ĐÁNH GIÁ MỘT SỐ CÁCH THỨC TÍNH XÁC SUẤT SPAM CỦA TOKEN ỨNG DỤNG TRONG PHÂN LOẠI THƯ RÁC

Nguyễn Tu Trung, Nguyễn Ngọc Hưng, Phạm Thanh Giang

Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

Tóm tắt: Phân loại thư rác là bài toán được quan tâm nghiên cứu từ rất lâu trên thế giới với nhiều hướng tiếp cận khác nhau. Tính năng phân loại thư rác được tích hợp vào module phân loại thư rác của Mail Server hay Mail Client. Hiện nay, khi mà các phương pháp truyền thống vẫn có những điểm yếu nhất định thì phương pháp phân loại dựa trên nội dung tỏ ra hiệu quả với việc sử dụng các kĩ thuật trong học máy thống kê. Trong đó, phân loại thư rác dựa trên Bayes với ưu điểm đơn giản, dễ sử dụng sử và tốc độ nhanh nên được cài đặt phổ biến trong các hệ thống Mail Server hay Mail Client. Bài báo này trình bày đánh giá về một số cách thức tính xác suất là Spam của các Token thông qua ứng dụng phân loại thư rác.

Từ khóa: Thư rác, phân loại thư rác, Bayes, học máy thống kê, Token, Spam, Ham.

I. MỞ ĐẦU

Một trong những dịch vụ mà Internet mang lại đó là dịch vụ thư điện tử, đó là phương tiện giao tiếp rất đơn giản, tiện lợi, rẻ và hiệu quả giữa mọi người trong cộng đồng sử dụng dịch vụ Internet. Tuy nhiên chính vì những lợi ích của dịch vụ thư điện tử mang lại mà số lượng thư trao đổi trên Internet ngày càng tăng, và đa số trong số hững thư đó là thư rác (spam). Thư rác (spam mail) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận. Thư rác thường được gửi với số lượng rất lớn, không được người dùng mong đợi, thường với mục đích quảng cáo, đính kèm virus, gây phiền toái hó chịu cho người dùng, làm giảm tốc độ truyền internet và tốc độ xử lý của email server, gây thiệt hại rất lớn về kinh tế.

Theo thống kê của kaspersky năm 2014 [12], Tỷ lệ thư rác trong lưu lượng truy cập email trong tháng Hai tăng 4.2% so với tháng trước, đạt trung bình 69.9%. Tuy nhiên, tỷ lệ này thấp hơn 1.2% so với tháng Hai năm 2013. Ba nguồn phát tán thư rác hàng đầu gồm có Trung Quốc (23%), Mỹ (19.1%) và Hàn Quốc (12.8%). Việt Nam đứng vị trí thứ 7 với 2.95%, giảm so với tháng Một chiếm 3.1%. Những kẻ lừa đảo thường nhắm mục tiêu đến các trang mạng xã hội (27.3%), dịch vụ thư điện tử (19.34%) và các tổ chức thanh toán trực tuyến (16.73%). Theo [13], về tình hình thư rác quý III 2015, tỷ lệ thư rác trong lưu lượng email đã giảm so với Quý II, nhưng các kỹ thuật lừa đảo người dùng và vượt qua bộ lọc email ngày càng trở nên tinh vi hơn. Trong Quý III 2015, tỷ lệ thư rác chiếm 54.2% toàn lưu lượng email, giảm 0,8% so với quý II. Đã có những thay đổi lớn trong top 3 các quốc gia là mục tiêu tấn công bằng email trong Quý III 2015. Đứng vị trí đầu là Đức chiếm 18.47 (giảm 1.12% so với quý II). Đứng vị trí thứ 2 là Brazil và thứ 3 là Nga với tỷ lệ 7.56% (tăng 2.82% so với quý II). Về nguồn gốc phát tán thư rác, Mỹ vẫn là quốc gia có nguồn thư rác lớn nhất chiếm 15.34%. Việt Nam đứng thứ hai với 8.42% (tăng 5.04% so với quý II). Xếp vị trí thứ 3 là Trung Quốc chiếm tỷ lệ 7.15%, không thay đổi so với quý II. Tiếp sau đó là các nước Nga (5.79%), Đức (4.39%), Pháp (3.32%).

Có nhiều phương pháp lọc thư rác khác nhau. Mỗi phương pháp đều có những ưu nhược điểm riêng. Trong đó, phương pháp lọc nội dung để phân loại thư rác đã và đang được quan tâm, nghiên cứu và ứng dụng nhiều nhất. Phương pháp này dựa vào nội dung và chủ đề bức thư để phân biệt thư rác và thư hợp lệ. Phương pháp này có ưu điểm đó là chúng ta có thể dễ dàng thay đổi bộ lọc để nó có thể lọc các loại thư rác cho phù hợp.

Trong phương pháp học dựa trên nội dung, lọc thư rác sử dụng các kĩ thuật học máy thống kê là một phương pháp có triển vọng với nhiều ứng dụng thương mại như Hotmail, Google, Yahoo. Các phương pháp học máy và xác suất thống kê cho phép phân loại cả những thư rác chưa từng xuất hiện trước đó. Trong [1], Awad đã trình bày một đánh giá, so sánh một số phương pháp học máy (Bayesian classification, k-NN, ANNs, SVMs...) cho vấn đề lọc thư rác. Trong [6], Shahar Yifrah và Guy Lev trình

Tác giả liên hệ: Nguyễn Tu Trung
Email: trungnt.sremis@gmail.com
Đến toàn soạn: 12/2017 , chỉnh sửa: 4/2018 , chấp nhận đăng: 8/2018

bày báo các về dự án xây dựng bộ lọc thư rác sử dụng các kỹ thuật học máy. Trong [10], các tác giả đã so sánh hiệu quả của các bộ lọc thư rác khác nhau sử dụng Naïve Bayes, SVM, và KNN. Các kết quả thử nghiệm cho thấy các bộ lọc sử dụng các kĩ thuật này đều cho độ chính xác rất cao.

Đặc thù của các kĩ thuật dựa trên nội dung là phải phân tích từ trong nội dụng và tính giá trị token hay đặc trưng. Một khi số lượng các token, các đặc trưng lớn thì các phương pháp như SVMs, ANNs có tốc độ huấn luyện rất chậm. Trong các kĩ thuật lọc thư rác dựa trên học máy thống kê, kĩ thuật Bayes tỏ ra đơn giản, hiệu quả, tốc độ thực thi rất nhanh, không những trong giai đoạn phân loại mà cả khi huấn luyện. Thuật toán Bayes đã được áp dụng vào chương trình lọc thư rác spambayes, và cho kết quả lọc khá hiệu quả. Có lẽ, đây là lý do mà bộ lọc sử dụng kĩ thuật này được cài đặt phổ biến trong các hệ thống Mail Server (Zimbra), Mail Client. Các phần mềm Mail Client như Outlook, Outlook Express, Thunderbird/Mozilla Mail & Newsgroups, Eudora, hay Opera Mail. Các thuật toán Naïve Bayes là những thuật toán kinh điển trong kĩ thuật Bayes. Naïve Bayes rất phổ biến trong các bộ lọc thư điện tử chống Spam nguồn mở [9]. Có nhiều phiên bản của Naïve Bayes. Trong [9], các tác giả đã thảo luận, thử nghiệm và đánh giá về hiệu quả lọc Spam của các phiên bản này. Trong [5], Phan Hữu Tiếp cùng các cộng sự trình bày quy trình lọc thưc rác tiếng Việt dựa trên thuật toán Naïve Bayes và việc xử lý tách câu tiếng Việt. Trong [7], Tianda và cộng sự đã trình bày một so sánh giữa bộ phân loại thư rác chỉ sử dụng kĩ thuật Naïve Bayes và bộ phân loại thư rác sử dụng bộ phân loại thư rác kĩ thuật và luật kết hợp. Trong [4], các tác giả thảo luận về quy trình lọc thư rác thống kê sử dụng kĩ thuật phân loại Naïve Bayes. Một cách thuận tiện, đơn giản để cài đặt thuật toán Bayes trong việc lọc thư rác là thuật toán của Paul Graham [8][4] và biến thể khác của Tim Peter. Các thuật toán này đều phân tích, đánh giá và đưa ra đề xuất về các cách tính xác suất là spam của các token. Trong đó, cải tiến của Paul Graham cho độ chính xác rất cao. Trong [2], Jialin và cộng sự đã thảo luận, đánh giá về phương pháp lọc SMS rác sử dụng SVM và MTM (message topic model).

Trong bài báo này, chúng tôi tập trung nghiên cứu về việc sử dụng kĩ thuật Bayes ứng dụng trong vấn đề lọc thư rác thông qua việc đánh giá một số cách thức tính xác suất là Spam của các token từ việc phân tích công thức tính xác suất Spam của Paul Graham. Nhiều nghiên cứu gần đây đánh giá hiệu quả của các phương pháp học máy trong việc phân loại thư rác thông thường chỉ so sánh giữa các kĩ thuật mới với thuật toán Naïve Bayes, mà không trực tiếp so sánh với cải tiến hiệu quả của Paul Graham. Đây cũng chính là một lý do mà nhóm chúng tôi viết bài báo này. Các phần tiếp theo được trình bày như sau. Phần 2 trình bày về vấn đề lọc thư rác dựa trên Bayes. Phần 3 trình bày một số cách thức tính xác suất là Spam khác nhau của các token. Các thử nghiệm được trình bày trong phần 4. Kết luận được trình bày trong phần 5.

II. PHÂN LOẠI THƯ RÁC DỰA TRÊN BAYES

A. *Lọc thư rác dựa trên Bayes*

Kĩ thuật phân loại thư rác dựa trên Bayes được trình bày trong [3][5].

Coi mỗi email được biểu diễn bởi một vectơ thuộc tính đặc trưng $\vec{x} = (x_1, x_2, \dots, x_n)$ với $(x_1, x_2, \dots, x_n)$ là các giá trị thuộc tính $X_1, X_2, \dots, X_n$ tương ứng trong không gian đặc trưng (space model). Ta sử dụng giá trị nhị phân 0 và 1 để mô tả email đó có đặc điểm X_i hay không, giả xử nếu email đó có đặc điểm X_i thì ta đặt thuộc tính $X_i = 1$, còn nếu email đó không có đặc điểm X_i thì ta có thuộc tính $X_i = 0$.

Từ thuyết xác suất của Bayes và xác suất đầy đủ chúng ta có công thức tính xác suất mail với vectơ $\vec{x} = (x_1, x_2, \dots, x_n)$ thuộc vào lớp c như sau:

$$P(C = c|\vec{X} = \vec{x}) = \frac{P(C=c)P(\vec{X}=\vec{x}|C=c)}{\sum_{k\in\{Spam,Ham\}} P(C=k)P(\vec{X}=\vec{x}|C=c)} \quad (1)$$

Để đơn giản khi tính $P(\vec{X}|C)$ ta phải giả sử $X_1, X_2, \dots, X_n$ là độc lập. Khi đó biểu thức (1) tương đương với biểu thức sau:

$$P(C = c|\ \vec{X} = \vec{x}) = \frac{P(C=c)\prod_{i=1}^{n} P(X_i=x_i|C=c)}{\sum_{k\in\{Spam,Ham\}} P(C=k)\prod_{i=1}^{n} P(X_i=x_i|C=k)} \quad (2)$$

Giá trị được sử dụng rất rộng rãi để đánh hạng cho thuộc tính là giá trị tương hỗ *MI* (mutual information), ta lấy những thuộc tính có giá trị *MI* lớn nhất. Ta có thể tính giá trị tương hỗ *MI* mà mỗi đại diện của *X* thuộc về loại *C* như sau:

$$MI = \sum_{x\in\{0,1\},c\in\{Spam,Ham\}} P(X = x|C = c) log \frac{P(X=x|C=c)}{P(X=x)P(C=c)} \quad (3)$$

Một email được coi là spam nếu:

$$\frac{P(C=Spam|\ \vec{X}=\vec{x})}{P(C=Ham|\ \vec{X}=\vec{x})} > \lambda \quad (4)$$

Với λ là ngưỡng cho trước để xem xét so sánh với tỉ lệ giữa xác suất là *Spam* hay *Ham* của một thư. Trong đó, *Spam*: thư rác, *Ham*: thư hợp lệ.

Giả sử các thuộc tính X_i là độc lập khi đó ta có:

$$P(C = Spam|\ \vec{X} = \vec{x}) = 1 - P(C = Ham|\ \vec{X} = \vec{x}) \quad (5)$$

Khi đó (4) tương đương với:

$$P(C = Spam|\ \vec{X} = \vec{x}) > t \text{ với } t = \frac{\lambda}{1+\lambda} \quad (6)$$

B. *Công thức của Paul Graham*

Theo [8][4], Paul Graham đề xuất một cách tính xác suất làm Spam của các token. Công thức của Paul Graham không rất đơn giản, thuận tiện cho việc cài đặt mà còn cho độ chính xác phân loại thư rác rất cao.

Công thức tính xác suất Spam của token *w* như sau:

$$P(S|w) = \frac{\frac{SA(w)}{STM}}{\frac{SA(w)}{STM} + 2\frac{HA(w)}{HTM}} \quad (7)$$

Trong đó,

SA(w): số lần xuất hiện của token *w* trong kho thư rác.

HA(w): số lần xuất hiện của token *w* trong kho thư hợp lệ.

STM: tổng số thư trong kho thư rác.

HTM: tổng số thư trong kho thư hợp lệ.

Hệ số "2" để tăng khả năng nhận được thư hợp lệ.

Bảng I. Bảng dữ liệu huấn luyện trong [4].

Token	Số lần xuất hiện		*P(S|w)*
	trong Spam	trong Ham	
A	165	1235	0.2512473
Advised	12	42	0.4177898
As	2	579	0.0086009
Chance	45	35	0.7635468
Clarins	1	6	0.2950775
Exercise	6	39	0.2787054
For	378	1829	0.3417015
Free	253	137	0.8226372
Fun	59	9	0.9427419
Girlfriend	26	8	0.8908609
Have	291	2008	0.2668504
Her	38	118	0.4471509
I	9	1435	0.0155078
Just	207	253	0.6726596
Much	126	270	0.5396092
Now	221	337	0.6222218
Paying	26	10	0.8671995
Receive	171	98	0.8142107
Regularly	9	87	0.2062346
Take	142	287	0.5541010
Tell	76	89	0.6820062
The	185	930	0.3331618
Time	212	446	0.5441787
To	389	1948	0.3340176
Too	56	141	0.4993754
Trial	26	13	0.8339739
Vehicle	21	58	0.4762651
Viagra	39	19	0.8375393
You	391	786	0.5554363
Your	332	450	0.6494897

Tập dữ liệu huấn luyện trong [4] gồm có 432 thư rác và 2170 thư hợp lệ [4].

Khi này, xác suất là Spam của một thư *E* được tính theo công thức:

$$P(S|E) = \frac{\prod_{i=1}^{n} P(S|w_i)}{\prod_{i=1}^{n} P(S|w_i) + \prod_{i=1}^{n} P(H|w_i)} \quad (8)$$

Trong đó,

$$P(H|w_i) = 1 - P(S|w_i) \quad (9)$$

III. MỘT SỐ CẢI TIẾN TRONG CÁCH TÍNH XÁC SUẤT SPAM CỦA TOKEN

Từ công thức (7), chúng ta có một số nhận xét sau:

1) Việc tính xác suất là Spam của mỗi token
   - Chỉ phụ thuộc vào số lần xuất hiện của token *w* và tổng số thư trong mỗi kho thư rác và thư hợp lệ.
   - Chưa xem xét tổng số tần suất của tất cả token,
   - Chưa xem xét số thư chứa token trong mỗi kho thư rác và thư hợp lệ. Khi này, không biết được token xuất hiện trong chỉ một thư hay nhiều thư.
   - Hệ số "2" tăng khả năng nhận nhầm thư rác thành thư hợp lệ.

Trong trường hợp số lần xuất hiện của một token nào đó xấp xỉ hoặc bằng tổng số thư trong kho thư rác và xuất hiện rất ít trong kho hợp lệ. Khi này, tỉ lệ "*SA(w)/STM*" sẽ gần tới hoặc bằng 1 trong khi tỉ lệ "*HA(w)/HTM*" dần tới 0. Ta có xác suất là Spam của token *w* theo đó sẽ gần tới hoặc bằng 1 (theo công thức 7). Từ đây, theo công thức (8), xác suất là Spam của bức thư chứa token này sẽ rất cao hoặc bằng 1. Nói cách khác, xác suất là Spam của bức thư chứa token này gần như chỉ bị ảnh hưởng bởi token này. Ví dụ, nếu một thư chỉ xuất hiện token này 1 lần, các token khác trong thư này có xác suất là spam rất không cao nhưng thư này bị cho là Spam rất cao. Điều này là bất hợp lý.

Dựa theo phân tích trên, chúng tôi nhận thấy như sau: Xác suất là Spam của mỗi token có thể phụ thuộc các yếu tố sau:

a) Số lần xuất hiện của token *w* trong mỗi kho thư rác và thư hợp lệ.

b) Tổng số thư trong mỗi kho thư rác và thư hợp lệ.

c) Tổng số tần suất của tất cả token.

d) Số thư chứa token trong mỗi kho thư rác và thư hợp lệ.

Ngoài ra, việc thay đổi hệ số "2" trong trường hợp khác nhau để tăng cường khả năng nhận biết thư rác hay thư hợp lệ.

Từ đây, chúng tôi đưa ra một số công thức tính xác suất là Spam của mỗi token như như sau.

- Phụ thuộc vào các yếu tố a-c, ta được các công thức:

$$P(S|w) = \frac{\frac{SA(w)}{STA}}{\frac{SA(w)}{STA} + \frac{HA(w)}{HTA}} \quad (10)$$

$$P(S|w) = \frac{\frac{SA(w)}{STA}}{\frac{SA(w)}{STA} + 2\frac{HA(w)}{HTA}} \quad (10.1)$$

$$P(S|w) = \frac{2\frac{SA(w)}{STA}}{2\frac{SA(w)}{STA} + \frac{HA(w)}{HTA}} \quad (10.2)$$

- Phụ thuộc vào các yếu tố a-b, ta được các công thức:

$$P(S|w) = \frac{\frac{SA(w)}{STA}}{\frac{SA(w)}{STM} + \frac{HA(w)}{HTM}} \quad (11)$$

$$P(S|w) = \frac{\frac{SA(w)}{STM}}{\frac{SA(w)}{STM} + 2\frac{HA(w)}{HTM}} \quad \text{(Paul Graham)} \quad (11.1)$$

$$P(S|w) = \frac{2\frac{SA(w)}{STM}}{2\frac{SA(w)}{STM}+\frac{HA(w)}{HTM}} \quad (11.2)$$

- Phụ thuộc vào các yếu tố b-d, ta được các công thức:

$$P(S|w) = \frac{\frac{STM(w)}{STM}}{\frac{STM(w)}{STM}+\frac{HTM(w)}{HTM}} \quad (12)$$

$$P(S|w) = \frac{\frac{STM(w)}{STM}}{\frac{STM(w)}{STM}+2\frac{HTM(w)}{HTM}} \quad (12.1)$$

$$P(S|w) = \frac{2\frac{STM(w)}{STM}}{2\frac{STM(w)}{STM}+\frac{HTM(w)}{HTM}} \quad (12.2)$$

- Phụ thuộc vào các yếu tố c-d, ta được các công thức:

$$P(S|w) = \frac{\frac{SA(w)}{STM(w)}}{\frac{SA(w)}{STM(w)}+\frac{HA(w)}{HTM(w)}} \quad (13)$$

$$P(S|w) = \frac{\frac{SA(w)}{STM(w)}}{\frac{SA(w)}{STM(w)}+2\frac{HA(w)}{HTM(w)}} \quad (13.1)$$

$$P(S|w) = \frac{2\frac{SA(w)}{STM(w)}}{2\frac{SA(w)}{STM(w)}+\frac{HA(w)}{HTM(w)}} \quad (13.2)$$

- Phụ thuộc vào các yếu tố a-b-d, ta được các công thức:

$$P(S|w) = \frac{\frac{SA(w)}{STM}\frac{STM(w)}{STM}}{\frac{SA(w)}{STM}\frac{STM(w)}{STM}+\frac{HA(w)}{HTM}\frac{HTM(w)}{HTM}} \quad (14)$$

$$P(S|w) = \frac{\frac{SA(w)}{STM}\frac{STM(w)}{STM}}{\frac{SA(w)}{STM}\frac{STM(w)}{STM}+2\frac{HA(w)}{HTM}\frac{HTM(w)}{HTM}} \quad (14.1)$$

$$P(S|w) = \frac{2\frac{SA(w)}{STM}\frac{STM(w)}{STM}}{2\frac{SA(w)}{STM}\frac{STM(w)}{STM}+\frac{HA(w)}{HTM}\frac{HTM(w)}{HTM}} \quad (14.2)$$

- Phụ thuộc vào các yếu tố a-b-c-d, ta được các công thức:

$$P(S|w) = \frac{\frac{SA(w)}{STA}\frac{STM(w)}{STM}}{\frac{SA(w)}{STA}\frac{STM(w)}{STM}+\frac{HA(w)}{HTA}\frac{HTM(w)}{HTM}} \quad (15)$$

$$P(S|w) = \frac{\frac{SA(w)}{STA}\frac{STM(w)}{STM}}{\frac{SA(w)}{STA}\frac{STM(w)}{STM}+2\frac{HA(w)}{HTA}\frac{HTM(w)}{HTM}} \quad (15.1)$$

$$P(S|w) = \frac{2\frac{SA(w)}{STA}\frac{STM(w)}{STM}}{2\frac{SA(w)}{STA}\frac{STM(w)}{STM}+\frac{HA(w)}{HTA}\frac{HTM(w)}{HTM}} \quad (15.2)$$

Nếu sử dụng nhóm các công thức 10-12-13-14-15 thì vấn đề trong nhận xét (2) có thể được khắc phục.

IV. THỬ NGHIỆM

Tập dữ liệu mẫu CSDMC2010_SPAM [11]. Tập dữ liệu huấn luyện bao gồm SpamTrain và HamTrain.

A. *Thử nghiệm 1*

HamTrain có 2808 thư hợp lệ, SpamTrain có 1238 thư rác. Tập dữ liệu test bao gồm HamTest (141 thư hợp lệ) SpamTest (140 thư rác). Các bảng 2, 3 và 4 thống kê độ chính xác phân loại Spam thông qua thống kê chỉ số Precision trong các trường hợp: không có hệ số "2", hệ số "2" để tăng cường nhận thư hợp lệ, hệ số "2" để tăng cường nhận thư rác.

Bảng II. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp không có hệ số 2

Công thức	SPAM	HAM
10.1	62.857	96.454
11.1	98.571	92.908
12.1	98.571	90.780
13.1	90.714	94.326
14.1	98.571	85.816
15.1	94.286	92.199

Từ bảng 2, chúng ta thấy độ chính xác nhận SPAM của các công thức 11.1, 12.1 và 14.1 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.1 là cao nhất.

Bảng III. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư hợp lệ

Công thức	SPAM	HAM
10.2	83.571	96.454
11.2	89.286	96.454
12.2	87.143	95.035
13.2	82.143	95.745
14.2	93.571	92.908
15.2	80.714	93.617

Từ bảng 3, chúng ta thấy độ chính xác nhận SPAM của các công thức 14.2 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.2 và 11.2 là cao nhất.

Bảng IV. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư rác

Công thức	SPAM	HAM
10.3	97.857	92.908
11.3	99.286	82.270
12.3	99.286	80.142
13.3	98.571	85.816
14.3	99.286	79.433
15.3	98.571	86.525

Từ bảng 4, chúng ta thấy độ chính xác nhận SPAM của các công thức 11.3, 12.3 và 14.3 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.3 là cao nhất.

B. *Thử nghiệm 2*

HamTrain có 2535 thư hợp lệ, SpamTrain có 1014 thư rác. Tập dữ liệu test bao gồm HamTest (414 thư hợp lệ) SpamTest (364 thư rác). Các bảng 5, 6 và 7 thống kê độ chính xác phân loại Spam thông qua thống kê chỉ số Precision trong các trường hợp: không có hệ số "2", hệ số "2" để tăng cường nhận thư hợp lệ, hệ số "2" để tăng cường nhận thư rác.

Bảng V. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp không có hệ số 2

Công thức	SPAM	HAM
10.1	59.066	98.068
11.1	98.077	95.652
12.1	98.626	93.720
13.1	89.835	96.135
14.1	98.901	87.923
15.1	93.132	93.237

Từ bảng 5, chúng ta thấy độ chính xác nhận SPAM của các công thức 14.1 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.1 là cao nhất.

Bảng VI. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư hợp lệ

Công thức	SPAM	HAM
10.2	78.571	97.826
11.2	86.813	98.068
12.2	88.736	96.618
13.2	77.747	97.826
14.2	90.659	93.720
15.2	77.473	94.686

Từ bảng 6, chúng ta thấy độ chính xác nhận SPAM của các công thức 14.2 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 11.2 là cao nhất.

Bảng VII. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư rác

Công thức	SPAM	HAM
10.3	95.879	94.686
11.3	99.725	84.541
12.3	99.725	82.126
13.3	98.626	87.923
14.3	99.725	81.159
15.3	98.077	89.855

Từ bảng 7, chúng ta thấy độ chính xác nhận SPAM của các công thức 11.3, 12.3 và 14.3 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.3 là cao nhất.

C. Thử nghiệm 3

HamTrain có 2448 thư hợp lệ, SpamTrain có 986 thư rác. Tập dữ liệu test bao gồm HamTest (501 thư hợp lệ) SpamTest (392 thư rác). Các bảng 8, 9 và 10 thống kê độ chính xác phân loại Spam thông qua thống kê chỉ số Precision trong các trường hợp: không có hệ số "2", hệ số "2" để tăng cường nhận thư hợp lệ, hệ số "2" để tăng cường nhận thư rác.

Bảng VIII. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp không có hệ số 2

Công thức	SPAM	HAM
10.1	58.929	98.204
11.1	98.469	95.808
12.1	98.469	93.613
13.1	90.051	96.407
14.1	98.980	88.224
15.1	91.837	92.814

Từ bảng 8, chúng ta thấy độ chính xác nhận SPAM của các công thức 14.1 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.1 là cao nhất.

Bảng IX. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư hợp lệ

Công thức	SPAM	HAM
10.2	78.571	98.004
11.2	85.459	98.204
12.2	87.500	96.607
13.2	76.786	98.004
14.2	90.051	93.413
15.2	75.765	94.810

Từ bảng 9, chúng ta thấy độ chính xác nhận SPAM của các công thức 14.2 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 11.2 là cao nhất.

Bảng X. Thống kê độ chính xác phân loại tập thư rác và thư hợp lệ trong trường hợp hệ số 2 để tăng nhận thư rác

Công thức	SPAM	HAM
10.3	95.918	94.611
11.3	99.745	85.030
12.3	99.745	82.236
13.3	98.724	87.625
14.3	99.745	82.036
15.3	97.959	89.820

Từ bảng 10, chúng ta thấy độ chính xác nhận SPAM của các công thức 11.3, 12.3 và 14.3 là cao nhất. Trong khi đó, độ chính xác nhận HAM của các công thức 10.3 là cao nhất.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã thảo luận, phân tích về kĩ thuật lọc Spam sử dụng Bayes. Từ đó, đưa ra một số cách tính xác suất là Spam của token. Thử nghiệm cho thấy đó là những phương án thay tốt cho bộ lọc Spam dựa trên Bayes trong những trường hợp khác nhau.

Thông qua nhận xét trong các thử nghiệm, chúng tôi thấy rằng:

- Trong trường hợp không có hệ số "2", các công thức 11.1, 12.1 và 14.1 cho độ chính xác

nhận SPAM cao nhất; công thức 10.1 cho độ chính xác nhận HAM cao nhất.

- Trong trường hợp hệ số "2" để tăng cường nhận hợp lệ, các công thức 14.2 cho độ chính xác nhận SPAM cao nhất; công thức 11.2 cho độ chính xác nhận HAM cao nhất.
- Trong trường hợp hệ số "2" để tăng cường nhận rác, các công thức 11.3, 12.3 và 14.3 cho độ chính xác nhận SPAM cao nhất; công thức 10.3 cho độ chính xác nhận HAM cao nhất.

Như vậy, tùy vào mục đích cụ thể của ứng dụng: giữ loại HAM quan trọng hay loại bỏ SPAM nguy hiểm mà chọn công thức tương ứng.

Trong nghiên cứu tiếp theo, chúng tôi dự kiến để xuất công thức tính xác suất là Spam mới cho mỗi token sử dụng logic mờ.

TÀI LIỆU THAM KHẢO

[1] Awad W.A. and ELseuofi S.M., *Machine learning methods for spam e-mail classification*, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011, pp.173-184.

[2] Jialin ma, Yongjun zhang, Jinling liu, *Intelligent SMS spam filtering using topic model*, ieee international conference on intelligent networking and collaborative systems (incos), 2016.

[3] Johan Hovol, *Naïve Bayes Spam filtering using Word-Position-Based attributes*, Proceedings of the 15th NODALIDA conference, 2006, pp. 78–87.

[4] Paul Graham, *Better Bayesian filtering*. In Proceedings of the 2003 Spam Conference (http://spamconference.org/ proceedings2003.html), Cambridge, MA, 2003.

[5] Phan Hữu Tiếp, Vũ Đức Lung, Cao Nguyễn Thủy Tiên, Lâm Thành Hiển, *Phương pháp lọc thư rác tiếng việt dựa trên từ ghép và theo vết người sử dụng*, Hội thảo "*Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*", Cần Thơ, 2011.

[6] Shahar Yifrah và Guy Lev, *Machine Learning Final Project Spam Email Filtering*, ML Project, 2013.

[7] Tianda Yang, Kai Qian, Dan Chia-Tien Lo, *Spam filtering using Association Rules and Naïve Bayes Classifier*, IEEE International Conference on Progress in Informatics and Computing (PIC), 2015.

[8] Tianhao Sun, *Spam Filtering based on Naïve Bayes Classication*, May 2009.

[9] Vangelis Metsis, Ion And rout sopoulos and Georgios Paliouras, *Spam Filtering with Naïve Bayes–Which Naïve Bayes?*, CEAS2006-Third Conference on Email and Anti-Spam, Mountain View, California USA, July 27-28, 2006.

[10] Yun-Nung Chen, Che-An Lu, Chao-Yu Huang, *Anti-Spam Filter Based on Naïve Bayes, SVM, and KNN model*, AI term project group 14, 2009.

[11] http://csmining.org/index.php/spam-email-datasets-.html

[12] http://kaspersky.nts.com.vn/

[13] http://antoanthongtin.vn/

ASSESS SOME METHODS OF CALCULATING SPAM PROBABILITY OF TOKENS APPLIED IN SPAM EMAIL CLASSYFICATION

Abstract: Spam mail classification is interested in researching for long time in the world with many different approachs. Spam classification functions are intergrated in Mail Server or Mail Client. Currently, the traditional methods still have certain weaknesses, so statistical machine learning classification method based on the content has been proven more effective. Wherein, Bayes spam classification has some advantages such as simplicity, ease of use and short execution time, so it is implemented widely in Mail Server or Mail Client systems. This paper evaluates some Bayes spam classification methods based on token probability rules.

Keyword: Spam, Ham, Spam mail, Spam classification, Statistical machine learning, Tokens.

Nguyễn Tu Trung, Tốt nghiệp đại học trường ĐH Sư phạm Hà Nội 2 năm 2007, thạc sỹ tại trường ĐHCông Nghệ, ĐHQGHN năm 2011, tiến sĩ, Học viện Công nghệ Bưu chính Viễn thông năm 2018. Lĩnh vực nghiên cứu: Xử lý ảnh, xử lý tiếng nói, hệ thống thông tin, hệ thống nhúng.