BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Thanh Hà

NGHIÊN CỨU PHƯƠNG PHÁP XÁC ĐỊNH THỨ TỰ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

Chuyên ngành : Hệ thống thông tin
Mã số: 9.48.01.04

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - Năm 2023

Công trình được hoàn thành tại: Học viện Công nghệ Bưu chính viễn thông

Người hướng dẫn khoa học: PGS. TS. Trần Quang Anh, TS. Trần Hùng
(Ghi rõ học tên, chức danh khoa học, học vị)

Phản biện 1:……
…….

Phản biện 2:……
…….

Phản biện 3……
……..

Luận án được bảo vệ trước Hội đồng chấm luận cấp Học viện
họp tại:……..
……
Vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại thư viện:……..
(*ghi tên các thư viện nộp luận án*)

MỞ ĐẦU

1. GIỚI THIỆU

Thư điện tử là ứng dụng quan trọng, tiện lợi, nhiều ưu điểm. Sự phổ biến của thư điện tử làm xuất hiện vấn đề quá tải email. Nguyên nhân thứ nhất dẫn đến quá tải email là thư rác. Các bộ lọc thư rác được nghiên cứu để giải quyết vấn nạn này. Nguyên nhân thứ hai của vấn đề quá tải email là số lượng thư hợp lệ mà người dùng nhận được quá lớn, trong đó lượng thư quan trọng chỉ chiếm một phần nhỏ. Để giải quyết vấn đề này, bài toán tổng quát được đặt ra là xác định thứ tự ưu tiên của thư điện tử để từ đó gợi ý những bức thư có mức độ ưu tiên cao cho người dùng.

SpamAssassin là hệ thống lọc thư rác có hiệu năng tốt đang được áp dụng phổ biến. Nghiên cứu về lọc thư rác được chia thành các nhóm phương pháp khác nhau, trong đó có một hướng nghiên cứu về lọc thư rác trên nền tảng SpamAssassin.

2. TÍNH CẤP THIẾT CỦA LUẬN ÁN

Thư rác chiếm tỷ lệ cao trong tổng số thư điện tử và gây ra nhiều thiệt hại về kinh tế, xã hội. Cùng với sự phát triển của các bộ lọc thư rác thì kỹ thuật phát tán thư rác cũng không ngừng được cải tiến. Lưu lượng sử dụng thư điện tử ngày càng cao và nội dung thư rác không ngừng thay đổi.

Tình trạng quá tải thư điện tử đang trở nên ngày càng nghiêm trọng và gây tốn thời gian cho người dùng thư điện tử, dẫn đến giảm năng suất làm việc. Trong khi đó, các bộ lọc thư rác không giải quyết được vấn đề quá tải email gây ra bởi thư hợp lệ.

3. MỤC TIÊU CỦA LUẬN ÁN

Các mục tiêu của luận án được đưa ra dựa trên *ba vấn đề* chưa được giải quyết. *Thứ nhất*, đã có nhiều phương pháp xây dựng tập luật được đề xuất dành cho SpamAssassin, nhưng việc lựa chọn luật và gán điểm số cho luật vẫn được thực hiện tách rời nhau, dẫn đến tập luật tìm được chưa thực sự tối ưu. *Thứ hai*, nền tảng SpamAssassin tuy được sử dụng rộng rãi trên các máy chủ thư điện tử nhưng chưa có tính năng dự đoán hành động cho người dùng. Nghiên cứu tính năng dự đoán hành động cho SpamAssassin sẽ giúp triển khai tính năng này trên những hệ thống máy chủ thư điện tử hiện tại trở nên dễ dàng hơn. *Thứ ba*, những nghiên cứu trước đó về xếp hạng thư điện tử đạt được độ chính xác chưa cao. Từ các vấn đề này, luận án đặt ra *ba mục tiêu* chính:

1. Nghiên cứu và đề xuất phương pháp tự động sinh tập luật lọc thư rác cho nền tảng SpamAssassin
2. Nghiên cứu và đề xuất phương pháp dự đoán hành động người dùng dựa trên nền tảng SpamAssassin.
3. Nghiên cứu và đề xuất phương pháp xếp hạng thư điện tử với năm mức độ ưu tiên, có độ chính xác dự đoán cao hơn so với các phương pháp cũ.

4. PHƯƠNG PHÁP NGHIÊN CỨU

Luận án vận dụng các phương pháp nghiên cứu cơ sở lý thuyết, kế thừa kết quả nghiên cứu, phân tích thực nghiệm và so sánh, đối chứng kết quả thí nghiệm. Trước tiên, luận án tham khảo các kiến thức nền tảng có liên quan đến đối tượng nghiên cứu là thư điện tử tiếng Việt. Các tài liệu tham khảo tập trung vào các bài toán và phương pháp phân loại và xác định thứ tự ưu tiên của thư điện tử đã công bố. Từ đó rút ra các kết quả nghiên cứu có giá trị và các vấn đề còn tồn đọng. Tiếp đó, luận án kế thừa kết quả của các nghiên cứu được tham khảo đồng thời đề xuất các phương pháp mới để giải quyết các vấn đề còn tồn đọng. Các thí nghiệm được thực hiện đối với các phương pháp đề xuất và kết quả thực nghiệm được phân tích để rút ra được các kết luận. Kết quả thí nghiệm trên phương pháp đề xuất sẽ được đánh giá, so sánh về mặt định lượng cũng như về mặt định tính với những nghiên cứu đã công bố có liên quan.

5. CÁC ĐÓNG GÓP CỦA LUẬN ÁN

Đóng góp thứ nhất là phương pháp sinh luật lọc thư rác cho SpamAssassin dựa trên mạng nơ-ron. Phương pháp đề xuất cho phép lựa chọn và điều chỉnh tập đặc trưng ngay trong quá trình huấn luyện mô hình. Mục tiêu là tìm ra tập luật và tập điểm số với hiệu quả phát hiện thư rác cao nhất.

Đóng góp thứ hai là phương pháp dự đoán hành động người dùng dựa trên nền tảng SpamAssassin. Ưu điểm của phương pháp là dễ triển khai và có tốc độ xử lý nhanh. Các mô hình dự đoán hành động người dùng đã được xây dựng dựa trên sự kết hợp nhiều tập luật SpamAssassin.

Đóng góp thứ ba là phương pháp xếp hạng thư điện tử với năm mức độ ưu tiên dựa trên phương pháp học sâu nhằm giải quyết vấn đề quá tải thư điện tử. Phương pháp khai thác đồng thời nhóm đặc trưng nội dung và đặc trưng xã hội từ dữ liệu của người dùng. Phương pháp đề xuất có độ chính xác cao hơn các phương pháp cũ và có thể được áp dụng để xây dựng ứng dụng xếp hạng thư điện tử độc lập.

6. BỐ CỤC CỦA LUẬN ÁN

Chương 1 – Tổng quan về thư điện tử và xác định thứ tự ưu tiên của thư điện tử.

Chương 2 – Phát hiện thư rác.

Chương 3 – Dự đoán hành động người dùng thư điện tử.

Chương 4 – Xếp hạng thư điện tử.

CHƯƠNG 1 – TỔNG QUAN VỀ THƯ ĐIỆN TỬ VÀ XÁC ĐỊNH THỨ TỰ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

1.1. HỆ THỐNG THƯ ĐIỆN TỬ

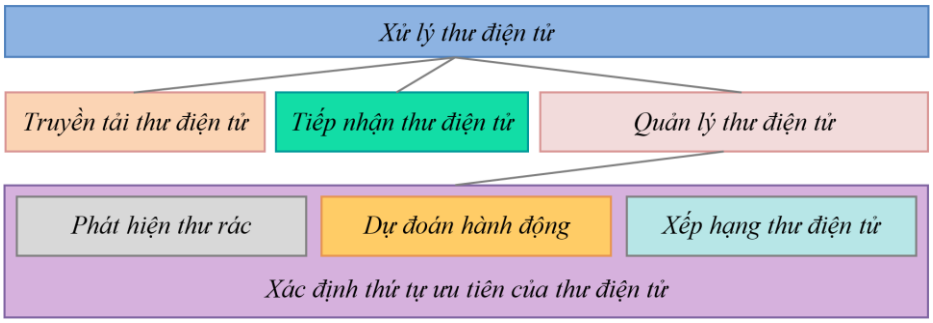
1.1.1. Sơ lược về thư điện tử

Thư điện tử có lịch sử phát triển từ năm 1972 tới nay. Từ khi ra đời, nhiều hệ thống, giao thức thư điện tử đã ra đời và được cải tiến. Thư điện tử được sử dụng rộng rãi trên thế giới và khối lượng sử dụng thư điện tử vẫn còn tiếp tục tăng.

1.1.2. Cấu trúc của một bức thư điện tử

Một bức thư gồm các trường header và nội dung thư, trong đó nội dung thư có thể là văn bản thuần túy hoặc siêu văn bản (HTML), có thể đính kèm nhiều dạng tập tin. Một số trường header có ý nghĩa với nghiên cứu về thư điện tử là: From, To, In-Reply-To, References, Cc, Bcc, Subject, Date.

1.1.3. Mô hình xử lý thư điện tử



Hình 1.1: Mô hình xử lý thư điện tử tổng quát

Ba khâu chính trong hệ thống xử lý email là truyền tải (transport), tiếp nhận (delivery) và quản lý email (management). Trong đó, các nghiên cứu trong luận văn tập trung ở khâu quản lý email. Trong mỗi khâu lại có nhiều giao thức được xây dựng để quy định việc định dạng văn bản và giao tiếp qua mạng. Tác vụ lọc thư rác thường được thực hiện ở bước truyền tải và tiếp nhận thư điện tử bởi số lượng thư rác rất lớn, cần phải được loại bỏ trước khi thư rác được truyền đến hòm thư của người sử dụng. Tác vụ dự đoán hành động và xếp hạng thư điện tử thường được thực hiện ở bước quản lý thư điện tử vì mục tiêu của hai bài toán này là sắp xếp, bố trí lại hòm thư của người dùng. Cả ba bài toán lọc thư rác, dự đoán hành động người dùng và xếp hạng email đều nằm trong bài toán tổng quát là bài toán xác định thứ tự ưu tiên của thư điện tử.

1.1.3. Sơ lược về thư rác

Hành động phát tán thư rác là khai thác thư điện tử trên quy mô lớn nhằm phục vụ mục đích thương mại, quảng cáo, và một số mục đích xấu khi không được sự đồng tình của người sử dụng. Có quan điểm coi thư rác là những thư quảng cáo không được yêu cầu, có quan điểm rộng hơn cho rằng thư rác bao gồm thư quảng cáo, thư quấy rối, và những thư có nội dung không lành mạnh không được người dùng mong muốn và được gửi với số lượng lớn. Nhìn chung, thư rác là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận. Với thư rác, người gửi thư thường không có quan hệ với người nhận.

1.2. CÁC BÀI TOÁN XÁC ĐỊNH THỨ TỰ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

Xác định thứ tự ưu tiên của thư điện tử là tên gọi chung để chỉ tất cả các bài toán có mục tiêu phân biệt thư điện tử theo mức độ quan trọng, có 3 bài toán con là *lọc thư rác, dự đoán hành động người dùng* và *xếp hạng thư điện tử.*

1.2.1. Lọc thư rác

Lọc thư rác là bài toán xác định thứ tự ưu tiên của thư điện tử với hai mức độ: *thư rác* và *thư hợp lệ*. Thư rác xuất hiện chủ yếu do nhu cầu quảng cáo. Ngoài ra thì thư rác còn có mục đích quấy rối, lừa đảo. Nhìn

chung, thư rác là thư không được mong muốn và được gửi với số lượng lớn, trong đó người gửi không có quan hệ với người nhận.

1.2.2. Dự đoán hành động của người dùng thư điện tử

Dự đoán hành động người dùng của thư điện tử có mục tiêu gợi ý hành động cần thực hiện đối với một bức thư cho người dùng.

Dự đoán hành động người dùng là một dạng của bài toán xác định thứ tự ưu tiên của thư điện tử, được phân biệt với các bài toán khác ở đặc điểm kết quả dự đoán là một hành động của người dùng. Mục tiêu bài toán này là gợi ý một trong một số hữu hạn hành động được định nghĩa sẵn mà người dùng cần thực hiện đối với một bức thư nhận. Bài toán này nằm trong nhóm các bài toán phân loại, không đặt ra ràng buộc về quan hệ tương đối giữa các hành động. Bài toán được định nghĩa cụ thể như sau.

Gọi tập hợp tất cả các bức thư mà người dùng nhận được là tập E. Gọi tập hợp các hành động là A = {a1, a2, an}, ($n \geq 2$). Ta cần tìm một hàm dự đoán hành động sao cho đầu vào là một bức thư và đầu ra là hành động cần làm với bức thư đó:

$$f(m): E \rightarrow A$$

Phương pháp chung để giải bài toán này bằng kỹ thuật học máy là định nghĩa một tập dữ liệu huấn luyện M = {m1, m2, …, mn}. M là tập con của E. Mỗi bức thư trong tập M được người dùng gán cho một hành động phù hợp nhất với nó từ tập A. Đó là quá trình gán nhãn cho tập dữ liệu. Tập M đã được gán nhãn được dùng để huấn luyện một mô hình học máy. Kết quả huấn luyện là một mô hình có chức năng gần giống với chức năng của hàm f (m) nói trên. Một số dạng của bài toán dự đoán hành động đã được nghiên cứu đó là gợi ý trả lời thư [44], dự đoán một trong ba hành động phổ biến (trả lời, đọc, xóa) [51], và phát hiện hành động (lưu trữ, trả lời) [25]. Với cách tiếp cận của nghiên cứu [25], cả hai hành động có thể đồng thời xảy ra, nghĩa là người dùng có thể thực hiện các hành động trả lời và lưu trữ trên cùng một bức thư.

1.2.3. Xếp hạng thư điện tử

Bài toán xếp hạng thư điện tử có mục tiêu chính là đánh giá tầm quan trọng của thư điện tử, nhằm sắp xếp các bức thư theo thứ tự tầm quan trọng. Việc này giúp cải thiện hiệu quả sử dụng thư điện tử của người dùng, từ đó giải quyết vấn đề quá tải thư. Để làm được điều này, bài toán cần phải đánh giá tầm quan trọng của từng bức thư. Có hai hướng tổng quát để dự đoán tầm quan trọng của một bức là phân loại và hồi quy. Ta cần tìm một hàm dự đoán có dạng:

$$g(m): E \rightarrow P$$

Phương pháp phân loại giả thiết rằng các mức độ quan trọng là rời rạc, có thể có hoặc không có quan hệ tương đối (lớn hơn, nhỏ hơn) với nhau. Trong trường hợp này, P là một tập hợp hữu hạn các giá trị rời rạc. Kích thước của tập P là 2 đối với bài toán phát hiện thư rác. Đối với bài toán dự đoán hành động người dùng, tập P có thể chứa từ hai hành động [44] hoặc nhiều hơn hai hành động [51]. Ngược lại, phương pháp hồi quy giả thiết rằng các mức độ quan trọng là liên tục và có quan hệ tương đối với nhau. Khi đó, P là một tập số thực liên tục. Một nghiên cứu về xếp hạng thư điện tử vào năm 2005 [11] là ví dụ về giải quyết bài toán theo hướng hồi quy.

Khác với bài toán phân loại thư điện tử, bài toán xếp hạng thư điện tử tập trung vào mô phỏng mức độ quan trọng của các bức thư đối với người dùng. Trong khi đó, tầm quan trọng của bức thư không làm ảnh hưởng tới việc lựa chọn nhãn trong bài toán phân loại theo thư mục.

Hiện tại, một số giải pháp đã được đưa ra để xếp hạng thư điện tử với các thuật toán và tiêu chí đánh giá kết quả khác nhau [11, 40, 49]. Tuy vậy, nhìn chung bài toán xếp hạng thư điện tử vẫn chưa được giải quyết triệt để. Thí nghiệm của bài báo [11] cho thấy trung bình sai số của phương pháp xếp hạng đối với 236 bức

thư dùng để thử nghiệm là 33.1 với độ lệch chuẩn là 29.1. Những con số này cho thấy mức độ sai số trong xếp hạng còn lớn và giá trị độ lệch chuẩn cao cho thấy có những bức thư được xếp hạng sai lệch xa so với thứ tự thực tế của chúng. Từ kết quả thí nghiệm của nghiên cứu [49], sai số trung bình thấp nhất đạt được là khoảng 0.8. Với 5 mức độ ưu tiên được sử dụng nghiên cứu này, tuy chỉ số accuracy không được công bố nhưng ta có thể tính được khoảng giá trị của accuracy dựa theo sai số đó. Trường hợp có chỉ số accuracy tệ nhất: có khoảng 80% bức thư được dự đoán sai với sai số là 1. Trường hợp có chỉ số accuracy tốt nhất: có 20% bức thư được dự đoán sai với sai số là 4. Vậy, chỉ số accuracy trong trường hợp tốt nhất tương ứng với trung bình sai số 0.8 là 80% và trong trường hợp xấu nhất là 20%. Với việc sử dụng thêm nhiều đặc trưng xã hội, nghiên cứu [46] cho thấy hiệu quả cao hơn, với trung bình sai số đạt được khoảng 0.67, trên tập dữ liệu nhỏ hơn so với tập dữ liệu trong [49]. Giá trị trung bình sai số này tương ứng với accuracy tối đa là 83.25%. Các kết quả nói trên cho thấy bài toán xếp hạng cần được tiếp tục nghiên cứu.

1.3. TỔNG QUAN NGHIÊN CỨU VỀ XÁC ĐỊNH THỨ TỰ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

1.3.1. Nghiên cứu về lọc thư rác

Nhiều phương pháp khác nhau đã được đề xuất dành cho bài toán lọc thư rác. Các phương pháp lọc thư rác có thể được phân loại dựa theo đặc trưng được sử dụng (tiêu đề, nội dung, header…) hoặc theo phương pháp (lọc theo danh sách, xác thực người gửi, lọc theo luật, học máy có giám sát, học máy không giám sát…)

1.3.1.1. Lọc thư rác trên nền tảng SpamAssassin

```
body        MONEY_BACK        /money back guarantee/i
describe    MONEY_BACK        Money back guarantee
score       MONEY_BACK        1.887
```

Hình 1.1: Một luật từ khóa của SpamAssassin áp dụng với phần body của thư điện tử.

SpamAssassin là nền tảng lọc thư rác dựa trên phương pháp luật có trọng số. Bộ bộ lọc thư rác dựa trên SpamAssassin có dạng tập luật có chứa nhiều luật, trong đó mỗi luật là một đặc trưng của email và mỗi luật được gắn với một trọng số. Dựa vào tập luật và các đặc trưng xuất hiện trong bức thư, ta có thể tính toán được điểm số của bức thư. Điểm số nói trên được so sánh với giá trị ngưỡng T để kết luận bức thư có phải là thư rác hay không.

```
This mail is probably spam. The original message has been attached
along with this report, so you can recognize or block similar unwanted
mail in the future. See http://spamassassin.org/tag/ for more details.

Content analysis details: (5.03 points, 5 required)
FREE               (1.872 points)
DEAR_FRIEND        (0.732 points)
MONEY_BACK         (1.887 points)
BODY_BEST          (0.539 points)
```

Hình 1.2: Nội dung bức thư bị SpamAssassin đánh dấu là thư rác, bao gồm báo cáo về các luật được áp dụng và bức thư gốc dưới dạng tệp đính kèm.

Để xây dựng tập luật SpamAssassin, ta cần thực hiện hai bước: xác định tập luật (lựa chọn đặc trưng) và gán trọng số cho luật (tối ưu tập luật). Trước đây, việc xây dựng tập luật SpamAssassin được thực hiện thủ công hoặc tự động, hoặc tự động một phần. Hai phương pháp đầu tiên đã áp dụng thuật toán GA và SGD [17] để tự động điều chỉnh trọng số cho tập luật trong khi việc lựa chọn luật vẫn được thực hiện thủ công. Nghiên cứu [28] hoàn thiện quy trình tự động sinh tập luật bằng cách bổ sung một số phương pháp tự động lựa chọn luật từ dữ liệu. Nghiên cứu [62] tiếp tục cải thiện phương pháp lựa chọn luật bằng cách bổ sung thêm đặc trưng (luật ham) và thử nghiệm thêm thuật toán HPSOWM để gán điểm số cho tập luật.

1.3.1.2. Lọc thư rác bằng phương pháp học máy thống kê

Học máy là phương pháp phổ biến nhất để giải quyết bài toán lọc thư rác. Bộ lọc Bayes [14] và các bộ lọc dựa trên thống kê tương tự có hiệu phát hiện thư rác cao nhưng có hạn chế đó là có thể bị đánh bại bởi kỹ thuật statistical poisoning [21]. Bộ lọc SVM [9] không đòi hỏi cao về lựa chọn thuộc tính nhưng phụ thuộc nhiều vào tính đại diện của dữ liệu huấn luyện. Mạng nơ-ron cũng đã được thử nghiệm với bài toán lọc thư rác với hiệu quả cao, tuy nhiên hiệu quả chưa được chứng minh là ổn định [43]. Các phương pháp tự động sinh tập luật cho SpamAssassin có thể được phân loại vào nhóm phương pháp học máy bởi vì tập luật SpamAssassin có cơ chế hoạt động giống với mạng nơ-ron một lớp (perceptron).

1.3.1.3. Lọc thư rác bằng phương pháp mạng thư điện tử

Hướng nghiên cứu này khai thác đặc trưng xã hội (gắn với người gửi thư) của thư điện tử bằng cách xây dựng đồ thị người dùng thư điện tử. Các phương pháp trong nhóm này tính các chỉ số của mỗi người dùng email theo lý thuyết đồ thị: hệ số phân cụm [29], MailRank [27], hệ số phân cụm mở rộng và PageRank có trọng số [54].

1.3.1.4. Lọc thư rác dựa trên dấu hiệu

Nhóm phương pháp này là những phương pháp đơn giản, xuất hiện từ khi thư điện tử mới trở nên phổ biến. Các dấu hiệu để lọc thư rác có thể là các từ, cụm từ trong bức thư, cũng có thể là những luật phức tạp hơn, ví dụ như regular expression. Luật cũng có thể được gắn trọng số hoặc không có trọng số. Phương pháp này có các hạn chế: tốn nhiều thời gian để xây dựng bộ lọc, khả năng thích nghi với dữ liệu mới thấp, dễ dàng bị vô hiệu hóa bởi các kỹ thuật phát tán thư rác mới.

1.3.1.5. Lọc thư rác dựa trên cơ sở hạ tầng gửi thư

Giao thức SMTP được thiết kế đơn giản nhằm đảm bảo hiệu năng, và đó cũng là hạn chế mà kẻ phát tán thư rác có thể lợi dụng [35]. Có nghiên cứu đề xuất thay đổi giao thức SMTP bằng một giao thức khác [33]. Tuy nhiên đề xuất này thiếu thực tế vì việc triển khai quá tốn kém và gây ra ảnh hưởng với nhiều dịch vụ và hoạt động kinh doanh liên quan. Phương pháp [42] đề xuất phát hiện thư rác dựa trên chuỗi lệnh SMTP bất thường. Cách làm này có thể bị vô hiệu hóa nếu kẻ phát tán thư rác cố tình giả mạo chuỗi lệnh gửi thư giống với thư hợp lệ. Phương pháp chuỗi hỏi đáp ngăn chặn thư rác bằng cách yêu cầu người gửi trả lời một câu hỏi xác nhận trong lần gửi đầu tiên nhằm gây khó khăn cho việc gửi thư hàng loạt. Các phương pháp DomainKeys [48], SPF [60] và DMARC [68] đề xuất phương pháp ngăn chặn kỹ thuật giả mạo địa chỉ người gửi của kẻ phát tán thư rác. Những phương pháp này làm tăng độ khó của việc gửi thư rác, nhưng vẫn chưa thể triệt để loại bỏ thư rác vì kẻ phát tán có thể thông qua nhiều dịch vụ gửi thư lớn như Gmail, MailGun, SendGrid… để phát tán thư rác trà trộn cùng thư hợp lệ.

1.3.2. Nghiên cứu về dự đoán hành động người dùng

Phương pháp dự đoán hành động người dùng có mục tiêu gợi ý hành động phù hợp đối với các bức thư mà người dùng nhận được, giúp tiết kiệm thời gian xử lý email. Nghiên cứu [25] chỉ ra những yếu tố trong nội dung email có ảnh hưởng đến hành động của người dùng. Mô hình Bayes đã được áp dụng để dự đoán hành động người dùng [51] với ba hành động: *trả lời*, *đọc* và *xóa*. Mô hình Bayes gặp khó khăn trong việc phân biệt giữa *đọc* và *trả lời* bởi vì sự giống nhau giữa giống nhau hai hành động này. Dự đoán thư cần trả lời cũng là một dạng của bài toán dự đoán động người dùng. Phương pháp [44] tiếp cận bài toán này theo hướng sử dụng các quy tắc dựa trên kinh nghiệm (heuristics). Một nghiên cứu trên quy mô lớn [71] cũng đề xuất phương pháp dự đoán hành động người dùng với bốn hành động: *trả lời*, *đọc*, *xóa sau khi đọc* và *xóa khi chưa đọc*. Nghiên cứu này kết luận hành động người dùng và tầm quan trọng không có mối liên hệ rõ ràng. Nghiên cứu [74] đề xuất một số kỹ thuật trích chọn đặc trưng mới để giải quyết bài toán dự đoán hành động trả lời thư. Nghiên

cứu [82] đề xuất xây dựng hệ gợi ý để dự đoán hành động trả lời thư. Với một bức thư được gửi cho nhiều người, hệ gợi ý dự đoán hành động trả lời thư đối với từng người nhận dựa trên lịch sự nhận thư của người đó.

1.3.3. Nghiên cứu về xếp hạng thư điện tử

Bài toán xếp hạng thư điện tử có mục tiêu phân loại email dựa trên quan điểm cá nhân của người dùng về tầm quan trọng của bức thư thay vì hành động dành cho bức thư. Trong số những phương pháp đã công bố, các phương pháp xếp hạng email cá nhân chiếm đa số. Điều đó thể hiện nhu cầu lớn về xếp hạng email cá nhân và tính cá nhân hóa rõ rệt của bài toán. Nghiên cứu [40] sử dụng một số đặc trưng mạng xã hội kết hợp với nội dung thư để giải bài toán xếp hạng email với năm mức độ ưu tiên. Với phương pháp biểu diễn đặc trưng tf-idf và mô hình phân loại năm lớp dựa trên thuật toán SVM, độ chính xác đạt được của mô hình còn thấp. Phương pháp [45] định nghĩa tầm quan trọng của email là xác suất mà người dùng sẽ thực hiện bất kỳ hành động nào đối với nó. Nghiên cứu [45] áp dụng các máy phân loại đơn giản trên hòm thư cá nhân của người dùng để phát hiện những bức thư quan trọng. Nghiên cứu này cũng một lần nữa khẳng định tính cá nhân hóa của bài toán xếp hạng email. Nghiên cứu [49] tiếp tục thực hiện thí nghiệm trên bài toán xếp hạng email với năm mức độ dựa trên đặc trưng nội dung, trong đó phương pháp phân loại và hồi quy đã được so sánh. Thí nghiệm cho thấy phân loại là hướng tiếp cận phù hợp hơn dành cho dữ liệu email cá nhân.

1.3.4. Các tiêu chí đánh giá

1.3.4.1. Đánh giá mô hình phân loại nhị phân

Mô hình phân loại nhị phân được ứng dụng trong bài toán lọc thư rác. Có nhiều tiêu chí được dùng để đo hiệu quả của mô hình này. Recall, hay độ đo triệu hồi, thể hiện độ nhạy của bộ lọc cũng như độ hoàn thiện của kết quả dự đoán. Precision thể hiện độ tin cậy của kết quả dự đoán. Tiêu chí F_1 là sự cân bằng giữa hay tiêu chí được coi là đối lập nhau: recall và precision. Ngoài ra, tiêu chí đơn giản accuracy, thể hiện tỷ lệ dự đoán đúng trên toàn bộ số lần dự đoán, cũng được sử dụng trong nhiều nghiên cứu về cả ba bài toán.

1.3.4.2. Đánh giá mô hình phân loại đa lớp

Mô hình phân loại đa lớp được ứng dụng cho bài toán dự đoán hành động người dùng và bài toán xếp hạng email. Bởi vì số lượng lớp phân loại lớn hơn 2 nên các tiêu chí recall, precision và F_1 không áp dụng được cho những mô hình này. Thay vào đó là biến thể của các tiêu chí nói trên: recall_m, precision_m, Fscore_m, recall_μ, precision_μ, Fscore_μ.

1.4. TẬP DỮ LIỆU THƯ ĐIỆN TỬ

1.4.1. Tập dữ liệu Enron

Enron là tập dữ liệu email tiếng Anh lớn gồm có thư của các nhân viên từ tập đoàn Enron, Hoa Kỳ. Tập dữ liệu gồm 200,399 của 158 người dùng. Tập dữ liệu Enron chưa được gán nhãn. Hầu hết các nghiên cứu sử dụng tập dữ liệu này đều tiến hành gán nhãn cho một phần của nó. Cho tới nay, tập dữ liệu Enron được sử dụng cho các bài toán lọc thư rác và phân loại email.

1.4.2. Tập dữ liệu TREC

Hội nghị TREC cung cấp một số tập dữ liệu tiếng Anh và tiếng Trung về lọc thư rác. Số lượng thư của các tập dữ liệu TREC được thể hiện trong Bảng 1.1.

Bảng 1.1: Số lượng thư trong các tập dữ liệu mẫu về thư rác của hội thảo TREC

Tập mẫu	Số lượng spam	Số lượng ham	Tổng số
TREC 2005	52790	39399	92189
TREC 2006	24912	12910	37822
TREC 2006 (tiếng Trung)	42854	21766	64620
TREC 2007	50199	25220	75419

Các tập dữ liệu TREC chỉ gồm thư gửi đến, không có thư gửi đi nên không phù hợp để áp dụng cho các nghiên cứu với đặc trưng xã hội.

1.4.3. Các tập dữ liệu khác

Một số tập dữ liệu công khai khác về email là UC, SRI và LingSpam. UC và LingSpam là các tập dữ liệu lọc thư rác kích thước nhỏ. SRI đã từng được sử dụng [28], tuy nhiên hiện tại không còn được duy trì.

1.4.4. Tập dữ liệu thư điện tử tiếng Việt

Chưa có tập dữ liệu thư điện tử tiếng Việt được công bố công khai. Do khan hiếm dữ liệu phục vụ nghiên cứu, luận án đề xuất xây dựng một tập dữ liệu thư điện tử có nội dung tiếng Việt. Tập dữ liệu thô bao gồm 37,003 bức thư đến từ 7 tình nguyện viên là những người dùng Gmail. Dữ liệu được tải về từ Google Takeout, được xử lý trích xuất thông tin bằng thư viện Mail::Mbox::MessageParser và MIME::Parser của ngôn ngữ Perl.

1.4.4.1. Loại bỏ thư có nội dung trùng lặp

Những bức thư trùng lặp được loại bỏ bằng một phương pháp bán tự động, kết hợp giữa công cụ tính mức độ tương tự của các bức thư và xét duyệt thủ công. Mức độ tương tự được tính dựa trên khoảng cách Euclidean. Các bức thư có mức độ giống nhau trên 75% được xem xét và loại bỏ thủ công.

1.4.4.2. Loại bỏ thư tiếng nước ngoài

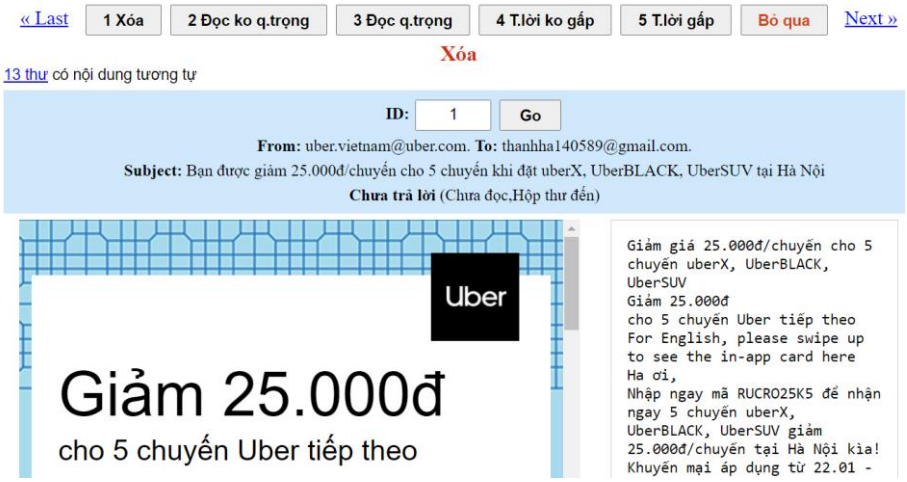
Một trong những vấn đề khi phát hiện thư có nội dung tiếng Việt là không dễ để phân biệt văn bản tiếng Việt không dấu và văn bản tiếng Anh. Luận án đã sử dụng từ điển 10,000 từ tiếng Anh phổ biến và tự điển từ đơn tiếng Việt gồm 5,847 từ để phát hiện thư có nội dung tiếng Việt. Luận án đề xuất chia các bức thư thành 3 nhóm: (1) thư tiếng Anh gồm > 50% số từ thuộc từ điển tiếng Anh, (2) thư tiếng Việt bao gồm các bức thư có dưới 50% số từ thuộc từ điển tiếng Anh và trên 35% thuộc từ điển tiếng Việt, (3) các bức thư chưa xác định ngôn ngữ. Những bức thư chưa xác định ngôn ngữ có khả năng thuộc về các ngôn ngữ khác và được gán nhãn ngôn ngữ một cách thủ công.

1.4.4.3. Các bước tiền xử lý khác

Ngoài ra, những bức thư quá ngắn, những bức thư chỉ bao gồm tệp đính kèm… cũng được loại bỏ. Luận án không nghiên cứu các bức thư mà toàn bộ nội dung được đưa vào tệp đính kèm hoặc trong hình ảnh. Trên thực tế, có thể đưa những bức thư không được phân loại vào một thư mục riêng để người dùng tự xử lý.

1.4.4.4. Gán nhãn cho tập dữ liệu

Sau các bước xử lý, tập dữ liệu còn lại 12,118 bức thư tiếng Việt. Tập dữ liệu được tiến hành gán nhãn thủ công bởi các tình nguyện viên. Trước tiên, tập dữ liệu được gán nhãn cho bài toán lọc thư rác với 02 lớp: *thư rác*, *thư hợp lệ*. Có 2,527 bức thư được gán nhãn *thư rác* và 9,591 bức thư được gán nhãn *thư hợp lệ* được gán nhãn cho bài toán lọc thư rác. Sau đó, những bức thư *hợp lệ* được chia ra thành hai nhãn *đọc* hoặc *trả lời*, thu được 7,966 bức thư cần *đọc* và 1,625 bức thư cần *trả lời*. Từ đây cho thấy trên thực tế những bức thư quan trọng có số lượng nhỏ so với những bức thư ít quan trọng hơn. Tiếp theo, 03 nhãn dữ liệu có sẵn sẽ được tiếp tục chia thành các mức độ ưu tiên nhỏ hơn. Mức độ *xóa* được coi là mức độ 1. Mức độ *đọc* được chia thành hai mức độ nhỏ là mức độ 2, *đọc không quan trọng*, và mức độ 3, *đọc quan trọng*. Mức độ *trả lời* được chia thành hai mức độ nhỏ là mức độ 4, *trả lời không gấp* và mức độ 5, *trả lời gấp*. Các bức thư được gãn nhãn mức độ quan trọng dựa trên quan điểm cá nhân về mức độ ưu tiên của người gãn nhãn, đồng thời là chủ sở hữu của dữ liệu.



Hình 1.3: Công cụ gán nhãn thư với chức năng phát hiện thư tương tự.

1.4.4.5. Bảo mật thông tin cá nhân

Để công bố tập dữ liệu nhằm phục vụ những nghiên cứu liên quan cũng như mô phỏng lại kết quả nghiên cứu của luận án, đồng thời không làm cho nội dung thư của tình nguyện viên bị tiết lộ, một phương pháp ánh xạ đơn giản đã được áp dụng để *nạc danh hóa* tập dữ liệu. Phương pháp này sử dụng một bảng ánh xạ từ các từ ngữ tiếng Việt thành các ký hiệu ngẫu nhiên. Để dịch ngược từ tập dữ liệu nạc danh về tập dữ liệu gốc, ta cần có bảng ánh xạ nói trên.

1.4.4.6. Các thông số của tập dữ liệu

Số lượng thư thuộc các mức độ ưu tiên được mô tả trong Bảng 1.2.

Bảng 1.1: Phân bổ thư theo nhãn của tập dữ liệu xếp hạng thư điện tử tiếng Việt.

Mức độ ưu tiên	**Số lượng thư**
1 – Thư cần xóa	2,527
2 – Đọc không quan trọng	2,179
3 – Đọc quan trọng	5,787
4 – Trả lời không gấp	970
5 – Trả lời gấp	655
Tổng số	12,118

1.5. KẾT LUẬN CHƯƠNG 1

Chương 1 giới thiệu tổng quan về thư điện tử và các bài toán xác định thứ tự ưu tiên của thư điện tử. Các cách tiếp cận để giải quyết các bài toán nói trên cũng được khảo sát và phân tích để chỉ ra những thành tựu đã đạt được cùng các vấn đề chưa được giải quyết. Thông qua tổng quan tài liệu, nhiều điểm hạn chế của các phương pháp xác định thứ tự ưu tiên của thư điện tử đã được nêu ra. Trong Chương 1 cũng đã trình bày về việc thu thập và xây dựng tập dữ liệu thư điện tử tiếng Việt phục vụ các nghiên cứu trong các chương tiếp theo của luận án.

CHƯƠNG 2: PHÁT HIỆN THƯ RÁC

2.1. MỞ ĐẦU

2.1.1. Đặc điểm của thư rác

Thư rác được phát tán để phục vụ các mục đích xấu. Người gửi không biết người nhận là ai và ngược lại. Những người cùng nhận thư rác thường không liên lạc với nhau. Sự trao đổi diễn ra một chiều, số lượng thư nhận được của kẻ phát tán thư rác nhỏ hơn rất nhiều so với số thư gửi đi. Địa chỉ gửi thư thường được làm giả. Thư rác được gửi với số lượng lớn, có nội dung quảng cáo, quấy nhiễu hoặc phát tán mã độc, đường dẫn giả mạo nhằm mục đích lừa đảo, đánh cắp thông tin cá nhân, phá hoại dữ liệu của người dùng.

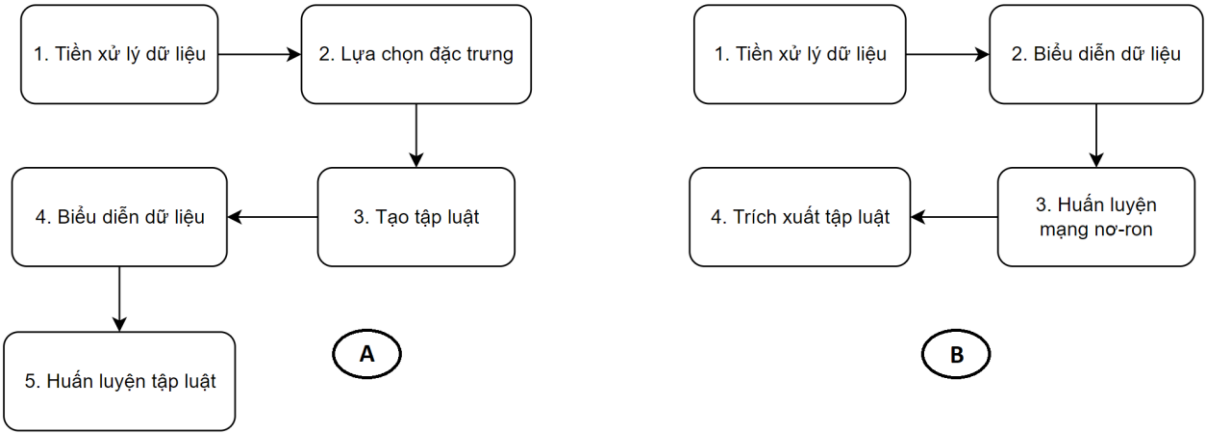
2.1.2. Những vấn đề còn tồn tại

SpamAssassin được sử dụng phổ biến trên các máy chủ thư điện tử, có tốc độ xử lý nhanh, đáp ứng xử lý thư điện tử trong thời gian thực. Tuy vậy, bộ lọc SpamAssassin không có sẵn một hệ thống tự động sinh tập luật. Các phương pháp tự động sinh tập dựa trên mô hình học máy [17, 28, 62] dành cho SpamAssassin cũng đã được đề xuất nhưng vẫn còn tồn tại một số vấn đề. Thứ nhất, khâu lựa chọn luật được thực hiện tách rời với khâu gán điểm số, làm giới hạn không gian tìm kiếm của thuật toán huấn luyện. Thứ hai, chưa có phương pháp để cân bằng độ nhạy và tỷ lệ lọc nhầm của tập luật SpamAssassin sao cho lợi ích đạt được là tối ưu.

2.2. ỨNG DỤNG MẠNG NƠ-RON ĐỂ TỰ ĐỘNG LỰA CHỌN ĐẶC TRƯNG CHO BÀI TOÁN SINH LUẬT SPAMASSASSIN

Trong mục này, luận án đề xuất giải pháp cho vấn đề tồn tại thứ nhất của bài toán lọc thư rác. Phương pháp có mục tiêu cải thiện tập đặc trưng cho mô hình lọc thư rác. Phương pháp được đặt tên là SD_1 trong luận án. Mục tiêu bài toán là xác định một tập luật (đặc trưng) hữu hạn và gán điểm số cho tập luật đó.

2.2.1. Quy trình xây dựng tập luật SpamAssassin với mạng nơ-ron



Hình 2.1: (a) Quy trình tự động sinh tập luật SpamAssassin truyền thống; (b) Quy trình tự động sinh tập luật SpamAssassin dựa trên mạng nơ-ron

Quy trình thông thường (Hình 2.2a) để sinh tập luật SpamAsassin gồm 5 bước: tiền xử lý dữ liệu, lựa chọn đặc trưng, biểu diễn dữ liệu, tạo tập luật và huấn luyện tập luật. Phương pháp đề xuất thực hiện lựa chọn đặc trưng đồng thời với huấn luyện tập luật. Quy trình đề xuất (Hình 2.2b) gồm 4 bước: tiền xử lý dữ liệu, biểu diễn dữ liệu, huấn luyện mạng nơ-ron, trích xuất tập luật. Phương pháp đề xuất hướng tới: *lựa chọn* tập luật tốt nhất, *gán điểm số tối ưu* cho tập luật và *giới hạn số lượng luật* nhằm đảm bảo hiệu năng xử lý cho bộ lọc SpamAssassin.

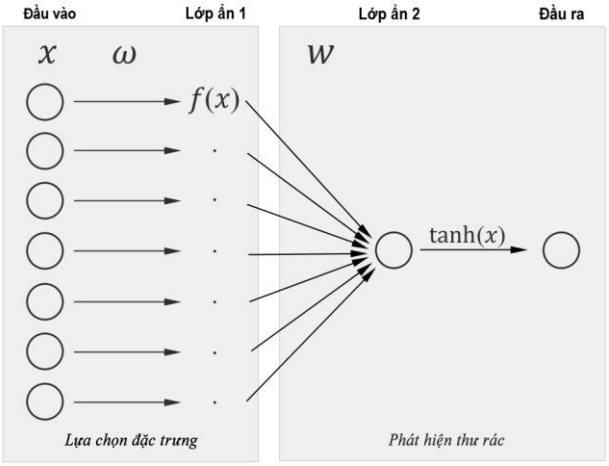
Mô hình mạng nơ-ron nhiều lớp được đề xuất để thực hiện mục tiêu nói trên. Mô hình gồm hai thành phần nối tiếp có chức năng lựa chọn đặc trưng và tối ưu điểm số. Mô hình được xây dựng trên giả thiết những đặc trưng có trị tuyệt đối của trọng số lớn là những đặc trưng quan trọng.

2.2.2. Tiền xử lý và biểu diễn dữ liệu

Phương pháp vnTokenizer [34] được sử dụng để tách từ tiếng Việt từ tiêu đề và nội dung email. Một bức thư được biểu diễn dưới dạng vector nhị phân. Các từ ngữ trong tiêu đề thư được tách riêng với các từ ngữ trong nội dung thư nhằm tạo ra hai loại đặc trưng Header và Body.

2.2.3. Mô hình mạng nơ-ron

Mạng nơ-ron được đề xuất (Hình 2.3) gồm hai thành phần chính: (1) một lớp mạng có nhiệm vụ lựa chọn đặc trưng; (2) mạng perceptron mô phỏng bộ lọc thư rác SpamAssassin, gồm một lớp ẩn. Mạng nơ-ron nói trên được huấn luyện bằng một thuật toán được xây dựng từ thuật toán SGD với mini-batch, sử dụng hàm tổn thất MSE làm cơ sở để điều chỉnh các trọng số. Ba siêu tham số α, β và ε được sử dụng bởi thành phần lựa chọn đặc trưng. α là số lượng luật mục tiêu, β là số lượng luật được lựa chọn ở thời điểm hiện tại, ε là độ lớn của luật được lựa chọn. α là tham số cố định được đặt bởi người dùng trước khi huấn luyện. β và ε thay đổi trong quá trình huấn luyện nhằm điều chỉnh số lượng đặc trưng được lựa chọn sao cho $\beta \leq \alpha$.



Hình 2.3: Cấu trúc mạng nơ-ron với hai phần lựa chọn đặc trưng (*FS*) và phát hiện thư rác (*P*)

Hàm kích hoạt *tanh* được sử dụng thay cho hàm kích hoạt *sigmoid* vì nó đối xứng tại trục tọa độ, cho phép mạng nơ-ron huấn luyện nhanh hơn [52].

2.2.4. Tạo tập luật SpamAssassin

Sau khi hoàn thành huấn luyện mô hình, những đặc trưng được chọn và trọng số tương ứng sẽ được sử dụng để sinh tập luật SpamAssassin. Mỗi đặc trưng tương đương với một từ khóa. Những từ khóa được chọn sẽ trở thành một luật trong tập luật. Nếu đặc trưng thuộc về tập Vs, luật được tạo ra sẽ là kiểu luật HEADER. Nếu đặc trưng thuộc về tập Vb, luật tạo ra sẽ có kiểu là BODY. Trọng số của đặc trưng sẽ được dùng để làm điểm số của luật. Trong cơ chế phân loại của mạng perceptron, đầu ra là tổng có trọng số của đầu vào được cộng thêm giá trị độ lệch bias (**Error! Reference source not found.**). Đầu ra này được so sánh với ngưỡng 0 để đưa ra kết quả dự đoán. Điều đó tương đương với việc lấy tổng có trọng số của đầu vào và so sánh với ngưỡng -bias. Cơ chế phát hiện thư rác của SpamAssassin là so sánh tổng có trọng số của đầu vào (**Error! Reference source not found.**) với ngưỡng T để đưa ra kết quả dự đoán. Nếu trực tiếp sử dụng trọng số của các đặc trưng để làm điểm số cho luật, ta có thể lấy giá trị -bias để làm ngưỡng cho SpamAssassin. Một cách khác để gán điểm số cho tập luật trong khi vẫn giữ nguyên giá trị ngưỡng T = 5.0 là sử dụng công thức (**Error! Reference source not found.**) để tính điểm số của luật từ trọng số

2.3. ỨNG DỤNG TỐI ƯU HÓA ĐA MỤC TIÊU ĐỂ XÁC ĐỊNH ĐIỂM SỐ CHO TẬP LUẬT SPAMASSASSIN

Trong mục này, luận án đề xuất giải pháp cho vấn đề tồn tại thứ hai của bài toán lọc thư rác. Phương pháp có mục tiêu tối ưu các trọng số của tập luật SpamAssassin theo nhiều mục tiêu (tối ưu hóa đa mục tiêu), hướng tới cân bằng các chỉ số quan trọng của bộ lọc. Phương pháp được đặt tên là SD_2 trong luận án.

2.3.1. Ứng dụng tối ưu hóa đa mục tiêu để sinh tập luật SpamAssassin

Hai tiêu chí phổ biến để đánh giá bộ lọc thư rác là độ chính xác *recall* và tỷ lệ lọc nhầm FAR. Mô hình có tiêu chí *recall* càng cao thì càng tốt, trong khi FAR càng thấp thì càng tốt. Tuy nhiên, khi mô hình được tối ưu cho *recall* thì FAR cũng tăng, và khi tối ưu FAR thì kéo theo *recall* giảm. Từ đó dẫn đến nhu cầu cân bằng hai tiêu chí nói trên để đạt được lợi ích tối đa từ một bộ lọc thư rác. Để tối ưu hóa đồng thời cả hai tiêu chí, ta cần giải quyết bài toán tối ưu đa mục tiêu.

Coi bài toán gán điểm số cho tập luật SpamAssassin là một bài toán tối ưu hóa đa mục tiêu trong đó ta cần tìm ngưỡng T và tập điểm số sao cho *recall* và FAR là tối ưu. Do giữa *recall* và FAR có sự phụ thuộc lẫn nhau, trên thực tế không thể tìm được phương án toàn vẹn. Thay vào đó, ta tìm tập các phương án thỏa hiệp, còn được gọi là các phương án tối ưu Pareto [22] để tối ưu hóa lợi ích mà tập luật mang lại.

2.3.2. Ứng dụng phương pháp tối ưu hóa Pareto

Một phương án được gọi là phương án tối ưu Pareto nếu nó không bị vượt trội bởi bất cứ phương án nào khác trong không gian phương án. Tuy nhiên việc tìm tập tối ưu Pareto đầy đủ thường không khả thi. Do đó, ta tìm tập Pareto được biết tốt nhất: (1) Là tập con của tập tối ưu Pareto; (2) Các phương án phân bố đều và đa dạng trên đường biên Pareto; (3) Cấc phương án biểu thị được toàn cảnh của đường biên Pareto.

2.3.3. Các giải thuật tiến hóa đa mục tiêu

Các giải thuật tiến hóa đa mục tiêu (MOEA) như NSGA hay SPEA là công cụ phù hợp để tìm tập Patero được biết tốt nhất. Điểm khác biệt chủ yếu giữa các MOEA nằm ở cách tính độ thích nghi, cách duy trì quần thể ưu tú và phương pháp để đa dạng hóa quần thể. Xếp hạng Pareto là một phương pháp thường dùng để tính độ thích nghi của cá thể. Có hai chiến lược thường dùng để hiện thực việc duy trì quần thể ưu tú: (i) lưu trữ các cá thể ưu tú trong chính quần thể và (ii) lưu trữ các cá thể ưu tú trong một danh sách thứ cấp bên ngoài quần thể và đưa chúng trở lại quần thể. Phương pháp chia sẻ độ thích nghi được dùng để đa dạng hóa quần thể.

2.3.4. Ứng dụng SPEA-II để giải quyết bài toán

Trong phương pháp đề xuất, giải thuật tiến hóa đa mục tiêu SPEA-II được lựa chọn để giải bài toán. SPEA-II có các điểm chính:

Biểu diễn nhiễm sắc thể: dùng phương pháp mã hóa số thực (real-coded method) [63].

Tính toán giá trị hàm mục tiêu: các giá trị *recall* và FAR được tính toán bằng cách trực tiếp áp dụng tập luật SpamAssassin trên dữ liệu huấn luyện. Giá trị 1-*recall* được dùng thay cho *recall* để hai tiêu chí có cùng mục tiêu là tối thiểu hóa.

Cơ chế chọn lọc: cơ chế chọn lọc dựa trên đấu loại trực tiếp (binary tournament selection) bằng cách so sánh giá trị hàm mục tiêu của hai cá thể ngẫu nhiên.

Phép toán lai tạo: phép toán lai tạo giả nhị phân (simulated binary crossover).

Phép toán đột biến: phép đột biến đa thức (polynomial mutation operator).

Gán độ thích nghi: phương pháp xếp hạng Pareto.

Duy trì quần thể ưu tú: phương án sử dụng một danh sách thứ cấp.

Đa sạng hóa quần thể: giữa hai cá thể có cùng thứ tự xếp hạng, cá thể có khoảng cách mật độ lớn hơn sẽ được chọn làm cha mẹ của thế hệ tiếp theo.

2.4. THỰC NGHIỆM

2.4.1. Thí nghiệm ứng dụng mạng nơ-ron để sinh tập luật SpamAssassin

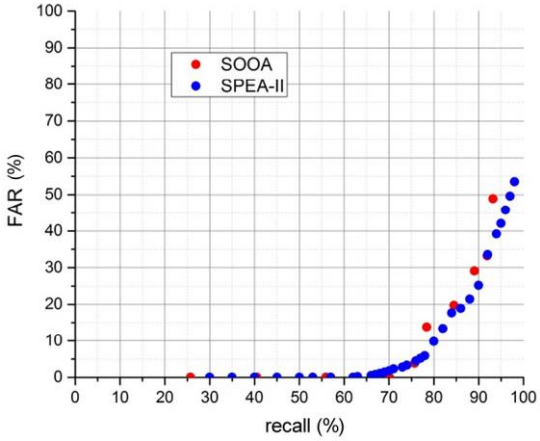
Bảng 2.1: Kết quả thí nghiệm so sánh một số
phương pháp sinh tập luật SpamAssassin dành cho tiếng Việt

Phương pháp	Precision			F_1		
	250 luật	500 luật	750 luật	250 luật	500 luật	750 luật
Cơ sở	0.8156	0.8716	0.8581	0.8578	0.8616	0.8807
Phương pháp [62]	0.9206	0.9372	0.9224	0.9235	0.9498	0.9517
Phương pháp đề xuất SD_1	0.9516	0.9621	0.9633	0.9699	0.9535	0.9693

Thí nghiệm so sánh 3 phương pháp sinh luật SpamAssassin trên tập dữ liệu lọc thư rác tiếng Việt nói trên. Phương pháp cơ sở (baseline) là một phương án sinh tập luật SpamAssassin đơn giản bao gồm việc lựa chọn các từ khóa phổ biến nhất trong tập dữ liệu để xây dựng tập luật và huấn luyện tập luật bằng phương pháp SGD [17]. Phương pháp thứ hai là phương pháp [62], một phương pháp sinh tập luật SpamAssassin cho tiếng Việt trong đó tập đặc trưng và thuật toán huấn luyện được cải thiện. Phương pháp thứ ba là phương pháp đề xuất. Các phương pháp được đánh giá bởi 2 tiêu chí: precision và F_1. Phương pháp mới cho thấy hiệu quả của tập luật được duy trì tốt hơn với số lượng luật nhỏ. Trong tất cả các kịch bản, phương pháp mới cho thấy hiệu quả ngang bằng hoặc cao hơn so với các phương pháp còn lại.

2.4.2. Thí nghiệm về cân bằng SDR và FAR dựa trên tối ưu hóa đa mục tiêu

Thí nghiệm đánh giá hiệu quả của phương pháp tối ưu đa mục tiêu SPEA-II so với phương pháp sinh tập luật SpamAssassin dựa trên tối ưu đơn mục tiêu (SOOA) [16]. Thí nghiệm cho thấy SPEA-II cho các kết quả tốt hơn so với SOOA trên cả hai phương diện tối ưu hóa FAR hay SDR. Hơn nữa, phương pháp đề xuất cho phép lựa chọn giải pháp phù hợp với yêu cầu trong số một tập các phương án Pareto. Tương tự với các phương pháp sinh tập luật SpamAssassin trước đó, khi số lượng luật tăng cao hơn thì SPEA-II cũng tìm được những kết quả tốt hơn.



Hình 2.8: Kết quả kịch bản thí nghiệm 2 với bộ lọc 100 luật

Luận án trình bày nhiều kịch bản thí nghiệm, trong đó kịch bản thí nghiệm 2 có số lượng thư huấn luyện và thử nghiệm lần lượt là 500 và 250. Kết quả của kịch bản thí nghiệm này với bộ lọc 100 luật được thể hiện trong Bảng 2.8.

2.5. KẾT LUẬN CHƯƠNG 2

Chương 2 trình bày hai phương pháp lọc thư rác dựa trên xây dựng tập luật SpamAssassin. Phương pháp thứ nhất tập trung cải thiện chất lượng của tập đặc trưng bằng cách đồng bộ hóa khâu lựa chọn đặc trưng với khâu điều chỉnh trọng số luật. Phương pháp thứ hai tập trung cải thiện hiệu quả của khâu tối ưu hóa điểm số

cho tập luật SpamAssassin bằng phương pháp tối ưu hóa đa mục tiêu. Nội dung trình bày trong chương này là kết quả các công trình nghiên cứu số 2 và số 4 của tác giả.

CHƯƠNG 3: DỰ ĐOÁN HÀNH ĐỘNG NGƯỜI DÙNG THƯ ĐIỆN TỬ

3.1. MỞ ĐẦU

Vấn đề quá tải email không chỉ biểu hiện ở số lượng thư rác trong toàn bộ lưu lượng thư điện tử, mà còn ở số lượng thư hợp lệ mà một người dùng nhận được hằng ngày. Hệ thống dự đoán hành động người dùng gợi ý một hành động phù hợp để thực hiện đối với mỗi bức thư mà người dùng nhận được.

3.1.1. Những khó khăn, tồn tại

Hiệu quả dự đoán được công bố bởi các nghiên cứu về dự đoán hành động người dùng còn hạn chế. Tiêu chí đánh giá của các phương pháp trong hướng nghiên cứu này chưa thống nhất. Mỗi người dùng có cách lựa chọn riêng đối với hành động cần thực hiện trên một bức thư, cho nên bài toán có tính cá nhân hóa. Các nghiên cứu về dự đoán hành động người dùng gặp phải tình trạng khan hiếm dữ liệu.

3.1.2. Hướng tiếp cận giải quyết bài toán

Nghiên cứu [25] cho thấy nội dung thư và mối quan hệ giữa người gửi và người nhận có ảnh hưởng đến hành động của người dùng đối với bức thư. Nhận thấy nền tảng SpamAssassin có tốc độ xử lý nhanh, đáp ứng yêu cầu lọc thư rác trên thời gian thực, luận án đề xuất xây dựng hệ thống dự đoán hành động người dùng dựa trên nội dung thư dành cho nền tảng này. Đồng thời, phương án giải quyết vấn đề tồn tại về dữ liệu nghiên cứu cũng như về tính cá nhân hóa của bài toán cũng được trình bày.

3.2. DỰ ĐOÁN HÀNH ĐỘNG NGƯỜI DÙNG VỚI TẬP LUẬT SPAMASSASSIN

Phương pháp đề xuất ở phần này được đặt tên là **UAP$_1$**. Một tập luật SpamAssassin hoạt động tương tự một máy phân loại nhị phân. Một số kỹ thuật phổ biến để kết hợp nhiều máy phân loại nhị phân trở thành máy phân loại đa lớp đó là OVA, OVO và DAG.

3.2.1. Xây dựng máy phân loại nhị phân

Tập luật SpamAssassin được dùng trong phương pháp này được xây dựng dựa trên phương pháp được đề xuất trong [28]. Tập dữ liệu huấn luyện được chia thành hai phần: tập D1 bao gồm thư rác và tập D2 bao gồm thư hợp lệ. Phương pháp [34] được áp dụng để tách từ ngữ từ tiêu đề và nội dung thư. Những từ ngữ có tần số quá thấp được loại bỏ. Trong số các từ khóa được giữ lại, một tập luật được lựa chọn ra dựa theo công thức conditional probability (1.4) trong lý thuyết xác suất Bayes. Tập luật này được huấn luyện với tập dữ liệu để tìm ra tập điểm số tối ưu bằng thuật toán huấn luyện tập luật SpamAssassin dựa trên SGD [17].

3.2.2. Xây dựng máy phân loại đa lớp

3.2.2.1. OVA (One vs. All)

N máy phân loại nhị phân C_i (i = 1, 2…N) được xây dựng, mỗi máy phân loại C_i có khả năng phân biệt dữ liệu thuộc về một lớp X_i với dữ liệu thuộc về (N – 1) lớp còn lại.

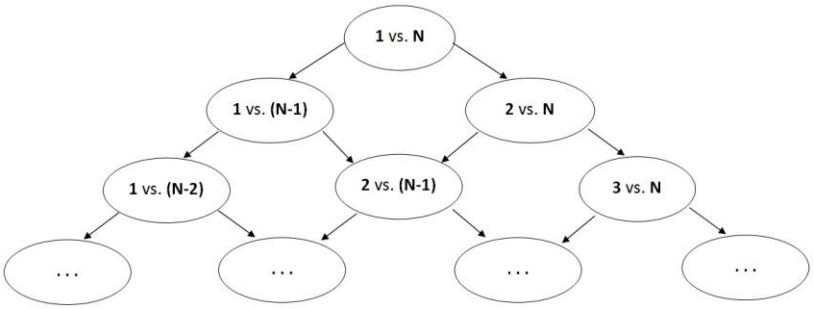
3.2.2.2. OVO (One vs. One)

Trong phương án OVO, ta cần xây dựng một máy phân loại $C_{i,j}$ để phân loại dữ liệu thuộc về hai lớp khác nhau bất kỳ X_i và X_j ($i \neq j$). Một bức thư đầu vào được xử lý bởi tất cả các máy phân loại, sau đó các kết quả phân loại được tổng hợp để đưa ra dự đoán cuối cùng. Có nhiều phương án tổng hợp kết quả dành cho OVO: max sum (lấy kết quả có tổng điểm số bình chọn cao nhất), majority voting (lấy kết quả được đa số máy phân loại lựa chọn) và most confident (lấy kết luận của máy phân loại tự tin nhất). Với số lượng lớp là N, số lượng máy phân loại nhị phân cần được xây dựng là $N_R = N \times (N - 1) \div 2$

3.2.2.3. DAG (Directed Acyclic Graph)

Giống như OVO, phương án DAG yêu cầu một máy phân loại nhị phân cho mỗi cặp hai lớp X_i và X_j khác nhau. Như vậy, số lượng tập luật cần xây dựng cho phương án DAG tương tự là $N_R = N \times (N - 1) \div 2$. Tuy nhiên, DAG giảm thiểu số lần dự đoán của các máy phân loại nhị phân xuống còn (N – 1) lần dự đoán bằng

cách sử dụng một cây quyết định nhị phân (Hình 3.6). Các máy phân loại được sắp xếp theo thứ tự trong cây nhị phân và một trong hai máy phân loại (nhánh trái hoặc nhánh phải) sẽ được chọn làm máy phân loại kế tiếp phụ thuộc vào kết quả của mỗi lần dự đoán.



Hình 3.6: Mô hình dự đoán dựa trên cây nhị phân của phương án DAG.

3.3. ÁP DỤNG LUẬT HAM ĐỂ CẢI THIỆN TẬP LUẬT SPAMASSASSIN TRONG BÀI TOÁN DỰ ĐOÁN HÀNH ĐỘNG NGƯỜI DÙNG

Phương pháp đề xuất ở phần này được đặt tên là **UAP_2**. Phương pháp UAP_2 là phương pháp cải tiến dựa trên phương pháp UAP_1. Các phương án phân loại đa lớp OVA, OVO và DAG trong UAP_1 được kế thừa trong UAP_2. Trong phương pháp UAP_1, các đặc trưng trong dữ liệu thư hợp lệ chưa được sử dụng để xây dựng các tập luật SpamAssassin. Phương pháp UAP_2 để xuất phương án tự động gán nhãn hành động cho dữ liệu và bổ sung luật ham để tăng hiệu quả của mô hình dự đoán.

3.3.1. Tự động gán nhãn cho dữ liệu

Gán nhãn một cách thủ công cho tập dữ liệu huấn luyện là một công việc tốn nhiều thời gian. Nghiên cứu về dự đoán hành động người dùng thường gặp khó khăn khi tìm dữ liệu thí nghiệm. Luận án đề xuất kịch bản xây dựng mô hình dự đoán hành động người dùng có thể áp dụng trên thực tế để mang lại lợi ích cho cả người sử dụng và người làm nghiên cứu.

Một công cụ gán nhãn dữ liệu tự động và bộ công cụ xây dựng và huấn luyện mô hình được gửi đến người sử dụng email. Người dùng email thực hiện thí nghiệm trên dữ liệu cá nhân với bộ công cụ đó và gửi các kết quả đầu ra về cho chuyên gia. Chuyên gia dựa vào kết quả thử nghiệm và những thông tin được ghi lại trong quá trình xây dựng mô hình để điều chỉnh các tham số cho phù hợp, rồi gửi lại bộ cấu hình mới cho người dùng để họ lặp lại việc xây dựng mô hình. Người sử dụng nhận được mô hình dự đoán hành động đối với thư điện tử. Nhà nghiên cứu thu thập được kết quả thí nghiệm trên dữ liệu thực tế. Trong toàn bộ quy trình nói trên, dữ liệu email cá nhân của người dùng được bảo mật.

Để gán nhãn tự động, ta cần phân tích các trường header của bức thư. Những bức thư có nhãn "unread" được bỏ qua không gán nhãn. Phân tích những bức thư trong thư mục thư đã gửi, trích xuất trường "In-Reply-To" để tìm ra những bức thư đã được trả lời và gãn nhãn "trả lời" cho chúng. Những bức thư trong thư mục thư đã xóa được gán nhãn "xóa". Những bức thư còn lại được gán nhãn "đọc".

3.3.2. Sinh tập luật SpamAssassin với luật Ham

Luật ham giúp cho điểm số dự đoán của một tập luật SpamAssassin thể hiện chính xác hơn mức độ tự tin trong kết quả dự đoán. Điểm này có ý nghĩa quan trọng trong các phương án kết hợp nhiều máy phân loại nhị phân như OVA, OVA, DAG.

Mỗi tập luật đơn lẻ trong phương pháp UAP_2 được xây dựng theo quy trình tượng tự như trong phương pháp UAP_1. Sau đây là cải tiến trong phương pháp đề xuất. Việc lựa chọn các từ ngữ để xây dựng tập luật được thực hiện hai lần để chọn ra các đặc trưng tốt nhất trong các bức thư rác (luật spam) và các bức thư hợp lệ (luật ham). Số lượng luật ham và luật spam có thể được điều chỉnh bởi người sử dụng. Tham khảo thí nghiệm

của nghiên cứu [17], trong đó hiệu quả của tập luật cao nhất khi luật spam và luật ham có tỷ lệ cân bằng 1:1, thí nghiệm phương pháp UAP_2 trong luận án sử dụng 500 luật spam và 500 luật ham. Tập luật được biểu diễn dưới dạng một mô hình mạng nơ-ron một lớp (perceptron) trong đó các trọng số của mạng nơ-ron đại diện cho các điểm số của tập luật. Thuật toán huấn luyện SGD được sử dụng để gán điểm số cho tập luật được lựa chọn.

3.4. ỨNG DỤNG PHƯƠNG PHÁP SD_1 TRONG MÔ HÌNH DỰ ĐOÁN HÀNH ĐỘNG NGƯỜI DÙNG

Phương pháp đề xuất ở phần này được đặt tên là **UAP_3**. Phương pháp UAP_3 là phương pháp cải tiến dựa trên phương pháp UAP_1 và UAP_2. Các phương án phân loại đa lớp OVA, OVO và DAG được kế thừa từ UAP_1. Tuy nhiên, tập luật SpamAssassin trong UAP_3 được xây dựng bằng phương pháp SD_1 đã đề xuất ở Chương 2.

3.4.1. Cải tiến máy phân loại nhị phân trong mô hình phân loại đa lớp

Hướng tiếp cận của phương pháp UAP_1 và UAP_2 là kết hợp máy phân loại nhị phân để giải quyết bài toán dự đoán hành động người dùng. Trong phương án UAP_3, các máy phân loại nhị phân trong phương pháp UAP_1 và UAP_2 được thay thế bằng các tập luật được sinh ra bằng phương pháp SD_1.

3.4.2. Cải thiện trong khâu tiền xử lý dữ liệu

Hiện tượng văn bản tiếng Việt sử dụng các bảng mã tiếng Việt khác nhau, sử dụng cách bỏ dấu khác nhau, gây mất tính nhất quán trong dữ liệu. Trong phương pháp UAP_3, các bảng mã và cách bỏ dấu khác nhau của văn bản tiếng Việt được chuyển về cùng một dạng thống nhất. Tuy vẫn áp dụng phương pháp tách từ [34] giống như các phương pháp UAP_1 và UAP_2, trong phương pháp UAP_3, chỉ những từ ngữ có nghĩa và các dấu chấm câu được giữ lại. Những đơn vị văn bản không có nghĩa hoặc có tính chất cụ thể như số điện thoại, địa chỉ email, đường đẫn trang web và các giá trị số cũng được bỏ qua không sử dụng làm đặc trưng.

3.4.3. Sinh tập luật SpamAssassin dựa trên mạng nơ-ron

Mô hình mạng nơ-ron được áp dụng để xây dựng các máy phân loại nhị phân trong phương pháp này được lấy từ phương pháp SD_1. Mạng nơ-ron bao gồm hai thành phần chính: lớp mạng *FS* (có tác dụng lựa chọn đặc trưng) và lớp mạng *P* (có tác dụng phân loại bức thư đầu vào với các đặc trưng được chọn). Lớp mạng *P* mô phỏng cơ chế phát hiện thư rác sử dụng luật có trọng số của SpamAssassin.

Mô hình được huấn luyện bằng thuật toán SGD với mini-batch, là thuật toán cân bằng các ưu điểm và hạn chế của hai thuật toán SGD và GD. Phương pháp UAP_3 được đưa vào thí nghiệm so sánh trong phần sau để đánh giá hiệu quả khi thay thế các máy phân loại nhị phân cho bài toán dự đoán hành động người dùng.

3.5. THỰC NGHIỆM

3.5.1. Tiêu chí đánh giá

Ba tiêu chí đánh giá là *accuracy*, *precision vĩ mô* P_m (1.23) và FPR_{del} (3.1) được sử dụng để đánh giá hiệu quả của ba phương pháp trong thí nghiệm. Tiêu chí FPR_{del} có ý nghĩa vì nó thể hiện tỷ lệ thư quan trọng bị dự đoán nhầm là thư cần xóa, là lỗi mang lại thiệt hại lớn nhất cho người dùng. Tiêu chí precision vĩ mô (P_m) cho phép hai hành động với số lượng thư nhỏ là *xóa* và *trả lời* có sức ảnh hưởng ngang bằng đến kết quả đánh giá so với hành động *đọc*.

3.5.2. Thí nghiệm

Bảng 3.1 tổng hợp kết quả của thí nghiệm so sánh ba phương pháp dự đoán hành động người dùng. Các phương án phân loại đa lớp được sử dụng là OVA, OVO-MS, OVO-MV, OVO-MC và DAG. Thí nghiệm so sánh cho thấy phương pháp UAP_2 giúp giảm tỷ lệ gợi ý nhầm đối với hành động xóa thư trong khi phương pháp UAP_3 giúp tăng độ chính xác chung của các gợi ý so với phương pháp UAP_1. Phương pháp UAP_2 có tiêu chí *accuracy* không cao hơn rõ rệt so với phương pháp UAP_1 nhưng lại tiêu chí FPR_{del} thấp hơn đáng kể so với phương pháp UAP_1 nhờ vào việc bổ sung các luật ham. Tuy phương pháp UAP_3 có tỷ lệ dự đoán cao hơn, tỷ lệ FPR_{del} cũng cao hơn đáng kể cho với phương pháp UAP_2.

Bảng 3.1: Kết quả thí nghiệm so sánh các phương pháp UAP_1, UAP_2 và UAP_3 theo ba tiêu chí *accuracy*, P_m và FPR_{del} (%).

	Phương pháp	**UAP_1**			**UAP_2**			**UAP_3**		
	Tiêu chí	*Acc.*	P_m	FPR_{del}	*Acc.*	P_m	FPR_{del}	*Acc.*	P_m	FPR_{del}
Phương án	OVA	0.822	0.766	6.245	0.812	0.760	**0.855**	0.854	0.817	1.585
	OVO-MS	0.795	0.740	3.128	0.805	0.791	1.658	**0.865**	**0.839**	3.045
	OVO-MV	0.769	0.737	2.148	0.767	0.734	1.001	0.821	0.789	2.169
	OVO-MC	0.755	0.735	1.731	0.819	0.793	1.877	0.814	0.796	3.107
	DAG	0.804	0.806	1.345	0.761	0.713	0.675	0.802	0.839	2.940

3.6. KẾT LUẬN CHƯƠNG 3

Chương 3 đã đề xuất dự đoán hành động bằng phương pháp phân loại, áp dụng trên nền tảng lọc thư rác SpamAssassin. Tập dữ liệu lọc thư rác tiếng Việt đã được sử dụng trong các thí nghiệm. Nội dung được trình bày trong chương này là tổng hợp kết quả từ công trình nghiên cứu số 1, số 2 và số 5 của tác giả. Ba phương pháp dự đoán hành động người dùng là UAP_1, UAP_2 và UAP_3 đã được đề xuất, trong đó UAP_2 và UAP_3 có những ưu điểm khác nhau so với UAP_1. Các phương pháp đề xuất có thể được áp dụng nhanh chóng trên các hệ thống máy chủ thư điện tử đang cài đặt bộ lọc thư rác SpamAssassin.

CHƯƠNG 4: XẾP HẠNG THƯ ĐIỆN TỬ

4.1. MỞ ĐẦU

Vấn đề quá tải email gây nhiều tác hại cho người dùng. Hệ thống xếp hạng email là một hướng giải quyết vấn đề quá tải email, giúp người dùng tiết kiệm thời gian xử lý thư, đem lại lợi ích thiết thực. Mục đích của hệ thống xếp hạng thư điện tử là dự đoán tầm quan trọng của các bức thư trong hòm thư của người dùng và sắp xếp lại những bức thư chưa được xử lý theo thứ tự giảm dần mức độ quan trọng. Trong chương này, phương pháp xếp hạng thư điện tử dựa trên phân loại được trình bày trong chương này. Trong phương pháp đề xuất, mô hình phân loại có đầu ra gồm 5 mức độ quan trọng là *xóa*, *đọc không quan trọng*, đ*ọc quan trọng*, *trả lời không gấp* và *trả lời gấp*.

4.1.1. Những khó khăn và tồn tại

Khó khăn thứ nhất nằm ở định nghĩa về tầm quan trọng của bức thư. Khó khăn thứ hai là độ khó phân loại đến từ số lượng lớp phân loại cao hơn so với những dạng khác của bài toán xác định thứ tự ưu tiên của thư điện tử. Khó khăn thứ ba, cũng là vấn đề chung với bài toán dự đoán hành động người dùng, là sự khan hiếm dữ liệu email cá nhân dành cho nghiên cứu.

4.1.2. Hướng tiếp cận của bài toán

Với bài toán này, phương pháp phân loại có hiệu quả hơn so với phương pháp hồi quy [49]. Trong luận án, bài toán xếp hạng email được giải quyết theo hướng phân loại với 5 mức độ ưu tiên. Ngoài ra, các phương pháp học sâu đã thể hiện hiệu quả vượt trội trong các bài toán xử lý văn bản và lọc thư rác, nhưng chưa được áp dụng cho bài toán xếp hạng email. Luận án tập trung nghiên cứu phương pháp phân loại dựa trên học sâu để giải quyết bài toán. Nhiều nghiên cứu về email đã cho thấy hiệu quả của những đặc trưng mạng xã hội đối với phân loại email. Trong luận án, các đặc trưng xã hội của email được đề xuất và kết hợp cùng đặc trưng nội dung nhằm đem lại hiệu quả cao nhất cho mô hình phân loại email.

4.2. XẾP HẠNG THƯ ĐIỆN TỬ BẰNG PHƯƠNG PHÁP HỌC SÂU

Phần này trình bày tóm tắt phương pháp xếp hạng email EP_2, Đặc trưng được sử dụng là sự kết hợp giữa đặc trưng văn bản biểu diễn dưới dạng vector từ ngữ *word2vec* và đặc trưng mạng xã hội được trích xuất từ dữ liệu. Một cấu trúc mạng nơ-ron sâu được trình bày để đáp ứng đầu vào kết hợp hai loại đặc trưng nói trên, trong đó sử dụng các đơn vị LSTM kết hợp với một số kỹ thuật thiết kế và huấn luyện của mạng nơ-ron.

4.2.1. Phương pháp học sâu trong xử lý thư điện tử

Mô hình auto-encoder xếp chồng [65] và mạng nơ-ron đa thể [75] đã được áp dụng thành công trong bài toán lọc thư rác. Mô hình MLP đã được thử nghiệm [80] với tập dữ liệu thư rác SpamBase và cho hiệu quả cao hơn phương pháp phân loại Bayes. Mạng LSTM cũng đã được áp dụng cho bài toán lọc thư rác [84], trong đó nội dung thư được biểu diễn bằng phương pháp *word2vec*.

4.2.2. Tiền xử lý dữ liệu

Để đáp ứng yêu cầu biểu diễn nội dung bằng phương pháp word2vec, nội dung thư được xử lý tách ra theo câu và theo từ. Phương pháp xử lý văn bản tiếng Việt VNCoreNLP [78] cùng với một số bước tiền xử lý văn bản đã được áp dụng để thực hiện tách từ và tách câu.

4.2.3. Biểu diễn đặc trưng mạng xã hội

Dựa trên giả thiết rằng một người sẽ tiếp tục gửi những bức thư có tầm quan trọng giống với những bức thư mà người đó đã gửi trong quá khứ, mỗi người gửi được biểu diễn bởi một tập hợp 5 số nguyên. Một kẻ phát tán thư rác thường nhận rất ít hoặc hoàn toàn không nhận thư, trong khi một người dùng email quan trọng thường nhận rất nhiều thư gửi về. Dựa trên giả thiết này, số lượng thư mà một người đã nhận cũng được đếm

để đưa vào vector đặc trưng xã hội của người đó. Vector đặc trưng xã hội gồm 6 phần tử được chuẩn hóa thành các giá trị số thực trong khoảng [0, 1).

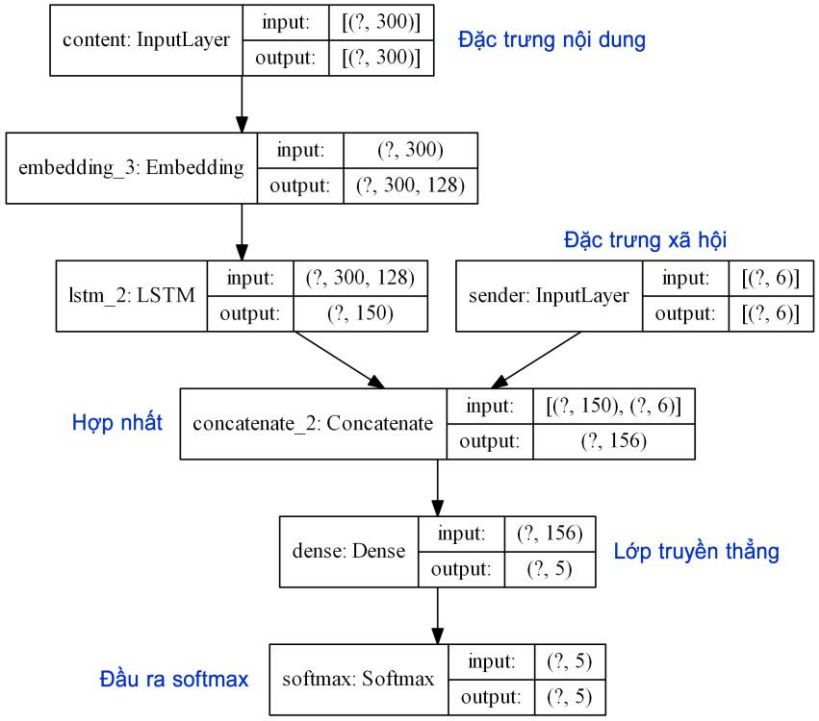
4.2.4. Biểu diễn đặc trưng nội dung

Có hai cách để sử dụng word embedding: word embedding huấn luyện sẵn (pre-trained) và word embedding trực tuyến (online). Word embedding huấn luyện sẵn là việc sử dụng một tập dữ liệu văn bản để tạo ra một bộ vector từ ngữ của tất cả các từ xuất hiện trong tập dữ liệu đó, dùng phương pháp học máy không giám sát. Sau đó, bộ vector từ ngữ đó được dùng làm trọng số của lớp Embedding trong mạng nơ-ron, các trọng số của lớp này không đổi khi huấn luyện mạng. Word embedding trực tuyến là việc khởi tạo trọng số ngẫu nhiên cho lớp Embedding và huấn luyện lớp này cùng với mạng nơ-ron. Luận án sử dụng cả hai phương án nói trên trong thí nghiệm. Với phương án word embedding huấn luyện sẵn, word2vec rong bộ công cụ Gensim được sử dụng để tạo bộ vector từ ngữ từ dữ liệu huấn luyện.

Kích thước của vector từ ngữ thường được lựa chọn là 300 trong nhiều nghiên cứu. Luận án khảo sát hiệu quả khi sử dụng vector từ ngữ có kích thước nhỏ hơn 300 dựa trên giả thiết rằng một tập dữ liệu nhỏ sẽ cần các vector từ ngữ nhỏ hơn. Các thí nghiệm sẽ so sánh các kích thước vector từ ngữ lần lượt là 128 và 300 để xác minh giả thiết nói trên.

4.2.5. Cấu trúc mạng nơ-ron

Cấu trúc mạng nơ-ron trong Hình 4.1 được đề xuất để kết hợp đặc trưng người gửi một cách hiệu quả với các đặc trưng nội dung.



Hình 4.1: Cấu trúc mạng nơ-ron đề xuất dành cho đầu vào kết hợp đặc trưng nội dung và đặc trưng xã hội

Đặc trưng nội dung được xử lý bởi cấu trúc mạng nơ-ron hồi quy LSTM. Đầu ra của lớp LSTM có dạng một vector số thực, được hợp nhất với các đặc trưng xã hội của người gửi để tạo thành vector đặc trưng của bức thư. Vector đặc trưng hợp nhất được dùng làm đầu vào cho cấu trúc mạng truyền thẳng ở cuối mạng nơ-ron, và kết quả dự đoán được thể hiện bởi lớp đầu ra Softmax.

4.2.6. Huấn luyện mạng nơ-ron

Thuật toán tối ưu đóng vai trò quan trọng cho sự thành công của việc huấn luyện mạng nơ-ron. Các thuật toán Adagrad [47], RMSProp [53], Adam 73] đều dựa trên phương pháp lan truyền ngược sai số và cơ chế leo

dốc của thuật toán cổ điển GD. Tuy nhiên, mỗi thuật toán đều có những cơ chế riêng để tương thích với các cấu trúc mạng và loại dữ liệu khác nhau.

Cross-entropy (hay còn gọi là *log loss* hoặc *logistic loss*) và *trung bình của bình phương lỗi* (MSE) là những hàm tổn thất thường dùng cho mô hình mạng nơ-ron cho bài toán phân loại. *Cross-entropy* được dùng để đánh giá mức độ tự tin của kết quả dự đoán đưa ra bởi một mô hình. Hàm MSE thể hiện khoảng cách giữa các dự đoán của mô hình và kết quả đúng.

Thuật ngữ *over-fitting* nói về hiện tượng mô hình học máy mô phỏng chính xác tập dữ liệu huấn luyện nhưng không mô phỏng chính xác tập dữ liệu cần dự đoán trên thực tế. Ngoài kỹ thuật dừng huấn luyện sớm (*early stopping*) thì *dropout* [58] cũng là một cách hiệu quả để tránh hiện tượng *over-fitting*.

4.3. XẾP HẠNG THƯ ĐIỆN TỬ DỰA TRÊN SPAMASSASSIN

Phần này trình bày tóm tắt phương pháp xếp hạng email EP_1, được xây dựng dựa trên phương pháp dự đoán hành động người dùng UAP_3 đã trình bày ở trên.

4.3.1. Xây dựng máy phân loại nhị phân

Nội dung thư được biểu diễn dưới dạng vector nhị phân *x* với đặc trưng được trích xuất từ nội dung thư. Mô hình mạng nơ-ron với hai lớp ẩn trong Hình 2.3 được huấn luyện dựa trên một tập dữ liệu. Khi áp dụng mô hình này với bài toán phát hiện thư rác, tập dữ liệu sẽ bao gồm thư rác và thư hợp lệ. Dữ liệu huấn luyện mô hình phân loại nhị phân luôn bao gồm hai lớp, thành phần của hai lớp dữ liệu tùy thuộc vào các phương án OVA, OVO, DAG. Ví dụ, trong phương án OVA, máy phân loại nhị phân dành cho mức độ *xóa* sẽ được huấn luyện như một bộ lọc thư rác trong đó thư *xóa* được coi là thư rác và 4 mức độ còn lại được coi là thư hợp lệ.

4.3.2. Các phương án phân loại đa lớp

Có ba cách phổ biến để xây dựng máy phân loại đa lớp từ nhiều máy phân loại nhị phân là OVA, OVO và DAG. Trong đó, phương án OVO có ba biến thể là OVO-MS, OVO-MV và OVO-MC. Với bài toán phân loại 5 lớp, phương án OVA cần xây dựng là 5 máy phân loại nhị phân. Số lượng cần xây dựng cho phương án OVO và DAG là 10 máy phân loại bởi vì ta cần một máy phân loại cho mỗi cặp hai mức độ ưu tiên khác nhau. Ở phương án OVA, mỗi máy phân loại M_i được gắn với lớp C_i. Khi dự đoán của máy phân loại M_i (một giá trị số thực) đối với bức thư *m* là cao nhất, ta khẳng định bức thư m thuộc về lớp C_i. Ở phương án OVO, các máy phân loại được ký hiệu là $M_{i,j}$ để thể hiện đó là máy phân loại để phân biệt một bức thư thuộc về lớp C_i hay C_j. Kết quả của tất cả 10 máy phân loại được tổng hợp lại theo ba cách khác nhau là MS, MV và MC. Phương án DAG tìm ra kết quả dự đoán theo một cây quyết định nhị phân (Hình 3.6).

4.4. THỰC NGHIỆM

Các thí nghiệm dưới đây sử dụng tập dữ liệu thư điện tử tiếng Việt được trình bày trong Chương 1 và được thống kê trong Bảng 1.2.

4.4.1. Tiêu chí đánh giá

Ba tiêu chí đánh giá là *accuracy*, *cross-entropy* và *điểm số F_1 vĩ mô* (macro F_1 score) [41] được lựa chọn sử dụng để đánh giá các mô hình xếp hạng email theo hướng phân loại trong các thí nghiệm của luận án. Những tiêu chí này phù hợp với bài toán phân loại đa lớp không đặt giả thiết về quan hệ tương đối giữa các lớp. Giá trị *cross-entropy* càng thấp càng thể hiện sự tự tin của mô hình với kết quả dự đoán.

4.4.2. So sánh các thuật toán tối ưu mạng nơ-ron (thí nghiệm 1)

Mục tiêu của thí nghiệm này là so sánh những thuật toán huấn luyện mạng nơ-ron. Thí nghiệm được thực hiện đối với ba thuật toán phổ biến: Adam, RMSProp và Adagrad. Các thí nghiệm riêng biệt được thực hiện với word embedding huấn luyện sẵn và word embedding trực tuyến, với kích thước vector cùng là 128. Kết quả của thí nghiệm 1 được trình bày trong Bảng 4.1.

Bảng 4.1: Kết quả so sánh ba thuật toán huấn luyện mạng nơ-ron

Thuật toán huấn luyện	**Accuracy**		**Macro F_1**		**CCE**	
	(a)	(b)	(a)	(b)	(a)	(b)
Adam	0.6641	0.9115	0.3769	**0.8641**	6.6992	**0.6650**
Adagrad	0.5209	0.6448	0.1374	0.5090	**1.7875**	1.0729
RMSProp	**0.7134**	**0.9126**	**0.5014**	0.8632	5.9510	0.7260
(a) Word embedding trực tuyến, m = 128 (b) Word embedding huấn luyện sẵn (word2vec), m = 128						

4.4.3. So sánh các phương án word embedding (thí nghiệm 2)

Trong thí nghiệm này, hai kích thước vector từ ngữ là 128 và 300 được so sánh đối với cả phương án word embedding huấn luyện sẵn và word embedding trực tuyến. Thuật toán RMSProp được sử dụng cho tất cả các mô hình trong thí nghiệm. Bảng 4.3 thể hiện kết quả của thí nghiệm 2.

Bảng 4.2: Kết quả thí nghiệm so sánh các cấu hình word embedding khác nhau.

Cấu hình embedding	**Accuracy**	**Macro F1**	**CCE**
Word2vec, m = 128	0.9126	0.8632	0.7260
Word2vec, m = 300	0.9185	0.8764	0.7146
Trực tuyến, m = 128	0.7134	0.5014	5.9510
Trực tuyến, m = 300	0.7900	0.5918	4.2800

4.4.4. So sánh một số phương pháp xếp hạng thư điện tử (thí nghiệm 3)

Thí nghiệm 3 có mục tiêu so sánh hiệu quả của các phương pháp xếp hạng email:

1) Phương pháp EP_1 được mô tả trong phần 4.3

2) Phương pháp xếp hạng thư điện tử dựa trên học sâu EP2 được trình bày trong phần 4.2

3) Phương pháp xếp hạng email dựa trên máy phân loại SVM được giới thiệu trong [49], tạm đặt tên là phương pháp YooEP.

Cả 3 phương pháp được thực hiện trên cùng tập dữ liệu xếp hạng email tiếng Việt được mô tả ở trên. Bảng 4.3 tổng hợp kết quả của thí nghiệm 3.

Bảng 4.3: So sánh phương pháp EP_2 với phương pháp EP_1 và YooEP [49]

Phương pháp	**Accuracy**	**Macro F_1**	**CCE**
OVA-EP_1	0.8219	0.7757	1.0036
EP_2*, word2vec, m = 300	**0.9185**	**0.8764**	0.7146
YooEP-OVA, epoch=50	0.7137	0.4529	0.7893
YooEP-OVA, epoch=100	0.7847	0.5550	0.6161
YooEP-OVA, epoch=150	0.8225	0.6360	**0.5207**
* EP_2 dùng thuật toán huấn luyện RMSProp, số lượng epoch = 15			

4.5. KẾT LUẬN CHƯƠNG 4

Chương 4 đã trình bày các đề xuất để giải quyết bài toán xếp hạng email, một hướng nghiên cứu mới và có ý nghĩa to lớn đối với người dùng email. Nội dung trình bày trong chương này được tổng hợp từ kết quả đã được công bố trong các công trình nghiên cứu số 1, số 2 và số 3 của tác giả. Phương pháp xếp hạng email dựa trên phân loại đa lớp EP_1 là sự kết hợp giữa mô hình dự đoán hành động người dùng trình bày trong nghiên cứu số 1 và phương pháp sinh tập luật SpamAssassin dựa trên mạng nơ-ron từ nghiên cứu số 2. Phương pháp xếp hạng email dựa trên mô hình học sâu EP_2 đã được công bố một phần trong nghiên cứu số 3.

KẾT LUẬN

Xác định thứ tự ưu tiên của thư điện tử là một hướng giải quyết tình trạng quá tải thư điện tử, một vấn đề đang ngày càng trở nên cấp thiết. Luận án tập trung nghiên cứu phương pháp xác định thứ tự ưu tiên của thư điện tử theo 03 hướng tiếp cận chính là lọc thư rác, dự đoán hành động người dùng và xếp hạng thư điện tử. Luận án thể hiện 03 đóng góp chính, một là đề xuất phương pháp tự động sinh tập luật mới cho SpamAssassin, trong đó bước lựa chọn luật và bước xác định trọng số của luật được tiến hành đồng thời. Hai là đề xuất phương pháp sử dụng nền tảng SpamAssassin kết hợp với các mô hình phân loại đa lớp để gợi ý hành động người dùng. Ba là đề xuất phương pháp học sâu để xếp hạng thư điện tử theo 5 mức độ ưu tiên khác nhau, sử dụng word embedding để biểu diễn nội dung thư kết hợp với đặc trưng mạng xã hội. Ngoài ra, để thực hiện thí nghiệm cho các đề xuất trong 03 hướng tiếp cận nói trên, luận án đã thu thập và xây dựng tập dữ liệu thư điện tửu tiếng Việt.

Đóng góp thứ nhất cho bài toán lọc thư rác của luận án là đề xuất phương pháp xây dựng tập luật SpamAssassin dựa trên mạng nơ-ron. Phương pháp có hiệu quả dự đoán cải thiện hơn so với những phương pháp cũ. Thông qua tổng quan tài liệu, luận án nhận thấy các phương pháp xây dựng tập luật SpamAssassin dựa trên học máy đều thực hiện tách rời hai khâu lựa chọn đặc trưng và huấn luyện trọng số. Cách làm này dẫn đến một hạn chế đó là chưa kiểm chứng được hiệu quả của tập đặc trưng được chọn trên dữ liệu bởi vì chỉ có một tập đặc trưng duy nhất được lựa chọn và không được so sánh với các tập đặc trưng tiềm năng khác. Mô hình mạng nơ-ron được đề xuất trong đóng góp thứ nhất có mục tiêu giải quyết vấn đề nói trên. Mô hình gồm hai lớp ẩn, một lớp có chức năng lựa chọn đặc trưng và lớp còn lại có chức năng điều chỉnh trọng số của đặc trưng, từ đó hợp nhất hai khâu lựa chọn luật và gán điểm số vốn tách rời trong các phương pháp sinh tập luật SpamAssassin trước đó, giúp nâng cao chất lượng của tập luật được xây dựng.

Đóng góp thứ hai cho bài toán lọc thư rác của luận án là một phương pháp khác để sinh tập luật lọc thư rác cho SpamAssassin, hướng tới mở rộng tác vụ sinh tập luật từ bài toán tối ưu đơn mục tiêu thành bài toán tối ưu đa mục tiêu, chú trọng cải thiện khâu gán điểm số cho tập luật. Phương pháp này giải quyết vấn đề quan trọng của bài toán sinh tập luật SpamAssassin đó là sự cân bằng giữa hai tiêu chí đối nghịch recall và FAR.

Với bài toán dự đoán hành động người dùng, luận án đã đề xuất phương pháp giải quyết với mô hình phân loại đa lớp trên nền tảng SpamAssassin. Kết quả từ các nghiên cứu đã công bố số 1, số 2 và số 5 đã được tổng hợp để đề xuất và cải tiến phương pháp dự đoán hành động. Luận án đã ứng dụng cách kỹ thuật khác nhau để kết hợp nhiều tập luật SpamAssassin thành máy phân loại đa lớp có tác dụng dự đoán hành động cho người dùng thư điện tử. Phương pháp này có tính ứng dụng cao trên thực tế bởi vì sự phổ biến của hệ thống SpamAssassin và tốc độ xử lý nhanh của cơ chế luật có trọng số. Luận án cũng trình bày hai phương án nhằm cải thiện hiệu quả của mô hình dự đoán hành động nói trên dựa trên cải thiện hiệu quả của các máy phân loại nhị phân thành phần, từ đó nâng cao hiệu quả của máy phân loại đa lớp. Cách thứ nhất là ứng dụng thêm luật ham cho tập luật SpamAssassin. Cách thứ hai là ứng dụng phương pháp sinh tập luật SpamAssassin dựa trên mạng nơ-ron từ nghiên cứu đã công bố số 2. Thí nghiệm so sánh cho thấy phương án thứ nhất giúp giảm tỷ lệ gợi ý nhầm đối với hành động xóa thư trong khi phương án thứ hai giúp tăng độ chính xác chung của các gợi ý.

Về bài toán xếp hạng thư điện tử, đóng góp của luận án là một mô hình phân loại dựa trên học sâu để giải quyết bài toán xếp hạng thư điện tử, là kết quả đã được công bố trong công trình nghiên cứu đã công bố số 3. Mô hình được đề xuất không chỉ tích hợp các kỹ thuật học sâu, trong đó nổi bật là cấu trúc mạng LSTM, mà còn sử dụng bộ đặc trưng nội dung kết hợp với đặc trưng xã hội. Thuật toán word2vec đã được sử dụng để biểu diễn thông tin ngữ nghĩa trong nội dung thư điện tử. Các chỉ số khác nhau liên quan đến người gửi thư đã

được trích xuất thành vector đặc trưng xã hội đại diện cho người gửi thư. Các thí nghiệm đã được thực hiện trên tập dữ liệu xếp hạng thư điện tử cá nhân do tác giả thu thập và xử lý. Phương pháp đề xuất đã thể hiện hiệu quả tốt hơn đáng kể so với phương pháp học máy truyền thống dựa trên máy phân loại SVM và bộ đặc trưng TF-IDF. Ngoài ra, so sánh đã được đưa ra giữa các cấu hình khác nhau của mô hình mạng nơ-ron, cụ thể là về kích thước vector từ ngữ và lựa chọn về thuật toán huấn luyện.

Trong khuôn khổ thời gian thực hiện nghiên cứu hạn chế, còn nhiều khía cạnh mà luận án chưa nghiên cứu một cách đầy đủ. Những vấn đề mà luận án chưa giải quyết được dưới đây sẽ là định hướng cho các nghiên cứu tiếp theo.

Các đề xuất trong luận án đã được thử nghiệm và so sánh với một số phương pháp khác nhưng số lượng phương pháp được thử nghiệm, so sánh còn hạn chế. Trong các nghiên cứu tiếp theo, những phương pháp đề xuất cần được thử nghiệm trên các tập dữ liệu khác. Đồng thời, cần thử nghiệm thêm nhiều phương pháp liên quan trên bộ dữ liệu mà luận án đã xây dựng. Những thí nghiệm nói trên có mục tiêu làm rõ hiệu quả của các đề xuất so với các phương pháp liên quan và tiếp tục kiểm chứng chất lượng của tập dữ liệu đã xây dựng.

Các đề xuất trong luận án chủ yếu sử dụng đặc trưng nội dung của thư điện tử và đặc trưng liên quan đến người gửi thư. Trong tương lai, nghiên cứu sẽ khai thác thêm các đặc trưng khác của thư điện tử như thời gian gửi/nhận thư, địa chỉ mạng của người gửi thư, và các thông tin từ những trường header khác. Ngoài ra, xác định thứ tự ưu tiên cho những bức thư có nội dung được mã hóa dưới dạng hình ảnh cũng là một trong những nội dung cần được khảo cứu trong các nghiên cứu tiếp theo.

Ngoài phương pháp xác định thứ tự ưu tiên của thư điện tử, còn những hướng khác để giải quyết tình trạng quá tải thư điện tử, ví dụ như tóm tắt nội dung thư hoặc trích xuất nội dung chính của thư điện tử. Đây cũng là những hướng nghiên cứu cần được xem xét trong tương lai.

Để cải thiện đóng góp cho bài toán xếp hạng thư điện tử của luận án, hướng nghiên cứu sau này sẽ áp dụng thêm những kỹ thuật biểu diễn nội dung mới hơn so với word2vec. Thử nghiệm thêm với các phương pháp biểu diễn từ ngữ phụ thuộc vào ngữ cảnh như ELMo [79] và BERT [83]. Về mặt thuật toán, các nghiên cứu tiếp theo có thể ứng dụng thêm những mô hình học sâu dành cho văn bản mà luận án chưa thử nghiệm. Về mặt đặc trưng, nghiên cứu tiếp theo có thể bổ sung thêm các đặc trưng xã hội khác nhau vào vector biểu diễn thư điện tử

DANH MỤC CÁC CÔNG TRÌNH CÓ LIÊN QUAN ĐẾN LUẬN ÁN

TẠP CHÍ KHOA HỌC

[1] Thanh, H. N., Dinh, Q. D., & Anh-Tran, Q. (2017). Personalized Email User Action Prediction Based on SpamAssassin. In *Cong Vinh P., Tuan Anh L., Loan N., Vongdoiwang Siricharoen W. (eds) Context-Aware Systems and Applications. ICCASA 2016. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (Vol. 193). Springer, Cham. https://doi.org/10.1007/978-3-319-56357-2_17

[2] Nguyễn, H. T., Đặng, Q. Đ., & Trần, A. Q. (2020). A neural network method for spamassasin rules generation. *Journal of Science and Technology on Information and Communications*, *1*(4A), 4-11.

[3] Ha, N. T., Quan, D. D., & Anh, T. Q. (2021). Combining content and social features in a deep learning approach to Vietnamese email prioritization. *REV Journal on Electronics and Communications*, *11*(3–4).

HỘI NGHỊ KHOA HỌC

[4] Nguyễn X. T., Trần Q. A., Trịnh B. N., & Nguyễn T. H. (2015). Ứng dụng tối ưu hóa đa mục tiêu trong bài toán tự động phân loại thư rác. *Hội thảo Quốc gia 2015 về Điện tử, Truyền thông và Công nghệ thông tin (REV-ECIT 2015)*, 30–35.

[5] Thanh, H. N., Dinh, Q. D., & Tran, Q. A. (2018). Predicting user's action on emails: Improvement with ham rules and real-world dataset. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. https://doi.org/10.1109/KSE.2018.8573330