TRANG THÔNG TIN LUẬN ÁN TIẾN SĨ

Tên đề tài luận án tiến sĩ**: Nghiên cứu các phương pháp phát hiện tin nhắn rác tiếng Việt.**

Chuyên ngành: **Hệ thống thông tin**

Mã số: **9.48.01.04**

Họ và tên NCS: **Vũ Minh Tuấn**

Người hướng dẫn khoa học:

1. PGS.TS. Trần Quang Anh

2. TS. Nguyễn Xuân Thắng

Cơ sở đào tạo: **Học viện Công nghệ Bưu chính Viễn thông**

NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:

1. Luận án đã xây dựng bộ dữ liệu tin nhắn rác và tin nhắn thường phục vụ cho mục đích nghiên cứu. Bộ dữ liệu bao gồm 69.192 tin nhắn tiếng Việt có dấu và không dấu, được thu thập từ nhiều nguồn khác nhau như do tổ chức cung cấp, từ bẫy spam và do các tình nguyện viên cung cấp. Bộ dữ liệu có vai trò quan trọng trong việc phân tích các đặc điểm, đặc trưng của tin nhắn rác tiếng Việt, thử nghiệm và đánh giá các mô hình phát hiện tin nhắn rác trong luận án.

2. 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. So với các mô hình sử dụng bộ luật thống kê, sử dụng các thuật toán học máy truyền thống thì mô hình sử dụng học sâu – cụ thể là mạng CNN, đã thể hiệu khả năng phát hiện tin nhắn rác nổi trội trong điều kiện tin nhắn bị giới hạn về nội dung. Sự khác biệt về hiệu quả của các mô hình được thể hiện rõ ràng khi có sự thay đổi về độ dài tin nhắn thử nghiệm với từng mô hình.

3. 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. Phương án 1 tiếp cận theo hướng chuyển đổi dữ liệu đầu vào từ dạng tiếng Việt đa biến thể về đơn thể (có dấu hoặc không dấu). Phương án 2 tiếp cận theo hướng giữ nguyên dữ liệu đầu vào ở dạng tiếng Việt đa biến thể.

CÁC ỨNG DỤNG, KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HOẶC NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU:

Kết quả nghiên cứu của đề tài là nguồn tài liệu tham khảo có giá trị cho các doanh nghiệp liên

quan đến lĩnh vực viễn thông và bảo mật để nâng cao khả năng kiểm soát và ngăn chặn tin nhắn rác, tin quảng cáo sai quy định. 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:

(1) Nâng cao hiệu suất phát hiện: Cải thiện kiến trúc mô hình, tăng cường khả năng phân loại và phát hiện tin nhắn rác.

(2) Mở rộng tập dữ liệu: Mở rộng và cải thiện tập dữ liệu huấn luyện để đảm bảo tính đa dạng và đại diện cho nhiều loại tin nhắn rác tiếng Việt khác nhau.

(3) Xử lý đặc điểm ngôn ngữ phức tạp: Tiếng Việt có nhiều đặc điểm ngôn ngữ phức tạp như biến thể từ ngữ, dấu thanh và ngữ cảnh phong phú. Tích hợp các kỹ thuật xử lý ngôn ngữ tự nhiên và đặc trưng ngôn ngữ tiếng Việt vào mô hình để cải thiện khả năng hiểu và phân loại các loại tin nhắn rác.

(4) Áp dụng trong các lĩnh vực khác: Mô hình phát hiện tin nhắn rác tiếng Việt có thể được áp dụng trong nhiều lĩnh vực khác nhau như tin nhắn quảng cáo hay tin tức, trong các ứng dụng khác nhau như phòng chống gian lận, quản lý dữ liệu, và phân loại nội dung.

Xác nhận của đại diện tập thể	**Nghiên cứu sinh**
Người hướng dẫn khoa học	
PGS.TS. Trần Quang Anh	**Vũ Minh Tuấn**