HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VŨ MINH TUẤN

NGHIÊN CỨU CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT

Chuyên ngành: **Hệ thống thông tin**

Mã số: **9.48.01.04**

TÓM TẮT LUẬN ÁN TIẾN SỸ HỆ THỐNG THÔNG TIN

HÀ NỘI - 2024

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. Nguyễn Xuân Thắng

Phản biện 1:.....

.....

Phản biện 2:.....

.....

Phản biện 3:.....

.....

Luận án được bảo vệ trước Hội đồng chấm luận cấp Học viện họp tại: Học viện Công nghệ Bưu chính Viễn thông.

Vào hồi: giờ ngày … tháng … năm ……

Có thể tìm hiểu luận án tại:

Thư viện Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. LÝ DO LỰA CHỌN ĐỀ TÀI

Trong bối cảnh toàn cầu hóa và kỹ thuật số hóa, tin nhắn SMS đã trở thành một phương tiện giao tiếp không thể thiếu, mang lại lợi ích đáng kể cho cả người dùng cá nhân và tổ chức doanh nghiệp. Từ việc duy trì liên lạc cá nhân đến việc tận dụng trong chiến lược marketing và quảng cáo, SMS cho thấy sức mạnh trong việc truyền tải thông điệp một cách nhanh chóng và rộng rãi.

Với mục tiêu góp phần cải thiện chất lượng dịch vụ tin nhắn và bảo vệ người dùng khỏi những thông tin không mong muốn cũng như nguy cơ an ninh thông tin, đề tài nghiên cứu này tập trung vào việc phát triển và cải tiến các kỹ thuật phát hiện tin nhắn rác, đặc biệt là cho tin nhắn tiếng Việt. Nghiên cứu dự kiến sẽ khám phá và đánh giá các phương pháp hiện tại, đồng thời đề xuất các giải pháp mới để xử lý hiệu quả hơn vấn đề tin nhắn rác. Qua đó, đề tài không chỉ nhằm mục đích học thuật mà còn hướng đến việc đóng góp thực tiễn cho ngành viễn thông, đồng thời cải thiện trải nghiệm và sự an toàn cho người dùng tin nhắn SMS. Chính vì vậy, NCS đã quyết định chọn đề tài "Nghiên cứu các phương pháp phát hiện tin nhắn rác tiếng Việt" cho luận án tiến sĩ.

2. MỤC TIÊU, ĐỐI TƯỢNG, PHẠM VI VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mục tiêu nghiên cứu

- Xây dựng tập dữ liệu tin nhắn rác và tin nhắn thường bằng tiếng Việt.
- Phân tích mức độ phụ thuộc của hiệu quả các mô hình phát hiện tin nhắn rác vào độ dài nội dung của tin nhắn.

- Đề xuất phương pháp phát hiện tin nhắn rác thích hợp cho đặc thù đa biến thể của ngôn ngữ tiếng Việt.

2.2. Đối tượng và phạm vi nghiên cứu

Đối tượng NCS tập trung nghiên cứu là các phương pháp phát hiện tin nhắn rác tiếng Việt. Với khách thể nghiên cứu chính là những tin nhắn rác được phát tán đến người sử dụng dịch vụ di động, NCS đã chọn đối tượng khảo sát trên phạm vi rộng là những tập dữ liệu tin nhắn rác mẫu tiếng Việt.

2.3. Phương pháp nghiên cứu

Nghiên cứu sinh (NCS) thực hiện đề tài nghiên cứu bằng cách áp dụng các phương pháp sau: thu thập và chuẩn bị dữ liệu thử nghiệm cho mô hình phát hiện tin nhắn rác tiếng Việt; nghiên cứu lý thuyết qua việc đọc và phân tích tài liệu; triển khai thực nghiệm dựa trên lý thuyết; và đánh giá kết quả thực nghiệm qua các tiêu chí như Accuracy, Precision, Recall, và F1 Score..

3. Ý NGHĨA CỦA ĐỀ TÀI

Nghiên cứu cung cấp kiến thức về đặc điểm của tin nhắn và tin nhắn rác tiếng Việt, hỗ trợ nghiên cứu sau này trong việc phát hiện, phân loại và ngăn chặn tin nhắn rác. Luận án trình bày phương pháp mới trong việc phát hiện tin nhắn rác, bao gồm một mô hình phù hợp với đặc thù tiếng Việt và kỹ thuật xử lý dữ liệu phức tạp.

4. NHIỆM VỤ NGHIÊN CỨU VÀ KẾT QUẢ ĐẠT ĐƯỢC

4.1. Nhiệm vụ nghiên cứu

Nhiệm vụ đầu tiên của luận án là xây dựng một tập dữ liệu chứa cả tin nhắn thường và tin nhắn rác trong ngôn ngữ tiếng Việt.

Nhiệm vụ tiếp theo liên quan đến việc phân tích mức độ phụ thuộc của hiệu quả các mô hình phát hiện tin nhắn rác vào độ dài nội dung của tin nhắn.

Cuối cùng, **nhiệm vụ thứ ba** đề xuất phương pháp tiếp cận để giải quyết bài toán phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt trên cơ sở kế thừa kết quả của nhiệm vụ thứ hai.

4.2. Các kết quả đạt được

- **Một là** luận án đã xây dựng bộ dữ liệu tin nhắn rác và tin nhắn thường phục vụ cho mục đích nghiên cứu.
- **Hai là** luận án phân tích mức độ phụ thuộc của hiệu quả các mô hình phát hiện tin nhắn rác vào độ dài nội dung của tin nhắn. Từ đó, đề xuất được mô hình phát hiện tin nhắn rác tiếng Việt có tính ổn định khi độ dài tin nhắn thay đổi.
- **Ba là** luận án đề xuất 2 hướng tiếp cận giải quyết bài toán phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt.

5. BỐ CỤC LUẬN ÁN

Phần Mở đầu: Tập trung làm rõ những lý do cơ bản để lựa chọn đề tài, xác định rõ mục tiêu, đối tượng, phạm vi và phương pháp nghiên cứu của đề tài.

Chương 1: Giới thiệu tổng quan về tin nhắn SMS. Chương này làm rõ các khái niệm tin nhắn SMS, cấu trúc và cơ chế hoạt động của tin nhắn SMS; giới thiệu về tin nhắn rác, bao gồm các khải niệm, quy định và đặc trưng của tin nhắn rác và các bài toán liên quan.

Chương 2: Nghiên cứu mức độ ảnh hưởng của độ dài tin nhắn tới hiệu quả của mô hình phát hiện tin nhắn rác tiếng Việt. Từ đó, lựa chọn và đề xuất mô hình phát hiện tin nhắn rác tiếng Việt thích ứng với điều kiện giới hạn về nội dung tin nhắn.

Chương 3: Đề xuất 2 phương án tiếp cận để giải quyết bài toán phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt.

Chương 1 TỔNG QUAN VỀ TIN NHẮN SMS VÀ PHƯƠNG PHÁP PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT

1.1. TIN NHẮN SMS

1.1.1. Khái niệm tin nhắn SMS

SMS là từ viết tắt của Short Message Service. Đó là một công nghệ cho phép gửi và nhận các tin nhắn giữa các điện thoại với nhau. SMS xuất hiện đầu tiên ở Châu âu vào năm 1992 và được sử dụng phổ biến đến tận ngày nay.

1.1.2. Cấu trúc tin nhắn SMS

Nội dung của một tin nhắn SMS khi được gửi đi sẽ được chia làm 5 phần được mô như sau: Instructions to air interface, Instructions to SMSC, Instructions to handset, Instructions to SIM (optional), Message body.

1.1.3. Cơ chế hoạt động cơ bản của tin nhắn SMS

Trung tâm dịch vụ SMS (SMSC) chịu trách nhiệm quản lý và luân chuyển tin nhắn SMS trong mạng không dây. Khi một tin nhắn được gửi, nó đầu tiên đến SMSC, sau đó được chuyển đến người nhận. Nếu điện thoại người nhận tắt, SMSC lưu trữ tin nhắn cho đến khi máy bật lại.

1.2. TIN NHẮN RÁC VÀ CÁC ĐẶC TRƯNG

1.2.1. Định nghĩa tin nhắn rác

Tin nhắn rác phổ biến và xuất hiện trong nhiều hình thức, bao gồm email, bình luận blog, diễn đàn và thậm chí cả kết quả tìm kiếm bị nhiễm độc. Tại Việt Nam, tin nhắn rác tại Nghị định số 90/2008/NĐ-CP được định nghĩa là “tin nhắn được gửi đến người nhận mà người

nhận đó không mong muốn hoặc không có trách nhiệm phải tiếp nhận theo quy định của pháp luật".

1.2.2. Đặc trưng của tin nhắn rác

1.2.2.1. Đặc trưng dựa trên nội dung

Những công nghệ dựa trên đặc trưng nội dung sử dụng trong lọc thư rác có thể rất tiềm năng khi áp dụng với tin nhắn rác bao gồm cả 2 kĩ thuật: Lọc nội dung trực tiếp và lọc nội dung cộng tác.

Đối với nhóm đặc trưng dựa trên nội dung, các nhà nghiên cứu đều có những đề xuất và phân tích riêng để tối ưu việc lọc và phát hiện tin nhắn rác.

Như vậy, có thể thấy, các đặc trưng nội dung là cơ sở vững chắc để phát triển các mô hình lọc và phát hiện tin nhắn rác – đang được các nhà nghiên cứu khai thác triệt để. Tuy nhiên, còn có những nhóm đặc trưng khác cũng góp phần không nhỏ để nhận diện tin nhắn rác.

1.2.2.2. Đặc trưng phi nội dung

Như đã đề cập ở phần trước, ngoài những đặc trưng dựa trên nội dung, nhóm đặc trưng phi nội dung cũng được đề cập và nghiên cứu trong rất nhiều mô hình phát hiện spam nói chung và tin nhắn rác nói riêng: *Nhóm đặc trưng tĩnh, Nhóm đặc trưng về mạng, Nhóm đặc trưng về thời gian, Nhóm đặc trưng khác.*

1.3. TẬP DỮ LIỆU TIN NHẮN TIẾNG VIỆT

1.3.1. Nghiên cứu về tập dữ liệu tin nhắn

Hiện nay, trên thế giới có một số tập dữ liệu tin nhắn phổ biến phục vụ cho việc nghiên cứu. Tuy nhiên, hầu hết các tập dữ liệu tin nhắn SMS phổ biến vừa đề cập chỉ cung cấp dữ liệu bằng ngôn ngữ tiếng Anh. Với ngôn ngữ tiếng Việt, theo tìm hiểu của NCS thì chưa có một tập dữ liệu tin nhắn SMS nào chính thức nào được công bố.

Việc không có một tập dữ liệu tin nhắn rác và tin nhắn thường khiến cho việc phát triển và kiểm chứng các mô hình và thuật toán phát hiện tin nhắn rác tiếng Việt gặp nhiều khó khăn. Bên cạnh đó, dữ liệu tin nhắn SMS thường là dữ liệu cá nhân, đòi hỏi sự đồng ý của người dùng trước khi được sử dụng cho mục đích nghiên cứu. Do đó, việc xây dựng một tập dữ tin nhắn tiếng Việt phục vụ trong phạm vi nghiên cứu của luận án là hết sức cần thiết.

1.3.2. Xây dựng tập dữ liệu tin nhắn tiếng Việt

1.3.2.1. Mục đích

Xây dựng một tập dữ liệu tin nhắn thường và tin nhắn rác tiếng Việt phục vụ cho nghiên cứu đòi hỏi sự đầu tư và chăm chỉ trong việc thu thập, xác thực và xử lý dữ liệu.

1.3.2.2. Phương pháp thu thập dữ liệu

a) Yêu cầu về thu nhập tin nhắn

Đối với tin nhắn thường, việc thu thập dữ liệu không giới hạn hay tập trung vào một chủ đề cụ thể nào. Việc này nhằm đảm bảo tính đại diện cho các nội dung thực tế của dữ liệu, đồng thời tăng tính đa dạng của tập dữ liệu tin nhắn thường.

Đối với tin nhắn rác, yêu cầu quan trọng nhất là xác định đúng tính chất của tin nhắn trước khi cung cấp cho luận án. Việc đánh giá một tin nhắn có bị coi là tin rác hay không phụ thuộc vào đánh giá chủ quan của người dùng căn cứ trên định nghĩa về tin nhắn rác.

b) Phương pháp và quy trình thu thập

i) Thu thập qua công cụ website

NCS xây dựng một biểu mẫu online để các tình nguyện viên có thể cung cấp nội dung tin nhắn cho luận án (cả tin nhắn rác và tin nhắn thường).

ii) Trích xuất trực tiếp từ điện thoại của cộng tác viên

Ngoài phương pháp đóng góp từng tin nhắn thông qua công cụ website, NCS còn hướng dẫn các tình nguyện viên trực tiếp trích xuất tin nhắn điện thoại cá nhân, chuyển qua máy tính cá nhân và gửi lại cho NCS.

iii) Bẫy tin nhắn

Để chủ động tăng số lượng tin nhắn rác trong tập dữ liệu, nghiên cứu sinh đã triển khai mạng lưới "bẫy bình mật" với 32 số điện thoại để bẫy spammer. Các số điện thoại này được đưa lên các diễn đàn, hội nhóm trên mạng xã hội và các nhóm chat khác nhau với mục tiêu là lọt vào danh sách khách hàng của các đối tượng phát tán tin nhắn rác quảng cáo.

iv) Các nguồn khác

Ngoài các phương pháp chủ động thu thập tin nhắn từ các tình nguyện viên, NCS được Trung tâm VNCERT/CC hỗ trợ cung cấp 24.467 tin nhắn rác tiếng Việt để phục vụ cho nghiên cứu.

1.3.2.3. Xử lý dữ liệu

Trước khi thực hiện huấn luyện bất cứ mô hình nào, việc thiết yếu là tiền xử lý dữ liệu.

a) *Làm sạch dữ liệu (Data cleaning)*
b) *Chuẩn hóa từ (Stemming)*
c) *Loại bỏ từ nối (Stop-word removal)*
d) *Tách từ (Word Segmentation)*

1.3.3. Mô tả và phân tích tập dữ liệu

1.3.3.1. Số lượng tin nhắn

Tập dữ liệu tin nhắn phục vụ nghiên cứu bao gồm hai loại: tin nhắn rác và tin nhắn thường. Tổng cộng, có 47.808 tin nhắn rác và 44,807 tin nhắn thường được thu thập và xử lý để phục vụ luận án.

1.3.3.2. Độ dài tin nhắn

Dữ liệu cho thấy tin nhắn thường thường có độ dài ngắn hơn, với phần lớn các tin nhắn rơi vào nhóm từ 5 đến 10 từ (18,506 tin nhắn) và dưới 5 từ (6,375 tin nhắn).

Ngược lại, tin nhắn rác lại có xu hướng dài hơn với sự tập trung chủ yếu vào các nhóm từ 20 đến 30 từ (14,456 tin nhắn) và trên 30 từ (14,642 tin nhắn).

Tính không đồng nhất về độ dài tin nhắn trong dữ liệu cũng có thể dẫn đến những khó khăn trong việc huấn luyện và triển khai mô hình phát hiện rác. Một mô hình nếu không được tối ưu hóa để xử lý cả hai loại tin nhắn với độ dài khác nhau có thể gặp phải tình trạng thiên lệch, hoạt động kém hiệu quả khi phải phân loại các tin nhắn không thuộc nhóm mà nó được huấn luyện nhiều nhất.

1.3.3.3. Biến thể ngôn ngữ

Trong tổng số 38,724 tin nhắn rác đã được xử lý, có 27,106 tin nhắn ở dạng tiếng Việt có dấu và 11,618 tin nhắn ở dạng không dấu. Ngược lại, trong 30,468 tin nhắn thường, phần lớn (28,944 tin nhắn) là tiếng Việt có dấu, trong khi chỉ có 1,524 tin nhắn ở dạng không dấu. Sự khác biệt này phản ánh rằng người gửi tin nhắn rác thường sử dụng cả hai dạng ngôn ngữ để tăng cường khả năng vượt qua các bộ lọc rác, trong khi tin nhắn thường chủ yếu sử dụng tiếng Việt có dấu, thể hiện tính tự nhiên và rõ ràng trong giao tiếp hàng ngày.

1.4. BÀI TOÁN PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT

1.4.1. Phân tích mức độ ảnh hưởng của độ dài tin nhắn đến hiệu quả của mô hình phát hiện tin nhắn rác

Cho một tập dữ liệu $\mathcal{D}$ gồm n tin nhắn x_i. Ta có tập $\mathcal{D} = \{x_1, x_2, x_n\}$ gồm các tin nhắn đã được phân loại thành tin nhắn rác (spam) và tin nhắn thường (ham). Mỗi tin nhắn x_i có độ dài l_i, được định nghĩa là số lượng từ hoặc ký tự trong tin nhắn. Bài toán đặt ra là phân tích mức độ ảnh hưởng của l_i đến hiệu quả của một mô hình phát hiện tin nhắn rác.

Cụ thể, NCS xây dựng một số mô hình phát hiện tin nhắn rác f, được huấn luyện và kiểm thử trên các tập dữ liệu con $\mathcal{D}_{short}$, $\mathcal{D}_{medium}$ *và* $\mathcal{D}_{long}$, được phân chia dựa trên độ dài tin nhắn. Hiệu quả của mô hình được đánh giá qua các chỉ số chính như Accuracy, Recall, Precision và F1 Score.

1.4.2. Xây dựng mô hình phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt

a) Hướng tiếp cận 1: Xây dựng mô hình phát hiện tin nhắn rác tiếng Việt dạng đơn thể (có dấu HOẶC không dấu).

Cho tập dữ liệu tin nhắn tiếng Việt đa biến thể $\mathcal{D}_{multi}$ bao gồm cả tin nhắn có dấu và không dấu. Nhiệm vụ là:

- Xử lý tập dữ liệu để tạo ra hai tập dữ liệu đơn thể $\mathcal{D}_{cd}, \mathcal{D}_{kd}$ (có dấu và khôgn dấu).
- Huấn luyện mô hình f trên hai tập dữ liệu $\mathcal{D}_{cd}, \mathcal{D}_{kd}$.
- Đánh giá và so sánh hiệu quả của mô hình f trên từng tập dữ liệu đơn thể dựa trên chỉ số Accuracy, Recall, Precision và F1 Score.

b) *Hướng tiếp cận 2: Lựa chọn mô hình và thực hiện tối ưu để phát hiện tin nhắn rác tiếng Việt dạng đa biến thể (hỗn hợp có dấu VÀ không dấu) hiệu quả nhất.*

Các tập dữ liệu: $\mathcal{D}_{cd}, \mathcal{D}_{kd}$ và $\mathcal{D}_t$ lần lượt là các tập dữ liệu bao gồm m tin nhắn tiếng Việt với các biến thể: tiếng Việt có dấu, tiếng Việt không dấu và tổng hợp cả tiếng Việt có dấu và không dấu. Mỗi tin nhắn trong tập dữ liệu này được phân loại thành rác (spam) hoặc không rác (ham).

Nhiệm vụ là xây dựng một mô hình phát hiện tin nhắn rác g có khả năng xử lý tốt các biến thể ngôn ngữ này. Mô hình g sẽ được huấn luyện trên tập dữ liệu $\mathcal{D}_{train}$ (bao gồm các tin nhắn từ tất cả các biến thể) và được kiểm thử trên các tập dữ liệu riêng biệt $\mathcal{D}_{cd}, \mathcal{D}_{kd}$ và $\mathcal{D}_t$.

1.5. NGHIÊN CỨU TỔNG QUAN VỀ PHƯƠNG PHÁP PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT

1.5.1. Nghiên cứu về mức độ ảnh hưởng của độ dài tin nhắn với hiệu quả của mô hình phát hiện tin nhắn rác

Mặc dù các nghiên cứu trên đã đạt được những kết quả nhất định trong việc xác định tin nhắn rác nhưng kết quả chỉ áp dụng cho ngôn ngữ tiếng Anh. Có thể thấy rằng mặc dù các nghiên cứu hiện tại đã góp phần đáng kể vào việc phát triển các mô hình phát hiện tin nhắn rác hiệu quả, nhưng vẫn còn thiếu những phân tích chi tiết về tác động của độ dài tin nhắn đến hiệu quả của các mô hình này. Điều này đặt ra nhu cầu cần thiết phải tiến hành các nghiên cứu chuyên sâu, tập trung vào việc phân tích mức độ ảnh hưởng của độ dài tin nhắn đối với từng nhóm thuật toán, từ đó lựa chọn ra thuật toán có tính ổn định cao nhất trên các tập dữ liệu có độ dài tin nhắn khác nhau.

1.5.2. Nghiên cứu về phương pháp phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt

Mặc dù nghiên cứu trước đây đã giới thiệu các mô hình sử dụng phương pháp nhúng từ để xác định tin nhắn rác ở nhiều ngôn ngữ khác nhau, vẫn chưa có những đánh giá chi tiết về mức độ ảnh hưởng của kĩ thuật biểu diễn từ ngữ với mô hình phát hiện tin nhắn rác tiếng Việt. Đặc biệt là tiếng Việt đa biến thể. Từ đó, đề xuất mô hình phù hợp cho bài toán phát hiện tin nhắn rác tiếng Việt đa biến thể. Nghiên cứu này tìm cách đánh giá hiệu quả của các phương pháp biểu diễn từ kết hợp với các phương pháp phân loại học máy và học sâu khác nhau.

1.6. CÁC ĐỘ ĐO ĐÁNH GIÁ PHÂN LỚP NHỊ PHÂN

Trong phạm vi nghiên cứu các phương pháp phát hiện tin nhắn rác tiếng Việt, NCS đã chọn ra một số độ đo đánh giá hệ thống phân loại nhị phân (spam – ham) để đánh giá các mô hình đề xuất trong luận án. Các độ đo bao gồm: Accuracy, Precision và Recall, F1 Score.

1.7. TỔNG KẾT CHƯƠNG

Chương 1 đã cung cấp một cái nhìn tổng quan về tin nhắn SMS, từ khái niệm cơ bản đến cấu trúc và cơ chế hoạt động của nó. Đồng thời, NCS cũng trình bày các nghiên cứu về tập dữ liệu tin nhắn, đặc biệt là tập dữ liệu tiếng Việt, nhằm xác định các phương pháp hiệu quả trong việc xây dựng mô hình phát hiện tin nhắn rác. Cuối cùng, vấn đề quan trọng về việc phát hiện tin nhắn rác trong tiếng Việt đã được đề cập, với phân tích về tác động của độ dài tin nhắn và việc phát triển mô hình có khả năng xử lý những thách thức đặc trưng đa biến thể của ngôn ngữ tiếng Việt.

Chương 2 PHÂN TÍCH MỨC ĐỘ ẢNH HƯỞNG CỦA ĐỘ DÀI TIN NHẮN TỚI HIỆU QUẢ CỦA MÔ HÌNH PHÁT HIỆN TIN NHẮN RÁC

2.1. MỞ ĐẦU

2.1.1. Vai trò của độ dài tin nhắn trong ngôn ngữ tiếng Việt

Khi phân tích ảnh hưởng của độ dài tin nhắn tới hiệu quả của mô hình phát hiện tin nhắn rác tiếng Việt, cần đặc biệt lưu ý đến các đặc điểm ngôn ngữ đặc thù của tiếng Việt. Trước hết, tiếng Việt là ngôn ngữ đơn âm tiết với các dấu thanh đặc trưng, và mỗi từ có thể mang nhiều nghĩa khác nhau tùy thuộc vào dấu thanh được sử dụng. Điều này làm cho các tin nhắn ngắn trở nên khó khăn hơn trong việc xử lý, vì chỉ cần một từ sai dấu cũng có thể làm thay đổi hoàn toàn ý nghĩa của câu. Khi xử lý các tin nhắn ngắn, mô hình phát hiện tin nhắn rác cần phải được thiết kế sao cho có khả năng nhận diện chính xác ngữ nghĩa ngay cả khi ngữ cảnh bị giới hạn.

2.1.2. Vấn đề tồn tại và hướng giải quyết bài toán

Một trong những thách thức lớn đối với các mô hình phát hiện tin nhắn rác tiếng Việt là việc xử lý hiệu quả các tin nhắn có độ dài khác nhau. Độ dài tin nhắn có thể tác động đáng kể đến khả năng phân loại chính xác giữa tin nhắn rác và tin nhắn thường. Sự phân bố không đồng đều về độ dài tin nhắn tạo ra những khó khăn cho các mô hình học máy. Cụ thể, theo phân tích ở mục 1.3.3, tin nhắn thường thường ngắn gọn, chủ yếu rơi vào nhóm từ 5 đến 10 từ, trong khi tin nhắn rác lại có xu hướng dài hơn, với nhiều tin nhắn thuộc nhóm từ 20 đến trên 30 từ. Điều này có thể khiến mô hình bị thiên lệch, chỉ hoạt động tốt với một nhóm độ dài nhất định và kém hiệu quả trên các nhóm độ dài khác.

Để giải quyết những khó khăn này, cần tiến hành phân tích chi tiết về mức độ ảnh hưởng của độ dài tin nhắn đến hiệu quả của từng nhóm thuật toán. Việc thử nghiệm với các thuật toán khác nhau sẽ giúp xác định nhóm nào hoạt động ổn định hơn trên dữ liệu có độ dài tin nhắn đa dạng. Sau quá trình phân tích và thử nghiệm, lựa chọn thuật toán có tính ổn định cao nhất là bước cần thiết để đảm bảo mô hình duy trì được hiệu quả phân loại cao, bất kể tin nhắn có ngắn gọn hay dài dòng. Cuối cùng, mô hình cần được tối ưu hóa và huấn luyện với dữ liệu phong phú, bao gồm các tin nhắn có độ dài khác nhau, để đảm bảo rằng nó có thể xử lý tốt mọi tình huống thực tế. Việc này không chỉ nâng cao hiệu quả tổng thể trong việc phát hiện tin nhắn rác mà còn giúp mô hình trở nên mạnh mẽ, giảm thiểu rủi ro phân loại sai trong môi trường ngôn ngữ đa dạng như tiếng Việt.

Xây dựng mô hình phát hiện tin nhắn rác tiếng Việt đã trở thành một thách thức đáng kể trong bối cảnh nội dung tin nhắn ngắn, có giới hạn về độ dài và mục đích sử dụng chủ yếu là truyền tải thông tin nhanh chóng. Các phương pháp học máy truyền thống đã chứng minh hiệu quả trong việc phân loại email, với nội dung thường dài và phong phú. Tuy nhiên, khi chúng ta áp dụng chúng vào tin nhắn SMS - với nội dung hạn chế chỉ 160 ký tự, những hạn chế mới nảy sinh rõ ràng.

Tóm lại, việc xây dựng mô hình phát hiện tin nhắn rác tiếng Việt thích nghi với điều kiện nội dung ngắn đòi hỏi một phương pháp hiệu quả, mà không chỉ dựa vào nội dung văn bản mà còn kết hợp các đặc trưng khác. Trong phần tiếp theo của luận án, NCS thực hiện thí nghiệm các mô hình khác nhau bao gồm: mô hình dựa trên tập luật, mô hình với các thuật toán học máy truyền thống và mô hình sử dụng học sâu với 3 tập dữ liệu tin nhắn có độ dài khác nhau với mục đích

tìm ra mô hình phát hiện tin nhắn rác hiệu quả nhất trong bối cảnh nội dung tin nhắn bị giới hạn về độ dài.

2.2. PHÂN TÍCH ẢNH HƯỞNG CỦA ĐỘ DÀI TIN NHẮN TỚI HIỆU QUẢ CỦA MÔ HÌNH PHÁT HIỆN TIN NHẮN RÁC

2.2.1. Dữ liệu thử nghiệm

Từ tập dữ liệu tin nhắn rác và tin nhắn thường đều sử dụng tiếng Việt chuẩn (có dấu đầy đủ) bao gồm: 27.106 tin nhắn rác và 28.944 tin nhắn thường, NCS đã chia thành 2 loại dữ liệu. Tập dữ liệu huấn luyện (training) bao gồm: 15.000 tin nhắn rác và 15.000 tin nhắn thường. Số lượng tin nhắn còn lại được sử dụng cho tập dữ liệu kiểm tra (testing), được chia thành 3 tập dữ liệu theo tiêu chí độ dài tin nhắn với 3 cấp độ: Ngắn, Trung bình và Dài.

2.2.2. Thiết kế thử nghiệm

2.2.2.1. Mô hình phát hiện tin rác với phương pháp thống kê

Trong mô hình này, luận án đã áp dụng SpamAssassin vào bài toán phát hiện tin nhắn rác.

Quá trình xây dựng bộ luật được thực hiện qua 03 bước:

Bước 1: Tách từ có nghĩa từ tin nhắn sử dụng bộ công cụ vnTokenizer.

Bước 2: Lựa chọn từ khóa để xây dựng bộ luật:

Bước 3: Cập nhật điểm cho bộ luật:

NCS đã xây dựng một phần mềm trên nền tảng Android dành cho các thiết bị di động của người dùng cuối. Phần mềm sử dụng trực tiếp bộ luật để phân tích và lọc tin nhắn rác ngay trên thiết bị di động của người dùng. Về mặt giao diện người dùng và chức năng, phần mềm

gồm có 03 chức năng chính: Tải tập luật từ máy chủ, Phát hiện và đánh dấu tin nhắn rác, Gửi mẫu tin nhắn rác về máy chủ.

Sau khi huấn luyện với 30.000 tin nhắn, bộ luật được sinh ra và áp dụng thí nghiệm trên 10.000 tin nhắn hỗn hợp thuộc nhóm thí nghiệm.

Để đánh giá mức độ ảnh hưởng của độ dài tin nhắn với hiệu quả của mô hình sử dụng phương pháp này, NCS đã thực hiện thí nghiệm với ngưỡng 1.25 của bộ luật trên 3 tập dữ liệu DS-Short, DS-Medium và DS-Long

Bảng 2-1 Kết quả mô hình sử dụng bộ luật với ngưỡng 1.25

Tập dữ liệu	***Bộ luật thống kê với Ngưỡng 1,25***			
	Acc.	***Rec.***	***Prec.***	***F1Score***
DS-Short	0,554	0,602	0,528	0,563
DS-Medium	0,756	0,743	0,748	0,745
DS-Long	0,853	0,866	0,841	0,853

Có thể thấy, độ dài tin nhắn có ảnh hưởng lớn đến hiệu quả của phương pháp phát hiện tin nhắn rác dựa trên bộ luật. Với những tin nhắn ngắn, hệ thống có thể gặp khó khăn trong việc phân loại do thiếu dữ liệu, trong khi đó tin nhắn dài hơn lại cung cấp đủ thông tin giúp tăng độ chính xác. Do đó, việc kết hợp bộ luật với các phương pháp khác, chẳng hạn như máy học, có thể sẽ cải thiện khả năng phân loại trên cả ba bộ dữ liệu, đặc biệt là với những tin nhắn ngắn.

2.2.2.2. Mô hình phát hiện tin rác với học máy truyền thống

Các thuật toán học máy truyền thống bao gồm Support Vector Machine (SVM), Naïve Bayes (NB) và k-Nearest Neighbor (k-NN). Để đánh giá hiệu quả của mô hình với tập dữ liệu tin nhắn có độ dài

khác nhau, NCS đã lần lượt thực hiện thí nghiệm các thuật toán với 3 tập dữ liệu DS-Short, DS-Medium và DS-Long để ghi nhận và đánh giá kết quả. Kết quả thí nghiệm thu được được trình bày trong bảng dưới đây.

Bảng 2-2 Kết quả thí nghiệm với học máy truyền thống với tập dữ liệu DS-Short

	Accuracy	**Precision**	**Recall**	**F1Score**
SVM	0.643	0.631	0.654	0.642
NB	0.603	0.598	0.629	0.613
k-NN	0.611	0.638	0.625	0.631

Bảng 2-3 Kết quả thí nghiệm với học máy truyền thống với tập dữ liệu DS-Medium

	Accuracy	**Precision**	**Recall**	**F1Score**
SVM	0.810	0.795	0.824	0.809
NB	0.760	0.748	0.793	0.773
k-NN	0.770	0.804	0.788	0.796

Bảng 2-4 Kết quả thí nghiệm với học máy truyền thống với tập dữ liệu DS-Long

	Accuracy	**Precision**	**Recall**	**F1Score**
SVM	0.891	0.875	0.906	0.890
NB	0.836	0.823	0.872	0.847
k-NN	0.847	0.884	0.866	0.875

Kết quả ghi nhận sau thí nghiệm cho thấy sự chênh lệch rõ ràng khi thực hiện kiểm tra với bộ dữ liệu tin nhắn có độ dài khác nhau.

2.2.2.3. Mô hình phát hiện tin rác với học sâu

Theo kết quả thu được trong các thí nghiệm với thuật toán học máy truyền thống ở phần trên, rõ ràng hiệu quả của các thuật toán bị ảnh hưởng bởi độ dài của tin nhắn. Trên cơ sở nghiên cứu để đề xuất mô hình phát hiện tin nhắn rác thích ứng với điều kiện nội dung giới hạn, luận án đã phân tích, tìm hiểu và thực hiện thí nghiệm mô hình phát hiện tin nhắn rác với các thuật toán học sâu: mạng nơ-ron tích chập CNN (Convolutional Neural Network), mạng LSTM (Long Short-Term Memory) để đánh giá khả năng giải quyết mục tiêu của bài toán.

2.3. SO SÁNH VÀ ĐÁNH GIÁ KẾT QUẢ

Sau khi thực hiện thí nghiệm với 3 mô hình trên 3 tập dữ liệu bao gồm 3 nhóm tin nhắn có độ dài khác nhau, kết quả thu được đã phản ánh mức độ ảnh hưởng của độ dài tin nhắn tới hiệu quả của mỗi mô hình.

Từ kết quả trên, có thể đi đến kết luận mô hình sử dụng thuật toán học sâu CNN là mô hình phát hiện tin nhắn rác tiếng Việt thích ứng với điều kiện nội dung giới hạn có hiệu quả cao nhất trong số các đại diện được đưa vào thực nghiệm.

2.4. TỔNG KẾT CHƯƠNG

Trong chương này, luận án đã thực nghiệm 03 mô hình phát hiện tin nhắn rác tiếng Việt để đánh giá mức độ ảnh hưởng của độ dài tin nhắn tới hiệu quả phát hiện tin nhắn rác của các mô hình. Trong chương này, sự khác biệt của nghiên cứu được nhấn mạnh thông qua

việc áp dụng các phương pháp LSTM và CNN vào xử lý tin nhắn rác tiếng Việt, trong khi các phương pháp này đã được sử dụng rộng rãi trong các bài toán xử lý văn bản khác.

Chương 3 PHƯƠNG PHÁP PHÁT HIỆN TIN NHẮN RÁC VỚI ĐẶC ĐIỂM ĐA BIẾN THỂ CỦA TIẾNG VIỆT

3.1. MỞ ĐẦU

3.1.1. Đặc điểm đa biến thể của tiếng Việt

Trong tiếng Việt, chúng ta có hai biến thể chính là "tiếng Việt có dấu" và "tiếng Việt không dấu". Tiếng Việt có dấu sử dụng các dấu thanh như dấu sắc, dấu huyền, dấu hỏi, dấu ngã, dấu nặng để biểu thị thanh điệu và phát âm của từng từ. Nhờ vào các dấu thanh này, tiếng Việt có dấu mang lại sự chính xác và rõ ràng hơn trong việc diễn đạt ý nghĩa.

Các kỹ thuật vector hóa biểu diễn văn bản đóng vai trò trọng yếu trong việc xây dựng các mô hình học sâu, như CNN và LSTM, cho bài toán phát hiện tin nhắn rác tiếng Việt. Chúng giúp chuyển đổi văn bản thô thành định dạng số mà máy tính có thể xử lý được, mở đường cho việc phân tích và hiểu ngữ cảnh của từ ngữ. Điều này đặc biệt quan trọng trong tiếng Việt, nơi có sự đa dạng về cách viết có dấu và không dấu.

3.1.2. Vấn đề tồn tại và hướng giải quyết bài toán

Trong Chương 3, căn cứ trên kết quả đạt được ở Chương 2 về việc sử dụng mô hình học sâu (CNN) để phát hiện tin nhắn rác tiếng Việt, luận án đề xuất hai hướng tiếp cận để giải quyết bài toán phát hiện tin nhắn rác đa biến thể của tiếng Việt. Cụ thể như sau:

- *Hướng tiếp cận 1: Ứng dụng mô hình học sâu để phát hiện tin nhắn rác tiếng Việt dạng đơn thể (có dấu HOẶC không dấu).*

- *Hướng tiếp cận 2: Lựa chọn kĩ thuật vector hoá kết hợp với thuật toán học sâu để phát hiện tin nhắn rác tiếng Việt dạng đa biến thể (hỗn hợp có dấu VÀ không dấu) hiệu quả nhất.*

3.2. MÔ HÌNH SỬ DỤNG HỌC SÂU PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT ĐƠN THỂ

3.2.1. Giới thiệu mô hình

Đối với bài toán tiếng Việt đơn thể, mô hình học sâu CNN kết hợp với kĩ thuật vector hoá word2vec được giữ nguyên như đã đề xuất và thử nghiệm tại mục 2.2.2.3 của Chương 2.

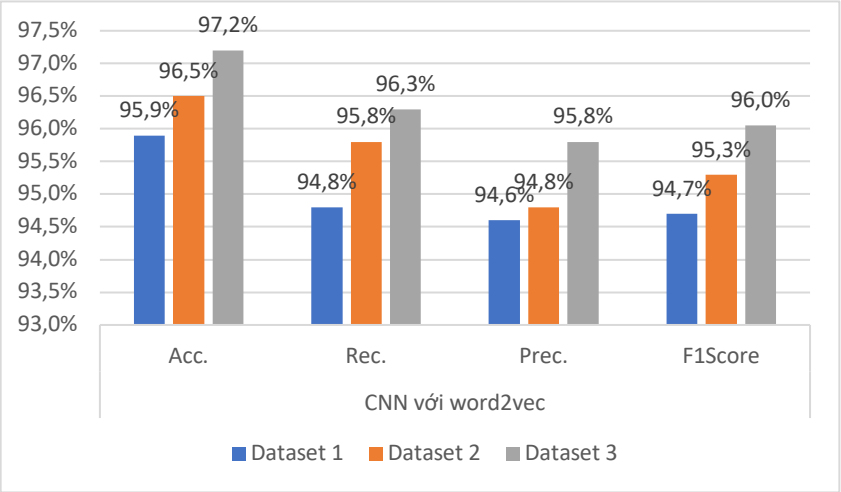
3.2.2. Thử nghiệm mô hình

3.2.2.1. Dữ liệu thử nghiệm

Bộ dữ liệu ban đầu trong mô hình này được đặt tên là dataset 1 – bao gồm tập dữ liệu tin nhắn rác và tin nhắn thường. Dataset 1 được giữ nguyên bản, chứa cả tiếng Việt có dấu và tiếng Việt không dấu. Để tạo ra các biến thể của bộ dữ liệu, hai phiên bản được tạo ra bằng cách khôi phục và loại bỏ dấu thanh của các tin nhắn gốc. Các thử nghiệm được tiến hành trên các biến thể của bộ dữ liệu để xác định hiệu quả của mô hình trong việc phát hiện tin nhắn rác tiếng Việt đơn thể. Đối với bộ dữ liệu không có dấu thanh (dataset 2). Đối với bộ dữ liệu với đầy đủ dấu thanh (dataset 3), các tin nhắn không có dấu thanh được khôi phục bằng cách thủ công bởi các tình nguyện viên để đảm bảo tính chính xác của quá trình chuyển đổi.

3.2.2.2. Thiết kế thử nghiệm

Mô hình phát hiện tin nhắn rác tiếng Việt thích ứng với điều kiện nội dung giới hạn được giữ nguyên và triển khai trên ba tập dữ liệu dataset 1, dataset 2 và dataset 3.



Hình 3-1 So sánh kết quả của mô hình CNN trên 3 tập dữ liệu

3.3. MÔ HÌNH HỌC SÂU KẾT HỢP KĨ THUẬT VECTOR HOÁ PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT ĐA BIẾN THỂ

3.3.1. Giới thiệu mô hình

3.3.1.1. Kết hợp CNN với các kỹ thuật vector hóa văn bản

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và học máy, việc kết hợp các kỹ thuật vector hóa văn bản như Word2Vec, GloVe, FastText, và PhoBERT với mạng nơ-ron tích chập (CNN) đã mở ra hướng tiếp cận mới và mạnh mẽ cho bài toán phát hiện tin nhắn rác tiếng Việt. Đặc biệt, trong bối cảnh đa biến thể của tiếng Việt, nơi mà ngôn ngữ xuất hiện cả ở dạng có dấu và không dấu, sự kết hợp này không chỉ cung cấp khả năng hiểu ngữ cảnh và biểu diễn từ ngữ một cách chính xác mà còn giúp nắm bắt và phân tích các đặc trưng cấu trúc quan trọng trong tin nhắn.

3.3.1.2. Kết hợp LSTM với các kỹ thuật vector hóa văn bản

Trong bối cảnh của ngôn ngữ tiếng Việt với đặc điểm đa biến thể, việc phát hiện tin nhắn rác đòi hỏi sự hiểu biết sâu sắc về ngữ cảnh cũng như ngữ nghĩa. Kết hợp các kỹ thuật vector hóa văn bản như Word2Vec, GloVe, FastText, và PhoBERT với mô hình Long Short-Term Memory (LSTM) trong học sâu có thể tạo ra một phương pháp mạnh mẽ và hiệu quả để giải quyết thách thức này.

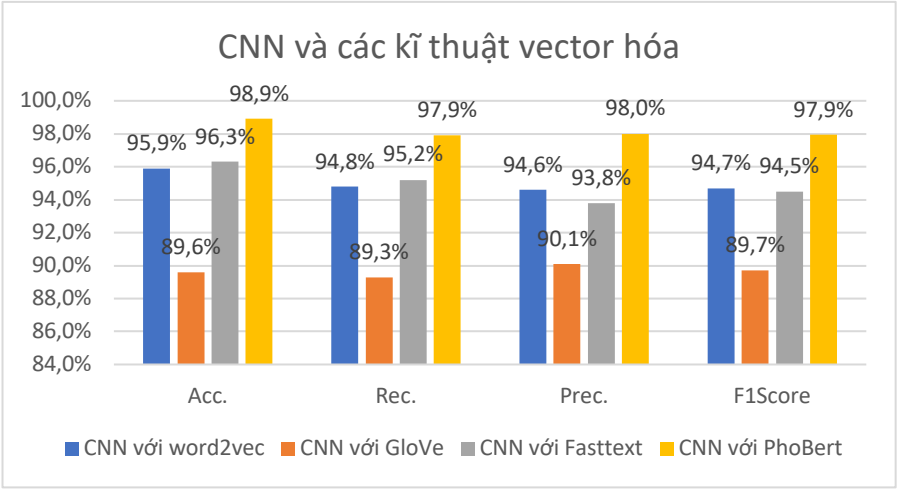
3.3.2. Thử nghiệm mô hình

3.3.2.1. Dữ liệu thử nghiệm

Trong thử nghiệm này, tập dữ liệu tin nhắn rác và tin nhắn thường ở dạng tiếng Việt đa biến thể được sử dụng để đánh giá tính hiệu quả của mô hình CNN và mô hình LSTM kết hợp với các kĩ thuật vector hóa văn bản. Đây cũng chính là tập dữ liệu dataset 1 được mô tả trong mục 3.2.2.1.

3.3.2.2. Thiết kế thử nghiệm

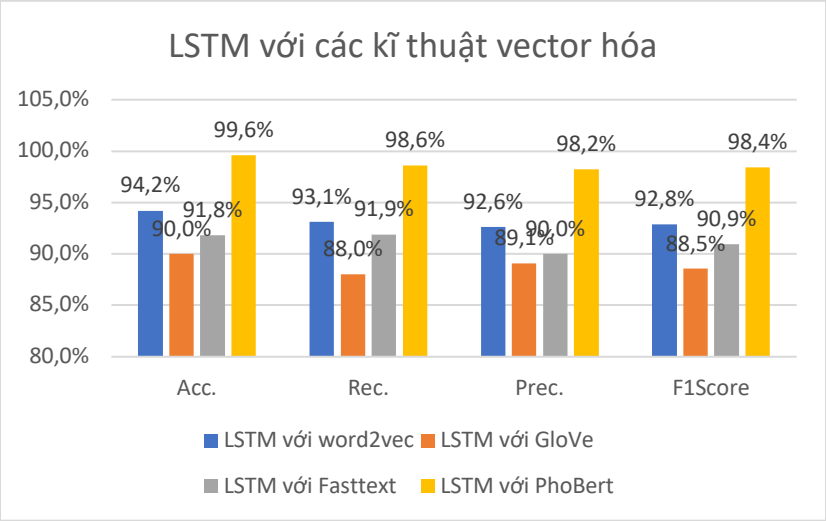
a) Mạng CNN và các kĩ thuật vector hóa văn bản



Hình 3-2 So sánh kết quả mô hình CNN khi kết hợp với các phép vector hoá

Kết quả thí nghiệm cho thấy sự kết hợp giữa các kỹ thuật vector hóa văn bản và mạng CNN đã tạo ra những mô hình có hiệu suất cao trong việc phát hiện tin nhắn rác tiếng Việt, đồng thời cũng thể hiện sự khác biệt rõ rệt giữa các mô hình dựa trên kỹ thuật vector hóa khác nhau khi kết hợp với mạng CNN để phát hiện tin nhắn rác tiếng Việt.

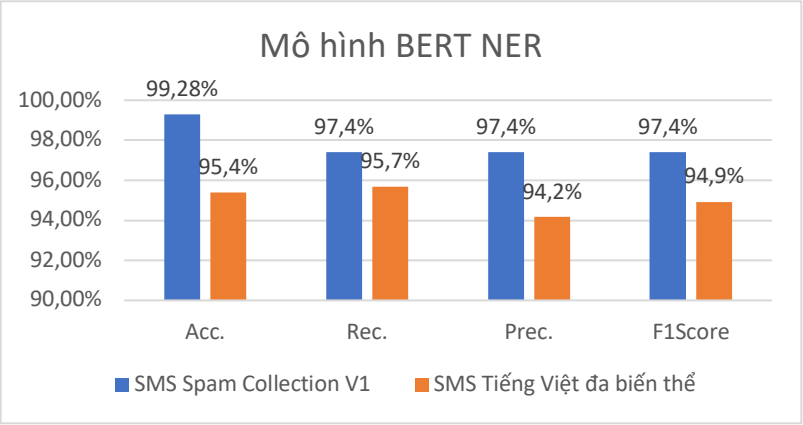
b) Mạng LSTM và các kĩ thuật vector hóa văn bản



Hình 3-3 So sánh kết quả mô hình LSTM khi kết hợp với các phép vector hoá

Kết quả thí nghiệm từ biểu đồ dữ liệu cho thấy rằng việc kết hợp các kỹ thuật vector hóa văn bản với mạng LSTM đã mang lại hiệu suất đáng chú ý trong việc phát hiện tin nhắn rác tiếng Việt. Đặc biệt, việc áp dụng PhoBERT đã vượt trội hơn hẳn so với các kỹ thuật khác.

c) Mô hình BERT NER



Hình 3-4 So sánh kết quả mô hình BERT NER chạy trên tập dữ liệu tiếng Anh và tiếng Việt đa biến thể

Trong thí nghiệm này, NCS tái hiện lại mô hình phát hiện spam được giới thiệu trong một nghiên cứu gần đây. NCS tạm gọi tên phương pháp này là BERT NER theo tên nhóm tác giả đã đề xuất. Thí nghiệm được tái hiện với tập dữ liệu tin nhắn tiếng Việt đa biến thể. Kết quả của thí nghiệm sẽ được so sánh với các nhóm thí nghiệm trên để đảm bảo tính khách quan của mô hình mà luận án đề xuất.

3.4. ĐÁNH GIÁ VÀ SO SÁNH KẾT QUẢ

Sau khi thực hiện thí nghiệm kết hợp thuật toán LSTM và CNN với lần lượt các kĩ thuật vector hoá, NCS đã chọn ra nhóm kết hợp có hiệu suất cao nhất từ mỗi nhóm để so sánh kết quả: CNN kết hợp với PhoBERT, LSTM kết hợp với PhoBERT và BERT NER trên tập dữ liệu tiếng Việt đa biến thể.

Như vậy, với hai hướng tiếp cận đề xuất để giải quyết bài toán phát hiện tin nhắn rác đa biến thể của tiếng Việt, luận án đã thực hiện những thử nghiệm để thu về kết quả, đưa ra kết luận.

3.5. TỔNG KẾT CHƯƠNG

Trong Chương 3 của luận án, NCS đã đề xuất ứng dụng mô hình học sâu kết hợp với học chuyển giao và lựa chọn mô hình học sâu phù hợp, kết hợp với học chuyển giao để phát hiện tin nhắn rác tiếng Việt đa biến thể.

KẾT LUẬN

A. Kết quả đạt được của luận án

Với bố cục bao gồm 3 chương, các kết quả chính đạt được của luận án có thể được tóm tắt như sau:

- Đóng góp bộ dữ liệu tin nhắn rác và tin nhắn thường tiếng Việt được thu thập từ các nguồn khác nhau, phục vụ cho nghiên cứu các phương pháp phát hiện tin nhắn rác tiếng Việt.
- Phân tích mức độ ảnh hưởng của độ dài tin nhắn tới hiệu quả của mô hình phát hiện tin nhắn rác. Từ đó lựa chọn mô hình có hiệu quả ổn định nhất trong điều kiện nội dung tin nhắn thay đổi.
- Đề xuất 2 phương án tiếp cận để giải quyết bài toán phát hiện tin nhắn rác với đặc điểm đa biến thể của tiếng Việt.

B. Những khó khăn tồn tại của luận án

Mặc dù đã đề xuất một số mô hình và hướng tiếp cận phát hiện tin nhắn rác tiếng Việt hiệu quả nhưng luận án đã gặp một số khó khăn và vẫn còn những tồn tại như sau:

- *Độ phức tạp ngôn ngữ tiếng Việt*
- *Tốn kém về tài nguyên tính toán*.

C. Định hướng phát triển

Dựa trên một số kết quả đạt được trong việc giải quyết bài toán, luận án có thể được phát triển và mở rộng theo một số hướng sau:

- Nâng cao hiệu suất phát hiện
- Mở rộng tập dữ liệu
- Xử lý đặc điểm ngôn ngữ phức tạp
- Sử dụng các kỹ thuật khác nhau

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

[CT1] Vu Minh Tuan, Nguyen Xuan Thang, Tran Quang Anh, "*Vietnamese SMS Spam Detection with Deep Learning and Pre-trained Language Model*", Journal of Science and Technology on Information and Communications: Vol. 1 No. 2 (2022).

[CT2] Vu Minh Tuan, Nguyen Xuan Thang, Tran Quang Anh, "*Evaluating the Efficiency of Vietnamese SMS Spam Detection Techniques*", Journal of Science and Technology on Information Security: Vol. 1 CS(18) 2023, ISSN 2615-9570.

[CT3] Vũ Minh Tuấn, Đặng Đình Quân, Nguyễn Thanh Hà, Trần Quang Anh, "*Lọc tin nhắn rác với Spam-Assassin*", Tạp chí Khoa học Công nghệ thông tin và Truyền thông, Số 3 – 4 (CS.01) (2016).

[CT4] Vu Minh Tuan, Do Thuy Duong, Tran Quang Anh, "*Evaluation of Word Embedding Techniques for the Vietnamese SMS Spam Detection Model*", Journal of Science and Technology on Information and Communications: Vol. 1 No. 2 (2023).

[CT5] Vu Minh Tuan, Do Thuy Duong, Tran Quang Anh, "*Proposing Appropriate SMS Spam Detection Approaches for Variations of the Vietnamese Language*", in 2023 RIVF International Conference on Computing and Communication Technologies, Hanoi, Vietnam (2023).