PHƯƠNG PHÁP LỌC THƯ RÁC TIẾNG VIỆT DỰA TRÊN TỪ GHÉP VÀ THEO VẾT NGƯỜI SỬ DỤNG

Phan Hữu Tiếp[1], Vũ Đức Lung[2], Cao Nguyễn Thủy Tiên[1], Lâm Thành Hiển[1]

[1] Đại học Lạc Hồng

[2] Đại học Công nghệ thông tin, Đại học Quốc Gia Tp.Hồ Chí Minh

Tóm tắt báo cáo. "Lọc thư spam" là bài toán đang được các nhà nghiên cứu quan tâm và đã xuất hiện nhiều hướng tiếp cận để xây dựng các hệ thống lọc cho hiệu quả cao. Tuy nhiên, có những vấn đề khó khăn thách thức khác đối với bài toán này: xây dựng bộ lọc thư spam tiếng Việt. Trong bài báo này, chúng tôi đề xuất mô hình áp dụng thuật toán Naïve Bayes để lọc thư spam tiếng Việt thông qua việc xử lý ngôn ngữ tiếng Việt.

Từ khóa: Lọc thư rác; anti-spam; spam tiếng Việt.

1. Giới thiệu

Tách từ là vấn đề quan tâm nhất khi lọc thư rác tiếng Việt do tiếng Việt có các đặc trưng riêng mặc dù tiếng Việt cũng dùng ký tự latinh như tiếng Anh. Tiếng Việt có 2 thành phần cơ bản [1]: tiếng và từ. Một số mối liên quan giữa từ và tiếng như sau.

Về ngữ pháp, tiếng là đơn vị cấu tạo của từ. Từ là đơn vị nhỏ nhất để tạo câu, hình thức và ý nghĩa của từ độc lập với cú pháp. Có 2 loại từ phổ biến: từ một tiếng (từ đơn) và từ *n* tiếng trở lên ($n<5$) gọi là từ phức. Trong đặt câu tiếng Việt, sử dụng từ chứ không sử dụng tiếng.

Trong tiếng Anh, từ được định nghĩa như sau*: "Từ là một nhóm ký tự có nghĩa, được phân cách bởi ký tự khoảng trắng trong câu"* (từ điển Webter). Ví dụ: "*I am a student*" sẽ tách được 4 từ: *I, am, a, student*. Trong tiếng Việt, ví dụ: "*Tôi là học sinh*" sẽ tách được 3 từ: *tôi, là, học sinh*. Trong đó từ ghép "*học sinh*" là từ được hình thành bởi 2 tiếng: "*học*", "*sinh*". Do sự khác biệt này, khi tách một từ ghép trong các thư rác thành các từ đơn thì lại được dùng phổ biến trong các thư tốt . Cụ thể, từ "*khuyến mãi*" là từ thường đ ược dùng trong thư rác nhưng khi tách ra thành từ "*khuyến*" và từ "*mãi*" thì những từ này lại được sử dụng nhiều trong các thư tốt . Như vậy, đối với thư rác tiếng Việt hướng tiếp cận phân tích dựa vào từ ghép hay từ có nghĩ a chứ không phải dựa vào từ đơn như trong tiếng Anh . Vấn đề hàng đầu đặt ra là chưa có bộ từ tiếng Việt nào hoàn hảo cho việc làm trên.

Trong bài báo này , chúng tôi giới thiệu một kỹ thuật lọc thư rác tiếng Việt đó là áp

dụng thuật toán Naïve Bayes tiếng Việt. Đồng thời, cũng đưa ra một giải pháp tách từ tiếng Việt hoàn toàn mới là dựa vào tần số xuất hiện của từ mà không quan tâm đến ngữ nghĩa của từ. Phần tiếp theo sẽ trình bày: phương pháp tiếp cận, quy trình thực hiện lọc thư & kết quả thử nghiệm, cuối cùng là kết luận.

2. Phương pháp tiếp cận

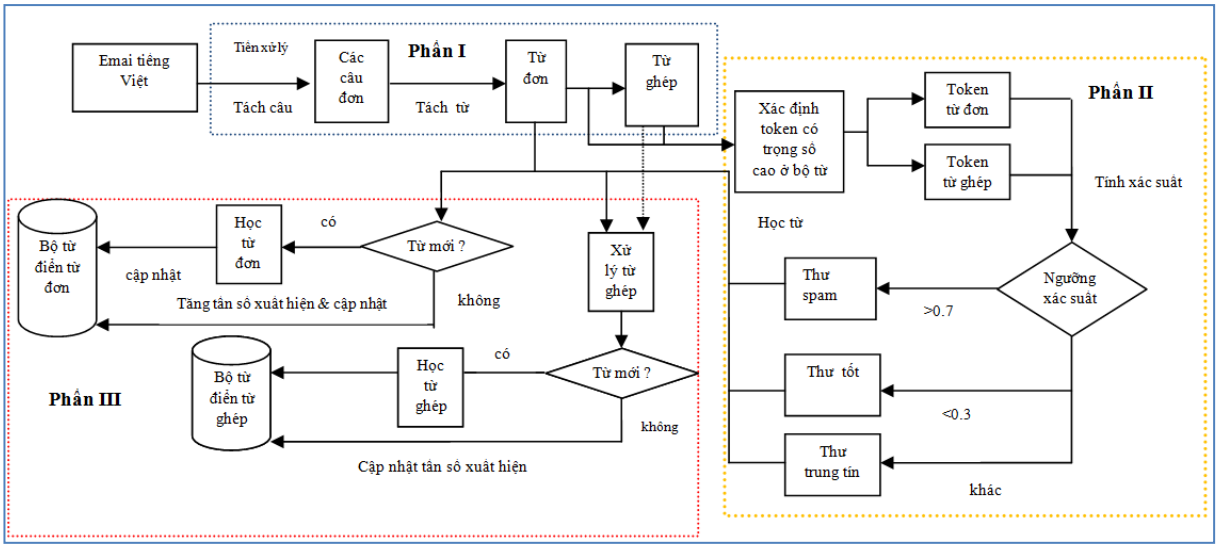
Trong tiếng Việt, tùy theo lĩnh vực, chủ đề khác nhau nên có nhiều từ, tiếng khác nhau về mặt phát âm cũng như ý nghĩa. Trong bài bào báo này, chỉ tập trung vào lĩnh vực thư rác tiếng Việt nên có sự giới hạn về số lượng về từ và tiếng sử dụng. Bài báo không tập trung vào mặt ý nghĩa cũng như những đặc trưng phức tạp của tiếng Việt như từ đồng nghĩa, từ láy, sự nhập nhằng ngữ nghĩa … mà chỉ xác định tần số của từ đơn, từ ghép tiếng Việt xuất hiện trong thư rác nên hướng tiếp cận khác biệt với các phương pháp xác định ngữ nghĩa từ tiếng Việt.

Hiện tại, chưa có một thống kê chính xác nào để xác định những đặc điểm chung của thư rác tiếng Việt. Theo khảo sát tổng quát, đa phần thư rác tiếng Việt tập trung vào quảng cáo, rao vặt mua bán và mời tham gia các diễn đàn, mạng xã hội. Phần dưới sẽ trình bày những mục tiêu chính của phương pháp tiếp cận này.

2.1. Mục tiêu chính

Xét một văn bản u gồm n tiếng $t=s_1s_{2..}s_n$. Mục tiêu chính là phân tích văn bản u thành m câu đơn $t=z_1z_2... z_m$ với $z_k= s_i... s_j$ $(1\leq k\leq m, 1\leq i, j\leq n)$ có thể chứa từ đơn hay từ phức. Ứng với mỗi câu, phân tích thành từng từ đơn thể. Đây là bước đầu tiên để xây dựng một danh sách các từ ghép được sử dụng nhiều trong thư rác tiếng Việt, theo ưu tiên xét tần số xuất hiện của từ. Từ đó, sử dụng thuật toán Naïve Bayes dựa trên tập hợp các từ vừa tìm được để tiến hành phân loại thư.

Quy trình lọc thư rác tiếng Việt cho cả quá trình huấn luyện và nhận dạng có thể được cụ thể hóa bằng mô hình như hình dưới. Trong mô hình thể hiện rõ tiến trình từ khi nhận thư, xử lý và phân loại bức thư nhận được, đồng thời cũng cập nhật lại tập huấn luyện cho việc học từ



Hình 1 –Mô hình tổng quát lọc thư spam tiếng Việt

Mô hình gồm 3 tiến trình nhỏ. Tiến trình 1 làm nhiệm vụ tiền xử lý và phân tích từ đơn, từ ghép có trong mỗi thư tiếng Việt truyền vào, trong tiến trình 2 áp dụng thuật toán Naïve Bayes dựa trên danh sách các từ đơn lẫn từ ghép đã phân tích trong tiến trình 1 để xác định tần số xuất hiện của các từ, qua đó phân lớp bức thư thuộc lớp thư bình thường, thư rác hay thư trung tín.

Trong tiến trình cuối cùng, các từ ghép, từ đơn mới sẽ tự động được học và cập nhật vào trong tập huấn luyện cơ sở, còn các từ đã tồn tại sẽ thay đổi tần số xuất hiện trong thư rác, thư bình thường và thư trung tín. Quy trình học từ này diễn ra một cách tự động. Số lượng từ học được phải qua quy trình kiểm tra để xác định là từ có trọng số đáng tin cậy hay không. Phần tiếp theo sẽ mô tả rõ quy trình 1 trong mô hình đề xuất.

2.2. Tiền xử lý và tách câu tiếng Việt

Trong mô hình trên, tiến trình 1 gồm 2 giai đoạn tiền xử lý và tách thành từng câu đơn của hệ thống. Tiến trình này có thể khái quát như sau:

Đưa vào tập T_s gồm những tài liệu huấn luyện, trong đó mỗi tài liệu $T_i \in T_s$ $(1 \leq i \leq s)$ thuộc về một trong ba lớp: thư rác, thư bình thường hay thư trung tín. Tài liệu huấn luyện này được chọn trong giai đoạn khởi tạo và được cập nhật trong giai đoạn phân lớp thành công một bức thư đầu vào (tiến trình thứ 3, học từ đơn và từ ghép trong mô hình).

Với mỗi tài liệu $T \in T_s$, một vector hỗ trợ V_t của quan hệ tần suất từ sẽ được xây dựng dựa vào các bước sau đây:

+ Xử lý loại bỏ các định dạng của ngôn ngữ HTML có trong bức thư.

+ Xử lý loại bỏ những từ phổ biến như "*thì*", "*là*", "*mà*", "*các*", "*những*",… và các từ dùng để nối câu như "*tuy nhiên*", "*mặc dù*", "*vì thế*", "*không những*", "*mà còn*",… những ký tự đặc biệt như "@", "#", "$", "?", "&",… để làm tăng tốc độ xử lý của việc tách từ do những từ loại này xuất hiện nhiều trong các tài liệu huấn luyện, đồng thời sự xuất hiện của các từ này không làm ảnh hưởng đến quá trình phân loại thư.

+ Chuyển toàn bộ văn bản thành các câu đơn chuẩn, mỗi từ trong câu đơn chuẩn cách nhau bởi một khoảng trắng duy nhất. Để tăng tốc độ xử lý có thể thay thế các dấu câu như dấu hỏi (*?*), dấu chấm than (*!*), dấy nháy… thành dấu chấm câu (.). Do không xét đến nội dung từ mà chỉ xét số lượng từ tìm được và xác định tần số xuất hiện của chúng có trong nội dung thư nên phần thay đổi này không làm mất đi tính chất của bức thư cần lọc. Sau giai đoạn tiền xử lý và tách nội dung thư, ta sẽ tiến hành phân tích từ đơn, từ ghép trong nội dung thư.

2.3. Phân tích từ đơn

Sau quá trình trên, mỗi tài liệu T_i thuộc tập tài liệu T_s được chuẩn hóa thành tập S_n câu đơn chuẩn, ứng với mỗi câu đơn S_j $(1 \leq j \leq n)$ sẽ chứa k từ đơn, mỗi từ đơn W_m $(1 \leq m \leq k)$ và W_{m+1} $(1 \leq m \leq k)$ được phân cách nhau bởi một ký tự khoảng trắng. Dựa vào đặc tính này, dễ dàng xây dựng được cơ sở dữ liệu các từ đơn chuẩn và tần số xuất hiện của chúng trong từng bức thư của tập huấn luyện. Do tiếp cận theo hướng không đề cập đến ý nghĩa của từ đơn, nên để tăng độ tin cậy của từ đơn trong thư , chúng tôi xét tần số xuất hiện của từ đơn theo hai cơ chế:

+ Học từ vựng bình thường: tần số xuất hiện của từ đơn trên toàn bộ tập huấn luyện được tính bằng số lần xuất hiện của chính từ đó, có phân biệt trong một thư xuất hiện bao nhiêu lần.

+ Học từ vựng cho quá trình lọc spam: tần số xuất hiện của từ đơn được tính trên từng bức thư, mỗi lần xuất hiện trong thư được tính là xuất hiện 1 lần, nếu trong thư, từ đó xuất hiện nhiều lần thì cũng tính là 1 lần.

Cụ thể hóa, trong câu đơn "*Học sinh học sinh học*" sẽ được tách làm 2 từ đơn : "*học*", "*sinh*" với tần số xuất hiện tính theo hai cơ chế trên lần lượt là "*học*" (3 lần), "*sinh*" (2 lần) và "*học*" (1 lần), "*sinh*" (1 lần).

Quá trình học từ đơn này lần lượt diễn ra trên hai tập huấn luyện thư rác và thư bình thường. Kết thúc quá trình phân tích từ đơn, sẽ hình thành được một tập hợp gồm nhiều từ đơn, mỗi từ đơn sẽ có 01 mã định danh (*id*) nhất định trong cơ sở dữ liệu. Ứng với mỗi

định danh *id* trên mỗi tập huấn luyện sẽ có 2 tần số xuất hiện: tần số tổng trên tập huấn luyện và tần số trên từng bức thư thuộc tập huấn luyện như đã trình bày như cách tính trên.

2.4. Phân tích từ ghép

Trong tiếng Việt, bên cạnh từ đơn còn có từ gồm 2 tiếng trở lên. Hiện tại, do chưa có từ điển chuẩn nào cho việc xử lý ngôn ngữ tiếng Việt, nên chúng tôi quyết định dựa vào bảng thống kê của bộ từ điển sử dụng bên dưới (http://dict.vietfun.com) để bắt đầu quá trình phân tích từ ghép từ tập hợp các từ đơn đã tìm được trong phần cuối giai đoạn 1. Do tính chất phức tạp của từ ghép về độ dài có thể gồm 2 tiếng, 3 tiếng, 4 tiếng… nên để thuận tiện cho quá trình nghiên cứu, đã thống kê dựa trên website http://dict.vietfun.com, số lượng từ ghép dựa vào số tiếng như bảng 1

Độ dài từ	Thông số	
	Tần số	Tỉ lệ %
1	8933	12.2
2	48995	67.1
3	5727	7.9
4	7040	9.7
>=5	2001	3.1
Tổng cộng	72994	100

Bảng 1 - Thống kê độ dài của từ trong từ điển (http://dict.vietfun.com)

Dựa vào bảng trên, hơn 67.1% từ trong từ điển có độ dài là 2 tiếng, khoảng 20% là từ đơn và từ có độ dài gồm 3-4 tiếng. Các từ dài hơn chỉ chiếm khoảng 3% trong tự điển. Qua đó, thấy rõ so với từ đơn và các từ ghép có độ dài lớn hơn thì từ ghép 2 tiếng chiếm số lượng khá lớn. Vì vậy, để đơn giản vấn đề, ban đầu tập trung vào việc phân tích từ ghép có 2 tiếng nhưng không xét về mặt nghĩa của từ. Quy trình phân tích từ ghép có thể khái quát hóa như sau:

+ Xét trong 1 câu tiếng Việt *S (Sentence)* sẽ gồm $W_1, W_2, W_3, \ldots W_n$ từ, mỗi từ W_i $(1 \leq i \leq n)$ là một từ đơn tiếng Việt. Do việc phân tích chỉ tập trung từ ghép có 2 tiếng nên mỗi từ ghép *CW (Compound Word)* được tạo bởi hai từ đơn đứng gần nhau W_i , W_{i+1} $(1 \leq i \leq n)$ và được cách nhau bởi 1 khoảng trắng.

+ Do không xét mặt ngữ nghĩa của từ nên trong quá trình tạo từ ghép theo cách trên sẽ dẫn đến các từ vô nghĩa. Cụ thể, xét trong 1 câu đơn "*Khuyến mãi cao*" sẽ tách được các

bộ từ : "*khuyến mãi*" và "*mãi cao*", như vậy từ ghép "*khuyến mãi*" có giá trị, còn từ "*mãi cao*" không có giá trị trong quá trình lọc thư rác.

Để giải quyết vấn đề này, qua kết quả quá trình thực nghiệm tách từ, đã sử dụng ngưỡng α dùng để đánh giá độ chính xác của từ ghép tìm được. Ngưỡng α được định nghĩa bởi người sử dụng. Mỗi từ ghép đều có riêng một ngưỡng α. Khi ngưỡng α thay đổi giá trị thì độ chính xác của từ ghép cũng bị thay đổi theo.

Để giảm thời gian lọc thư spam, chúng tôi đã xây dựng bộ từ điển các từ ghép theo cách trên. Giả sử có tập thư spam *SD (Spam Document),* mỗi thư $D_i \in SD$ sẽ có tập các câu đơn S_n. Trong mỗi câu đơn $S_i \in S_n$ $(1 \leq i \leq n)$ sẽ gồm các từ đơn $W_1, W_2, W_3, \ldots W_n$. Vận dụng cơ chế tách từ ghép nêu trên thỏa mỗi từ ghép *CW* chứa 1 bộ gồm 2 từ đơn $\{W_j, W_{j+1}\}$ $(1 \leq j \leq m)$, trong đó W_j và W_{j+1} là hai từ đơn liên tiếp đứng gần nhau và cách nhau bởi dấu khoảng cách. Ứng với mỗi từ ghép *CW* tìm được sẽ được đưa vào tập từ ghép nếu từ ghép chưa tồn tại trong tập từ ghép và tăng tần số xuất hiện nếu từ ghép tìm được đã tồn tại trong tập từ ghép.

Kết quả của quá trình tiền xử lý nêu trên, sẽ có được 1 tập từ ghép chứa cả từ có giá trị sử dụng và những từ 2 tiếng không có ý nghĩa. Mỗi từ trong tập từ này sẽ có 1 tần số *k* biểu diễn tần số xuất hiện của từ trong tập huấn luyện. Tần số *k* thể hiện tổng số lần xuất hiện của từ trên toàn bộ tập huấn luyện, mỗi lần từ xuất hiện thì tăng trọng số *k* lên 1 đơn vị.

Tính giá trị của ngưỡng α của mỗi từ *CW* trong bộ từ ghép

$$\alpha = \frac{k}{TotalMessage} \tag{1}$$

Trong đó *k* là tần số xuất hiện của từ ghép *CW* trong tập huấn luyện.

Dựa vào kết quả thử nghiệm tách từ, ngưỡng α lớn hơn 0.2 thì độ chính xác của từ có thể chấp nhận được. Những từ có ngưỡng α nằm ngoài khoảng cận trên được xếp vào tập các từ cần được huấn luyện tiếp tục.

2.5. Quy trình cập nhật từ vựng tiếng Việt

Trong mô hình lọc thư rác đã trình bày ở trên (Hình 1), sau khi đã phân lớp thư thuộc thư rác hay thư bình thường, quy trình học từ tự động được tiến hành. Đối với những từ đơn hay từ ghép mới chưa có trong bộ tự điển sẽ được cập nhật vào. Ngược lại, đối với những từ đã có, hệ thống sẽ cập nhật tần số xuất hiện của từ đó, đồng thời thay đổi tỷ lệ spam, ham của các từ đó.

Với quá trình tự học này, ứng với số lượng thư tiếng Việt càng lớn thì số lượng từ trong bộ tự điển càng cao, đồng thời sẽ tăng độ chính xác cho việc tính xác suất thư rác hay thư bình thường, hỗ trợ rất nhiều khi áp dụng công thức Naïve Bayes.

Phần trên, chúng tôi đã đề xuất phương pháp tiếp cận việc tách từ trong tiếng Việt. Phần tiếp theo, chúng tôi sẽ đưa ra quy trình lọc thư rác tiếng Việt dựa vào thuật toán Naïve Bayes.

3. Quy trình lọc thư rác tiếng Việt

3.1.Áp dụng thuật toán Naïve Bayes

Dựa trên công thức Naïve Bayes, áp dụng nguyên tắc tính xác suất cho các *id* từ đơn ở phần (2.3) hay từ ghép (2.4) bằng thuật toán Naïve Bayes như sau:

Giả sử nội dung của mỗi bức thư điện tử là: *content*

Lớp thư rác ký hiệu là: *spam*

Lớp thư hợp lệ ký hiệu là: *ham*

Xác suất để một thư điện tử là thư rác: *P(spam | content)*

$Word_1$, $Word_2$, $Word_3$, ... *Wordm* là các từ đặc trưng xuất hiện trong *content*.

$$P(spam \mid content) = \frac{P(content \mid spam) * P(spam)}{Total} \quad (2)$$

Trong đó *Total* được xác định bằng

$$Total = P(content \mid spam) * P(spam) + P(content \mid ham) * P(ham) \quad (3)$$

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

$$P(content \mid ham) = \prod P(word_i \mid ham) \quad (4)$$

$$P(content \mid spam) = \prod P(word_i \mid spam) \quad (5)$$

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

$$P(spam) = \frac{TotalSpam}{TotalMessage} \quad (6)$$

$$P(ham) = \frac{TotalHam}{TotalMessage} \quad (7)$$

Trong quá trình phân lớp thư, ngoài lớp thư rác và thư hợp lệ, nếu xác suất spam là *>0.7* sẽ được phân vào lớp thư spam, nếu xác suất spam là *<0.3* thì được phân vào thư

bình thường, còn trong trường hợp ngược lại thì sẽ được đưa vào phân lớp thứ ba: lớp thư trung tín. Những thư thuộc lớp này sẽ chờ người duyệt thư quyết định phân loại là thư hợp lệ hay thư rác. Xác suất xác định thư rác có thể thay đổi để làm tăng độ tin cậy cho quá trình lọc thư spam, những tỉ lệ nêu trên được xác định trong quá trình thử nghiệm.

Trong mô hình đã đề cập ở trên, trong phần thứ 2, sau khi có danh sách từ đơn và từ ghép, áp dụng thuật toán Naïve Bayes dựa trên danh sách các từ để tìm các token có giá trị tốt nhất trong danh sách. Thử nghiệm của đề tài dựa trên các dạng token các nhau: token toàn từ đơn, token toàn từ ghép và token vừa từ đơn và từ ghép. Dưới đây là ví dụ áp dụng công thức tính tỉ lệ spam và tỉ lệ ham theo công thức Bayes

Từ đơn	Tần số xuất hiện		
	Ham	Spam	Total
All messages	400	600	1000
With *"bán"*	300	100	400
With *" mua"*	10	90	100

Bảng 2 - Ví dụ minh họa phân tích từ đơn

Áp dụng công thức tính

$$P(spam \mid token) = \frac{P(spam) * P(token \mid spam)}{P(token)} \quad (8)$$

Thu được các giá trị sau đây

P(spam|"bán") = P(600/1000) * P(300/600) / P(400/1000) = 0.6*0.5/0.4=0.75=75%

P(ham|"bán") = P(400/1000) * P(100/400)/P(400/1000) = 0.4*0.25/0.4=0.25=25%

P(spam|"mua") = P(600/1000) * P(90/600) / P(100/1000) = 0.6*0.15/0.1=0.9=90%

P(ham|"mua") = P(400/1000) *P(10/400) /P(100/1000) = 0.4*0.025/0.1=0.1=10%

3.2. Kết quả thực nghiệm

Để việc lọc thư rác tiếng Việt đạt hiệu quả cao, việc tách từ chiếm một trí trí rất quan trọng. Tuy nhiên, việc đánh giá độ chính xác của việc tách từ rất phức tạp, đặc biệt đối với từ ghép. Do đó bài báo này thực hiện các thử nghiệm sau đây:

Tách câu, tách từ (cả từ đơn lẫn từ ghép) dựa trên một tập huấn luyện gồm nhiều thông tin thuộc nhiều lĩnh vực khác nhau trên mạng Internet.

Phân loại thư spam áp dụng thuật toán Naïve Bayes dựa trên tập hợp từ đơn, từ ghép và dựa trên từ đơn lẫn từ ghép. Ngoài ra, bộ lọc thư spam còn có chức năng theo vết người

sử dụng, nghĩa là nếu người dùng đăng nhập sau một số lần nào đó thì những email không đọc sẽ được gán là thư spam và tự động chuyển sang hộp Spam. Nói một cách khác, nếu 1 email nằm trong hộp Inbox sau bao nhiêu lần check mail mà người dùng không mở ra xem thì mặc định email đó sẽ chuyển sang hộp Spam mà không cần hỏi người sử dụng, giảm thời gian check mail của người dùng.

Thử nghiệm lọc thư rác tiếng Việt bằng Naïve Bayes, sử dụng tập huấn luyện là bộ từ đơn và từ ghép đã nêu trên: dữ liệu thử nghiệm là 01 tập hợp gồm nhiều email tiếng Việt $D=\{d_1, d_2,..., d_n\}$ trong đó mỗi email sẽ thuộc vào một trong ba loại: thư rác, thư bình thường và thư trung tín. Với mỗi tài liệu d_i $(1 \leq i \leq n)$, sau qua các phương pháp xử lý nêu trên, kết quả cuối cùng di được biểu diễn $d_i = g_1g_2...g_m$ với g_k $(1 \leq k \leq m)$ là từ đơn hay từ ghép đã xử lý.

Chúng tôi xây dựng tập dữ liệu huấn luyện để thực hiện các thí nghiệm trên. Đối với thử nghiệm đầu tiên, đã thu thập gần 800 dữ liệu để triển khai và cho kết quả như sau

Loại từ	Thông số	
	Số lượng	Tỉ lệ từ đúng
Từ đơn	4506	85%
Từ ghép	11980	80%

Bảng 3 - Kết quả tách từ trên 800 dữ liệu mẫu

Đối với thử nghiệm 2 và 3, chúng tôi xây dựng xây dựng tập huấn luyện để thực hiện. Do tập huấn luyện phải là thư tiếng Việt nên chúng tôi phải sử dụng thống kê trê Internet, một mặt tìm email tiếng Việt, mặt khác xin sự giúp đỡ của các diễn đàn để thu thập email tiếng Việt. Để tiến trình huấn luyện được thuận lợi, chúng tôi chia dữ liệu thu thập được thành 2 loại: thư rác và thư bình thường. Tổng dữ liệu thử nghiệm gồm 384 thư rác và 500 thư bình thường để bắt đầu tiến trình huấn luyện. Với tập huấn luyện như trên, chúng tôi đã tách được 1042 từ đơn và 5914 từ ghép

Lĩnh vực nghiên cứu tiếng Việt phong phú như kinh tế, khoa học, xã hội, sức khỏe, thể thao… nên việc nghiên cứu ngữ nghĩa các từ, các câu sẽ rất phức tạp và để xử lý chính xác cũng mất nhiều thời gian. Ngoài ra, theo thống kê trong bảng 1 cho thấy từ ghép tiếng Việt chủ yếu là loại từ có độ dài 2 tiếng, do vậy việc tách từ chúng tôi cũng chỉ thực hiện cho từ ghép có độ dài tối đa 2 tiếng. Trong giới hạn đó, kết quả thực nghiệm phân loại 100 thư tiếng Việt bằng cách dựa vào tập huấn luyện từ đơn và từ ghép được thể hiện bằng bảng thống kê bên dưới.

Thử nghiệm trên	Kết quả phân loại		Độ chính xác	
	Spam	Ham	Spam	Ham
Từ đơn	79/100	90/100	79%	90%
Từ ghép	94/100	92/100	94%	92%
Vừa từ đơn vừa từ ghép	85/100	80/100	85%	80%

Bảng 4 - Kết quả phân loại thư rác

Như vậy, dựa vào bảng kết quả trên, chứng tỏ việc lọc thư rác tiếng Việt theo từ đơn có xác suất thấp hơn so với từ ghép (79 % so với 94%). Như vậy nếu dùng theo phương pháp Naïve Bayes cho tiếng Anh thì đối với tiếng Việt không hiệu quả .

Qua kết quả thử nghiệm cho thấy khả năng lọc thư tiếng Việt theo cơ chế tách từ đơn và từ ghép sẽ cho kết quả chính xác hơn với thời gian thực hiện chấp nhận được.

4. Kết luận

Bài báo này đã đề xuất việc sử dụng phương pháp tách từ đơn , từ ghép dựa trên bộ huấn luyện thư, đồng thời áp dụng thuật toán Naïve Bayes để tiến hành lọc thư spam tiếng Việt. Điểm mới của bài báo này là đề xuất phương pháp lọc thư rác tiếng Việt sử dụng thuật toán Bayes không phải chỉ dựa trên các từ đơn như đối với tiếng Anh mà còn dựa trên cả từ ghép theo tiếng Việt.

Kết quả thực nghiệm cho thấy hướng tiếp cận của bài báo đạt được độ chính xác cao hơn khi phân loại thư rác tiếng Việt so với phương pháp Bayesian cổ điển chỉ dùng cho các từ đơn tiếng Việt.

Thư spam tiếng Việt đang trong giai đoạn phát triển, do vậy, vấn đề khó khăn lớn là thu thập tập huấn luyện thư rác và thư bình thường bằng tiếng Việt. Với tập huấn luyện càng lớn thì độ chính xác của việc học từ đơn và từ ghép càng được nâng cao, góp phần rất lớn trong việc tính xác suất theo công thức Naïve Bayes. Do đó chúng tôi sẽ tiếp tục thu thập để có được bộ huấn luyện lớn hơn nhằm nâng cao độ chính xác của phương pháp này.

Ở Việt Nam hiện nay, thư spam rất phức tạp, đôi khi người dùng nhận được thư spam bằng tiếng Anh, đôi khi bằng tiếng Việt, đôi khi chứa cả tiếng Anh lẫn tiếng Việt. Vì vậy, hướng nghiên cứu tiếp theo của chúng tôi là đưa ra phương pháp lọc thư rác thích hợp cho cả tiếng 2 thứ tiếng.

Tài liệu tham khảo (References)

[1] Dinh Dien, Author, "*Tu Tieng Viet*", Proceeding of ICMLC 2002 Conference, Beijing, November 2002, pp. 111-116.

[2] Dinh Dien, Hoang Kiem, Nguyen Van Toan, Author, "*Vietnamese Word Segmentation*", The sixth Natural Language Processing Pacific Rim Sympossium, Tokyo, Japan 2001, pp 749-758.

[3] Foo S, Li H, Author, "*Chinese Word Segmentation and Its Effect on Information Retrieval*", Information Processing & Management : Anh International Journal 40(1), 2004, pp 161-190.

[4] Le An Ha, Author "*A method for word segmentation in Vietnamese*", Proceedings of Corpus Linguistics 2003, Lancaster, UK, 2003.

[5] H Nguyen, T.Vu, N.Tran, K.Hoang, Author "*Internet and Genertics Algorithm-base text Categorization for Document in Vietnamese*", Research, Innovation and Vision of the Future, the 3rd International Conference in Computer Science, (RIVF 2005), Can Tho, Viet Nam 2005.