trên nghịch đảo tần số văn bản (Inverse Document Frequency - IDF).

***Phương pháp dựa trên tần số từ khóa (Term Frequency - TF):***

Trọng số của từ khóa trong văn bản được tính dựa trên số lần xuất hiện của từ khóa trong văn bản. Gọi  $tf_{ij}$  là tần số xuất hiện của từ khóa  $t_i$  trong văn bản  $d_j$ , khi đó trọng số  $w_{ij}$  của từ khóa trong văn bản được tính theo một trong các công thức sau:

$$w_{ij} = \sqrt{tf_{ij}}$$

$$w_{ij} = 1 + \log (tf_{ij})$$

$$w_{ij} = tf_{ij}$$

***Phương pháp dựa trên nghịch đảo tần số văn bản (Inverse Document Frequency - IDF):***

Phương pháp này dựa trên lập luận, một từ quá thông dụng (xuất hiện nhiều trong văn bản) sẽ có độ qua trọng kém hơn từ chỉ xuất hiện trong một văn bản hoặc một tập nhỏ các văn bản. Công thức tính trọng số  $w_{ij}$  như sau:

$$w_{ij} = \log \frac{m}{df_i}$$

Với  $df_i$  là số lượng văn bản có chứa từ khóa  $t_i$  trong tập  $m$  văn bản đang xét.

***Phương pháp TF-IDF:*** Là phương pháp đánh trọng số kết hợp từ hai phương pháp TF và IDF. Công thức tính trọng số  $w_{ij}$  theo phương pháp này như sau:

$$w_{ij} = \begin{cases} [1 + \log (tf_{ij})] \log \left( \frac{m}{df_i} \right) & \text{nu } tf_{ij} \geq 1. \\ 0 & \text{nu } tf_{ij} = 0. \end{cases}$$

Trong đó:

- $w_{ij}$  là trọng số của từ khóa thứ  $t_i$  trong văn bản  $d_j$ .

- $tf_{ij}$  là số lần xuất hiện của từ khóa  $t_i$  trong văn bản  $d_j$ .
- $df_i$  là số lần xuất hiện của từ khóa  $t_i$  trong tập  $m$  văn bản đang xét.
- $m$  là số lượng văn bản trong tập đang xét.

### 3.1.3 Phương pháp tổng hợp quan điểm dựa vào từ điển

Trong [KCL06], Ku và Liang đã đưa ra phương pháp phát hiện và tổng hợp quan điểm tin tức dựa vào từ điển với độ chính xác cao, và tốt hơn so với các phương pháp phát hiện quan điểm sử dụng học máy SVM, hay Cây quyết định. Chi tiết phương pháp của Ku và Liang được mô tả trong [KCL06]. Trong nội dung khóa luận, chúng tôi mô tả thuật toán quyết định xu hướng quan điểm của Ku và Liang đề xuất để từ đó đưa ra thuật toán áp dụng cho pha tổng hợp quan điểm trong mô hình đề xuất.

Ku và Liang thực hiện thuật toán quyết định xu hướng quan điểm của một câu bằng chiến lược Bottom-up: Phát hiện từ quan điểm trong câu, dựa vào trọng số quan điểm của từ quan điểm để quyết định xu hướng quan điểm của câu:

- Thuật toán:

1. Đối với mỗi câu

2. Đối với mỗi từ nhận định trong câu này

3. Nếu một toán tử phủ định xuất hiện trước thì đảo ngược xu hướng nhận định.

4. Quyết định xu hướng quan điểm của câu này bằng hàm số của các từ nhận định và người đưa ra quan điểm như sau.

$$S_p = S_{\text{opinion-holder}} \times \sum_{j=1}^n S_{w_j}$$

Trong đó  $S_p$ ,  $S_{\text{opinion-holder}}$  và  $S_{w_j}$  là điểm nhận định của câu  $p$ , độ quan trọng của người đưa ra quan điểm<sup>5</sup>, và  $S_{w_j}$  điểm nhận định của từ  $w_j$ , và  $n$  là tổng số các từ nhận định trong  $p$ .

<sup>5</sup> Khi người đưa ra quan điểm là chuyên gia về lĩnh vực đó thì sẽ được nhân một trọng số tương ứng.

## 3.2 Mô hình thống kê áp dụng tổng hợp quan điểm cho văn bản tin tức tiếng Việt

### 3.2.1 Phân tích mô hình và đề xuất

Mô hình và phương pháp của Sushant Kumar và cộng sự có nhiều ưu điểm, nhưng khi áp dụng vào bài toán có một số vấn đề như sau:

1. **Trong pha trích xuất văn bản:** Các tác giả đề xuất phương pháp lấy các đoạn mô tả trả về (snippet) từ máy tìm kiếm để tiến hành tổng hợp quan điểm. Nhưng khi áp dụng cho miền dữ liệu tiếng Việt, các đoạn snippet trả về có chất lượng thấp, thường là các đoạn trong bài báo có đề cập tới từ khóa thay vì là các quan điểm đánh giá. Do đó trong mô hình đề xuất chúng tôi không sử dụng các snippet này.
2. **Trong pha tổng hợp quan điểm:** Trong pha tổng hợp, các tác giả có bước xử lý kiểm tra các tổng hợp, nếu có tổng hợp nào vượt quá số lượng từ cho phép thì tiến hành tóm tắt để đưa ra tổng hợp cuối cùng. Chúng tôi nhận thấy tổng hợp quan điểm dựa trên truy vấn người dùng cần thể hiện được đa dạng nhất các quan điểm, đánh giá của người dùng trên truy vấn. Do đó, chúng tôi không có bước xử lý giới hạn tóm tắt các tổng hợp quan điểm.

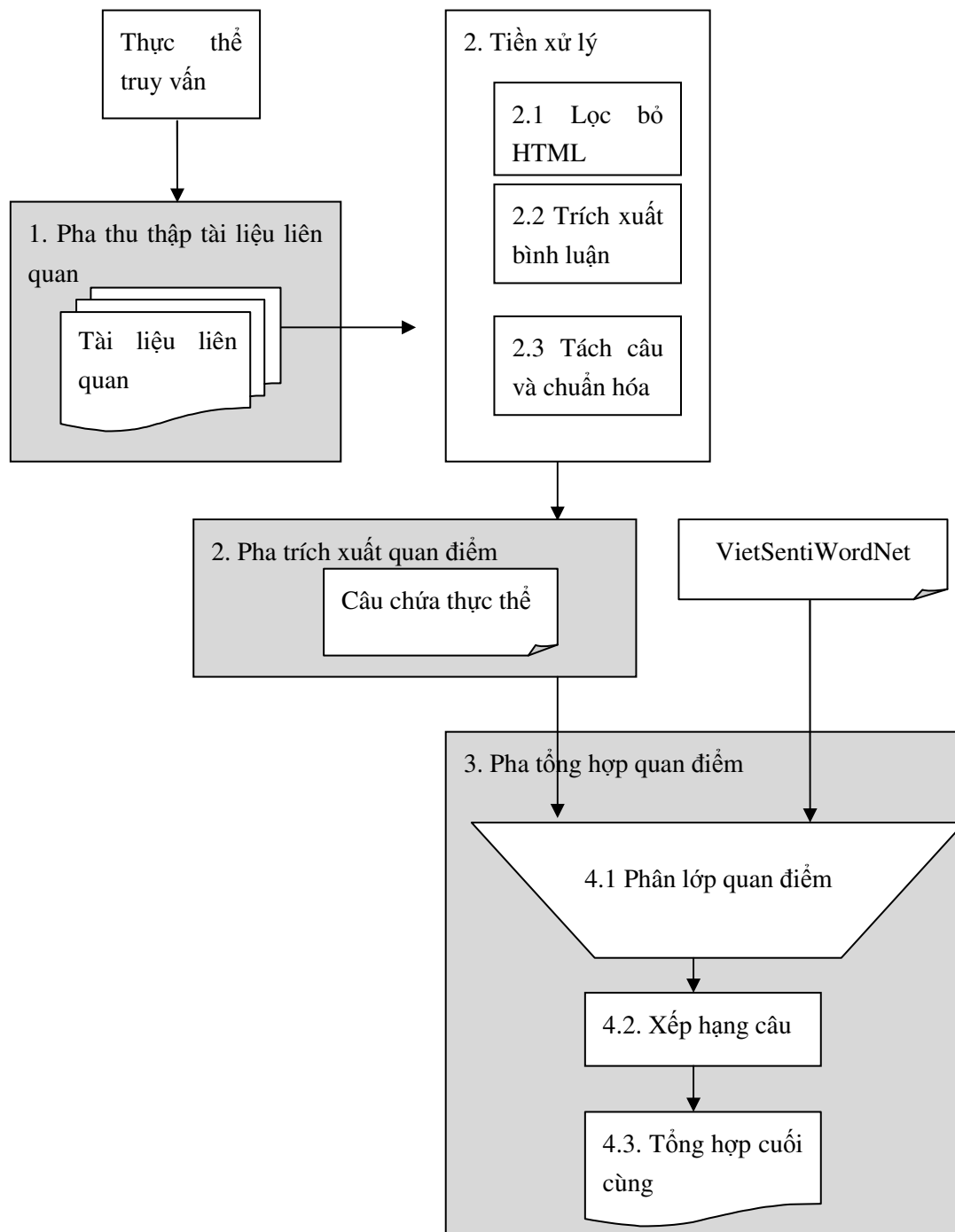
Mô hình áp dụng hệ thống học máy FastSum vào hệ thống tóm tắt quan điểm dựa vào truy vấn người dùng của Jack G. Conrad và cộng sự có nhược điểm khi áp dụng vào bài toán như sau:

1. **Pha câu hỏi nhận định và phân tích mục tiêu:** Pha có nhiệm vụ nhận câu hỏi đầu vào của người dùng và phân tích mục tiêu mà người dùng hướng tới. Pha này để thực hiện trong tiếng Việt là bài toán khó, do phương pháp phân tích câu hỏi cần áp dụng nhiều kỹ thuật và phương pháp để tìm mục tiêu câu hỏi của người dùng, bao gồm các phương pháp trích xuất đặc trưng cho câu hỏi. Công việc này cần xây dựng một tập dữ liệu câu hỏi lớn mất rất nhiều thời gian và công sức. Do đó, trong mô hình đề xuất, chúng tôi không có pha nhận câu hỏi và phân tích mục tiêu, thay vào đó truy vấn đầu vào của hệ thống đề xuất là các danh từ chỉ thực thể như tên người, sự kiện, địa điểm, địa danh... xác định.

- 2. Loại bỏ dư thừa:** Bước xử lý này nhằm loại bỏ các nhận định không đúng với từ khóa truy vấn. Do đặc điểm miền dữ liệu mô hình áp dụng là trang tin tức VnExpress.Net, theo đó các bình luận trong một trang tin đều được kiểm duyệt bởi biên tập viên nên các bài bình luận đều tập trung vào chủ đề. Vì vậy trong mô hình đề xuất, chúng tôi không có bước xử lý loại bỏ dư thừa.

## Mô hình đề xuất:

Trên việc phân tích hai mô hình, chúng tôi đề xuất mô hình cho bài toán như sau:



Hình 5. Mô hình tổng hợp quan điểm dựa trên phương pháp thống kê



Mô hình đề xuất có ba pha chính:

- *Pha thu thập tài liệu liên quan:* Nhận đầu vào là một danh từ chỉ tên thực thể truy vấn của người dùng, gửi truy vấn tới máy tìm kiếm và lấy các kết quả trả về.
- *Pha trích xuất quan điểm:* Từ tập các trang lấy về từ máy tìm kiếm liên quan tới quan điểm, pha này thực hiện trích xuất ra các tài liệu, các đoạn bình luận có liên quan tới từ khóa truy vấn phục vụ cho pha tổng hợp phía sau.
- *Pha tổng hợp quan điểm:* Từ tập các tài liệu và đoạn bình luận liên quan tới truy vấn, pha này tiến hành tổng hợp các quan điểm và đưa ra tổng hợp chia theo năm mức cho người dùng.

### 3.2.2 Phân tích phương pháp và đề xuất

Phương pháp của Sushant Kumar và cộng sự có nhiều ưu điểm, nhưng đối với miền dữ liệu tiếng Việt khi áp dụng có một số vấn đề khó khăn:

**1. Câu truy vấn đầu vào:** Trong phương pháp các tác giả đưa ra, truy vấn đầu vào là dưới dạng câu hỏi. Tuy nhiên với đặc trưng dữ liệu tiếng Việt, để biểu diễn câu truy vấn đầu vào cho hệ thống là một bài toán khó. Do đó, chúng tôi giới hạn truy vấn đầu vào của hệ thống là những danh từ, là tên riêng, tên tổ chức, địa điểm... mà người dùng muốn tìm kiếm quan điểm.

**2. Thuật toán tổng hợp:** Thuật toán tổng hợp của tác giả sử dụng phương pháp tính độ tương tự unigram giữa các câu để xếp hạng. Do đặc trưng tiếng Việt, chúng tôi thực hiện tổng hợp và xếp hạng câu dựa vào từ điển VietSentiWordNet được xây dựng cho miền dữ liệu tin tức tiếng Việt.

Trên cơ sở phương pháp của các tác giả Ấn Độ, kết hợp với phương pháp của Jack G. Conrad và cộng sự chúng tôi đề xuất phương pháp giải quyết cho bài toán như sau:

Về tổng quan, hệ thống có ba pha chính là các pha:

1. Pha trích xuất văn bản
2. Pha trích xuất quan điểm
3. Pha tổng hợp quan điểm

**Pha trích xuất văn bản:** Pha này có nhiệm vụ thu thập dữ liệu liên quan tới truy vấn phục vụ cho pha trích xuất quan điểm.

Pha này thực hiện thu thập các tài liệu trả về từ máy tìm kiếm Google. Với truy vấn đầu vào là danh từ chỉ tên người, tên tổ chức, địa điểm... được đưa vào máy tìm kiếm Google với mẫu truy vấn sẽ là “từ khóa” and “Ý kiến bạn đọc” site:vnexpress.net.

Với truy vấn đầu vào, tiến hành lấy về tập các trang có liên quan tới truy vấn. Với mẫu truy vấn này, nếu có các kết quả trả về, thì các kết quả sẽ là các trang trên VnExpress.Net có chứa các thông tin bình luận của người đọc.

Ví dụ: từ khóa truy vấn là “Rùa Hồ Gươm”, thì từ khóa đưa vào máy tìm kiếm Google sẽ là “Rùa Hồ Gươm” and “Ý kiến bạn đọc” site:vnexpress.net



Hình 6. Truy vấn máy tìm kiếm lấy các trang liên quan

Kết quả truy vấn máy tìm kiếm trả về chứa rất nhiều thông tin bình luận của người đọc. Qua khảo sát và thử với các truy vấn khác nhau. Chúng tôi cho thấy với mẫu truy vấn như trên, dữ liệu bình luận trên trang tin VnExpress.Net là giàu thông tin và với định dạng trả về là HTML thuần, không sử dụng JavaScript giúp dễ dàng trích xuất thông tin.

**Bước tiếp xử lý:** tập các tài liệu trả về từ máy tìm kiếm là các bài trên VnExpress.Net được thực hiện trích chọn lấy ra các thông tin:

- ✓ Tiêu đề bài báo
- ✓ Nội dung bài báo
- ✓ Bình luận

- Tiêu đề bình luận
- Nội dung bình luận
- Người bình luận

Phản bình luận của độc giả được tiến hành tách câu, tách từ phục vụ cho các bước xử lý phía sau. Dữ liệu sau khi trích xuất được lưu như trong phụ lục.

### **Phatrích xuất quan điểm:**

**Thuật toán:** Trên cơ sở áp dụng thuật toán trích xuất quan điểm của Sushant Kumar và cộng sự và có một số thay đổi cho phù hợp với đặc trưng tiếng Việt như sau:

❖ **Trích xuất quan điểm liên quan tới truy vấn:** Do đặc trưng dữ liệu áp dụng là các bài bình luận trên trang tin VnExpress.Net. Qua đánh giá, chúng tôi thấy rằng các bài bình luận của người dùng trong một bài báo đều về một chủ đề xác định. Do trên VnExpress mỗi một bình luận được đăng đều qua bước duyệt của biên tập viên trang báo, theo đó những nội dung bình luận không đúng chủ đề đều được loại bỏ. Đây là một thuận lợi lớn cho chúng tôi áp dụng mô hình. Do đó chúng tôi xác định những câu bình luận liên quan tới truy vấn là:

- Tất cả bình luận trong bài báo nếu từ khóa truy vấn có trong tiêu đề bài báo
- Tất cả các bình luận trong bài báo nếu từ khóa truy vấn là từ đặc trưng cho bài tin.
- Các đoạn bình luận nếu như trong đoạn bình luận có chứa từ khóa truy vấn.

Thuật toán trích xuất các câu quan điểm tập trung vào truy vấn, với thuật toán như sau:

- **Bước 1:** Với mỗi bài báo
- **Bước 2:** Trích ra các danh từ đặc trưng cho tài liệu (top 10) sử dụng trọng số TF-IDF
- **Bước 3:** Lấy các bình luận của bài báo là bình luận về từ khóa truy vấn:
  - Lấy toàn bộ bình luận trong bài báo nếu:
    - Tiêu đề tài liệu có chứa từ khóa
    - Từ khóa là một trong các từ đặc trưng của bài báo
  - Lấy các đoạn bình luận là bình luận về từ khóa nếu trong đoạn bình luận có chứa từ khóa
- **Bước 4:** Tiến hành tách câu và tách từ cho các đoạn bình luận
- **Bước 5:** Với mỗi câu, kiểm tra tất cả các từ trong câu. Nếu có từ nhận định ở vị trí thứ  $k$

- **Bước 6:** Kiểm tra các từ ở vị trí  $(k-2)$  đến  $k$ . Nếu có từ ở trong khoảng này nằm trong danh sách từ phủ định thì tiến hành nhân với trọng số tương ứng.
- **Bước 7:** Kiểm tra các từ ở vị trí  $(k-2)$  đến  $(k+2)$ . Nếu có từ ở trong khoảng này nằm trong danh sách từ nhân thì nhân với trọng số tương ứng.
- **Bước 8:** Tiếp theo tính độ phân cực trung bình của câu bằng hàm số các từ nhận định như theo phương pháp của Ku và Liang được nêu trong [KLC06], kết hợp với đặc trưng tiếng Việt, chúng tôi đưa ra công thức tính quan điểm của câu bằng hàm số của các cụm từ chứa từ quan điểm trong câu như sau:

$$S_p = \frac{1}{n} \sum_{j=1}^n S_{w_j}$$

Trong đó  $S_p$  là điểm nhận định của câu  $p$ ,  $S_{w_j}$  điểm nhận định của cụm từ chứa từ nhận định  $w_j$  và  $n$  là tổng số các từ nhận định trong  $p$ .

Ở bước 6 và 7, bằng thống kê vị trí xuất hiện của các từ thể hiện phủ định và thể hiện độ mạnh, chúng tôi đã đưa ra được nhận xét: các từ phủ định luôn đứng trước từ quan điểm nằm trong khoảng từ  $k-2$  đến  $k$  với  $k$  là vị trí của từ quan điểm; các từ thể hiện độ mạnh của quan điểm có thể xuất hiện trước hoặc sau từ quan điểm, do đó chúng tôi xét khoảng xuất hiện từ thể hiện độ mạnh là từ  $k-2$  đến  $k+2$  với  $k$  là vị trí của từ quan điểm.

Trong phương pháp tính điểm nhận định mức câu, chúng tôi có sử dụng ba từ điển: VietSentiWordNet, Negdict và Strengthdict do Vũ Xuân Sơn và cộng sự xây dựng và phát triển như đã giới thiệu ở phần 3.2.

### **Pha tổng hợp quan điểm:**

Phương pháp tổng hợp của Sushan Kumar và cộng sự là phương pháp tổng hợp dành cho đầu vào truy vấn là dạng câu hỏi, do đó pha tổng hợp của Sushan Kumar là không phù hợp với mô hình đề xuất. Để thực hiện pha tổng hợp quan điểm, chúng tôi dựa trên phương pháp tổng hợp quan điểm được Ku và Liang nêu trong [KLC06], từ đó đề xuất phương pháp tổng hợp mức đoạn bình luận dựa vào phương pháp tổng hợp ở mức câu:

### **Thuật toán tổng hợp:**

1. Đối với mỗi đoạn bình luận

2. Quyết định xu hướng quan điểm của đoạn phụ thuộc vào tính toán xu hướng quan điểm mức câu bên trong như sau:

$$S = \sum_{j=1}^m S_p$$

Trong đó  $S$  và  $S_p$  là điểm nhận định của đoạn và của câu  $p$ ,  $m$  là số lượng câu quan điểm.

3. Dựa vào phương pháp tổng hợp các đoạn bình luận. Kết quả tổng hợp quan điểm cuối cùng đối với mỗi thực thể truy vấn được tổng hợp đưa vào năm lớp (gọi  $S$  là phân cực quan điểm trung bình của mỗi đoạn):

- Rất tích cực: chứa tập đoạn bình luận có trọng số  $S > 1$
- Tích cực: chứa tập đoạn bình luận có trọng số  $0,3 < S < 1$
- Trung lập: chứa tập đoạn bình luận có trọng số  $-0,3 < S < 0,3$
- Tiêu cực: chứa tập đoạn bình luận có trọng số  $-1 < S < -0,3$
- Rất tiêu cực: chứa tập đoạn bình luận có trọng số  $S < -1$

### **Xếp hạng đoạn:**

Các đoạn quan điểm về thực thể truy vấn được xếp hạng theo trọng số quan điểm của đoạn. Theo đó những đoạn quan điểm từ cùng một tài liệu tin tức được ưu tiên xếp cạnh nhau.

### **Tóm tắt chương 3**

Trong chương này, dựa trên phân tích phương pháp tổng hợp quan điểm dựa vào mô hình thống kê và mô hình học máy SVM, khóa luận đã đưa ra được mô hình phù hợp cho tổng hợp tin tức trên miền dữ liệu tiếng Việt.

Trong chương sau, khóa luận mô tả quá trình thực nghiệm mô hình và đánh giá kết quả đạt được. Kết quả thực nghiệm cho thấy mô hình đề xuất là hoàn toàn khả thi.

## Chương 4: Thực nghiệm và đánh giá

Dựa vào mô hình đề xuất ở chương 3, khóa luận tiến hành thu thập dữ liệu, trích xuất đặc trưng cho từng văn bản tin tức và tổng hợp quan điểm liên quan tới truy vấn.

### 4.1. Môi trường và các công cụ sử dụng thực nghiệm

#### *Cấu hình phần cứng*

*Bảng 10. Cấu hình hệ thống thử nghiệm*

Thành phần	Chỉ số
CPU	2.0 GHz Dual Core Intel
RAM	2GB
OS	Windows 7 Pro
Bộ nhớ ngoài	250GB

#### *Các phần mềm sử dụng*

*Bảng 11. Công cụ phần mềm sử dụng*

STT	Tên phần mềm	Tác giả	Nguồn
1	Eclipse-SDK-3.5-win32		<a href="http://www.eclipse.org/downloads">http://www.eclipse.org/downloads</a>
2	JvnTextPro 2.0	N.C.Tú-P.X.Hiếu- N.T.Trang	<a href="http://jvntextpro.sourceforge.net/">http://jvntextpro.sourceforge.net/</a>
3	LingPie 4.0.1		<a href="http://alias-i.com/lingpipe/web/download.html">http://alias-i.com/lingpipe/web/download.html</a>

## 4.2 Dữ liệu thử nghiệm

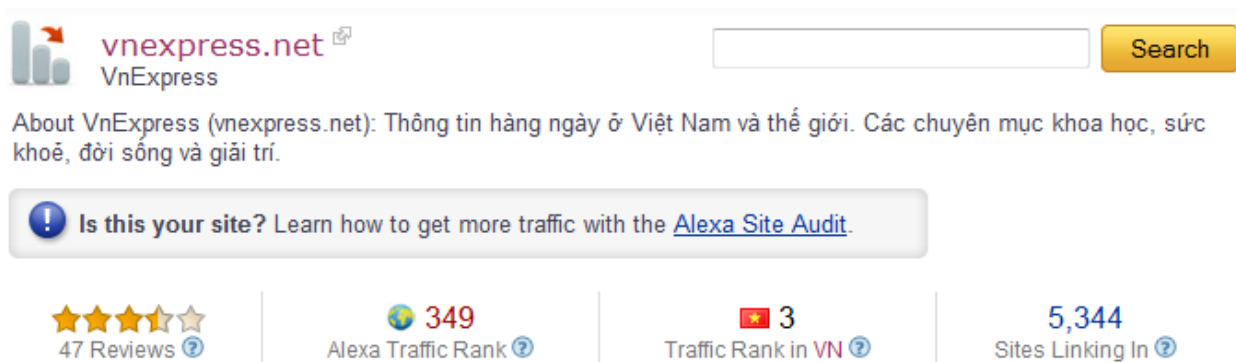
### 4.2.1 Đặc trưng trang tin tức VnExpress

Trong nội dung khóa luận, chúng tôi thực nghiệm trên dữ liệu từ trang tin tức VnExpress.Net. Trong phần này, chúng tôi giới thiệu cấu trúc bài viết và bình luận của người dùng trên trang tin VnExpress.Net.

#### Giới thiệu về VnExpress.Net:

VnExpress được thành lập bởi tập đoàn FPT vào ngày 26/2/2011 và được Bộ Thông tin và Truyền thông cấp giấy phép số 511/GP-BVHTT ngày 25/11/2002.

VnExpress là tờ báo điện tử đầu tiên tại Việt Nam không có phiên bản báo giấy. Tính tới thời điểm viết báo cáo này, theo bảng xếp hạng của Alexa, VnExpress luôn có số người truy cập lớn nhất trong số hơn mười tờ báo điện tử tại Việt Nam và nằm trong top 400 website được truy cập nhiều nhất thế giới:



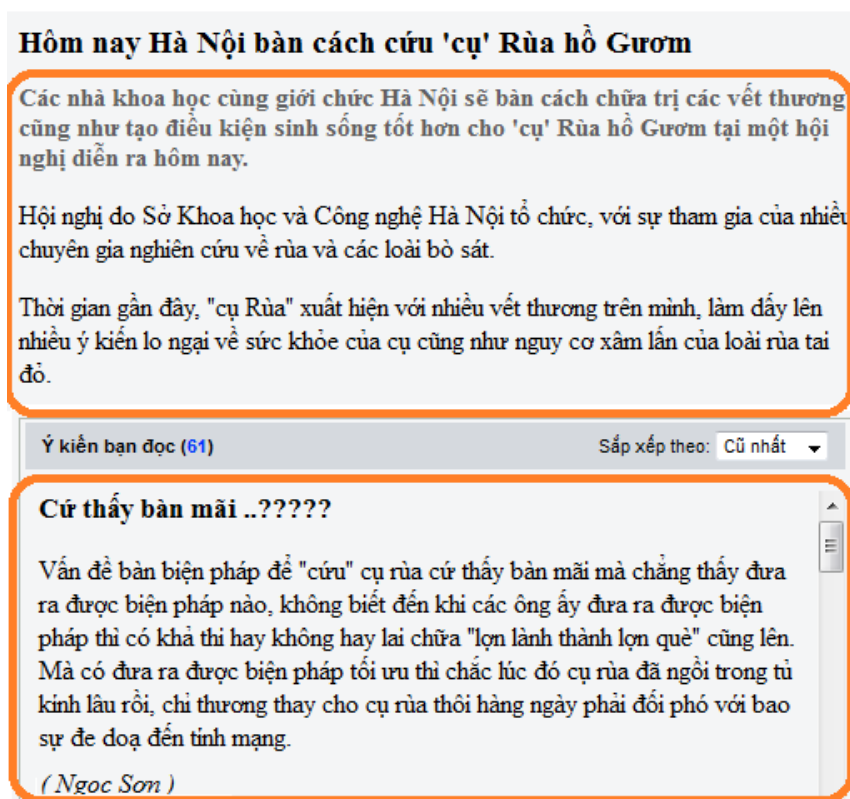
Hình 7: Bảng xếp hạng của VnExpress.Net trên Alexa

Hình 7 cho thấy thứ hạng của VnExpress.Net đứng thứ 349 trên thế giới, và đứng thứ 3 Việt Nam.

Bài viết đa dạng, cùng số lượng lớn các bình luận của người đọc về tất cả các lĩnh vực đời sống xã hội, trang tin VnExpress là kho dữ liệu tốt cho chúng tôi thực hiện thử nghiệm mô hình hệ thống đề xuất. Kết quả ở phần đánh giá thực nghiệm càng khẳng định sự lựa chọn trang tin VnExpress cho việc thử nghiệm mô hình là hoàn toàn đúng đắn.

## Cấu trúc bài tin trên VnExpress:

Một bài tin tức trên VnExpress.Net sử dụng font chữ **Time New Roman** với kích thước font chữ là **11.8pt**. Ví dụ một bài tin tiêu đề “*Hôm nay Hà Nội bàn cách cứu ‘cụ’ Rùa hồ Gươm*”<sup>6</sup> được trình bày như hình dưới:



Hình 8: Một bài tin trên trang VnExpress.Net

Các bài tin trên trang VnExpress.Net có các phần quan trọng:

- ✓ Tiêu đề bài báo
- ✓ Nội dung bài báo
- ✓ Bình luận của người đọc
  - Tiêu đề bình luận
  - Nội dung bình luận
  - Người bình luận

Bảng dưới cho thấy việc tổ chức các thành phần trong bài tin của VnExpress là hoàn toàn có cấu trúc và dễ dàng cho việc trích chọn thông tin:

<sup>6</sup><http://vnexpress.net/gl/khoa-hoc/2011/02/hom-nay-ha-noi-ban-cach-cuu-cu-rua-ho-guom/>



Bảng 12: Thành phần trong bài tin và định dạng HTML

Thành phần	Định dạng HTML
Tiêu đề bài báo	<code>&lt;h1 class="Title"&gt;Tiêu đề bài báo &lt;/h1&gt;</code>
Nội dung bài báo	<code>&lt;div cpms_content="true" style="overflow:hidden"&gt; Nội dung tin tức &lt;/div&gt;</code>
Tiêu đề bình luận	<code>&lt;p class="Title"&gt;Tiêu đề bình luận&lt;/p&gt;</code>
Nội dung bình luận	<code>&lt;p class="Normal"&gt;Nội dung bình luận &lt;/p&gt;</code>
Người bình luận	<code>&lt;p class="Normal"&gt; &lt;i&gt;Người bình luận &lt;/i&gt; &lt;/p&gt;</code>

**Trang tin VnExpress.Net có những ưu điểm:** Nguồn dữ liệu bài tin phong phú, cùng với số lượng lớn các bình luận của người dùng, và đặc biệt là định dạng thông tin có cấu trúc giúp dễ dàng cho nhiệm vụ trích chọn thông tin. Với các đặc điểm trên, trang tin VnExpress là trang tin điển hình để chúng tôi tiến hành thực nghiệm mô hình trên miền dữ liệu tin tức tiếng Việt.

#### 4.2.2 Thu thập dữ liệu

Dữ liệu thử nghiệm được lấy về từ trang báo điện tử <http://vnexpress.net>. Sử dụng phần mềm IDM Grabber để lấy dữ liệu với link đầu vào là trang chủ <http://vnexpress.net/>:

- ✓ Trong Tasks của công cụ download IDM, chọn Run Site Grabber. Mỗi Site Grabber gồm 4 bước:
  - Chọn địa chỉ trang web cần download: Thực nghiệm chọn trang <http://vnexpress.net/>
  - Chọn nơi lưu dữ liệu download về.
  - Thiết lập bộ lọc link levels: Thực nghiệm chọn duyệt độ sâu trang tới mức 2 và không lấy từ links bên ngoài trang.
  - Thiết lập điều kiện lọc nâng cao: Bằng khảo sát dữ liệu, chúng tôi thấy các trang sau khi tải về nội dung html có kích thước < 20k đều là những trang có ít thông tin. Do đó, chúng tôi thiết lập ràng buộc các trang lấy về có kích thước > 20kb.
  - Thiết lập bộ lọc file cần download: Thực nghiệm chọn download file \*.html, \*.htm.
- ✓ Sau khi thiết lập các thông số. IDM sẽ tự động download dữ liệu từ trang web đã thiết lập.

Dữ liệu các trang web lấy về sau khi loại bỏ dữ liệu nhiễu (là các trang web ít thông tin) có tổng số là 1.548 bài báo trong đó có 214 file có bình luận của người đọc chiếm 13.83%.

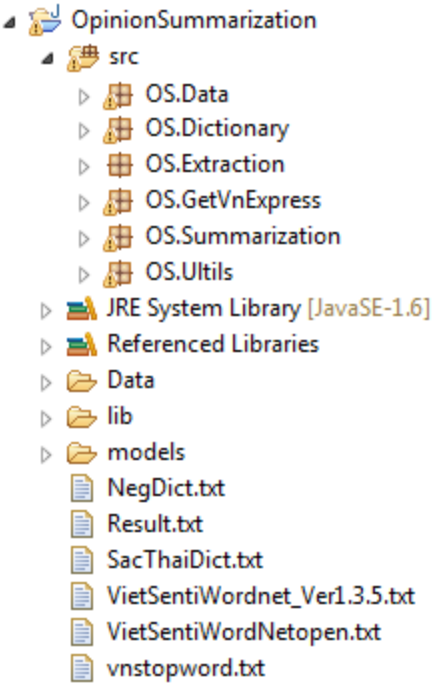
### 4.3 Thực nghiệm

Trong phần này, khóa luận đưa ra một số kết quả thực nghiệm để chứng minh cho tính đúng đắn và tính thực tiễn của mô hình đề xuất. Thực nghiệm được xây dựng theo đúng mô hình đề xuất.

#### 4.3.1 Mô tả cài đặt chương trình

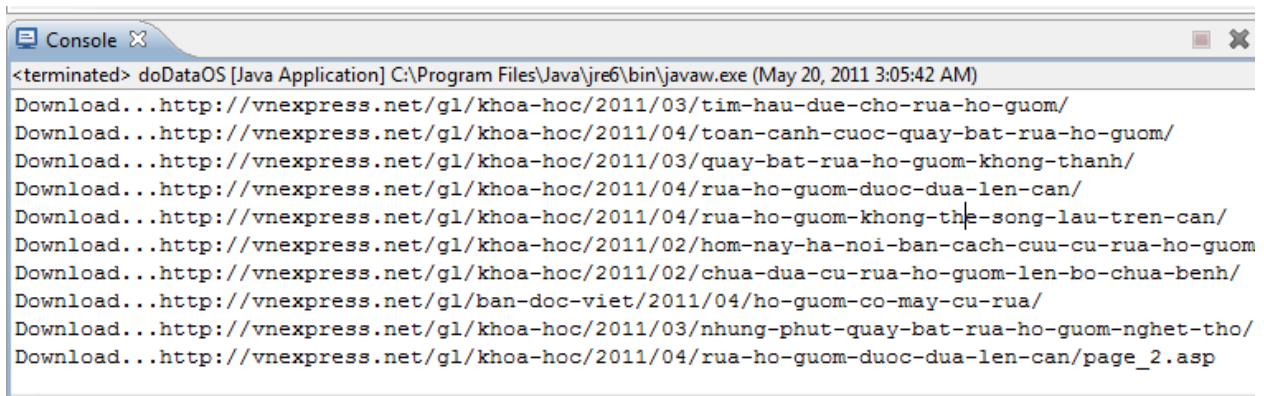
Chương trình được tổ chức thành 5 gói:

*Bảng 13: Các gói cài đặt trong thực nghiệm*

<b>OS.Data:</b> thực hiện các thao tác với tài liệu: chuyển các tài liệu thành các đối tượng, thao tác với đối tượng...	
<b>OS.GetVnExpress:</b> thực hiện truy vấn và lấy các tài liệu liên quan tới truy vấn được trả về từ Google.	
<b>OS.Extraction:</b> thực hiện trích xuất các tài liệu liên quan tới từ khóa truy vấn	
<b>OS.Summarization:</b> thực hiện tổng hợp quan điểm và đưa ra output cho người dùng.	
<b>OS.Dictionary:</b> thực hiện các thao tác với từ điển VietSentiWordNet để khai phá nhận định sử dụng từ điển	
<b>OS.Utills:</b> gói này chứa các thư viện hỗ trợ cho các bước xử lý bên trên.	

#### 4.3.2 Thực nghiệm hệ thống

**Thực nghiệm pha thu thập tài liệu liên quan:** với truy vấn đầu vào là “Rùa Hồ Gươm”, pha này gửi truy vấn lên máy tìm kiếm Google và lấy về được tập 100 trang web trả về từ máy tìm kiếm:



Hình 9: Thực nghiệm pha thu thập tài liệu liên quan

**Thực nghiệm bước tiền xử lý dữ liệu:** Dữ liệu lấy về là tập các trang web với định dạng html được trích xuất ra các thành phần:

- ✓ Tiêu đề bài báo
- ✓ Nội dung bài báo
- ✓ Bình luận
  - Tiêu đề bình luận
  - Nội dung bình luận
  - Người bình luận

Ví dụ một tài liệu sau khi được tiền xử lý:

```
.news>
→
<content>
→→<title>Rùa tai đỏ xâm nhập khắp hồ Guom</title>
→→<body_news>
Dù Hà Nội đã có văn bản yêu cầu thu gom, tiêu hủy rùa tai đỏ, loài xâm
Quốc Tử Giám, loài vật này ngày một nhiều.
→→</body_news>
→</content>
<comment>
→<title_cm>Nên có biện pháp tiêu hủy loài rùa này</title_cm>
→<body_cm>
→→Tôi mong là ai biết về loài này sẽ nói lại với những ai mua rùa
bán và việc thu gom tiêu hủy sẽ có hiệu quả hơn.
→</body_cm>
→<holder>(Thuy Linh)</holder>
</comment>
<comment>
→<title_cm>Quản lý thiếu chặt chẽ</title_cm>
→<body_cm>
→→Tôi cũng không hiểu tại sao quản lý kiểu gì mà lại để cho rùa t
→</body_cm>
→<holder>(hung)</holder>
</comment>
```

Hình 10: Ví dụ một tài liệu sau bước tiền xử lý

**Thực nghiệm pha trích xuất quan điểm:** Thử nghiệm với tập dữ liệu đã thu thập được, với truy vấn đầu vào là “Rùa Hồ Gươm”, kết quả có 53 đoạn bình luận liên quan tới từ khóa “Rùa Hồ Gươm”.

```

Console X
<terminated> OpinionSummari [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (May 20, 2011 7:57:26 AM)
rùa hồ gươm
<br>Sao các cụ không nghĩ đến nhân bản vô tính cho nó đơn giản<br>
( toi )
Tìm hậu duệ chi cụ rùa hồ Gươm
<br>Theo em thì có thể nhân bản vô tính!<br>    |
( Đặng Đình Trí )
Gửi ông Nguyễn tiên hung
tìm hậu duệ cho cụ rùa
<br>Ông Hưng ơi ông nghĩ j mà đưa ra ý tưởng rùa máy? Rùa Hồ Gươm là báu vật quốc gia
( Lê Văn Linh )
Cần chữa trị cho Cụ trước!
<br>Mình thấy các bạn đều có ý kiến riêng của mình...nhưng có một số bất ổn<br>
( lehoang1993 )
Có tất cả 53 đoạn bình luận liên quan tới từ khóa rùa hồ gươm
  
```

Hình 11: Thực nghiệm pha trích xuất quan điểm với từ khóa “Rùa Hồ Gươm”

Một số đoạn bình luận trong kết quả được trích ra:

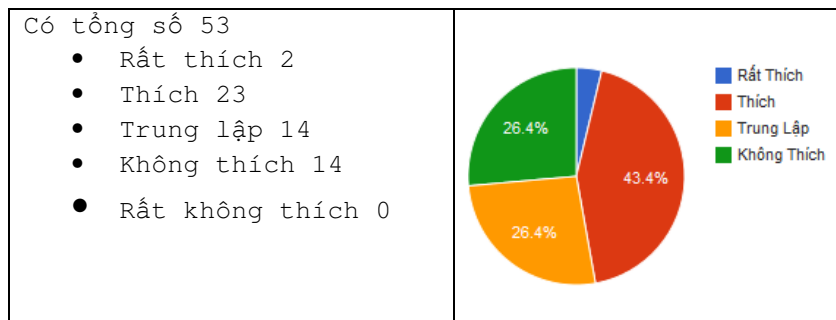
Bảng 14: Một số đoạn bình luận liên quan tới từ khóa “Rùa Hồ Gươm”

Tiêu đề bình luận	Nội dung bình luận	Người bình luận
Nước hồ Gươm	Trong khi chờ đợi đưa cụ Rùa lên bờ để chữa trị vết thương cho cụ, tôi xin góp ý kiến sa83n dịp này nên xử lý nước dưới hồ gươm luôn...	Phan Tấn Lộc
Thương "Cụ Rùa" vì chỉ còn một mình	Chỉ còn mình Cụ trong Hồ mà lâu nay đã trở thành huyền thoại, được đài báo và các cơ quan nói đến nhiều...	Tùng Lê
Gửi Bạn Tấn Lộc - Nước Hồ Gươm	Cụ rùa đã sống ở Hồ Gươm rất lâu nên đã quen với môi trường sống ở đây...	Đỗ Minh Ngọc
Cụ Rùa!	Thật đáng thương cho Cụ .Vì nước Hồ Gươm nên xem lại ,chúng ta cần bảo tồn,và xem	Lê Hiền

	lại rửa tai đỏ! Chúc Cụ mau lành bệnh...	
Nhất trí ý kiến của bạn	Mình xin chia sẻ ý kiến của bạn. Nhân dịp cụ Rửa lên bờ chúng ta cần triệt để làm sạch nước hồ và tiêu diệt bọn rửa tai đỏ khó chịu...	Bùi Phạm Tú Trang

**Thực nghiệm pha tổng hợp quan điểm:** Với từ khóa truy vấn là “Rửa Hồ Gươm” hệ thống cho ra kết quả:

*Bảng 15: Kết quả tổng hợp quan điểm với từ khóa truy vấn “Rửa Hồ Gươm”*



#### 4.3.3 Đánh giá kết quả thực nghiệm

Chúng tôi thực hiện phương pháp đánh giá tổng hợp của hệ thống như [DK08]. Hai nhà báo Phạm Thị Hồng Anh<sup>7</sup> và Nguyễn Thị Nguyệt<sup>8</sup> được chọn làm chuyên gia về tin tức (cả hai đều là nhà báo và nhà biên tập có kinh nghiệm trong lĩnh vực tin tức) để cho điểm đánh giá kết quả tổng hợp theo hai độ đo.

1. Khả năng trả lời truy vấn của hệ thống (Độ\_đo\_1)
2. Chất lượng ngôn ngữ của kết quả trả về (độ\_đo\_2)

Hai độ đo này được xây dựng trên thang điểm 5. Với tiêu chí điểm cho mỗi độ đo như trong bảng 16 và 17:

<sup>7</sup>Phóng viên báo Người Đại Biểu Nhân Dân, cơ quan ngôn luận của Quốc Hội Việt Nam. Email: [anhph@qh.gov.vn](mailto:anhph@qh.gov.vn)

<sup>8</sup> Biên tập viên báo Việt Báo. Email: [nguyennt@vietbao.vn](mailto:nguyennt@vietbao.vn)

*Bảng 16: Thang điểm đánh giá khả năng trả lời câu hỏi của hệ thống đề xuất*

Bậc	Ý nghĩa	Mô tả
5	Rất tốt	Kết quả tổng hợp tập trung vào từ khóa truy vấn, các câu có chứa phân cực quan điểm về từ khóa
4	Tốt	Kết quả tổng hợp có liên quan tới từ khóa, tuy nhiên quan điểm không tập trung, có sự phân cực quan điểm
3	Trung bình	Hơi liên quan tới từ khóa và có sự phân cực quan điểm
2	Kém	Kết quả tổng hợp bị trùng lặp và có phân cực quan điểm
1	Rất kém	Không tập trung vào câu hỏi, phân cực về một phía (chỉ có một trong các quan điểm tích cực, tiêu cực hoặc trung lập)

Với **Độ đo 1** (độ đo khả năng trả lời câu hỏi của hệ thống): dựa vào cột mô tả, các chuyên gia tiến hành phân loại cho kết quả trả về theo mức đoạn theo các bậc. Bậc này tương ứng với số điểm của từng đoạn. Điểm độ đo 1 cho toàn bộ tổng hợp được tính bằng công thức:

$$\text{Đ\_đo\_1} = \frac{\text{Tng điểm đ đo 1 cho tt c các đơn}}{\text{Tng s đơn}}$$

*Bảng 17: Thang điểm đánh giá chất lượng ngôn ngữ học*

Điểm	Tiêu chí	Điểm quan tâm
+1	Ngữ pháp	Định dạng bài viết, các thiếu sót, lỗi chính tả, ...
+1	Không có thông tin dư thừa	Không có sự lặp lại nội dung, sự kiện, cụm danh từ...
+1	Câu viết rõ ràng	Dễ dàng nhận dạng đại từ và cụm danh từ...
+1	Tính tập trung	Quan điểm tập trung, rõ ràng, thông tin đầy đủ...
+1	Cấu trúc	Bài viết có cấu trúc, các câu có sự

		liên quan tới nhau..
--	--	----------------------

**Với Độ\_đo\_2**(*độ đo chất lượng ngôn ngữ học*): các chuyên gia tiến hành phân tích chất lượng ngôn ngữ của kết quả tổng hợp theo mức đoạn. Với mỗi tiêu chí đánh giá đạt được thì cộng thêm 1 điểm. Điểm tổng hợp độ đo hai kết quả tổng hợp được tính bằng công thức:

$$\text{Đ đo 2} = \frac{Tng \text{ đim đ đo 2 cho tt c các đơn}}{Tng s \text{ đơn}}$$

Điểm đánh giá tổng hợp cuối cùng cho một truy vấn được tính bằng cách lấy trung bình điểm *Độ\_đo\_1* và *Độ\_đo\_2*.

Tiến hành đánh giá kết quả tổng hợp với năm truy vấn đầu vào theo hai độ đo với các tiêu chí chấm điểm như trên, ta có bảng tổng hợp kết quả đánh giá như sau:

*Bảng 18: Kết quả đánh giá thực nghiệm với 5 truy vấn*

Truy vấn đầu vào	Số lượng đoạn bình luận liên quan tới truy vấn	Điểm độ đo 1	Điểm độ đo 2	Điểm trung bình
Cụ Rùa	231	3	4	3.5
Nữ sinh	163	5	3	4
Clip	280	5	4	4.5
CSGT	320	4	3	3.5
Uyên Linh	190	4	4	4
Trung bình trung toàn hệ thống với 5 truy vấn				3.9

Từ bảng đánh giá kết quả của hệ thống cho thấy với mức điểm là trên 3. Kết quả đánh giá cho thấy mô hình hệ thống đề xuất là khả quan, có khả năng áp dụng vào thực tế.

#### **Tóm tắt chương 4**

Trong chương này, chúng tôi đã tiến hành thực nghiệm, xem xét và đánh giá kết quả của quá trình thử nghiệm mô hình tổng hợp quan điểm dựa phương pháp thống kê áp dụng cho văn bản tin tức tiếng Việt. Qua phân tích và đánh giá thực nghiệm đã cho thấy tính đúng đắn của phương pháp sử dụng trong khóa luận.

## Kết luận và định hướng phát triển

### Kết quả đạt được của khóa luận:

Đã cài đặt, thử nghiệm ban đầu trên một tập dữ liệu là các trang tin VnExpress. Với mô hình và phương pháp đề xuất, hệ thống hoàn có thể mở rộng sang tất cả các văn bản tin tức trên các trang báo điện tử khác với các bổ xung trong việc tiền xử lý văn bản đầu vào. Kết quả đánh giá mô hình cho thấy hệ thống có khả năng phát triển và ứng dụng trong thực tế.

Hiện tại, bài toán khai phá quan điểm trên văn bản tin tức tiếng Việt còn, với mô hình và phương pháp đề xuất bước đầu tiếp cận và định hướng phát triển khai phá quan điểm tin tức tiếng Việt.

**Các vấn đề chưa đạt được:** bên cạnh các kết quả đạt được, do hạn chế về mặt thời gian và kiến thức, khóa luận vẫn còn các hạn chế sau:

- **Truy vấn đầu vào hệ thống:** hệ thống vẫn còn hạn chế khi truy vấn đầu vào của người dùng bắt buộc phải là các danh từ chỉ tên thực thể xác định. Với giới hạn này, hệ thống chưa thể tổng hợp theo hướng quan điểm người dùng quan tâm tức là người dùng chỉ muốn biết các quan điểm về một khía cạnh nào đó của thực thể chứ không quan tâm tới toàn bộ quan điểm về thực thể.

#### Ví dụ:

Khi người dùng muốn chỉ muốn tìm các quan điểm cụ thể về “*Nữ Sinh đánh nhau*” thay vì toàn bộ quan điểm về “*Nữ sinh*” như hiện tại.

- **Phương pháp tổng hợp quan điểm:** phương pháp tổng hợp quan điểm hiện tại của hệ thống còn hạn chế do từ điển VietSentiWordNet chưa bao quát hết miền dữ liệu tin tức.

### Định hướng tương lai:

sẽ tiến hành phát triển và đưa thêm một số pha xử lý để hệ thống có thể nhận truy vấn đầu vào dưới dạng ngôn ngữ tự nhiên thay vì danh từ chỉ tên thực thể xác định như hiện tại.



Mở rộng từ điển cho phương pháp tổng hợp đạt kết quả chính xác hơn và có thể áp dụng cho nhiều miền dữ liệu khác nhau.

Triển khai hệ thống vào thực tế với đầy đủ các pha xử lý như một máy tìm kiếm quan điểm đầu tiên cho tin tức tiếng Việt.

## Tài liệu tham khảo

### Tài liệu tiếng Việt

[THST09] Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2009). *Giáo trình khai phá dữ liệu Web*, Nhà xuất bản giáo dục Việt Nam, 2009.

[SHH11] Vũ Xuân Sơn, Trần Trung Hiếu, Lê Thu Hà, Đào Thủy Ngân. “Xây dựng từ điển VietSentiWordNet ứng dụng khai phá quan điểm trên tin tức”. Công trình SVNCKH năm 2011, Đại Học Công Nghệ, ĐHQGHN.

### Tài liệu tiếng Anh

[ADSB10] Amitava Das, Sivaji Bandyopadhyay, “Topic-Based Bengali OpinionSummarization”, 2010

[EHM10] Elena, Horacio, Manuel. “Experiments on Summary-based Opinion Classification”, 2010

[TWU10] Thelwall, M., Wilkinson, D. & Uppal, S. Data mining emotion in social network communication: Gender differences in MySpace, *Journal of the American Society for Information Science and Technology*, 61(1), 190-199.

[BO09] Bruno Ohana. “Opinion mining with the SentWordNet lexical resource”, 2009

[JGR09] Jackie, Giuseppe, Raymond. “Optimization-based Content Selection for Opinion Summarization”, 2009

[KSR09] Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the European Association for Computational Linguistics (EACL 2009)*, Athens, Greece, 2009. ACL

- [DK08] Hoa Trang Dang and Karolina Owczarzak. “Overview of the TAC 2008 Update Summarization Task”, 2008
- [FRJJ08] Frank Schilder, Ravikumar Kondadadi, Jochen L. Leidner, and Jack G. Conrad. Thomson Reuters at TAC 2008: Aggressive filtering with FastSum for update and opinion summarization. *In Proceedings of the First Text Analysis Conference (TAC 2008)*, pages 396–405, Gaithersburg, MD, 2008. NIST
- [JJLF08] Jack G. Conrad, Jochen L. Leidner, Frank Schilder, Ravi Kondadadi. “Query-based Opinion Summarization for Legal Blog Entries”, 2008
- [BoLee08] Bo Pang, Lillian Lee. “Opinion Mining and Sentiment Analysis”, 2008
- [AMT08] Aurélien Bossard, Michel Génèreux and Thierry Poibeau. “CBSEAS, a Summarization System Integration of Opinion Mining Techniques to Summarize Blogs”. *TAC 2008*
- [PSS08] Prof. Sudeshna Sarkar. “Multi-Document Update and Opinion Summarization”, 2008
- [SD08] Sushant Kumar and Diptesh Chatterjee. “Statistical Model for Opinion Summarization”, 2008
- [BL07] Bing Liu. “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”. *Chapter II*, 2007
- [AF06] Andrea Esuli, Fabrizio Sebastiani. “SentiWordNet: A public available lexical resource for opinion mining”. LREC’06
- [VC06] Veselin, Claire. “Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning”, 2006
- [KLC06] Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen. “Opinion extraction, summarization and tracking in news and blog corpora”. AAAI 2006
- [HL04] Minqing Hu and Bing Liu. “Mining and Summarizing Customer Reviews”. SIGKDD 2004, pages 168-177. 2004.
- [KH04] Soo-Min Kim and Eduard Hovy. “Determining the Sentiment of Opinions”. Coling, pages 1367-1373. 2004.

- [JR03] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", First International Conference on. Machine Learning, 2003.
- [PLV02]B. Pang, L. Lee and S. Vaithyanathan. "Thumbs up?Sentiment classification using machine learning techniques". *Proceedings of the 2002 Conference on EMNLP*, pages 79-86. 2002.
- [JCD01] John M. Conroy and Dianne P. O’Leary. Text summarization via hidden markov models. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 406–407, New York, NY, USA, 2001. ACM.
- [HM97] Hatzivassiloglou, V. and McKeown, K. Predicting the Semantic Orientation of Adjectives. ACL- EACL’97, 1997
- [KD] KerstinDenecke. "AreSentiWordNetScoresSuitedforMulti-DomainSentimentClassification?"
- [MKG] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai. Sentiment Strength Detection in Short Informal Text