BỘ GIÁO DỤC ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG
***

LƯƠNG QUỐC SƠN

NGHIÊN CỨU XÂY DỰNG BỘ LỌC THƯ RÁC HỖ TRỢ SONG NGỮ ANH - VIỆT

Luận văn thạc sỹ công nghệ thông tin

Đồng Nai, 2012

BỘ GIÁO DỤC ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG
****

LƯƠNG QUỐC SƠN

NGHIÊN CỨU XÂY DỰNG
BỘ LỌC THƯ RÁC
HỖ TRỢ SONG NGỮ ANH - VIỆT

Chuyên ngành: Công nghệ thông tin

Mã số: 60,48,02.01

Luận văn thạc sỹ công nghệ thông tin

Người hướng dẫn khoa học:

TS. VŨ ĐỨC LUNG

Đồng Nai, 2012

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn thạc sỹ công nghệ thông tin "nghiên cứu xây dựng bộ lọc thư rác hỗ trợ song ngữ Anh - Việt" là kết quả của quá trình học tập, nghiên cứu khoa học độc lập, nghiêm túc.

Các số liệu trong luận văn là trung thực, có nguồn gốc rõ ràng, được trích dẫn và có tính kế thừa, phát triển từ các số liệu, tạp chí, các công trình nghiên cứu đã được công bố, trên các website.

Các phương pháp nêu trong luận văn được rút từ những cơ sở lý luận và quá trình nghiên cứu tìm hiểu.

Đồng Nai, tháng 6 năm 2012

Tác giả

Lương Quốc Sơn

LỜI CẢM ƠN

Lời đầu tiên tôi xin chân thành gửi lời cám ơn sâu sắc đến TS.Vũ Đức Lung đã tận tình giúp đỡ tôi trong suốt thời gian học tập vừa qua, đặc biệt là hướng dẫn tôi hoàn thành đề tài này.

Tôi chân thành cám ơn các thầy cô Trung Tâm Thông Tin Tư Liệu, trường Đại Học Lạc Hồng, nơi tôi công tác và nghiên cứu đã tạo điều kiện và hỗ trợ tôi trong suốt thời gian qua.

Tôi cũng xin chân thành cám ơn các thầy cô khoa công nghệ thông tin đã tận tình giảng dạy, chỉ bảo và cung cấp cho tôi những kiến thức hết sức cần thiết trong suốt thời gian học, và cũng xin gởi lời cám ơn chân thành đến những người thân, bạn bè và đồng nghiệp đã giúp đỡ và động viên tôi trong suốt thời gian học tập cũng như trong thời gian thực hiện luận văn.

Chân thành cám ơn !

Biên Hòa, ngày 05 tháng 06 năm 2012

Lương Quốc Sơn

MỞ ĐẦU

1.Tóm lược đề tài:

Thư rác (spam) là thư điện tử được gửi hàng loạt với nội dung mà người nhận không mong đợi, không muốn xem, hay chứa những nội dung không liên quan đến người nhận và thường được sử dụng để gửi thông tin quảng cáo. Do có giá thành tương đối thấp so với các phương pháp quảng cáo khác, thư rác hiện chiếm một tỷ lệ lớn và ngày càng tăng trong tổng số thư điện tử được gửi qua Internet. Sự xuất hiện và gia tăng thư rác không những gây khó chịu và làm mất thời gian của người nhận mà còn ảnh hưởng tới đường truyền Internet và làm chậm tốc độ xử lý của máy chủ thư điện tử, gây thiệt hại lớn về kinh tế.

Để loại bỏ hoặc giảm thiểu ảnh hưởng của thư rác, nhiều cách tiếp cận khác nhau đã được nghiên cứu và sử dụng. Giải pháp đấu tranh với thư rác rất đa dạng, bao gồm từ các cố gắng về pháp lý trong việc xây dựng luật ngăn chặn phát tán thư rác cho tới những giải pháp kỹ thuật nhằm phát hiện và ngăn chặn thư rác trong những giai đoạn khác nhau của quá trình tạo và phát tán thư. Trong số giải pháp được sử dụng, lọc thư theo nội dung đang là một trong những giải pháp được sử dụng rộng rãi và có triển vọng nhất. Lọc thư theo nội dung là phương pháp phân tích nội dung thư để phân biệt thư rác với thư bình thường, kết quả phân tích sau đó được sử dụng để quyết định chuyển tiếp thư đến người nhận hay không (trong phạm vi nghiên cứu này, nội dung thư được giới hạn là những nội dung trình bày dưới dạng văn bản).

Do việc lọc theo nội dung đòi hỏi phân tích phần văn bản chứa trong tiêu đề hay nội dung thư, thuật toán lọc nội dung cần được xây dựng phù hợp với ngôn ngữ mà thư sử dụng. Hiện nay, nhiều thuật toán lọc nội dung hiệu quả đã được nghiên cứu và sử dụng cho thư viết bằng tiếng Anh.

Trong vòng vài năm gần đây, việc sử dụng Internet nói chung và thư điện tử nói riêng ngày càng phổ biến tại Việt nam. Một trong những hệ quả của sự phát triển này là ngày càng có nhiều thư rác gửi tới các tài khoản thư điện tử tại Việt nam (tài khoản có đuôi .vn). Những thư rác này bao gồm cả thư viết bằng tiếng Anh và thư viết bằng tiếng Việt. Việc xuất hiện ngày càng nhiều thư rác tiếng Việt đặt ra yêu cầu cấp thiết phải có những phương pháp lọc thư có thể xử lý được thư rác loại này.

Do các thuật toán lọc thư thông dụng mới chỉ được nghiên cứu và thử nghiệm cho tiếng Anh, để có thể sử dụng giải pháp lọc nội dung cho thư tiếng Việt cần nghiên cứu làm rõ hiệu quả của thuật toán khi phân tích nội dung thư viết bằng tiếng Việt. Bên cạnh đó cần thực hiện những cải tiến cho phù hợp khi chuyển từ phân loại nội dung tiếng Anh sang phân loại nội dung tiếng Việt. Để giải quyết những vấn đề vừa nêu, trong phạm vi đề tài này, chỉ chú trọng nghiên cứu một số giải pháp lọc nội dung cho thư rác tiếng Việt và tiếng Anh. Nội dung nghiên cứu bao gồm thử nghiệm làm rõ khả năng lọc thư tiếng Việt, đề xuất và phân tích so sánh các cải tiến với thuật toán, thử nghiệm trên dữ liệu thực. Sau khi thử nghiệm so sánh, giải pháp lọc thư có hiệu quả cao sẽ được cài đặt trong một bộ lọc thư có khả năng tích hợp vào máy chủ thư điện tử.

2. Mục tiêu đề tài

Nghiên cứu tổng quan các phương pháp lọc thư rác thông dụng hiện nay, từ đó đề xuất mô hình và xây dựng chương trình thử nghiệm lọc các thư rác được viết bằng tiếng Anh hoặc tiếng Việt.

Bên cạnh đó, cũng nghiên cứu kỹ thuật tách câu, tách từ đơn, từ ghép trong tiếng Việt mà chỉ xét về mặt tồn tại của từ, không xét về mặt ý nghĩa của từ.

3. Nội dung thực hiện đề tài

Tìm hiểu về thư spam: các loại thư spam, đặc điểm thư spam… Đặc biệt, tìm hiểu về thư spam tiếng Việt.

Nghiên cứu các kỹ thuật đang sử dụng hiện nay để lọc thư spam.

Nghiên cứu các bộ lọc spam truyền thống hiện được sử dụng.

Áp dụng thuật toán cho việc lọc thư spam Anh – Việt.

Nghiên cứu xây dựng bộ lọc spam cải tiến từ các bộ lọc đã có hoặc bộ lọc spam mới phù hợp hơn với các thư spam đặc biệt là tiếng Việt.

4. Phạm vi ứng dụng

Đề tài "**NGHIÊN CỨU XÂY DỰNG BỘ LỌC THƯ RÁC HỖ TRỢ SONG NGỮ ANH - VIỆT**" có thể được ứng dụng trong các máy chủ mail, và giúp cho người sử dụng loại bỏ được thư có nội dung spam song ngữ Anh – Việt, giảm thiểu thời gian lãng phí của người sử dụng khi phải tự mình kiểm tra nội dung từng bức thư.

CHƯƠNG 1: NGHIÊN CỨU TỔNG QUAN VỀ THƯ RÁC

1.1. Giới thiệu về thư rác

1.1.1. Lịch sử

Có thể chia lịch sử của thư rác thành 3 giai đoạn sau:

1.1.1.1. Giai đoạn thứ nhất – những năm đầu của thư rác

Nhiều ý kiến cho rằng thư rác đầu tiên được phát tán trên mạng diện rộng là vào năm 1978, đó là một quảng cáo từ Digital Equipment Corporation (DEC) [5]. Do dịch vụ thư điện tử lúc này chưa tiên tiến nên người phát tán thư rác (spammer) này phải đánh thủ công các địa chỉ thư điện tử muốn gửi và chỉ có khoảng 320 trong tổng số các địa chỉ thư điện tử mà spammer muốn gửi nhận được thư rác này trong lần phát tán đầu tiên. Vào 1988 xuất hiện kiểu thư rác khác là thư rác lừa đảo (như lừa đảo làm việc từ thiện, lừa đảo về kiếm tiền).

1.1.1.2. Giai đoạn thứ hai – thư rác được gửi thông qua phần mềm

Đầu thập niên 1990, với sự phát triển của Internet mang đến vấn nạn là số lượng thư rác tăng lên nhanh chóng. Lúc này các spammer dùng các phần mềm để tự động việc gửi thư điện tử đến một danh sách các địa chỉ. Ví dụ về thư rác phát tán qua các phần mềm tự động là thư rác Jesus và thư rác Cantel và Siegel.

Vào 1995 Jeff Slaton – tự nhận mình là "vua thư rác" , ông là một trong những người đầu tiên kiếm lợi nhuận từ các thư rác mà ông gửi đi, ông còn ép buộc các nạn nhân của mình trả phí nếu không muốn nhận thư rác. Việc làm của ông tạo ra ý tưởng cho các công ty thương mại là thuê những người như Jeff Slaton để phát tán thư điện tử với mục đích là quảng cáo giúp họ.

1.1.1.3. Giai đoạn thứ ba – phần mềm chống thư rác chống lại các phần mềm gửi thư rác

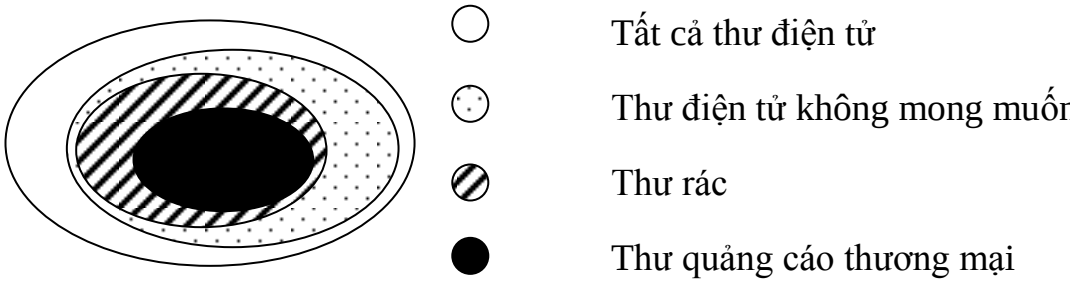
Vào 1996 xuất hiện các phần mềm chống thư rác đầu tiên như Spamblock, Internet Death Penalty, tuy nhiên vẫn không làm giảm sự phát triển của thư rác. Các địa chỉ thư điện tử của người dung được rao bán cho các công ty, tổ chức muốn thực hiện quảng cáo trên thư điện tử. Và từ 1997 đến nay sự phát triển của thư rác đã vượt quá sự kiểm soát, một thống kê cho thấy 97% tổng số thư điện tử được gửi trên mạng là các thư không mong muốn nhận từ người dùng. [9]

1.1.2. Định nghĩa

Có nhiều tranh cãi về việc đâu là định nghĩa chính xác của thư rác (spam email), bởi vì thư rác mang tính cá nhân hóa nên khó mà nói lên được hết ý nghĩa của thư rác. Nhiều ý kiến cho rằng thư rác là những "thư điện tử (email) không mong muốn". Định nghĩa này cũng không thực sự chính xác, như một nhân viên nhận những thư điện tử về công việc từ sếp của họ, đây là những thư điện tử người nhân viên không mong muốn nhưng chúng không phải là thư rác. Lại có ý kiến khác cho rằng thư rác là những "thư điện tử thương mại không được yêu cầu từ phía người nhận" - những thư này bao gồm các thư điện tử quảng cáo về các sản phẩm và thư điện tử lừa gạt. Nhưng định nghĩa này cũng không thực sự chính xác, nó làm mọi người nghĩ rằng thư rác giống như là thư đáng bỏ đi (junk mail).

Sau đó có ý kiến cho rằng thư rác là "số lượng lớn thư điện tử không yêu cầu" và trong số đó các thư điện tử quảng cáo, thương mại chiếm đa số, đây có lẽ là định nghĩa gần đúng với ý nghĩa của thư rác nhất. [7]

Hình vẽ sau sẽ thể hiện rõ định nghĩa của thư rác:



Hình 1.1: So sánh thư rác với các thư điện tử khác.

1.1.3. Mục đích chính gửi thư rác

Thư rác được gửi với các mục đích chính như sau:

- Quảng cáo sản phẩm, dịch vụ, … của tổ chức, công ty thương mại nào đó.
- Lợi dụng sự cả tin của người dùng để lừa gạt họ, như các hình thức kiếm tiền trực tuyến, …
- Gửi kèm virus trong tập tin kèm theo của thư điện tử, từ đó đưa virus vào máy nạn nhân và hệ thống mạng mà nạn nhân sử dụng. Sau đó lấy cắp các thông tin quan trọng của nạn nhân và hệ thống.
- Nói xấu, xuyên tạc ai đó, tuyên truyền những điều sai trái về chính trị.

…

1.1.4. Các đặc tính của thư rác

Thư rác chứa các đặc tính cơ bản sau:

- Thư rác mang tính tương đối vì thư mang tính cá nhân, có thể một thư điện tử này là vô bổ với người này nhưng với người khác lại có ích. Ví dụ một thư điện tử quảng cáo/ rao vặt cho một sản phẩm cụ thể có thể được một số người quan tâm nhưng những người còn lại xem đó là rác.
- Tính bất biến trong một thư rác thể hiện ở những từ cụm từ hầu như không thay đổi trong những lần spam. (Ví dụ: Tên người, tên công ty,

tên sản phẩm, mã sản phẩm, tên website của sản phẩm, địa chỉ lưu trữ/ mua bán sản phẩm, …).

- Đặc tính phần header của thư rác [9]
  - Địa chỉ thư điện tử của người nhận sẽ không thể hiện ở trường "To:" hoặc "Cc:", vì địa chỉ này sẽ được ẩn trong trường "Bcc:", spammer thực hiện hành động này để giấu số lượng lớn các địa chỉ thư điện tử mà spammer muốn gửi thư rác.
  - Để nội dung trống hoặc thiếu trường "To:".
  - Trường "To:" thể hiện một địa chỉ thư điện tử không hợp lệ.
  - Nội dung trường "From:" giống trường "To:".
  - Thiếu trường "From:".
  - Định danh - ID của thư điện tử bị thiếu hoặc là ID giả.
  - Trường "Bcc:" có tồn tại, vì ở các thư điện tử thông thường trường này thường không xuất hiện.
  - Trường "X-mailer" – là trường thể hiện tên phần mềm dùng để gửi thư điện tử, nếu trường này bao gồm tên của phần mềm gửi thư rác quen thuộc thì có thể xác định được là thư rác hay không.
  - X-UIDL header: là một định danh duy nhất được sử dụng bởi các giao thức POP để lấy thư điện tử từ một máy chủ mail. Nó thường được thêm vào giữa các máy chủ mail của người nhận và phần mềm thư điện tử của người nhận, nếu thư đến tại các máy chủ mail mà xuất hiện trường này thì là thư rác.
  - Tồn tại các dòng mã lệnh hoặc khoảng trắng tuần tự. Ví dụ như thêm mã lệnh trên chủ đề của thư và dùng khoảng trắng để giấu.
  - Tồn tại các dòng mã HTML không đúng quy tắc.
- Nội dung của thư chứa các từ thường xuất hiện trong thư rác (kiếm tiền, giàu nhanh, chọn nhanh,...).

- Sự giống nhau ở kích thước/ loại tập tin/ tên tập tin đính kèm thư rác ở các lần spam.

1.1.5. Các kỹ thuật tạo thư rác

Chỉnh sửa phần header của thư rác:

- Nhập địa chỉ của các người nhận thư rác vào trường "Bcc:" thay vì trường "To:" hoặc "Cc:".
- Thể hiện ở trường "To:" địa chỉ thư điện tử không hợp lệ để đánh lừa người nhận.
- Dùng mã HTML và khoảng trắng để che dấu thông tin nhằm mục đích đánh lừa người nhận thư rác.

Chỉnh sửa phần nội dung của thư rác:

- Gửi cùng một văn bản thư rác nhiều lần mà không thay đổi gì hết.
- Đảo một số đoạn trong văn bản thư rác cho lần gửi kế tiếp.
- Xóa bớt một số đoạn trong văn bản thư rác cho lần gửi kế tiếp.
- Thêm một số đoạn trong văn bản thư rác cho lần gửi kế tiếp.
- Thay đổi cách dùng từ nhưng ý nghĩa văn bản thư rác vẫn không đổi.
- Thêm các tag HTML vào văn bản thư rác để vượt qua các bộ lọc email spam.
- Dùng hình ảnh thay cho văn bản để tránh các bộ lọc thư rác thông qua văn bản. (biến dạng chữ để tránh nhận dạng ký tự quang học).

Tổ hợp của các cách trên.

1.2. Các kỹ thuật phát hiện và ngăn chặn thư rác

1.2.1. Kỹ thuật blacklisting

1.2.1.1. Giới thiệu

Một blacklist là một danh sách chứa thông tin các địa chỉ thư điện tử hay địa chỉ IP bị cho là địa chỉ phát tán thư rác. Blacklist còn được gọi là danh sách blackhole.

Trên thế giới có nhiều tổ chức chuyên về lĩnh vực thu thập và cung cấp blacklist của các máy chủ mail được kẻ phát tán thư rác sử dụng. Một số danh sách blacklist được cung cấp miễn phí còn một số khác thì phải mua. Các cơ sở dữ liệu blacklist được phần lớn các nhà cung cấp dịch vụ Internet (ISPs) và các nhà cung cấp dịch vụ băng thông rộng sử dụng để lọc thư rác được gửi vào mạng của họ hay những người dùng dịch vụ của họ.

Có nhiều loại danh sách blackhole khác nhau (IP blacklist, DNS blacklist, email blacklist) đưa đến nhiều mức độ lọc khác nhau trong cộng đồng mạng, cho các ISP tự do lựa chọn chính sách lọc thư rác phù hợp với mình. Mỗi blackhole có một tập luật và điều kiện khác nhau để xác định thư rác. Một vài danh sách quá khắt khe và quá nhiều điều kiện dẫn đến rủi ro các thư điện tử hợp lệ bị mất rất cao. (Chỉ nên dùng cho những địa chỉ biết chắc là nơi phát tán thư rác). Các danh sách blackhole có 2 yếu điểm quan trọng:

- Đầu tiên là thời gian lan truyền [7]. Các danh sách blackhole sẽ thêm các địa chỉ mạng vào danh sách của nó chỉ khi mạng đó được dùng để phát tán thư rác. Trước đây việc thêm các mạng đó vào danh sách làm việc tốt do kẻ phát tán thư rác khá bị động. Nhưng ngày nay kẻ phát tán thư rác có thể đánh cắp tài khoản dialup, sử dụng các open relays (Máy trung gian giúp gửi mail) tạo ra các host mới để gửi thư rác trước khi chúng được thêm vào danh sách blackhole. Nhiều danh sách đã bắt đầu blacklist không gian địa chỉ người dùng dialup và ISDN để chống lại các host phát tán thư rác mới này. Tuy nhiên nỗ lực này gặp phải vấn đề lớn là không gian địa chỉ này thường xuyên thay đổi.

- Thứ hai là chất lượng duy trì các danh sách blackhole [7]. Ngày nay nhiều danh sách blackhole được duy trì kém. Kết quả là một vài mạng hợp lệ bị thêm vào blacklist không bao giờ bị xóa, hay chậm xóa. Những vấn đề này làm cho một số blacklist rất không được tin cậy do chúng khóa cả những thư điện tử hợp lệ.

1.2.1.2. Ưu – khuyết điểm

Ưu điểm

- Dễ cài đặt.
- Dễ dành chia sẻ danh sách này cho người khác sử dụng.

Khuyết điểm

- Cần thời gian lan truyền để cập nhật danh sách nên có thể để lọt các thư rác từ những host sử dụng tài khoản dialup bị đánh cắp, open relays hay proxy server.
- Tốn nhiều công sức để duy trì danh sách blacklist.

1.2.1.3. Ghi chú

Chỉ nên dùng các blacklist tin cậy được cập nhật thường xuyên.

Chỉ nên blacklist các địa chỉ biết chắc là nơi phát tán thư rác.

1.2.2. Kỹ thuật whitelisting

1.2.2.1. Giới thiệu

Whitelist là một danh sách các địa chỉ thư điện tử hay địa chỉ IP được coi là không phát tán thư rác. Các danh sách whitelist thường được sử dụng trong các ứng dụng thư điện tử để cho phép người dùng tạo ra danh sách những người mà họ muốn nhận thư điện tử. Danh sách này sẽ ghi đè lên bất cứ danh sách blacklist nào, và nó cho phép thư điện tử được gửi vào inbox của người dùng mà không cần phải lọc như thư rác.

Whitelisting ngược với blacklisting, nó sử dụng một danh sách tin cậy. Theo mặc định mọi người sẽ bị blacklist trừ khi họ có tên trong danh sách whitelist.

Điểm khác biệt lớn nhất giữa kỹ thuật whitelisting và các kỹ thuật lọc nội dung là các kỹ thuật lọc nội dung được dùng để xác định thư rác, còn whitelisting được dùng để xác định người gửi. Hầu hết các whitelist được quản lý riêng bởi mỗi người dùng vì số lượng thư điện tử hợp lệ rất là lớn.

Kỹ thuật whitelisting có độ chính xác 100%, chủ yếu là vì nó chỉ cho phép những địa chỉ rõ ràng đi qua. Điều này là một lợi thế lớn, nhưng cũng có một ý bất lợi. Bởi vì tất cả thư điện tử của người lạ đều bị loại bỏ nên các thư điện tử hợp lệ từ những người muốn liên lạc với một người dùng nào đó cũng sẽ bị loại bỏ [7]. Người dùng đó không hề biết là có người đã cố gắng liên lạc với mình. Có vài cách để khắc phục nhược điểm này. Tạo ra whitelist các địa chỉ thư điện tử và một địa chỉ mail đặc biệt dùng để gửi tới người gửi chưa được whitelist. Một cách khác liên quan đến việc điều tiết người gửi (giới hạn tốc độ và số lượng thông điệp một người chưa được whitelist có thể gửi) và gửi đi một challenge/response (đây là một kỹ thuật khác sẽ được đề cập ở những phần sau).

Nhiều hệ thống whitelisting chỉ tạo danh sách whitelist dựa trên địa chỉ thư điện tử trong phần thông tin của trường “From:”. Điều này giúp phần lớn người dùng dễ dàng thêm các địa chỉ thư điện tử những người bạn của họ vào danh sách whitelist. Trường “From:” được xem là trường tin cậy, nhưng mà trong thực tế nó rất dễ bị giả mạo do bên nhận không chứng thực người gửi. Khi kẻ phát tán thư rác giả mạo một địa chỉ trong whitelist của người dùng, nếu người nhận xóa địa chỉ đó khỏi whitelist thì các thư điện tử từ người thực sự có địa chỉ đó sẽ bị khóa. Ngược lại nếu giữ lại địa chỉ đó thì người nhận sẽ nhận được tất cả các thư rác từ người gửi giả mạo địa chỉ đó. Không có giải

pháp trung gian cho vấn đề này, whitelisting chỉ có thể làm việc hoặc không làm việc.

Nhiều bộ lọc dựa trên nội dung sử dụng kỹ thuật whitelisting trước khi lọc nội dung để tăng cường độ chính xác.

1.2.2.2. Ưu – khuyết điểm

Ưu điểm

- Kết quả rất chính xác.
- Không phải dựa trên việc học nội dung thông điệp.

Khuyết điểm

- Có thể giả mạo địa chỉ trong danh sách whitelist.
- Tất cả người dùng phải được tin cậy mới có thể gửi email vào inbox được.
- Người dùng cần phải cấu hình danh sách whitelist một cách thủ công.

1.2.2.3. Ghi chú

Phù hợp cho những người dùng cần độ chính xác cao mà không bận tâm đến rủi ro có thể mất các email mang lại cơ hội nghề nghiệp hay cơ hội kinh doanh.

1.2.3. Kỹ thuật heuristic filtering

1.2.3.1. Giới thiệu

Phương pháp lọc mail Heuristic được phát triển vào cuối năm 1990. Phương pháp này sử dụng một tập các luật thông dụng nhằm nhận dạng tính chất của thư rác cụ thể nào đó. Các tính chất này có thể nằm trong nội dung hoặc có được do quan sát cấu trúc cụ thể đặc thù của thư rác. Không giống như các bộ lọc nguyên thủy, bộ lọc heristic có các luật để phát hiện cả thư rác lẫn thư hợp lệ. Các thông điệp chỉ có một ít tính chất là thư rác có thể được xem là thư hợp lệ nếu ta không thiết lập cảnh báo cho trường hợp này.

Heristic filtering làm việc dựa trên hàng ngàn luật được định nghĩa trước [4]. Mỗi luật đều được gán một điểm số để biết xác suất thông điệp có phải là thư rác không. Kết quả cuối cùng của biểu thức gọi là Spam Score. Spam score để đo mức độ của thư rác (thấp, trung bình hay cao). Thiết lập mức độ càng cao thì càng lọc được nhiều thư rác, tuy nhiên tỉ lệ false-positive (không phải là thư rác nhưng cho là thư rác) cũng sẽ tăng do các thư điện tử hợp lệ bị coi là thư rác cũng nhiều hơn. Dựa vào Spame Score và một ngưỡng xác định thì các thông điệp được phân lớp thành thư rác, thư hợp lệ và thư chưa xác định. Tuy nhiên cũng có ngoại lệ cho luật này:

- Các thông điệp từ người gửi trong whitelist không bao giờ bị coi là thư rác
- Các thông điệp từ người gửi trong blacklist luôn luôn bị coi là thư rác.

Heristic filtering có hai điểm yếu nghiêm trọng làm giảm hiệu quả của nó:

- Điểm yếu chính xuất phát từ lý do tập luật được thiết kế để mọi người sử dụng. Do đó cần phải cắt giảm một số luật để tránh một số lỗi false-positive quan trọng (các thư hợp lệ bị coi là thư rác). Kết quả là, phiên bản đầu tiên của SpamAssasin có một tỉ lệ lỗi là 1/10 thông điệp, các phiên bản sau này cải thiện chỉ còn 1/20 thông điệp, đạt độ chính xác khoảng 95%.
- Nhưng điểm quan trọng hơn là mọi người sử dụng chung một tập các luật, cho nên kẻ phát tán thư rác có thể học và thích nghi với các luật để vượt qua bộ lọc [7]. Bởi vì các tập luật và các cơ chế gán điểm số hầu như không thay đổi, những kẻ phát tán thư rác có thể tải công cụ heristic phiên bản mới nhất và chạy thử thư rác của chúng. Khi chúng đã xác định được các phần trong thư rác của mình tạo ra đã nằm trong tập luật của phần mềm thì chúng có thể

thay đổi thông điệp đó để qua mặt các luật. Sau khi được chỉnh sửa xong thông điệp sẽ được gửi đi và nó sẽ lọt qua các phần mềm sử dụng cùng tập luật ở trên. Kết quả là độ chính xác giảm nghiêm trọng, một vài nhà quản trị hệ thống cho biết trong một số trường hợp nó có thể giảm xuống 40% [7]. Độ chính xác sẽ tăng khi tác giả bộ lọc thêm các luật mới nhưng cũng sẽ nhanh chóng giảm khi những kẻ phát tán thư rác thích nghi với các luật này.

Các vấn đề cần quan tâm trong kỹ thuật này:

- Vấn đề duy trì [7]: mặc dù nhiều bộ lọc heristic rất hiệu quả trong việc giảm 85% thư rác hoặc hơn nữa, nhưng các tập luật cũng cần phải cập nhật liên tục do sự tiến hóa của thư rác. SpamAssassin sử dụng khoảng 900 đến 950 luật heristic khác nhau, và tập luật mới xuất hiện chỉ có thể duy trì độ chính xác trong khoảng thời gian ngắn. Người quản trị hệ thống không có thời gian để theo dõi 900 luật, vì thế trách nhiệm duy trì tập luật được giao cho những nhà duy trì phần mềm, và chúng ta cần phải cập nhật mỗi lần các luật mới được thêm.
- Vấn đề gán điểm số [7]: một khuyết điểm nữa của cách tiếp cận heristic là mỗi luật được gán một điểm số riêng, điểm số xác định độ quan trọng của luật trong việc phân tích thông điệp. Tuy nhiên, đối với mỗi người dùng độ quan trọng của mỗi luật khác nhau, các điểm số chỉ định nghĩa cho phần lớn cá nhân. Khi thư rác tiến hóa, các điểm số khác có thể tốt hơn, do đó cần nhà quản trị hệ thống điều chỉnh lại ngưỡng xác định thư rác của bộ lọc. Nhưng có lẽ một vấn đề mơ hồ hơn là các điểm số đó không thể hiện một điều gì đó cụ thể, chúng chỉ là các con số, và chúng không dựa vào bất kỳ một biểu thức toán học hay thống kê nào.

1.2.3.2. Ưu – khuyết điểm

Ưu điểm

- Độ chính xác cao hơn các phương pháp lọc thô sơ.
- Chúng ta có thể dễ dàng phân phối các tập luật.

Khuyết điểm

- Các tập luật cần được duy trì thường xuyên.
- Độ chính xác không tốt bằng các bộ lọc thống kê mới hơn.
- Những kẻ phát tán thư rác có thể sử dụng các tập luật để qua mặt bộ lọc.

1.3.3.3. Ghi chú

Phương pháp này phù hợp với những nhà quản trị hệ thống có thể chấp nhận tỉ lệ lỗi lớn hơn 5% với độ chính xác thường xuyên thay đổi.

1.2.4. Kỹ thuật challenge/ response

1.2.4.1. Giới thiệu

Challenge/response [7] là cách tiếp cận tương tự với kỹ thuật whitelisting. challenge/ response sẽ tự động gửi một thông điệp challenge tới người gửi thư. Trong thông điệp này, người gửi được yêu cầu làm một vài thao tác (như ấn vào một liên kết) để thông điệp đầu tiên được tới người nhận đồng thời người gửi được đưa vào danh sách whitelist, nếu không thông điệp sẽ không được gửi. Challenge/ response đã đẩy trách nhiệm duy trì whitelist cho người gửi thông điệp, rất nhiều người không thích điều này vì nó khiến họ phải làm công việc của bộ lọc thư rác. Nhiều người rất khó chịu khi phải trả lời các thông điệp challenge dẫn đến khuynh hướng họ sẽ không muốn giao tiếp với những người yêu cầu họ phản hồi thư điện tử challenge nữa.

Các vấn đề cần quan tâm đối với kỹ thuật challenge/ response:

- Phần lớn các lỗ hổng được tìm thấy trong whitelisting cũng có trong challenge/ response. Việc giả mạo vẫn thực hiện được dễ

dàng và làm vấn đề tồi tệ hơn, chính những người sử dụng challenge/ response có thể thêm địa chỉ của họ vào danh sách người gửi tin cậy trong whitelist của người nhận.

- Một điểm cần nói đến lưu lượng thư điện tử mà challenge/ response phát sinh ra. Thay vì giúp duy trì các tài nguyên, challenge/ response lại sử dụng thêm các tài nguyên do gửi các thư điện tử xác thực. Kết quả là hàng ngày một lượng thư điện tử lớn được gửi ra để xác thực các địa chỉ thư điện tử (trong đó có rất nhiều địa chỉ giả mạo). Có thể lên đến hàng triệu thư điện tử mỗi ngày cho một ISP cỡ nhỏ.
- Mọi người thường phàn nàn là challenge/ response làm trì trệ thư điện tử của họ. Ví dụ, nếu một người gửi trả lời một challenge thất bại, nhưng thư điện tử của họ lại là thư điện tử khẩn thì thư điện tử sẽ bị trì hoãn cho tới khi người gửi kiểm tra lại thư điện tử của họ và gửi lại.

1.2.4.2. Ưu – khuyết điểm

Ưu điểm

- Rất chính xác
- Không dựa trên việc học nội dung của thông điệp

Khuyết điểm

- Làm việc gửi thư điện tử bị chậm lại.
- Phía người gửi cần phải xác thực địa chỉ của mình một cách thủ công.
- Khiến cho nhiều người không muốn gửi thư điện tử tới chúng ta.
- Đường truyền chịu tải cao do lượng thư điện tử phát sinh lớn.
- Có thể bị giả mạo địa chỉ.

1.2.4.3. Ghi chú

Phù hợp cho người dùng muốn người gửi phải được xác thực trước khi giao tiếp và không quan tâm tới việc có thể mất các thư điện tử mang đến cơ hội nghề nghiệp hay những người dùng muốn giới hạn số lượng người họ muốn giao tiếp.

1.2.5. Kỹ thuật throttling

1.2.5.1. Giới thiệu

Throttling có thể xem là một trong những cách để chống thư rác nhạy cảm nhất đối với nhà cung cấp dịch vụ ở tầm nhỏ và trung bình, bởi vì nó không ngăn bất kỳ thư hợp lệ nào đi vào mạng. Thay vào đó, nó chỉ giảm lưu lượng mà một mạng hay một host có thể gửi. Kỹ thuật này sẽ bảo vệ các tài nguyên quan trọng đang bị kẻ phát tán thư rác sử dụng và làm cho lượng thư rác đi vào đường mạng ít hơn.

Throttling được sử dụng để dò và bảo vệ lưu lượng ra (outbound) vào (inbound) ở nhiều ISP. Điểm tốt của throttling là nó duy trì tài nguyên mà không tác động nhiều lên các thư hợp lệ và nó cũng khiến những kẻ phát tán thư rác phải tốn nhiều thời gian xử lý nhất.

Nguyên lý của phương pháp throttling là một lần phân phối (server phân phối đến các client) các thư hợp lệ sẽ không bao giờ gửi quá một ngưỡng lưu lượng xác định đến một mạng cụ thể nào đó. Ví dụ một danh sách thư điện tử hợp lệ có thể gửi ra ngoài một số lượng lớn thư điện tử, nhưng mỗi thông điệp đến người nhận khác nhau trên các mạng khác nhau. Hầu như, chỉ một số ít thông điệp gửi ra ngoài đi trực tiếp đến một mạng nào đó. Nói cách khác thì kẻ phát tán thư rác có thể dùng các đoạn script dùng để tấn công (bombard) một mạng bằng thư. Một công cụ điều tiết tốt sẽ xác định chính xác liệu người gửi có đang lợi dụng mạng hay không và giới hạn lượng băng thông người gửi có thể sử dụng.

Nhiều công cụ throttling hiện nay được thiết kế để điều tiết lưu lượng sử dụng của mọi người dựa vào tổng thông lượng, số lượng thông điệp, và các điều kiện khác. Thuận lợi của các công cụ này hơn các giải pháp throttling khác là chúng không dựa trên bộ lọc thư rác, nhưng sẽ là bất lợi nếu chính sách điều tiết của chúng quá khắt khe . Ví dụ nhiều thư điện tử hợp lệ có thể bị chặn nếu lưu lượng đã vượt quá ngưỡng lưu lượng. Chẳng hạn một CEO muốn gửi thư điện tử tới tất cả nhân viên của anh ta, những người không dùng chung máy chủ ISP với anh ta thì các thư điện tử có thể bị chậm do lượng thư điện tử gửi ra ngoài lớn.

1.2.5.2. Ưu – khuyết điểm

Ưu điểm

- Giúp duy trì các tài nguyên, giảm đáng kể lượng thư rác lưu thông trên mạng.

Khuyết điểm

- Không phải là một giải pháp chống thư rác thực sự.
- Có thể khiến người sử dụng hợp pháp bối rối khi thư điện tử của họ bị chậm.

1.2.5.3. Ghi chú

Phù hợp cho những nhà cung cấp dịch vụ và các công ty lớn cần duy trì các tài nguyên.

1.2.6. Kỹ thuật address obfuscation

1.2.6.1. Giới thiệu

Kỹ thuật address obfuscation [7] là kỹ thuật làm rối địa chỉ thư điện tử nhằm ẩn địa chỉ đó đối với kẻ phát tán thư rác. Kỹ thuật này được dùng chống lại các con bot (một chương trình máy tính nhỏ) chuyên thu thập địa chỉ thư điện tử mới trên các trang web để đưa là danh sách của những kẻ phát tán thư rác.

Khái niệm của address obfuscation khá là đơn giản. Thay vì hiển thị địa chỉ thư điện tử như là sieunhan.fit@khtn.edu.vn , bạn có thể nhìn thấy "sieunhan dot fit [at] khtn dot edu dot vn". Tuy nhiên, cách tiếp cận này thật sự không làm việc tốt như mọi người thường nghĩ vì các con bot thu thập địa chỉ ngày càng thông minh hơn, nó có thể lắp ráp lại địa chỉ thư điện tử trên. Những kẻ phát tán thư rác cũng nhận thông tin địa chỉ thư điện tử của người dùng từ những nơi khác ngoài Web. Ví dụ nhiều ISP và các công ty thẻ tín dụng bán danh sách các địa chỉ cho những kẻ phát tán thư rác.

Address obfuscation có thể giữ tên của một vài người khỏi một vài danh sách, nhưng nó cũng không phải là giải pháp thực sự để chống thư rác. Điều mà chúng ta cần là một giải pháp để chống lại các con bot thu thập địa chỉ (havest bot) hơn là làm rối địa chỉ thư điện tử của mình.

1.2.6.2. Ưu – khuyết điểm

Ưu điểm

- Có thể giữ cho địa chỉ không nằm trong một vài danh sách thư điện tử của những kẻ phát tán thư rác.

Khuyết điểm

- Địa chỉ thư điện tử trở lên không thân thiện và phức tạp hơn, không giải quyết được vấn đề chống thư rác.

1.2.6.3. Ghi chú

Phù hợp cho những người muốn chống thư rác có nhiều thời gian để làm rối thư điện tử bằng tay.

1.2.7. Kỹ thuật collaborative filtering

1.2.7.1. Giới thiệu

Collaborative filtering (CF) [7] là kỹ thuật lọc thông tin dựa trên sự hợp tác của nhiều agent, nhiều nguồn dữ liệu...

Các bộ lọc thư rác đã bắt đầu cài đặt CF để cho phép những cá nhân trong các nhóm tin cậy chia sẻ các thông điệp thư rác với nhau làm nhân tố chống lại một loại thư rác cụ thể nào đó.

Collaborative filtering làm cho vài cơ chế lọc thư rác đang tồn tại tăng khả năng lọc thư rác bằng cách cung cấp cho chúng thời gian (hoặc tài nguyên hoặc cả hai) để thích nghi với các loại thư rác mới. CF giúp cho nhiều người không phải nhận những thư rác mới mà người dùng khác đã nhận.

Điểm yếu của collaborative filtering lại nằm ở chính cộng đồng tham gia nó. Trong các cộng đồng lớn, thì có maintenance loop [7] (xảy ra khi nhiều người trong nhóm ghi vào tập dữ liệu chia sẻ một thư rác giống nhau) và tỉ lệ false-positive cao. Các mạng lớn hơn (nhóm có số người lớn hơn) thường có maintenance loop cao hơn do độ trễ trong cập nhật cơ sở dữ liệu cao hơn, cũng giống như độ trễ lan truyền (propagation delay [7]) trong kỹ thuật blacklisting. Những mạng tự động có thể đang hoạt động dựa trên thông tin sai đang được lan truyền hoặc thông tin sai bị kẻ xấu đưa vào mạng. Những mạng nhỏ hơn có độ chính xác cao hơn và cập nhật nhanh hơn nhưng thiếu khả năng bao phủ hết những thư rác mới đi vào.

1.2.7.2. Ưu – khuyết điểm

Ưu điểm

- Chống lại các loại thư loại mới.

Khuyết điểm

- Cần phải xem xét kỹ độ tin cậy, độ trễ trong việc lan truyền thông tin.

1.2.7.3. Ghi chú

Phù hợp cho việc thiết lập thêm một lớp bảo vệ trong các bộ lọc thư rác.

1.2.8. Kỹ thuật dùng máy học:

Học máy (Machine Learning – ML) là một lĩnh vự c nghiên cứu của Trí tuệ nhân tạo (Artificial Intelligence – AI)

Các định nghĩa về học máy:

- Một quá trình nhờ đó một hệ thống cải thiện hiệu suất (hiệu quả hoạt động) của nó [Simon, 1983]
- Một quá trình mà một chương trình máy tính cải thiện hiệu suấ t của nó trong một công việc thông qua kinh nghiệ m [Mitchell, 1997]
- Việc lập trình các máy tính để tối ưu hóa một tiêu chí hiệu suất dựa trên các dữ liệu ví dụ hoặc kinh nghiệm trong quá khứ [Alpaydin, 2004]

Biểu diễn một bài toán học máy [Mitchell, 1997]

Học máy = cải thiện hiệu quả một công việc thông qua kinh nghiệm

- Một công việc (nhiệm vụ) T
- Đối với các tiêu chí đánh giá hiệu năng P
- Thông qua (sử dụng) kinh nghiệm E

Bài toán học máy lọc thư rác (Email spam filtering)

- T : Dự đoán (để lọc) những thư điện tử nào là thư rác (spam email)
- P : % của các thư điện tử gử i đến được phân loại chính xác
- E : Một tập các thư điện tử (emails) mẫu, mỗi thư điện tử được biểu diễn bằng một tập thuộc tính (vd: tập từ khóa) và nhãn lớp (thư thường/thư rác) tương ứng.

Ưu điểm

- Khả năng thích nghi (học) cao với sự tiến hóa của thư rác rất nhanh.
- Thể hiện tính cá nhân hóa mạnh mẽ do mỗi người dùng có thể có một tập dữ liệu riêng. Chính điều này làm cho độ chính xác đối với từng người dùng tăng lên đáng kể.

Khuyết điểm

- Phải mất một khoảng thời gian đầu huấn luyện cho bộ lọc.

1.3. Phân tích và định hướng phát triển ứng dụng thử nghiệm

Bảng 1.1. Các phần mềm chống thư rác [12]

	SPAM fighter Pro	Cloudmark DesktopOne Pro	MailWasher Pro 2010	ChoiceMail One	iHateSpam	Clean Mail Home	Spam Bully	Spam Bully	SpamEater Pro	Spam Buster
Blacklisting	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Whitelisting	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Heristic Filtering	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Statistical Filtering	✓	✓								
Challenge/Response				✓			✓			

Qua bảng thống kê 1.1 và các kỹ thuật chống thư rác được giới thiệu ở trên, chúng ta đều thấy được ưu điểm và khuyết điểm của từng kỹ thuật. Đa số các kỹ thuật chống thư rác trên đều lọc dựa vào phần header của thư hoặc ngăn chặn ngay từ kẻ phát tán thư rác (blacklist, whitelist) mà ít quan tâm đến phần nội dung của thư.

Các kỹ thuật giới thiệu trên ngoại trừ kỹ thuật sử dụng máy học không có quá trình huấn luyện để "học" sự thay đổi của thư rác theo thời gian, chính vì thế khiến cho thư rác vượt qua các bộ lọc sử dụng các kỹ thuật trên khá lớn. Trong các kỹ thuật đã giới thiệu, kỹ thuật heuristic là có thể "học" sự thay đổi của thư rác nhưng phải do nhà quản trị mạng liên tục cập nhật các luật giúp cho bộ lọc nhận ra loại thư rác mới. Tuy nhiên điều này làm tăng số lượng công việc mà nhà quản trị mạng phải thực hiện.

Chính vì thế tác giả thực hiện luận văn mong muốn phát triển một hệ thống dựa vào các kỹ thuật đang được chú trọng phát triển gần đây là thống kê và so khớp.

CHƯƠNG 2: CÁC PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN VÀ NHẬN DẠNG THƯ RÁC

2.1. Bối cảnh phân loại văn bản hiện nay

Phân loại văn bản tự động là một lĩnh vực được chú ý nhất trong những năm gần đây. Để phân loại người ta sử dụng nhiều cách tiếp cận khác nhau như dựa trên từ khóa, dựa trên ngữ nghĩa các từ có tần số xuất hiện cao, mô hình Maximum Entropy, tập thô … Tiếng Anh là một trong những ngôn ngữ được nghiên cứu sớm và rộng rãi nhất với kết quả đạt được rất khả quan. Một số lượng lớn các phương pháp phân loại đã được áp dụng thành công trên ngôn ngữ này : mô hình hồi quy [Fuhr et al,1991], phân loại dựa trên láng giềng gần nhất (k-nearest neighbors) [Dasarathy, 1991], phương pháp dựa trên xác suất Naïve Bayes [Joachims, 1997], cây quyết định [Fuhr et al,1991], học luật quy nạp [William & Yoram, 1996], mạng nơron (neural network)[Wiener et al, 1995], học trực tuyến[William & Yoram, 1996], và máy vector hỗ trợ (SVM-support vector machine) [Vapnik, 1995]. Hiệu quả của các phương pháp này rất khác nhau ngay cả khi áp dụng cho tiếng Anh. Việc đánh giá gặp nhiều khó khăn do việc thiếu các tập ngữ liệu huấn luyện chuẩn. Thậm chí đối với tập dữ liệu được sử dụng rộng rãi nhất, Reuter cũng có nhiều phiên bản khác nhau. Hơn nữa, có rất nhiều độ đo được sử dụng như recall, precision, accuracy hoặc error, break-even point, F-measure …Chương này giới thiệu các thuật toán phân loại được sử dụng phổ biến nhất đồng thời so sánh giữa các phương pháp sử dụng kết quả của [Yang, 1997].

2.2. Biểu diễn văn bản

Bước đầu tiên trong qui trình phân loại văn bản là thao tác chuyển văn bản đang được mô tả dưới dạng chuỗi các từ thành một mô hình khác, sao cho phù hợp với các thuật toán phân loại, thông thường người ta thường biểu diễn văn bản bằng mô hình vector. Ý tưởng của mô hình này là xem mỗi một văn

bản (Di) được biểu diễn theo dạng $D_i = (\vec{d}_i, i)$, trong đó i là chỉ số dùng để nhận diện văn bản này và $\vec{d}_i$ là vector đặc trưng của văn bản Di này , trong đó : $\vec{d}_i = (w_{i1}, w_{i2}, \ldots\ldots, w_{in})$, và n là số lượng đặc trưng của vector văn bản , w_{ij} là trọng số của đặc trưng thứ j , $j \in \{1,2,\ldots,n\}$.

Một vấn đề cần quan tâm khi biểu diễn văn bản theo vector đặc trưng chính là việc chọn lựa đặc trưng và số chiều cho không gian vector . Cần phải chọn bao nhiêu từ , là các từ nào , phương pháp chọn ra sao ? . Đây là câu hỏi chúng ta phải trả lời trong qúa trình chuyển văn bản sang thành vector , có nhiều cách tiếp cận khác nhau để trả lời cho câu hỏi này , tiêu biểu là sử dụng phương pháp Information Gain , phương pháp DF – Thresolding hay phương pháp Term Strength . Phương pháp Information Gain sử dụng độ đo MI (Mutual Information) để chọn ra tập từ khóa đặc trưng có độ đo MI cao nhất . Tuy nhiên , việc chọn lựa phương pháp nào thì tuỳ thuộc vào độ thích hợp , phù hợp của phương pháp , của độ đo mà phương pháp đó sử dụng so với bài toán mà chúng ta đang xem xét giải quyết , có thể là nếu văn bản là một trang web thì sẽ có phương pháp để chọn lựa đặc trưng khác so với các văn bản loại khác .

Các đặc trưng của văn bản khi biểu diễn dưới dạng vector :

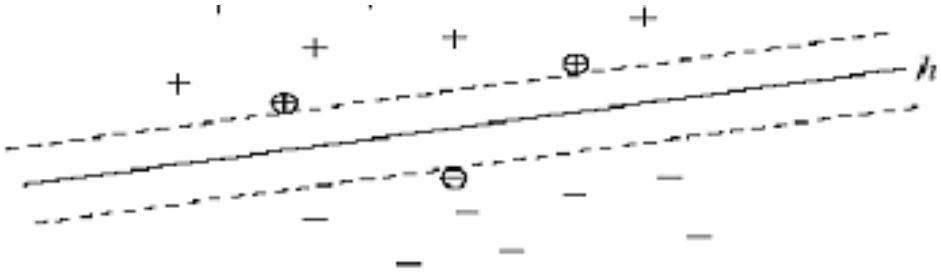
- Số nhiều không gian đặc trưng thường lớn .
- Các đặc trưng độc lập nhau.
- Các đặc trưng rời rạc : vector đặc trưng di có thể có nhiều thành phần mang giá trị 0 do có nhiều đặc trưng không xuất hiện trong văn bản di (nếu chúng ta tiếp cận theo cách sử dụng giá trị nhị phân 1, 0 để biểu diễn cho việc có xuất hiện hay không một đặc trưng nào đó trong văn bản đang được biểu diễn thành vector) , tuy nhiên

nếu đơn thuần cách tiếp cận sử dụng giá trị nhị phân 0, 1 này thì kết qủa phân loại phần nào hạn chế là do có thể đặc trưng đó không có trong văn bản đang xét nhưng trong văn bản đang xét lại có từ khóa khác với từ đặc trưng nhưng có ngữ nghĩa giống với từ đặc trưng này , do đó một cách tiếp cận khác là không sử dụng số nhị phân 0, 1 mà sử dụng giá trị số thực để phần nào giảm bớt sự rời rạc trong vector văn bản.

2.3. Support vector Machine (SVM)

SVM là phương pháp phân loại rất hiệu qủa được Vapnik giới thiệu năm 1995.

Ý tưởng của phương pháp là cho trước một tập huấn luyện được biểu diễn trong không gian vector, trong đó mỗi một văn bản được xem như một điểm trong không gian này. Phương pháp này tìm ra một siêu mặt phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng, tạm gọi là lớp + (cộng) và lớp – (trừ). Chất lượng của siêu mặt phẳng này được quyết định bởi một khoảng cách (được gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì càng có sự phân chia tốt các điểm ra thành hai lớp, nghĩa là sẽ đạt được kết qủa phân loại tốt. Mục tiêu của thuật toán SVM là tìm được khoảng cách biên lớn nhất để tạo kết qủa phân loại tốt .



Hình 2.1: Phân loại văn bản theo kỹ thuật Vector Machine.

Có thể nói SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán là tìm được một không gian H và siêu mặt phẳng quyết định h trên H sao cho sai số khi phân loại là thấp nhất, nghĩa là kết qủa phân loại sẽ cho kết qủa tốt nhất.

Phương trình siêu mặt phẳng chứa vector di trong không gian như sau:

$$\vec{d_i} \cdot \vec{w} + b = 0$$

$$h\left(\vec{d_i}\right) = \text{sign}\left(\vec{d_i} \cdot \vec{w}\right) = \begin{cases} +, \vec{d_i} \cdot \vec{w} + b > 0 \\ -, \vec{d_i} \cdot \vec{w} + b < 0 \end{cases} \quad (2.1)$$

Như thế vector h(di) biểu diễn sự phân lớp của vector di vào hai lớp. Gọi Yi mang giá trị +1 hoặc -1, khi đó Yi = +1 văn bản tương ứng với vector di thuộc lớp + và ngược lại nó sẽ thuộc vào lớp - . Khi này để có siêu mặt phẳng h ta sẽ giải bài toán sau :

Tìm Min $\left\|\vec{w}\right\|$ với $\vec{w}$ và b thỏa điều kiện : $\forall i \in 1,n : y_i(\text{sign}(d_i - w + b)) \geq 1$

Chúng ta thấy rằng SVM là mặt phẳng quyết định chỉ phụ thuộc vào các vector hỗ trợ có khoảng cách đến mặt phẳng quyết định là 1/wi. Khi các điểm khác bị xóa đi thì thuật toán vẫn cho kết qủa giống như ban đầu. Chính

đặc điểm này làm cho SVM khác với các thuật toán khác như kNN, LLSF, Nnet, NB vì tất cả dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

2.4. K–Nearest Neighbor (kNN)

kNN là phương pháp truyền thống khá nổi tiếng theo hướng tiếp cận thống kê đã được nghiên cứu trong nhiều năm qua. kNN được đánh giá là một trong những phương pháp tốt nhất được sử dụng từ những thời kỳ đầu trong nghiên cứu về phân loại văn bản .

Ý tưởng của phương pháp này đó là khi cần phân loại một văn bản mới, thuật toán sẽ xác định khoảng cách (có thể áp dụng các công thức về khoảng cách như Euclide, Cosine, Manhattan,…) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất ,gọi là k nearest neighbor – k láng giềng gần nhất, sau đó dùng các khoảng cách này đánh trọng số cho tất cả các chủ đề. Khi đó, trọng số của một chủ đề chính là tổng tất cả các khoảng cách ở trên của các văn bản trong k láng giềng có cùng chủ đề, chủ đề nào không xuất hiện trong k láng giềng sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo giá trị trọng số giảm dần và các chủ đề có trọng số cao sẽ được chọn làm chủ đề của văn bản cần phân loại.

Trọng số của chủ đề cj đối với văn bản x được tính như sau :

$$W\left(\vec{x}, c_j\right) = \sum_{\vec{d_i} \in \{kNN\}} sim\left(\vec{x}, \vec{d_i}\right).y\left(\vec{d_i}, c_j\right) - b_j \qquad (2.2)$$

Trong đó :

y (di, c) thuộc {0,1}, với :

- y = 0 : văn bản di không thuộc về chủ đề cj
- y = 1 : văn bản di thuộc về chủ đề cj

sim (x , d) : độ giống nhau giữa văn bản cần phân loại x và văn bản d. Chúng ta có thể sử dụng độ đo cosine để tính khoảng cách:

$$sim\left(\vec{x}, \vec{d_i}\right) = \cos\left(\vec{x}, \vec{d_i}\right) = \frac{\vec{x} \cdot \vec{d_i}}{\left\|\vec{x}\right\| \left\|\vec{d_i}\right\|} \qquad (2.3)$$

- - bj là ngưỡng phân loại của chủ đề cj được tự động học sử dụng một tập văn bản hợp lệ được chọn ra từ tập huấn luyện.

Để chọn được tham số k tốt nhất cho thao tác phân loại, thuật toán cần được chạy thử nghiệm trên nhiều giá trị k khác nhau, giá trị k càng lớn thì thuật toán càng ổn định và sai sót càng thấp.

2.5. Naïve Bayes (NB)

NB là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học [5] và nhiều lĩnh vực khác như trong các công cụ tìm kiếm, các bộ lọc mail…

Ý tưởng cơ bản của cách tiếp cận này là sử dụng xác suất có điều kiện giữa từ hoặc cụm từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Như thế NB không tận dụng được sự phụ thuộc của nhiều từ vào một chủ đề cụ thể. Chính giả định đó làm cho việc tính toán NB hiệu qủa và nhanh chóng hơn các phương pháp khác với độ phức tạp theo số mũ vì nó không sử dụng cách kết hợp các từ để đưa ra phán đoán chủ đề.

Mục đích chính là làm sao tính được xác suất Pr(Cj, d’), xác suất để văn bản d’nằm trong lớp Cj. Theo luật Bayes, văn bản d’ sẽ được gán vào lớp Cj nào có xác suất Pr(Cj, d’) cao nhất .

Công thức để tính Pr(Cj, d’) như sau :

$$H_{BAYES}(d') = \operatorname{argmax}\left(\frac{Pr(C_j) \cdot \prod_{i=1}^{|d'|} Pr(w_i | C_j)}{\sum_{c' \in c} Pr(c') \cdot \prod_{i=1}^{|d'|} Pr(w_i | C')} \right)_{c_j \in C} \quad (2.4)$$

Với :

- TF(wi, d') là số lần xuất hiện của từ wi trong văn bản d'
- |d'| là số lượng các từ trong văn bản d'
- wi là một từ trong không gian đặc trưng F với số chiều là |F|
- Pr(Cj) được tính dựa trên tỷ lệ phần trăm của số văn bản mỗi lớp tương ứng

$$Pr(C_j) = \frac{\|C_j\|}{\|C\|} = \frac{\|C_j\|}{\sum_{C' \in C} \|C'\|} \quad (2.5)$$

trong tập dữ liệu huấn luyện

$$Pr(w_i | C_j) = \frac{1 + TF(w_i, c_j)}{|F| + \sum_{w' \in |F|} TF(w', c_j)} \quad (2.6)$$

Ngoài ra còn có các phương pháp NB khác có thể kể ra như ML Naïve Bayes, MAP Naïve Bayes, Expected Naïve Bayes. Nói chung Naïve Bayes là một công cụ rất hiệu qủa trong một số trường hợp. Kết qủa có thể rất xấu nếu dữ liệu huấn luyện nghèo nàn và các tham số dự đoán (như không gian đặc trưng) có chất lượng kém.Nhìn chung đây là một thuật toán phân loại tuyến tính thích hợp trong phân loại văn bản nhiều chủ đề. NB có ưu điểm là cài đặt

đơn giản, tốc độ thực hiện thuật toán nhanh, dễ dàng cập nhật dữ liệu huấn luyện mới và có tính độc lập cao với tập huấn luyện .

2.6. Mạng Neural (Nnet)

Nnet được nghiên cứu mạnh trong hướng trí tuệ nhân tạo. Wiener là người đã sử dụng Nnet để phân loại văn bản, sử dụng 2 hướng tiếp cận : kiến trúc phẳng (không sử dụng lớp ẩn) và mạng nơron 3 lớp (bao gồm một lớp ẩn) [Wiener et al, 1995]

Cả hai hệ thống trên đều sử dụng một mạng nơron riêng rẽ cho từng chủ đề, NNet học cách ánh xạ phi tuyến tính những yếu tố đầu vào như từ, hay mô hình vector của một văn bản vào một chủ đề cụ thể.

Khuyết điểm của phương pháp NNet là tiêu tốn nhiều thời gian dành cho việc huấn luyện mạng nơron.

Ý tưởng của phương pháp này là mô hình mạng neural gồm có ba thành phần chính như sau: kiến trúc (architecture), hàm chi phí (cost function), và thuật toán tìm kiếm (search algorithm). Kiến trúc định nghĩa dạng chức năng (functional form) liên quan giá trị nhập (inputs) đến giá trị xuất (outputs).

Kiến trúc phẳng (flat architecture): Mạng phân loại đơn giản nhất (còn gọi là mạng logic) có một đơn vị xuất là kích hoạt kết quả (logistic activation) và không có lớp ẩn, kết quả trả về ở dạng hàm (functional form) tương đương với mô hình hồi quy logic. Thuật toán tìm kiếm chia nhỏ mô hình mạng để thích hợp với việc điều chỉnh mô hình ứng với tập huấn luyện. Ví dụ, chúng ta có thể học trọng số trong mạng kết quả (logistic network) bằng cách sử dụng không gian trọng số giảm dần (gradient descent in weight space) hoặc sử dụng thuật toán interated-reweighted least squares là thuật toán truyền thống trong hồi quy (logistic regression).

Kiến trúc mô dun (modular architecture): Việc sử dụng một hay nhiều lớp ẩn của những hàm kích hoạt phi tuyến tính cho phép mạng thiết lập các mối

quan hệ giữa những biến nhập và biến xuất. Mỗi lớp ẩn học để biểu diễn lại dữ liệu đầu vào bằng cách khám phá ra những đặc trưng ở mức cao hơn từ sự kết hợp đặc trưng ở mức trước.

Trong công trình của Wiener et al (1995) dựa theo khung của mô hình hồi quy, liên quan từ đặc trưng đầu vào cho đến kết quả gán chủ đề tương ứng được học từ tập dữ liệu. Do vậy, để phân tích một cách tuyến tính, tác giả dùng hàm sigmoid sau làm hàm truyền trong mạng neural:

$$P = \frac{1}{1+e^{n}} \quad (2.7)$$

Trong đó, $\eta = \beta^{T}x$ là sự kết hợp của những đặc trưng đầu vào và p phải thỏa

2.7. Phương pháp tách từ trong tiếng Việt

2.7.1. Tình hình nghiên cứu

Mặc dù giống tiếng Anh khi sử dụng ký tự latinh, tuy nhiên trở ngại lớn nhất là cấu trúc tiếng Việt khác biệt hoàn toàn so với cấu trúc tiếng Anh đã trình bày ở trên và đa phần các phương pháp thường dùng cách so khớp từ trực tiếp dựa trên bộ từ điển có sẵn và việc cập nhật bộ từ điển rất khó khăn, thường thực hiện bằng thao tác thủ công là chính.

Dựa trên các nghiên cứu trước, hướng tiếp cận dựa trên từ với mục tiêu tách được các từ hoàn chỉnh trong câu. Hướng tiếp cận này có thể chia làm 3 hướng chính: *dựa trên thống kê (statistics-based), dựa trên từ điển (dictionary-based)* và *hydrid (kết hợp nhiều phương pháp với hy vọng đạt được những ưu điểm của các phương pháp này)*

Hướng tiếp cận dựa trên thống kê (statistics-based): dựa trên các thông tin như tần số xuất hiện của từ trong tập huấn luyện ban đầu. Hướng tiếp cận này đặc biệt dựa trên tập dữ liệu huấn luyện, nhờ vậy nên hướng tiếp cận này tỏ ra rất linh hoạt và hữu dụng trong nhiều lĩnh vực riêng biệt.

Hướng tiếp cận dựa trên từ điển (dictionary-based): thường được sử dụng trong tách từ. Ý tưởng của hướng tiếp cận này là những cụm từ được tách ra từ văn bản phải khớp với các từ trong từ điển. Những hướng tiếp cận khác nhau sẽ sử dụng những loại từ điển khác nhau. Hướng tiếp cận "*full word/phrase*" cần sử dụng một bộ từ điển hoàn chỉnh để có thể tách được đầy đủ các từ hoặc ngữ trong văn bản, trong khi đó, hướng tiếp cận thành phần (component) lại sử dụng từ điển thành phần (component dictionary) [Wu &Tseng, 1993]. Từ điển hoàn chỉnh chứa tất cả các từ và ngữ được dùng trong tiếng Hoa, trong khi từ điển thành phần (component dictionarry) chỉ chứa các thành phần của từ và ngữ như hình vị và các từ đơn giản trong tiếng Hoa. Phần dưới sẽ trình bày các phương pháp tách từ trong ngôn ngữ tiếng Việt.

2.7.2. Một số phương pháp tách từ

2.7.2.1. Tách câu dựa trên Maximum Entropy

Phuong H.L. và Vinh H.T. [2] mô hình hóa bài toán tách câu dưới dạng bài toán phân lớp trên Maximum Entropy. Với mỗi chuỗi ký tự có thể là điểm phân cách câu (".", "?", hay "!"), ước lượng xác xuất đồng thời của ký tự đó cùng với ngữ cảnh xung quanh (biểu diễn bởi biến ngẫu nhiên c) và biến ngẫu nhiên thể hiện đó có thực sự là điểm phân tách câu hay không (*b \in {no, yes}*). Xác xuất mô hình được định nghĩa như sau

$$p(b,c) = л \prod_{j=1}^{k} \alpha_j^{f_j(b,c)} \qquad (2.8)$$

Ở đây: α_j là các tham số chưa biết của mô hình, mỗi a_j tương ứng với một hàm đặc trưng f_j. Gọi $B = \{no, yes\}$ là tập các lớp và C là tập của các ngữ cảnh. Các đặc trưng là các hàm nhị phân $f_j: B \times C \rightarrow \{0,1\}$ dùng để mã hóa thông tin cần thiết. Xác xuất để quan sát được điểm phân tách câu trong ngữ cảnh c được đặc trưng bởi xác xuất $p(yes, c)$. Tham số α_j được chọn là giá trị

làm cực đại hàm likehook của dữ liệu huấn luyện với các thuật toán GIS và IIS

Để phân lớp một ký tự tách câu tiềm năng vào một trong hai lớp *{yes, no}* – lớp *yes* nghĩa là đó thực sự là một ký tự phân tách câu, còn n*o* thì là ngược lại, dựa vào luật phân lớp như sau

$$p(yes|c) = p(yes,c)/p(c) = p\ (yes,c)/(p(yes,c) + p(no,c)) \quad (2.9)$$

Ở đây *c* là ngữ cảnh xung quanh ký tự tách câu tiềm năng đó và bao gồm cả ký tự đang xem xét. Sau đây là những lựa chọn hàm tiềm năng f_j để phân tách câu trong tiếng Việt.

- **Lựa chọn đặc trưng**

Các đặc trưng trong Maximum Entropy mã hóa các thông tin hữu ích cho bài toán tách câu. Nếu đặc trưng xuất hiện trong tập đặc trưng, trọng số tương ứng của nó dùng để hỗ trợ cho tính toán xác xuất *p(b|c)*.

Các ký tự tách câu tiềm năng được xác định bằng cách duyệt qua văn bản, xác định các chuỗi ký tự được phân cách bởi dấu cách (còn gọi là token) và chứa một trong các ký tự ".", "?", hay "!". Thông tin về token và thông tin ngữ cảnh về token liền trái, phải của token hiện tại được xác định xác xuất phần lớn.

Gọi các token chứa các ký tự kết thúc câu tiềm năng là "*ứng viên*". Phần ký tự đi trước ký tự kết thúc câu tiềm năng được gọi là "*tiền tố*", phần đi sau gọi là "*hậu tố*". Vị trí của ký tự kết thúc câu tiềm năng cũng được mô tả trong tập đặc trưng. Tập các ngữ cảnh được xem xét từ chuỗi ký tự được mô tả như dưới đây

1. Có/ không có 1 ký tự trống trước ký tự kết thúc câu tiềm năng.
2. Có/ không có 1 ký tự trắng sau ký tự kết thúc câu tiềm năng.
3. Ký tự kết thúc câu tiềm năng.
4. Đặc trưng tiền tố.

5. Độ dài tiền tố nếu nó có độ dài lớn hơn 0.
6. Ký tự đầu tiên của tiền tố là ký tự.
7. Tiền tố nằm trong danh sách các từ viết tắt.
8. Đặc trưng hậu tố.
9. Token đi trước token hiện tại.
10. Ký tự đầu tiên của token liền trước viết hoa/ không viết hoa.
11. Token liền trước nằm trong danh sách các từ viết tắt.
12. Token liền sau.
13. Token ứng viên được viết hoa/ không viết hoa.

Từ những ngữ cảnh trên, có thể rút ra tập ngữ cảnh từ tập dữ liệu (tập *C)*. Tập ngữ cảnh cùng với nhãn từ dữ liệu tạo ra một tập đặc trưng tương ứng. Xét ví dụ sau để làm rõ mối quan hệ giữa ngữ cảnh, đặc trưng:

"Những hacker máy tính sẽ có cơ hội chiếm giải thưởng trị giá 10.000 USD và 10.000 đola Singapore (5.882 USD) trong một cuộc tranh tài quốc tế mang tên "Hackers Zone" được tổ chức vào ngày 13/5/1999 tại Singapore."

Xem xét ký tự kết thúc câu tiềm năng "." Trong token "10.000 USD", từ vị trí này ta có thể rút ra một số ngữ cảnh sau:

1. Không có ký tự trắng trước ký tự ứng viên.
2. Không có ký tự trắng sau ký tự ứng viên.
3. Ký tự ứng viên là "."
4. Tiền tố: 10

Từ dữ liệu học này, có thể rút trích ra các đặc trưng như ví dụ dưới đây:

f{không có ký tự trắng trước ứng viên, no} = 1. Ý nghĩa của đặc trưng này là phát biểu: "*token không có ký tự trắng trước ứng viên và nhãn là no*" là đúng (đặc trưng nhận giá trị 1).

Sau khi ước lượng trọng số đặc trưng ta dựa vào các tham số đó để tính giá trị *p(yes|c)*. Nếu giá trị này *>50%*, nhãn tương ứng với ký tự ứng viên được ghi nhận là "yes" hay ký tự ứng viên thực sự là ký tự phân tách câu.

2.7.2.2. Phương pháp khớp tối đa (Maximum Matching)

Nội dung: Phương pháp khớp tối đa (*Maximum Matching*) [3], còn gọi là *Left Right Maximum (LRMM)*. Theo phương pháp này, sẽ duyệt một ngữ hoặc câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển, rồi cứ thế tiếp tục cho từ kế tiếp cho đến hết câu.

Dạng đơn giản: được dùng giải quyết nhập nhằng từ đơn. Giả sử có một chuỗi ký tự (tương đương với chuỗi tiếng trong tiếng Việt) $C_1, C_2,...,C_n$. Bắt đầu từ đầu chuỗi. Đầu tiên kiểm tra xem C_1, có phải là từ hay không, sau đó kiểm tra xem C_1C_2 có phải là từ hay không. Cứ tiếp tục tìm cho đến khi tìm được từ dài nhất. Từ có vẻ hợp lý nhất sẽ là từ dài nhất. Chọn từ đó, sau đó tiếp tục tìm như trên cho những từ còn lại cho đến khi xác định được toàn bộ chuỗi từ.

Dạng phức tạp: Quy tắc của dạng này là phân đoạn có vẻ hợp lý nhất là đoạn ba từ với chiều dài tối đa. Thuật toán bắt đầu như dạng đơn giản. Nếu phát hiện ra những cách tách từ gây nhập nhằng (ví dụ C_1 là từ và C_1C_2 cũng là từ), xem các chữ kế tiếp để tìm tất cả các đoạn ba từ có thể có bắt đầu với C_1 và C_1C_2. Ví dụ được những đoạn sau:

$C_1C_2\ C_3C_4$

$C_1C_2\ C_3C_4C_5$

$C_1C_2\ C_3C_4C_5\ C_6$

Chuỗi dài nhất sẽ là chuỗi thứ ba. Vậy từ đầu tiên của chuỗi thứ ba (C_1C_2) sẽ được chọn. Thực hiện lại các bước cho đến khi được chuỗi từ hoàn chỉnh

Ưu điểm của phương pháp trên có thể thấy rõ là đơn giản, dễ hiểu và chạy nhanh. Hơn nữa, chỉ cần một tập từ điển đầy đủ là có thể tiến hành phân đoạn văn bản, hoàn toàn không phải trải qua huấn luyện như các phương pháp được trình bày tiếp theo.

Nhược điểm của phương pháp này là nó không giải quyết được 2 vấn đề quan trọng nhất của bài toán phân đoạn từ tiếng Việt: thuật toán gặp phải nhiều nhập nhằng, hơn nữa nó hoàn toàn không có chiến lược gì với những từ chưa biết.

2.7.2.3. Phương pháp WFST (Weighted Finite – State Transducer)

Phương pháp *WFST* (*Weighted Finite – State Transducer*) [8], còn gọi là phương pháp chuyển dịch trạng thái hữu hạn có trọng số. Ý tưởng chính của phương pháp này áp dụng cho phân đoạn tiếng Việt là các từ được gán trọng số bằng xác suất xuất hiện của từ đó trong từ điển dữ liệu. Sau đó duyệt qua các câu, cách duyệt có trọng số lớn nhất được chọn là cách dùng để phân đoạn từ. Hoạt động của WFST có thể chia thành ba bước sau:

– Xây dựng từ điển trọng số: từ điển trọng số *D* được xây dựng như là một đồ thị biến đổi trạng thái hữu hạn có trọng số. Giả sử:
  + *H* là tập các tiếng trong tiếng Việt.
  + *P* là tập các loại từ trong tiếng Việt.
  + Mỗi cung của *D* có thể là:
    ++ Từ một phần tử của *H* tới một phần tử của *H*;
    ++ Từ phần tử ε (xâu rỗng) đến một phần tử của *P*.

Mỗi từ trong *D* được biểu diễn bởi một chuỗi các cung bắt đầu bởi một cung tương ứng với một phần tử của *H*, kết thúc bởi một cung có trọng số tương ứng với một phần tử của ε x *P*. Trọng số biểu diễn một chi phí ước lượng (estimated cost) cho bởi công thức:

$$C = -\log\left(\frac{f}{N}\right) \qquad (2.10)$$

Trong đó

f là tần số xuất hiện của từ

N là kích thước tập mẫu

Xây dựng các khả năng tách từ: Bước này thống kê tất cả các khả năng phân đoạn của một câu. Giả sử câu có *n* tiếng, thì có tới *2n*-1 cách phân đoạn khác nhau. Để giảm sự bùng nổ các cách phân đoạn, thuật toán loại bỏ ngay những nhánh phân đoạn mà chứa từ không xuất hiện trong từ điển.

Lựa chọn khả năng tách tối ưu: Sau khi liệt kê tất cả các khả năng phân đoạn từ, thuật toán chọn cách tách từ tốt nhất, đó là cách tách từ có trọng số bé nhất.

Ví dụ: "T*ốc độ truyền thông tin sẽ tăng cao*"

Từ điển trọng số:

"*tốc độ*" 8,68

"*truyền*" 12,31

"*truyền thông*" 12,31

"*thông tin*" 7,24

"*tin*" 7,33

"*sẽ*" 6,09

"*tăng*" 7,43

"*cao*" 6,95

Trọng số theo mỗi cách tách từ được tính là tổng các trọng số của từ theo từ điển trọng số

"*Tốc độ | tuyền thông | tin | sẽ | tăng | cao*"

"*Tốc độ | tuyền | thông tin | sẽ | tăng | cao*"

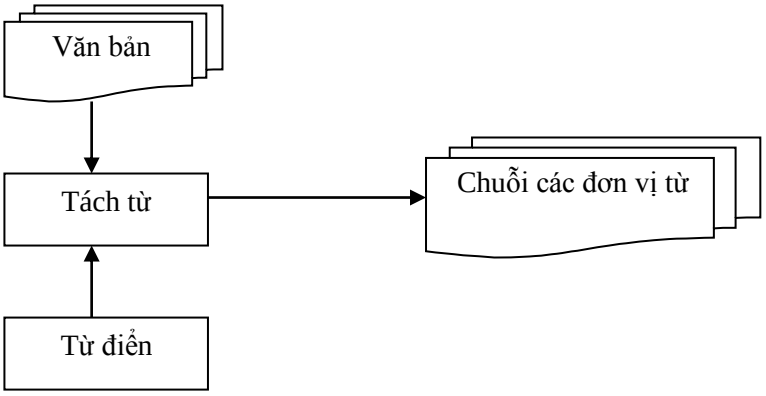
2.7.2.4. Bài toán tách từ và công cụ vnTokenizer

Ý tưởng: Cho một câu tiếng Việt bất kỳ, hãy tách câu đó thành những đơn vị từ vựng (từ), hoặc chỉ ra những âm tiết nào không có trong từ điển (phát hiện đơn vị từ vựng mới).

Giới thiệu công cụ vnTokenizer: công cụ tách từ tiếng Việt được nhóm tác giả Nguyễn Thị Minh Huyền, Vũ Xuân Lương và Lê Hồng Phương phát triển dựa trên phương pháp so khớp tối đa (Maximum Matching) với tập dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt.

Công cụ được xây dựng bằng ngôn ngữ Java, mã nguồn mở. Có thể dễ dàng sửa đổi nâng cấp và tích hợp vào các hệ thống phân tích văn bản tiếng Việt khác.

Quy trình thực hiện tách từ theo phương pháp khớp tối đa



Hình 2.2 – Quy trình tách từ theo vnTokenizer

- Đầu vào của công cụ tách từ vnTokenizer là một câu hoặc một văn bản được lưu dưới dạng tệp.
- Đầu ra là một chuỗi các đơn vị từ được tách.
- Các đơn vị từ bao gồm các từ trong từ điển cũng như các chuỗi số, chuỗi kí tự nước ngoài, các hình vị ràng buộc (gồm các phụ tố), các dấu câu và các chuỗi kí tự hỗn tạp khác trong văn bản (ISO, 2008). Các đơn vị từ không chỉ bao gồm các từ có trong từ điển, mà cả các từ mới hoặc các từ được sinh tự do theo một quy tắc nào đó (như phương thức thêm

phụ tố hay phương thức láy) hoặc các chuỗi kí hiệu không được liệt kê trong từ điển.

Công cụ sử dụng tập dữ liệu đi kèm là tập từ điển từ vựng tiếng Việt, danh sách các đơn vị từ mới bổ sung, được biểu diễn bằng ôtômat tối tiểu hữu hạn trạng thái, tệp chứa các biểu thức chính quy cho phép lọc các đơn vị từ đặc biệt (xâu dạng số, ngày tháng,…), và các tệp chứa các thống kê unigram và bigram trên kho văn bản tách từ mẫu.

Với các đơn vị từ đã có trong từ điển, khi thực hiện tách từ cũng được xử lý hiện tượng nhập nhằng bằng cách kết hợp với các thống kê unigram và bigram. Chẳng hạn trong tiếng Việt thường gặp các trường hợp nhập nhằng như:

- Xâu AB vừa có thể hiểu là 1 đơn vị từ, vừa có thể là chuỗi 2 đơn vị từ A-B.
- Xâu ABC có thể tách thành 2 đơn vị AB-C hoặc A-BC.

Đánh giá kết quả: Kết quả đánh giá của công cụ được cho là ổn định đối với nhiều loại văn bản/ văn phong khác nhau. Độ chính xác trung bình đạt được là khoảng 94%.

2.7.2.5. Phương pháp tách từ sử dụng n-grams

Nhiều phương pháp tách đặc trưng đã được thử nghiệm cho văn bản tiếng Việt. Trong số đó phải kể đến phương pháp sử dụng n-grams [1]. Phương pháp này coi mỗi đặc trưng là một cụm gồm n từ nằm liền nhau. Ưu điểm lớn nhất của phương pháp này là đơn giản và cho kết quả khá tốt. Tuy nhiên, nhóm tác giả nói trên lựa chọn ngay n = 1,2,3 và không so sánh với những giá trị n khác.

2.7.3. So sánh các phương pháp tách từ tiếng Việt

Nhìn chung, phương pháp dựa trên từ (word-base) cho độ chính xác khá cao (trên 95%) nhờ vào tập dữ liệu lớn, được đánh dấu chính xác, tuy nhiên

hiệu suất của thuật toán phụ thuộc hoàn toàn vào dữ liệu huấn luyện. Với các phương pháp cần phải sử dụng từ điển hoặc tập huấn luyện, ngoài việc tách từ thật chính xác, còn có thể nhờ vào các thông tin đánh dấu trong tập dữ liệu để thực hiện các mục đích khác cần đến việc xác định từ loại như dịch máy, kiểm tra lỗi chính tả, từ điển đồng nghĩa… Do vây, dù thời gian huấn luyện khá lâu, cài đặt phức tạp, chi phí tạo tập dữ liệu lớn rất tốn kém, nhưng kết quả mà hướng tiếp cận dựa vào từ mang lại cho mục đích dịch máy là rất lớn.

Hướng tiếp cận dựa trên ký tự (character-based) có ưu điểm dễ thực hiện, thời gian thực thi tương đối nhanh, tuy nhiên lại có độ chính xác không cao bằng phương pháp dựa trên từ. Hướng tiếp cận này thích hợp cho các mục đích nghiên cứu không cần đến độ chính xác tuyệt đối cũng như các thông tin về từ loại như phân loại văn bản, lọc spam, firewall… Nhìn trên tổng thể, hướng tiếp cận dựa trên từ có nhiều ưu điểm đáng kể trong việc định hướng nghiên cứu.

Dựa trên phần so sánh tổng thể các phương pháp và định hướng tách từ nêu trên cùng với mục tiêu chính của đề tài là phân loại nội dung web bằng tiếng Việt nên đề tài quyết định chọn hướng tiếp cận dựa trên "tiếng". Tuy nhiên, việc phân loại văn bản không yêu cầu việc tách từ phải có độ chính xác cao đến mức từng từ nên luận văn không tập trung vào mặt ý nghĩa cũng như những đặc trưng phức tạp của tiếng Việt như từ đồng nghĩa, từ láy, … mà chỉ xác định tần số của từ đơn, từ ghép tiếng Việt xuất hiện trong nội dung cần lọc nên hướng tiếp cận khác với các phương pháp xác định ngữ nghĩa từ tiếng Việt. Phần dưới sẽ trình bày những đặc điểm chính của phương pháp tiếp cận vấn đề.

Phương pháp tách từ sử dụng n-grams tuy không giải quyết được bài toán nhập nhằng về ngữ nghĩa từ nhưng có lợi thế khi áp dụng vào bài toán phân lớp văn bản do bộ từ điển từ dễ dàng cập nhật lượng từ đầy đủ phù hợp

với lớp văn bản mà đang muốn phân lớp mà không bị chi phối bởi các lớp khác do trong tiếng Việt có rất nhiều lĩnh vực mà tùy từng lĩnh vực, chủ đề khác nhau nên có nhiều từ, tiếng khác nhau về mặt phát âm cũng như ý nghĩa, đồng thời việc xử lý tốn một khoảng thời gian có thể chấp nhận được. Phần trên đã đưa ra các phương pháp tách từ trong tiếng Việt cũng như so sánh ưu nhược điểm của các phương pháp đó. Phần tiếp theo sẽ trình bày ứng dụng phương pháp tách từ để xây dựng bộ lọc thư rác tiếng Việt.

CHƯƠNG 3: XÂY DỰNG BỘ LỌC THƯ RÁC SONG NGỮ ANH – VIỆT DỰA TRÊN BAYES

3.1. Tổng quan mạng Bayes

3.1.1. Giới thiệu chung

Thuật toán Naïve Bayes [5] là một thuật toán phân tích thống kê, nó thực hiện trên dữ liệu số. Mô hình xác suất Naïve Bayes là phương pháp được sử dụng phổ biến nhất trong phân lớp tài liệu text. Ý tưởng của phương pháp Naïve Bayes là sử dụng các xác suất liên kết của các nhóm dựa trên một tài liệu. Sự đơn giản của nó là giả thiết các từ độc lập nhau.

Thuật toán Naïve Bayes trong bài toán lọc nội dung được thực hiện trên nguyên tắc coi một tài liệu text là được phát sinh bởi cách chọn ngẫu nhiên từ tất cả các từ có mặt trong nhóm. Các từ có cơ hội được bổ sung vào là tỉ lệ với xác suất tìm thấy từ trong nhóm đang được xem xét. Bộ phân lớp Naïve Bayes sau đó xác định khả năng nội dung cần đang được kiểm tra sẽ thuộc về nhóm nào. Naïve Bayes là một thuật toán đơn giản và nhanh, nó hoạt động tốt với các biểu diễn thống kê như là phương pháp túi từ (bag-of-words). Ngược lại với các phương pháp dựa trên luật, Naïve Bayes có thể được thực hiện tăng cường và cần thiết phải thực hiện bước tiền xử lý bổ sung để tạo vector đặc tính tần suất của từ với kích thước nhỏ. Vì kích thước của vector đặc tính có thể là khá lớn và do vậy cần có các bước bổ sung để giảm kích thước của nó.

3.1.2. Học Bayes (Bayes Learning)

Giả thiết rằng đã có một phân bố xác suất trước cho tất cả các biến cố. Giả thiết này sẽ là một phương pháp định lượng để đánh giá chứng cứ có được trong quá trình huấn luyện. Những phương pháp này cho phép xây dựng một ranh giới chi tiết hơn của các giả thiết luân phiên thay vì chỉ quan tâm đến tính ổn định của các giả thiết. Như vậy, các phương pháp Bayes cung cấp

các thuật toán học thực tế. Ngoài ra, nó còn được coi là một chuẩn để đánh giá các thuật toán học khác.

- **Xác suất điều kiện**

Giả sử rằng ta ấn định một hàm phân bố cho một không gian mẫu và sau đó học để nhận biết biến cố *E*. Cách thức ta thay đổi xác suất của các biến cố còn lại? Gọi xác suất mới của các biến cố *F* là xác suất điều kiện của *F* trên *E* và kí hiệu là *P(F|E)*.

Gọi $\Omega = \{w_1, w_2, w_3, \ldots, w_n\}$ là không gian mẫu gốc với hàm phân bố được gán là $m(w_j)$. Giả sử ta học thấy rằng biến cố *E* đã xảy ra. Ta muốn gán một hàm phân bố mới $m(w_j|E)$ tới Ω để phản ánh lại thực tế này. Rõ ràng là nếu một điểm mẫu w_j không có trong *E*, ta phải có $m(w_j|E) = 0$. Hơn nữa, khi không có thông tin trái ngược, có thể giả sử rằng xác suất cho w_k trong *E* sẽ có độ lớn tương tự đã có trước, khi học thấy *E* xảy ra. Vì lý do này, ta cần:

$$m(w_j|E) = cm(w_k) \tag{3.1}$$

Đối với tất cả w_k trong E, với c là hằng dương. Tuy nhiên ta cũng phải có

$$\left[\sum_E m(w_k \mid E) = c^{\sum_E m(w_k)} = 1\right] \tag{3.2}$$

Do đó

$$c = \frac{1}{\sum_E m(w_k)} = \frac{1}{P(E)} \tag{3.3}$$

Với giả thiết P(E)>0. Do vậy, sẽ định nghĩa

$$m(wk \mid E) = \frac{m(w_k)}{P(E)} \tag{3.4}$$

Cho w_k trong E. Phân bố mới này có tên là phân bố cho điều kiện E

Đối với biến cố F chung, có

$$P(F \mid E) = \sum_{F \cap E} m(w_k \mid E) = \sum_{F \cap E} \frac{m(w_{k)}}{P(E)} = \frac{P(F \cap E)}{P(E)} \tag{3.5}$$

Xác suất điều kiện là xác suất kết hợp với một biến cố F, dựa trên sự xuất hiện của một biến cố liên quan E. Biểu diễn xác suất điều kiện F dựa trên E là P(F|E). P(F|E) cũng có thể được phát biểu là xác suất xuất hiện của F khi E đã xảy ra, xác suất điều kiện được tính bằng công thức sau

$$P(F \mid E) = \frac{P(F \cap E)}{P(E)} \tag{3.6}$$

Có hai định lý quan trọng liên quan đến xác suất điều kiện

Đối với ba biến cố bất kỳ A1, A2 và A3 luôn có quan hệ như sau:

$$P(A1 \cap A2 \cap A3) = P(A1)P(A2) \mid A1)P(A3 \mid A1A2) \tag{3.7}$$

Nếu một biến cố A phải dẫn đến một trong những biến cố độc lập lẫn nhau A1, A2,…, An, khi đó

$$P(A) = P(A1)P(A \mid A1) + P(A2)P(A \mid A2) + ... + P(An)P(A \mid An) \tag{3.8}$$

Biến cố độc lập

Thực tế thường xảy ra trường hợp kiến thức mà một biến cố E nào đó xảy ra không tác động đến xác suất biến cố F khác xảy ra, nghĩa là P(F|E) = P(F). Ta muốn rằng trong trường hợp như thế này, công thức P(F|E) = P(F) cũng sẽ đúng.

Trong thực tế, công thức này bao hàm công thức kia. Nếu những công thức này là đúng, ta có thể nói rằng F là độc lập của E. Thí dụ như ta không mong muốn rằng kiến thức kết quả của việc đánh giá biến cố đầu tiên thay đổi xác suất ta muốn gán cho xác suất kết quả của việc đánh giá biến cố thứ hai, nghĩa là ta không muốn đánh giá thứ hai phụ thuộc vào đánh giá đầu tiên. Ý tưởng này được hình thức hóa thành định nghĩa biến cố độc lập như sau, từ định nghĩa xác suất điều kiện:

$$P(E \mid F) = \frac{P(F \cap E)}{P(E)} \tag{3.9}$$

$$P(E \cap F) = P(F)P(E \mid F) \tag{3.10}$$

Với hai biến cố *E* và *F* bất kì.

Nếu các biến cố *E* và *F* độc lập, sự xuất hiện của *F* không tác động đến sự xuất hiện của *E* và

$$P(E \mid F) = P(E) \tag{3.11}$$

Thay kết quả của công thức (3.12) vào công thức (3.11) ta có công thức cho các biến cố độc lập E và F:

$$P(E \cap F) = P(F)P(E) \tag{3.12}$$

Và ngược lại, nếu $P(E \cap F) = P(F)P(E)$, khi đó các biến cố E và F độc lập. Những phát biểu này có thể được tóm tắt lại như sau:

Các biến cố *E* và *F* độc lập nếu cả *E* và *F* có xác suất dương và nếu *P(E|F) = P(E) thì P(F|E) = P(F)*. Hay nói cách khác: nếu *P(E)> 0 và P(F) > 0*, khi đó *E* và *F* là độc lập với nhau nếu đối với bất kỳ tập con nào *{Ai, Aj,.., Am,}* của chúng, ta đều có:

$$P(Ai \cap Aj \cap ... \cap Am) = P(Ai)P(Aj)...P(Am) \tag{3.13}$$

3.1.3. Công thức Bayes

Cho kết xuất của trạng thái thứ hai trong thực nghiệm hai trạng thái tìm xác suất của kết xuất tại trạng thái đầu. Những xác suất này được gọi là xác suất Bayes. Giả sử rằng ta có tập biến cố {H1, H2, …, Hm} độc lập

$$\Omega = H1 \cup H2 \cup ... \cup Hm \tag{3.14}$$

Ta gọi những biến cố này là giả thuyết. Ta cũng có một biến cố *E* cung cấp một số thông tin về giả thuyết nào là đúng. Ta gọi những biến cố này là dữ liệu huấn luyện. Trước khi nhận dữ liệu huấn luyện, ta có tập xác suất trước $P(H_1)$, $P(H_2)$, …, $P(H_m)$ đối với các giả thuyết. Nếu ta biết giả thuyết đúng, ta biết được xác suất cho dữ liệu huấn luyện. Tức là, ta biết

P(E|H) với mọi i. Ta muốn tìm xác suất cho giả thuyết với dữ liệu huấn luyện đã cho, nghĩa là muốn tìm xác suất điều kiện $P(H_i|E)$. Những xác suất này gọi là xác suất sau.

Để tìm những xác suất này, ta viết chúng dưới dạng như công thức

$$P(Hi \mid E) = \frac{P(H_i \cap E)}{P(E)} \tag{3.15}$$

Ta có thể tính tử số từ thông tin đã cho bằng

$$P(Hi \cap E) = P(Hi)P(E \mid Hi) \tag{3.16}$$

Do chỉ có duy nhất một biến cố trong số các biến cố H_1, $H_2,\ldots,H_m$ là xảy ra, ta có thể viết xác suất của *E* như sau:

$$P(E) = P(H1 \cap E) + P(H2 \cap E) + (Hm \cap E) \tag{3.17}$$

Sử dụng công thức (2.17), công thức (2.18) có thể viết lại như sau

$$P(H1)P(E \mid H1) + P(H2)P(E \mid H2) + ... + P(Hm)P(E \mid Hm) \tag{3.18}$$

Từ (2.17), (2.18) và (2.20) có được công thức Bayes:

$$P(Hi \mid E) = \frac{P(H_i)P(E \mid H_i)}{\sum_{k=1}^{m} P(H_k)P(E \mid H_k)} \tag{3.19}$$

Công thức (2.21) cho phép ta tìm xác suất của các biến cố khác nhau H_1, $H_2,\ldots, H_n$ mà có thể là nguyên nhân làm cho biến cố *H* xảy ra.

Tâm điểm của định lý Bayes là tính hiển nhiên của một biến cố xác nhận khả năng xảy ra của một giả thuyết đúng với mức độ mà sự xuất hiện của tính hiển nhiên này sẽ là có khả năng xảy ra với giả sử của giả thuyết hơn là sự vắng mặt của nó. Biểu diễn hình thức của định lý Bayes trong trường hợp máy học như sau:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{3.20}$$

Trong đó

D là tập dữ liệu huấn luyện

h là một giả thuyết

P(h|D) là xác suất sau (posterior probability), là xác suất điều kiện của *h* sau khi tập huấn luyện được biểu diễn (dựa trên D)

P(h) là xác suất trước (prior probabiltity) của giả thuyết *h*. Giá trị này thường được tìm bằng cách tìm kiếm trong dữ liệu quá khứ (trong tập huấn luyện)

P(D) là xác suất trước của tập dữ liệu huấn luyện *D*. Giá trị này thường là một hằng số

$$P(D) = P(D \mid h)P(h) + P(D \mid \neg h)P(\neg h) \tag{3.21}$$

Nó có thể được tính dễ dàng khi cho bằng 1

$$P(h \mid D) và P(\neg h \mid D) \tag{3.22}$$

P(D|h) xác suất điều của *D* dựa trên *h*, và được gọi là khả năng có thể xảy ra (likelihood). Giá trị này được gán bằng 1 khi *D* và *h* là nhất quán và được gán bằng 0 khi *D* và *h* không nhất quán.

Định lý Bayes mang tính tổng quát và có thể được áp dụng vào bất kỳ trạng thái nào để tính toán một xác suất điều kiện khi đã biết các xác suất trước. Tính tổng quát của nó được chứng minh qua nguồn gốc của nó, nó rất đơn giản. Nguồn gốc của định lý Bayes không có gì đặc biệt. Nguồn gốc này là ngắn gọn và chỉ sử dụng định nghĩa của xác suất điều kiện và thay thế kết hợp.

3.1.4. Các bước tiến hành lọc nội dung bằng mạng Bayes

– Xác định rõ các đặc trưng sử dụng. Yêu cầu này sẽ xem xét các nội dung website cần hiển thị và tìm các *"từ"* hoặc *"nhóm từ"* mà chúng là dấu hiệu của lành mạnh hay không lành mạnh, đây có thể coi là cơ sở dữ liệu cho bộ lọc. Đây là một phần quan trọng trong nhiệm vụ này và có thể lặp lại một vài lần.
– Sử dụng một số phương pháp lựa chọn đặc trưng để phân tích dữ liệu và chọn đặc trưng, sau đó có thể ước lượng xác suất điều kiện và sử dụng các

luật Bayes để ước lượng xác suất của một nội dung website có phải là lành mạnh hay không

– Xác định rõ ngưỡng để loại bỏ tất cả nội dung email mà xác suất của chúng lớn hơn xác suất này.
– Thử nghiệm hệ thống lọc nội dung thư rác và ước lượng hiệu quả trong thực tế.
– Hệ thống lọc thư rác khác nhiều so với các công việc của phân loại văn bản ở lý do sau: Việc phân loại nhầm một nội dung hợp lệ thành nội dung không hợp lệ sẽ phát sinh hậu quả nghiêm trọng hơn là phân loại nhầm theo chiều ngược lại. Đây là chất lượng khác nhau giữa các lớp mà nó cần được ghi chép lại trong quá trình tính toán.

3.2. Phân tích nội dung email

3.2.1. Phân loại nội dung email

Khi một nội dung email được yêu cầu hiển thị thì nội dung đó thuộc vào một trong hai dạng: tiếng Anh hoặc tiếng Việt. Tuy nhiên, hai ngôn ngữ này có những đặc thù khá riêng biệt ngoại trừ đặc điểm chung đều là ngôn ngữ Latinh, cụ thể như bảng bên dưới:

Bảng 3.1 – Sự khác biệt cơ bản giữa tiếng Anh và tiếng Việt

Đặc điểm của tiếng Việt	**Đặc điểm của tiếng Anh**
Được xếp là loại hình đơn lập (isolae) hay còn gọi là loại hình phi hình thái, không biến hình, đơn tiết	Là loại hình biến cách (flexion) hay còn gọi là loại hình khuất chiết
Từ không biến đổi hình thái, ý nghĩa ngữ pháp nằm ở ngoài từ Ví dụ: Chị ngã em nâng và Em ngã chị nâng	Từ có biến đổi hình thái ý nghĩa ngữ pháp nằm trong từ. Ví dụ: I see him và He sees me

Phương thức ngữ pháp chủ yếu: trật tự từ và hư từ Ví dụ: Gạo xay và Xay gạo	Phương thức ngữ pháp chủ yếu là phụ tố Ví dụ: studying và studied
Ranh giới từ không được xác định mặc nhiên bằng khoảng trắng	Kết hợp giữa các hình vị là chặt chẽ, khó xác định, được nhận diện bằng khoảng trắng hoặc dấu câu
Tồn tại loại từ đặc biệt " từ chỉ loại" (classifier) hay còn gọi là phó danh từ chỉ loại kèm theo với danh từ như: cái bàn, cuốn sách, bức thư..	Hiện tượng cấu tạo bằng từ ghép thêm phụ tố (affix) vào gốc từ là rất phổ biến Ví dụ: anticomputerizational
Có hiện tượng láy và nói láy trong tiếng Việt. Ví dụ: lấp lánh, lung linh	

Từ bảng so sánh trên, có thể thấy được những đặc trưng cơ bản của tiếng Việt cũng như là khó khăn gặp phải khi tách từ trong tiếng Việt.

3.2.2. Đặc trưng của ngôn ngữ tiếng Việt

– Đơn vị cấu tạo từ là tiếng, tức là những âm tiết được sử dụng trong thực tiễn ngôn ngữ Việt. Tiếng có thể có nghĩa đủ rõ, có thể mang nghĩa bị phai mờ và tiếng có thể tự mình không có nghĩa. Hơn nữa, 3 hiện tượng này có thể chuyển hóa lẫn nhau.

– Tính chất âm tiết (tiếng) là một trong những đặc điểm chi phối đặc tính loại hình của ngôn ngữ Việt. Xét ở mặt số lượng tiếng:

+ Từ chỉ chứa một tiếng, gọi là từ đơn, như: nhà, đã, …

+ Từ nhiều tiếng, phần lớn là 2 tiếng, gọi là từ phức, như: nhà cửa, sạch sẽ, …

Nếu xét ở số lượng từ tố (yếu tố nhỏ nhất tham gia cấu tạo từ) tham gia cấu tạo từ thì có sự phân chia như sau:

+ Từ chỉ chứa một từ tố, gọi là đơn tố, như: nhà, đủng đỉnh, ra đi ô, …

+ Từ đơn tố gồm nhiều tiếng và có hiện tượng hòa âm tạo nghĩa, gọi là từ láy. Nếu không thì nó thuộc loại ngẫu kết.

+ Từ chứa nhiều từ tố, gọi là từ đa tố, như: nhà cửa, xe đạp, sạch sẽ, …

+ Từ đa tố nếu có hiện tượng hòa âm phối ngữ âm tạo nghĩa thì thuộc kiểu láy. Nếu không thì thuộc loại từ ghép.

– Việc tiền xử lý văn bản (tách từ, tách đoạn, tách câu…) sẽ thêm phức tạp với phần xử lý các hư từ, phụ từ, từ láy…
– Phương thức ngữ pháp chủ yếu là trật tự từ nên nếu áp dụng phương pháp tính xác suất xuất hiện của từ có thể không chính xác như mong đợi.
– Ranh giới từ không được xác định mặc nhiên bằng khoảng trắng. Điều này khiến cho việc phân tích hình thái (tách từ) tiếng Việt trở nên khó khăn. Việc nhận diện ranh giới từ là quan trọng làm tiền đề cho các xử lý tiếp theo sau đó như: kiểm tra lỗi chính tả, gán nhãn từ loại, thống kê tần suất từ…
– Vì giữa tiếng Anh và tiếng Việt có nhiều điểm khác biệt nên không thể áp dụng y nguyên các thuật toán tiếng Anh vào tiếng Việt.

Chính vì những nguyên nhân đó phần tiếp theo sẽ đề xuất các phương pháp xử lý nội dung tiếng Việt và tiếng Anh.

3.2.3. Phương pháp xử lý nội dung email

Như đã trình bày ở trên, nội dung email đang được đề cập là tiếng Việt hay tiếng Anh. Dưới đây sẽ đề xuất các phương pháp xử lý nội dung email.

- Cách thứ nhất là phân chia nội dung thành tiếng Anh và tiếng Việt, sau đó tiến hành phân loại nội dung tiếng Anh và tiếng Việt riêng. Tất nhiên, có thể có trường hợp trong một nội dung có cả tiếng Việt và tiếng Anh nhưng tỷ lệ này không nhiều.
- Cách thứ hai là xây dựng một bộ phân loại chung cho cả tiếng Anh và tiếng Việt. Cách thứ hai đơn giản hơn nhưng có thể gặp vấn đề khi lựa chọn tham số k để tách các k-gram.
- Nếu sử dụng cách thứ nhất thì xuất hiện một vấn đề cần giải quyết là phân biệt nội dung tiếng Anh và tiếng Việt. Mặc dù có những giải pháp phức tạp hơn được đề xuất cho vấn đề này, ở đây đề xuất sử dụng một giải pháp rất đơn giản. Khi lựa chọn đặc trưng, các đặc trưng được đánh dấu riêng tiếng Việt hoặc tiếng Anh và lưu vào bảng băm. Khi một nội dung mới xuất hiện, 10 đặc trưng ngẫu nhiên trong nội dung sẽ được băm vào bảng tiếng Việt và tiếng Anh. Nếu số lượng băm trúng trong bảng tiếng Việt lớn hơn bảng tiếng Anh thì nội dung được coi là nội dung tiếng Việt và ngược lại. Tuy nhiên, đối với những nội dung sử dụng cả tiếng Việt và tiếng Anh việc kết luận nội dung thuộc một trong hai ngôn ngữ duy nhất có thể ảnh hưởng tới quá trình phân loại tiếp theo.

Sau khi phân biệt được nội dung tiếng Anh thì sẽ được lọc riêng. Hiệu quả phân loại chung sau đó được lấy bằng trung bình cộng của phân loại cho nội dung tiếng Việt và nội dung tiếng Anh. Để tăng độ chính xác trong quá trình phân tích nội dung, có thể chia nhỏ nội dung thành từng câu đơn thể nhằm tạo tiền đề cho việc tách từ tiếng Việt mang lại độ chính xác cao nhất.

3.2.4. Phân tích câu

Quan niệm câu là một chuỗi ký tự kết thúc bởi một dấu chấm (.), (?) hay (!) không thể loại trừ các nhập nhằng, trong đó dấu chấm câu không chỉ là ký hiệu kết thúc câu: một số dùng trong các từ viết tắt hoặc trong chuỗi số. Tuy

nhiên, phương pháp dựa trên kinh nghiệm cơ bản này cho kết quả không tồi: nhìn chung, khoảng 90% các dấu chấm là ký hiệu kết thúc câu. Tuy nhiên, cũng cần lưu ý các trường hợp: trong đó các ký hiệu khác có thể được coi là dấu hiệu kết thúc câu. Ví dụ: các dấu câu như hai chấm, dấu chấm phẩy và dấu ngang (“:” , “;” và “-”) có thể theo sau bởi một câu hoàn chỉnh.

Mục đích cơ bản của phân tích từ vựng là tách và xác định các đặc trưng của văn bản, bắt đầu với việc tách một thông điệp ra thành các bộ phận nhỏ hơn, thường là các từ đơn giản.Vì vậy, việc tách câu rất quan trọng hỗ trợ cho việc tách từ về sau. Vì thế dấu phân cách nên dùng là khoảng trắng, vì khoảng trắng thường dùng để tách các từ trong hầu hết các ngôn ngữ, sau đây là một số phân cách câu được dùng rộng rãi:

+ Dấu chấm (.)
+ Dấu phẩy (,)
+ Dấu chấm phẩy (;)
+ Dấu nháy đôi (“)
+ Dấu hai chấm (:)
+ Dấu ngoặc vuông [ ]
+ Dấu ngoặc nhọn { }
+ Dấu ngoặc đơn ( )
+ Các toán tử + - / * = <>

Hiện nay, việc tách câu thường dựa trên một số tiêu chí sau đây:

- Đặt điểm phân cách câu sau dấu đóng ngoặc kép (nếu có)
- Loại ra một điểm phân cách câu giả định (là dấu chấm) trong các trường hợp sau:
    - Nếu nó đi sau một từ viết tắt thường không xuất hiện ở cuối câu, nhưng thường đi trước một danh từ riêng, ví dụ: Prof hay vs

- Nếu nó đi sau một từ viết tắt đã biết và không đi trước một từ viết hoa. Trường hợp này có thể giải quyết đúng hầu hết các trường hợp viết tắt như etc. hoặc Jr. (những từ có thể xuất hiện ở giữa hoặc cuối câu).

- Loại một điểm phân cách câu giả định với ? hay ! nếu nó đi trước một từ không viết hoa.
- Xem xét tất cả các điểm phân cách câu giả định còn lại như các điểm phân cách câu thực sự.

3.3. Xây dựng bộ lọc thư rác song ngữ Anh – Việt

3.3.1. Ý tưởng đề xuất

Ý tưởng đề xuất là tìm cách xây dựng một bộ phân loại nhằm phân loại cho một mẫu mới bằng cách huấn luyện từ những mẫu có sẵn. Ở đây mỗi mẫu được xét đến chính là mỗi một nội dung của email, tập các lớp mà mỗi nội dung có thể thuộc về là *y={tốt, xấu}*

Khi nhận được 1 nội dung cần hiển thị, dựa vào đặc điểm hay thuộc tính nào đó của nội dung để tăng khả năng phân loại chính xác nội dung đó. Các đặc điểm của 1 nội dung như: tiêu đề, nội dung, … Càng nhiều những thông tin như vậy xác suất phân loại đúng càng lớn, tất nhiên còn phụ thuộc vào kích thước của tập mẫu huấn luyện.

Việc tính toán xác suất sẽ dựa vào công thức Naïve Bayes, Theo Bayes ta phải tính P(Spam/email) và P(Ham/email) sau đó so sánh 2 giá trị này, giá trị nào lớn thì email được phân vào lớp đó. Tuy nhiên, khi phân loại nội dung có hai lỗi: lỗi nhận một nội dung tốt thành xấu và lỗi cho qua một thư rác. Loại lỗi thứ nhất nghiêm trọng hơn, vì vậy thông thường khi P(Spam/email) > = P(Ham/email) + ngưỡng nào đó thì ta mới phân nó vào lớp Spam.

P(spam) >= P(ham) + x => thư rác

P(spam) < P(ham) + x => thư tốt

3.3.2. Hướng tiếp cận

Theo ý tưởng đề xuất trên thì vấn đề cần giải quyết là phân lớp một nội dung email vào một trong hai lớp tốt và xấu trong một khoảng thời gian chấp nhận được. Dựa vào các phương pháp tách từ nêu trên cùng với ưu nhược điểm, hướng tiếp cận của đề tài dựa vào phương pháp tách từ dựa vào tần số xuất hiện của từ mà không dựa vào ngữ nghĩa của từ kết hợp với thuật toán Naïve Bayes vì những lý do sau đây:

Nội dung xấu chỉ nằm trong phạm vi (quảng cáo, tài chính, lừa đảo…). Do đó, đặc điểm và số lượng từ sẽ chỉ nằm trong lĩnh vực nhất định.

Bộ từ điển từ thuộc lĩnh vực nêu trên sẽ tạo mới và cập nhật thuận lợi dễ dàng do đã giới hạn phạm vi, đồng thời, thời gian xử lý phải đảm bảo nhanh chóng nên nếu theo hướng tiếp cận xử lý ngữ nghĩa sẽ mất rất nhiều thời gian.

Theo các công trình đã công bố. Naïve Bayes cho hiệu quả cao trong các bài toán phân lớp văn.

Đề tài ngăn chặn từ khóa theo các hướng sau

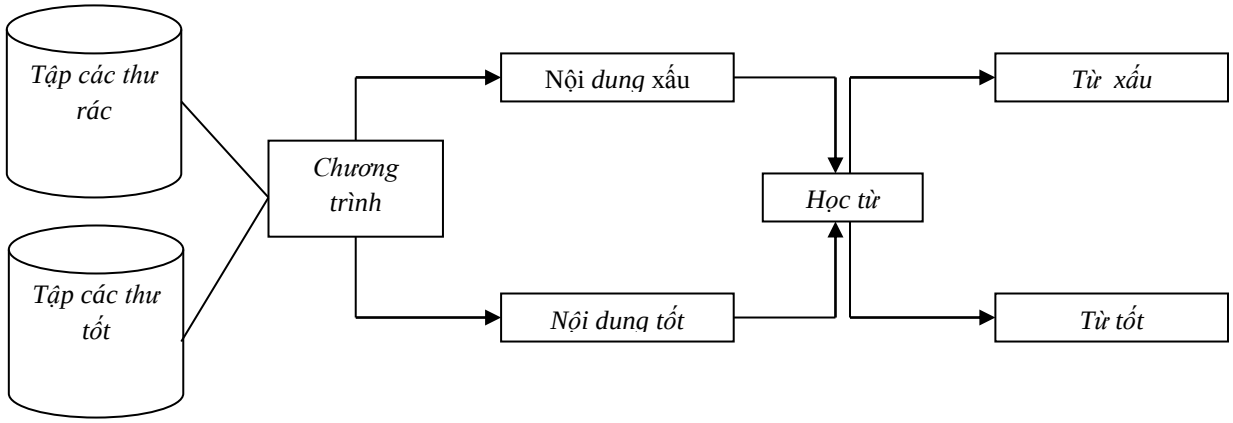
- Dựa vào nội dung trong tiêu đề của email (title)
- Dựa vào nội dung chính của email

3.3.3. Tiến trình thu thập nội dung

Ở tiến trình này sẽ làm nhiệm vụ thu thập dữ liệu.

Đầu tiên, đầu vào là tập các thư rác và tập các thư tốt, chương trình sẽ truy xuất trực tiếp vào nội dung toàn diện rồi tiến hành bóc tách. Sau quy trình khai thác nội dung sẽ độc lập với thư nguồn, được lưu trữ và tái sử dụng cho bước học từ.

Từ bước học từ ta đã xây dựng được bộ từ điển từ xấu và từ tốt



Hình 3.1 – Tiến trình học từ

- Từ xấu: Những từ thường xuất hiện trong thư spam hơn là trong thư ham.
- Từ tốt: Những từ thường xuất hiện trong thư ham hơn là trong thư spam.

3.3.4. Tiến trình phân loại ngôn ngữ Anh – Việt

Phân chia nội dung thành tiếng Anh và tiếng Việt, sau đó tiến hành phân loại nội dung tiếng Anh và tiếng Việt riêng. Tất nhiên, có thể có trường hợp trong một nội dung có cả tiếng Việt và tiếng Anh nhưng tỷ lệ này không nhiều.

Khi lựa chọn đặc trưng, các đặc trưng được đánh dấu riêng tiếng Việt hoặc tiếng Anh và lưu vào bảng băm. Khi một nội dung mới xuất hiện, chọn k đặc trưng ngẫu nhiên trong nội dung sẽ được băm vào bảng tiếng Việt và tiếng Anh. Nếu số lượng băm trúng trong bảng tiếng Việt lớn hơn bảng tiếng Anh thì nội dung được coi là nội dung tiếng Việt và ngược lại. Tuy nhiên, đối với những nội dung sử dụng cả tiếng Việt và tiếng Anh việc kết luận nội dung thuộc một trong hai ngôn ngữ duy nhất có thể ảnh hưởng tới quá trình phân loại tiếp theo.

Theo cách trên để xác định thư thuộc ngôn ngữ nào cần dựa vào số token ngẫu nhiên lựa chọn, nhưng lựa bao nhiêu là đủ là một vấn đề. Để lựa chọn được phù hợp, luận văn đã tiến hành thử nghiệm lựa chọn số token ngẫu nhiên và kết quả thử nghiệm trên 1500 thư và cho độ chính xác nhận diện một email là tiếng Anh hay Việt như trong bảng 3.2.

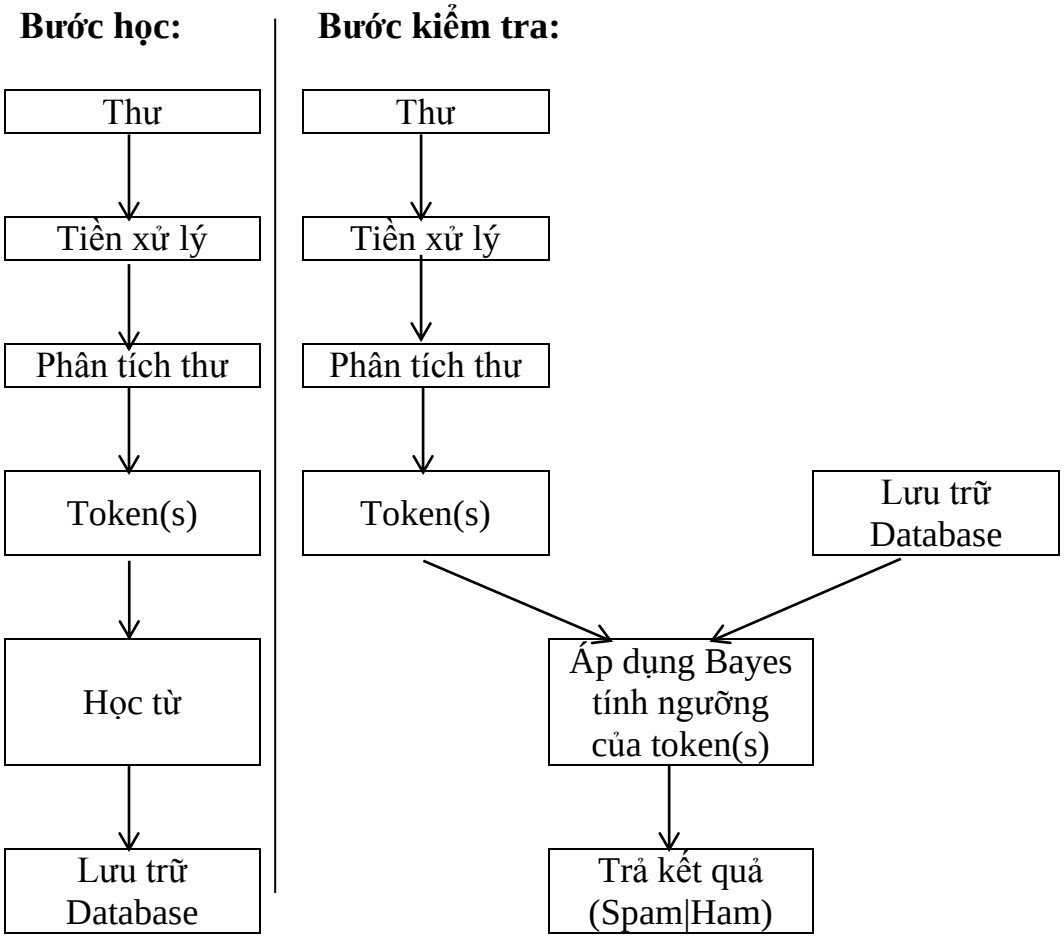
Bảng 3.2 – Bảng thực nghiệm độ chính xác phân loại Anh – Việt

K	**6 tokens**	**8 tokens**	**10 tokens**	**12 tokens**	**14 tokens**
Đặc thù Anh hoặc Việt	100%	100%	100%	100%	100%
Cả tiếng Anh và Việt	80%	86%	98%	98%	99%

Dựa vào kết quả thực nghiệm ta thấy chọn 10 tokens ngẫu nhiên vào việc nhận dạng văn bản là tiếng Anh hay Việt là phù hợp nhất.

3.3.5. Xây dựng bộ lọc thư rác song ngữ Anh – Việt:

3.3.5.1. Mô hình tổng quát:



Hình 3.2 – Mô hình tổng quát

Mô hình tổng quát bao gồm các công việc chính của hệ thống chia ra 2 phần rõ ràng là phần học và phần kiểm tra thư. Sau đây là mô tả chi tiết từng công đoạn.

3.3.5.2. Thư đầu vào:

Thu thập tập thư tốt và thư xấu, gồm 3000 thư tiếng anh (trong đó có 1500 thư tốt, 1500 thư xấu), 1200 thư tiếng việt (trong đó 600 thư tốt và 600 thư xấu). Lấy 1300 thư tốt tiếng anh, 1300 thư xấu tiếng anh, 500 thư tốt tiếng việt, 500 thư xấu tiếng việt để phục vụ cho mục đích học. Còn lại 200 thư tốt

tiếng anh, 200 thư xấu tiếng anh, 100 thư tốt tiếng Việt, 100 thư xấu tiếng Việt để phục vụ cho vấn đề kiểm tra.

3.3.5.3. Tiền xử lý:

Thực hiện nhiệm vụ chuẩn hóa lại nội dung thư bằng cách.

+ Loại bỏ các thẻ định dạng HTML trong thư.

+ Loại bỏ các từ nối câu và các từ không có ý nghĩa trong thư.

+ Các ký tự số vì không nói lên được ý nghĩa của bức thư.

+ Biến đổi toàn bộ nội dung thư thành các câu đơn phân biệt.

Sau khi chuẩn hóa xong nội dung bức thư, tiếng hành phân loại xem đó là thư tiếng Anh hay tiếng Việt bằng cách: chọn ra 10 tokens ngẫu nhiên trong bức thư và bỏ vô bảng từ điển tiếng Việt và tiếng Anh. Nếu số lượng tokens trong bảng tiếng Anh lớn hơn bảng tiếng Việt thì được xem là thư tiếng Anh, nếu số lượng tokens trong bảng tiếng Việt lớn hơn bảng tiếng Anh thì được xem là thư tiếng Việt

3.3.5.4. Phân tích nội dung thư:

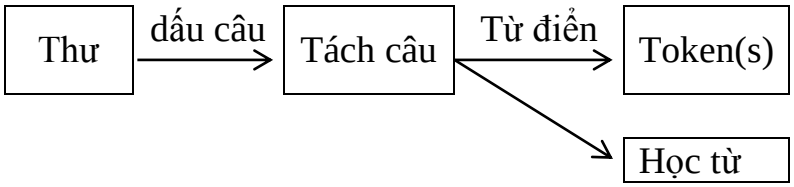
3.3.5.4.a. Thư tiếng Anh:

Trong hầu hết các nghiên cứu lọc thư rác tiếng Anh, đặc trưng được sử dụng là những từ riêng lẻ (word). Do đặc điểm của tiếng Anh nên việc xác định từ trong câu rất đơn giản, mỗi từ được phân cách với từ khác bằng dấu cách hoặc các dấu trắng khác.

3.3.5.4.b. Thư tiếng Việt:

Đối với tiếng Việt, từ có thể bao gồm nhiều tiếng, ví dụ từ “hàng hóa” bao gồm hai tiếng “hàng” và “hóa”. Trong khi có thể tách từng tiếng một cách dễ dàng thì việc xác định từ hoàn toàn không đơn giản.

Quy trình tách từ với thư tiếng Việt



Hình 3.3 – Quy trình tách từ với thư tiếng Việt

Từ bức thư ban đầu sau khi được tiền xử lý, nếu là thư tiếng Việt thì thực hiện quy trình trên. Tách bức thư ra làm nhiều câu đơn dựa vào dấu câu. Sau khi tách câu xong ta tiến hành thực hiện 2 bước sau:

Bước 1: Lấy ra các đặc trưng

Từ những câu đơn đó tiến hành lấy ra token(s) dựa vào bộ từ điển đã được xây dựng sẵn bằng phương pháp khớp tối đa [3]. Nếu có nhiều đặc trưng lồng nhau thì đặc trưng dài nhất sẽ được lựa chọn.

Bước 2: Học thêm từ vào bộ từ điển bằng cách

Sử dụng phương pháp n-grams [1]. Phương pháp này coi mỗi đặc trưng là một cụm gồm n từ nằm liền nhau. Ưu điểm lớn nhất của phương pháp này là đơn giản và cho kết quả khá tốt. Tuy nhiên, nhóm tác giả nói trên lựa chọn ngay n=1,2,3 và không so sánh với những giá trị n khác.

Giả sử như ta có 1 câu: “học sinh rất hài lòng với cách dạy”, trong từ điển của ta chỉ mới có từ “ hài lòng” thì khi đó bước học từ sẽ được mô tả như sau:

Khi đó các từ còn lại cần phải học là:

N=2: học sinh, sinh rất, với cách, cách dạy

N=3: học sinh rất, với cách dạy

Sau khi lấy ra được n-grams này cập nhật tần số xuất hiện của nó trong dữ liệu tạm thời trong cơ sở dữ liệu. Đến một ngưỡng α nào đó thì từ tạm đó sẽ được chuyển vô từ điển.

Ngưỡng α được xác định như sau:

$$\alpha = \frac{k}{Total\ message} \quad (3.23)$$

Trong đó:

- K: tần số xuất hiện của từ.
- Total Message: Tổng số thư.

Dựa vào kết quả thực nghiệm tách từ, thử nghiệm với khoảng 1000 từ đạt độ chính xác 94% nếu ngưỡng α >= 0.25 thì độ chính xác của từ có thể chấp nhận được. Những từ có ngưỡng α nằm ngoài ngưỡng trên được xếp vào tập các từ cần được huấn luyện tiếp tục.

3.3.5.5. Các bước tiếp theo

3.3.5.5a. Bước học:

Lưu trữ vào database: Sau khi lấy ra được Tokens(s).

Khi bức thư đã được phân tích token của nó sẽ được sử dụng để cập nhật tần số. Được dùng để theo dõi số lần xuất hiện của token đó trong thư rác và thư tốt. Mỗi token chỉ được tính xuất hiện một lần trong một bức thư.

Ví dụ: Từ "giáo viên" xuất hiện trong 50 thư rác và 230 thư tốt thì khi đó tần suất lưu trữ của từ "giáo viên" là "giáo viên"|50|230.

3.3.5.5.b. Bước kiểm tra:

- **Tính ngưỡng của token(s):**

Tính ngưỡng của các tokens: dựa vào token(s) được phân tích từ nội dung bức thư và tần số của các token(s) đã được lưu trữ ở cơ sở dữ liệu trước đó và công thức Naïve Bayes, áp dụng công thức tính xác suất cho các token(s) như sau:

+ Giả sử mỗi nội dung bực thư kiểm tra: content

+ Lớp thư rác: spam

+ Lớp thư tốt: ham

+ Word1, Word2, Word3, ... Wordn là các từ đặc trưng xuất hiện trong content.

Ta có

$$P(spam|content) = \frac{P(content|spam) * P(spam)}{Total\ message} \quad (3.24)$$

Trong đó total được xác định bằng

$$Total=P(content|spam)* P(spam)+ P(content|ham)* P(ham) \quad (3.25)$$

Với P(content|ham) và P(content|spam) được tính bằng

$$P(content|ham) = \prod_{1<i<n} P(word_i|\text{ham}) \quad (3.26)$$

$$P(content|spam) = \prod_{i<i<n} P(word_i|\text{spam}) \quad (3.27)$$

Cuối cùng, P(spam) và P(ham) được tính bởi công thức

$$P(spam) = \frac{Total\ spam}{Total\ message} \quad (3.28)$$

$$P(ham) = \frac{Total\ ham}{Total\ message} \quad (3.29)$$

- **Trả kết quả:**

Sau khi tính đc P(spam) và P(ham) theo Bayes ta sẽ so sánh 2 giá trị này nếu P(spam) > P(ham) thì ta quyết định đó là thư rác, ngược lại nếu P(spam) < P(ham) thì ta quyết định đó là thư bình thường.

Việc phân loại nhầm một thư hợp lệ thành thư rác sẽ phát sinh hậu quả nghiêm trọng hơn là phân loại nhầm một thư rác thành thư hợp lệ. Nên quá trình phân loại kết quả có thể cho người dùng tự quyết định giá trị sai số x nào đó tùy theo nhu cầu của người sử dụng

P(spam) > P(ham) + x => thư rác

P(spam) <= P(ham) + x => thư tốt

CHƯƠNG 4. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Dữ liệu thử nghiệm

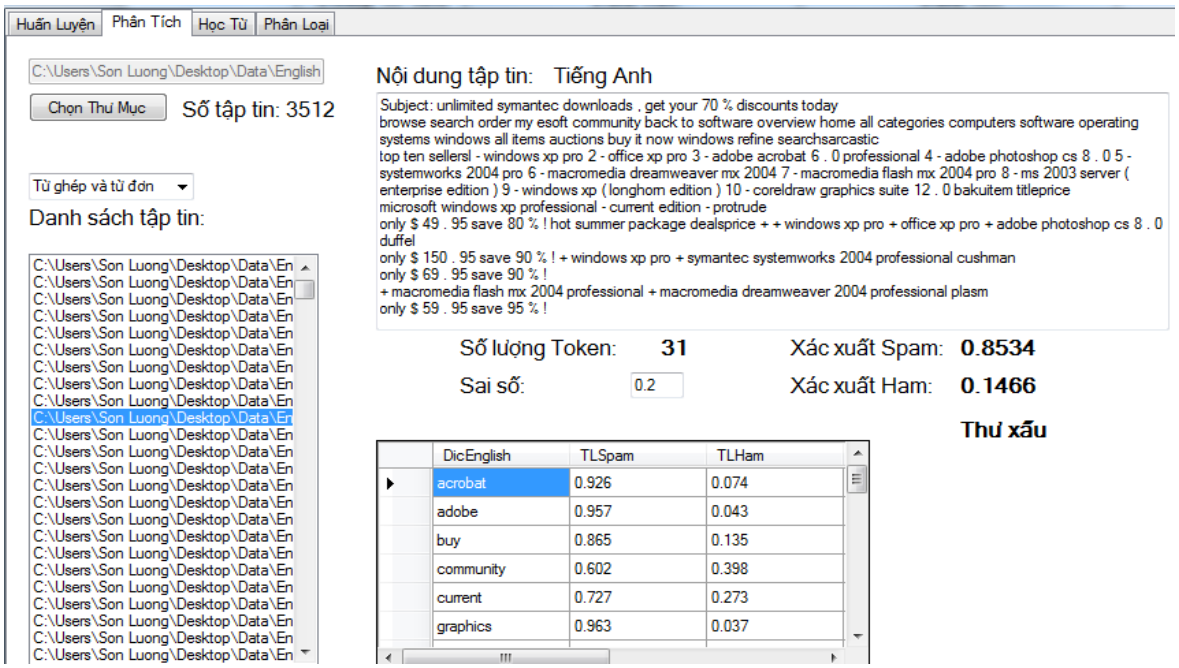
Một khó khăn khi thử nghiệm lọc thư là hiện nay chưa có những bộ dữ liệu mẫu chuẩn. Do vậy thư rác được thu thập được qua địa chỉ thư của mình và bạn bè tại Việt nam. Thư bình thường là những thư mà nhận được bao gồm cả thư tiếng Việt và tiếng Anh.

Đối với các thư bình thường nhận được chỉ giữ lại nội dung thư nhận được cuối cùng. Đối với những thư bao gồm cả văn bản và hình ảnh, chỉ có phần văn bản được sử dụng, phần hình ảnh bị bỏ qua không xem xét. Các thông số chính về bộ dữ liệu thử nghiệm được thống kê trong bảng 4.1.

Bảng 4.1 – Bộ dữ liệu thử nghiệm

Tổng số thư	Thư rác		Thư tốt	
600	Tiếng Việt	Tiếng Anh	Tiếng Việt	Tiếng Anh
	100	200	100	200

4.2. Thử nghiệm với thư tiếng Anh



Hình 4.1: Phân tích thư rác tiếng Anh

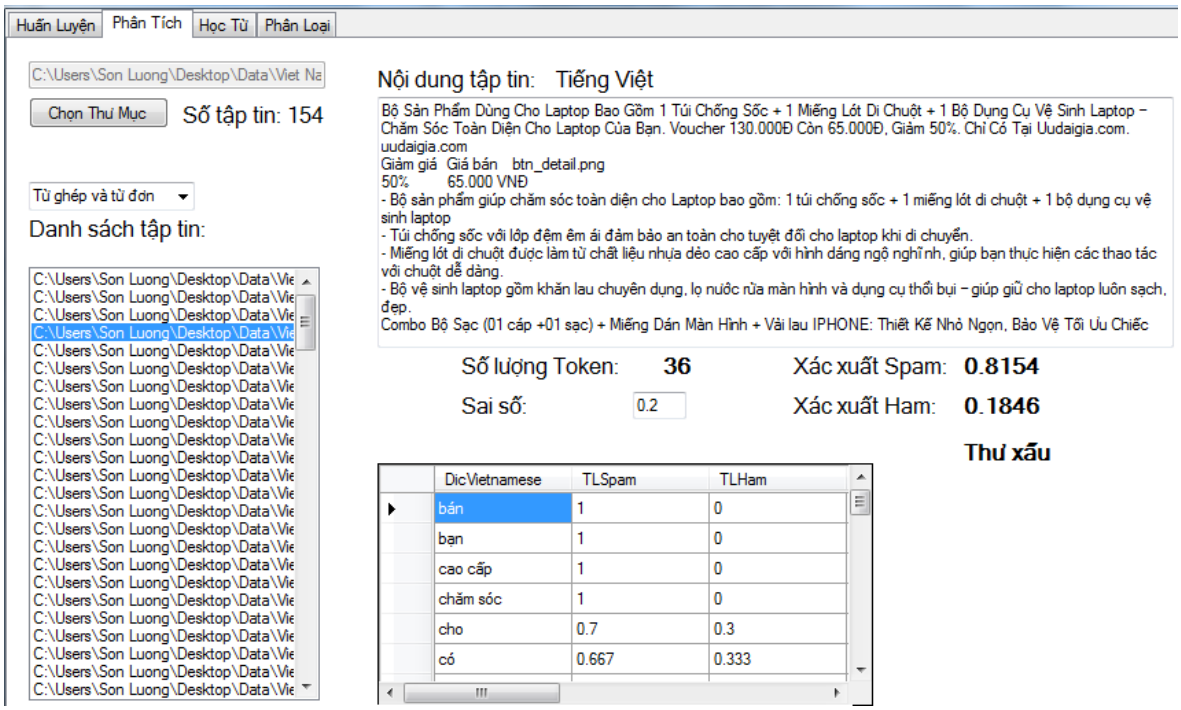
Để bắt đầu quá trình thử nghiệm lọc thư spam tiếng Anh, tác giả đã lấy 400 thư tiếng Anh ở bộ dữ liệu thư tiếng Anh ban đầu gồm 200 thư tốt và 200 thư rác tham gia quá trình lọc thư. Kết quả đạt được được thể hiện bằng bảng thống kê bên dưới.

Bảng 4.2 – Kết quả phân loại thư tiếng Anh

Loại thư	Số thư	Kết quả	Tỷ lệ
Thư tốt	200	193	96,5%
Thư rác	200	181	90,5%

Dựa vào kết quả của bảng thống kê cho thấy rõ việc tiếp cận của thuật toán Naïve Bayes đối với việc lọc thư tiếng Anh cho hiệu quả khả quan. Việc tiếp cận dựa trên phân tích từ này có thể làm tiền đề cho việc lọc thư spam tiếng Việt trên cơ sở từ vựng, bao gồm từ đơn và từ ghép

4.3. Thử nghiệm với thư tiếng Việt



Hình 4.2: Phân tích thư rác tiếng Việt.

Để bắt đầu quá trình thử nghiệm lọc thư spam tiếng Việt, tác giả đã lấy 200 thư tiếng Việt ở bộ dữ liệu thư tiếng Việt ban đầu gồm 100 thư tốt và 100 thư rác tham gia quá trình lọc thư. Kết quả đạt được được thể hiện bằng bảng thống kê bên dưới.

Bảng 4.3 – Kết quả phân loại thư tiếng Việt

Thư	**Thư tốt**	**Thư rác**	**Tỷ lệ**	
	100	100	**Thư tốt**	**Thư rác**
Từ đơn	94	83	94%	83%
Từ ghép	95	80	95%	80%
Từ ghép và từ đơn	97	88	97%	88%

Dựa trên kết quả thực nghiệm, cho thấy rõ việc lọc thư spam tiếng Việt theo từ ghép và từ đơn cho kết quả khả quan nhất (88%) so với kết quả lọc theo từ ghép (80%) và vừa từ đơn (83%). Điều này phần nào thể hiện hướng tiếp cận đúng của đề tài.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Đề tài đã đạt được những kết quả sau đây:

- Tìm hiểu các phương pháp lọc thống kê cũng như những điểm mạnh của các kỹ thuật phân loại văn bản nhằm áp dụng tốt vào quy trình lọc thư rác.
- So sánh các phương pháp tách từ trong tiếng Việt , từ đó lựa chọn phương pháp tối ưu nhất để giải quyết bài toán và xây dựng bộ từ điển hoàn chỉnh cho bài toán lọc thư rác.
- Nghiên cứu các thuận toán, đặc biệt là thuật toán Naïve Bayes ứng dụng vào quá trình phân lớp thư rác.
- Phân tích và xây dựng công cụ xác định một văn bản hay email là tiếng Anh hay tiếng Việt.
- Xây dựng được bộ lọc thư rác hỗ trợ cho 2 ngôn ngữ: Anh và Việt.
- Với tiếng Việt đã xây dựng được bộ từ điển tương đối đầy đủ gồm 4088 từ đơn và 7562 từ ghép về lĩnh vực thư rác, quá trình học từ vẫn tiếp tục học trong quá trình sử dụng bộ lọc.
- Đã thử nghiệm lọc thư spam tiếng Việt trên cả 03 cơ chế: từ đơn, từ ghép và cả từ đơn lẫn từ ghép. Có kết quả thực nghiệm để so sánh giữa 03 cơ chế trên

  → Kết quả thực nghiệm cho thấy hướng tiếp cận của đề tài khả quan cho độ chính xác cao trong một khoảng thời gian chấp nhận được. Tuy nhiên cần thu thập thêm bộ dữ liệu huấn luyện và thử nghiệm để có được kết luận chính xác nhất có thể.

Hướng phát triển:

– Cần tiếp tục nghiên cứu cải tiến khâu tiền xử lý văn bản, xây dựng các mẫu huấn luyện tiêu chuẩn cũng như điều chỉnh giải thuật để có thể nâng cao độ chính xác phân loại hơn nữa.

– Nâng khả năng lọc thư rác loại nội dung là hình ảnh, ký tự lạ, ….

– Giải quyết tốt hơn ở các định dạng tập tin đính kèm khác của thư rác.

– Xây dựng hệ thống Webmail và tích hợp bộ lọc vào hệ thống.

TÀI LIỆU THAM KHẢO

TIẾNG VIỆT

[1]. N. V. Cường, N. T. T. Linh, H. Q. Thuỵ, P. X. Hiếu. Bài toán lọc và phân lớp nội dung web tiếng Việt với hướng tiếp cận entropi cực đại. Hội thảo quốc gia một số vấn đề chọn lọc của công nghệ thông tin, Hải phòng, 2005.

[2]. Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Giáo trình Khai phá dữ liệu web, Nxb Giáo dục Việt Nam, 2009.

TIẾNG ANH

[3]. Chih-Hao Tsai, A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm, 1996.

[4]. Csaba Gulyás, Creation of a Bayesian network-based meta spam filter, using the analysis of different spam filters, 2006.

[5]. Goldszmidt D., Friedman, N.Geiger, Bayesian network Classifiers Machine Learning, 2006.

[6]. H. David D. Lewis, Ph.D.Ornarose, Inc. & David D. Lewis Consulting, Naive Bayes Text Classification for Spam Filtering ,2007

[7]. Jonathan A. Zdziarski, Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification, No Starch Press, 2005

[8]. Lafferty J., Conditional ramdom fields: probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, 2001.

[9]. Mike Spykerman, Typical spam characteristics, Red Earth Software, 2003.

WEBSITE:

[10]. BBC news, http://news.bbc.co.uk/2/hi/technology/7988579.stm, 2009

[11]. Brad Templeton, http://www.templetons.com/brad/spamterm.html

[12]. Top ten reviews, http://spam-filter-review.toptenreviews.com, 2012.