ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN**

**NĂM 2023**

**THIẾT KẾ HỆ THỐNG RPA TÍCH HỢP NLP ĐỂ TỰ ĐỘNG HÓA QUY TRÌNH CHUYỂN TIẾP EMAIL TRONG DOANH NGHIỆP**

**<MÃ SỐ ĐỀ TÀI>**

**Lĩnh vực khoa học: Công nghệ thông tin**

**Chuyên ngành: Công nghệ thông tin**

**Nhóm nghiên cứu:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TT** | **Họ tên** | **MSSV** | **Đơn vị** | **Nhiệm vụ** | **Điện thoại** | **Email** |
|  | Nguyễn Quang Học | K214060427 | Khoa Hệ thống thông tin | Nhóm trưởng | 0862124950 | hocnq21406c@uel.edu.vn |
|  | La Thế Anh | K214060389 | Khoa Hệ thống thông tin | Tham gia | 0985676445 | anhlt21406@st.uel.edu.vn |
|  | Vũ Lưu Hoàng Lan | K214060429 | Khoa Hệ thống thông tin | Tham gia | 0944772586 | lanvlh21406c@st.uel.edu.vn |
|  | Giả Ngọc Thảo Ly | K214060430 | Khoa Hệ thống thông tin | Tham gia | 0968868495 | lygnt21406c@st.uel.edu.vn |
|  | Phan Ngọc Bảo Tâm | K214061268 | Khoa Hệ thống thông tin | Tham gia | 0797955317 | tampnb21406c@st.uel.edu.vn |

ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN**

**NĂM 2023**

**THIẾT KẾ HỆ THỐNG RPA TÍCH HỢP NLP ĐỂ TỰ ĐỘNG HÓA QUY TRÌNH CHUYỂN TIẾP EMAIL TRONG DOANH NGHIỆP**

**Đại diện nhóm nghiên cứu** **Giảng viên hướng dẫn**  **Chủ tịch Hội đồng**

*(Ký, họ tên)* *(Ký, họ tên)* *(Ký, họ tên)*



**Lãnh đạo Khoa/Bộ môn/Trung tâm**

*(Ký, họ tên)*

LỜI CẢM ƠN

Trong quá trình nghiên cứu đề tài **“Thiết kế hệ thống RPA tích hợp NLP để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”**, nhóm nghiên cứu đã nhận được nhiều sự hỗ trợ, giúp đỡ, quan tâm cũng như sự chỉ bảo nhiệt tình của mọi người để có thể hoàn thành đề tài nghiên cứu. Nhóm xin gửi lời cảm ơn chân thành đến:

Khoa Hệ Thống Thông Tin, trường Đại học Kinh Tế - Luật đã tạo môi trường học tập và nghiên cứu rất tốt, giúp nhóm có thêm kiến thức và kĩ năng để thực hiện đề tài.

Tiến sĩ Nguyễn Thôn Dã - người trực tiếp hướng dẫn đề tài đã luôn dành nhiều thời gian, công sức hướng dẫn chúng em trong quá trình thực hiện và hoàn thành đề tài nghiên cứu. Thầy đã luôn động viên, quan tâm và hỗ trợ mỗi khi chúng em gặp khó khăn trong quá trình nghiên cứu.

Công ty TNHH Thương mại và Dịch vụ Song Linh đã hỗ trợ nhóm trong quá trình thu thập và tìm kiếm dữ liệu.

Cuối cùng, nhóm cảm ơn gia đình và bạn bè đã luôn động viên, khích lệ và đặc biệt là lời cảm ơn đến sự nỗ lực của tất cả thành viên trong nhóm để có thể hoàn thành đề tài một cách tốt nhất.

LỜI CAM ĐOAN CỦA NHÓM NGHIÊN CỨU

Nhóm xin cam đoan nghiên cứu về đề tài **“Thiết kế hệ thống RPA tích hợp NLP để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”** là công trình nghiên cứu khoa học độc lập của riêng nhóm, dưới sự hướng dẫn của Tiến sĩ Nguyễn Thôn Dã. Các số liệu sử dụng phân tích trong nghiên cứu có nguồn gốc rõ ràng, đã công bố theo đúng quy định và được nhóm trích dẫn hợp pháp. Các kết quả nghiên cứu của đề tài do nhóm chúng em tự tìm hiểu, phân tích một cách khách quan, trung thực và minh bạch. Các kết quả này chưa từng được công bố trong bất kỳ nghiên cứu nào khác. Nhóm xin chịu hoàn toàn trách nhiệm nếu có sự không trung thực về thông tin được sử dụng trong công trình nghiên cứu này.

MỤC LỤC

[**DANH MỤC BẢNG BIỂU…………………………………………………………... 4**](#_Toc129467467)

[DANH MỤC HÌNH ẢNH VÀ ĐỒ THỊ 5](#_Toc129467468)

[DANH MỤC NHỮNG TỪ VIẾT TẮT 7](#_Toc129467469)

[TÓM TẮT…. 9](#_Toc129467470)

[CHƯƠNG 1 TỔNG QUAN ĐỀ TÀI 10](#_Toc129467471)

[1.1 Tính cấp thiết của nghiên cứu 10](#_Toc129467472)

[1.2 Tổng quan tình hình nghiên cứu và khoảng trống nghiên cứu 11](#_Toc129467473)

[1.2.1 Tình hình nghiên cứu tại Việt Nam 11](#_Toc129467474)

[1.2.2 Tình hình nghiên cứu trên thế giới 12](#_Toc129467475)

[1.2.3 Đánh giá chung 13](#_Toc129467476)

[1.2.4 Khoảng trống nghiên cứu và những điểm mới của nhóm 14](#_Toc129467477)

[1.3 Mục tiêu nghiên cứu 14](#_Toc129467478)

[1.4 Đối tượng và phạm vi nghiên cứu 15](#_Toc129467479)

[1.5 Phương pháp nghiên cứu 15](#_Toc129467480)

[1.6 Ý nghĩa thực tiễn của nghiên cứu 16](#_Toc129467481)

[1.7 Cấu trúc nghiên cứu 16](#_Toc129467482)

[CHƯƠNG 2 CƠ SỞ LÝ LUẬN 18](#_Toc129467483)

[2.1 Tổng quan về thư điện tử (Email) 18](#_Toc129467484)

[2.1.1 Khái niệm về thư điện tử (Email) 18](#_Toc129467485)

[2.1.2 Tầm quan trọng của thư điện tử trong doanh nghiệp hiện nay 19](#_Toc129467486)

[2.2 Lý thuyết xử lý ngôn ngữ tự nhiên (NLP) 23](#_Toc129467487)

[2.2.1 Khái niệm về xử lý ngôn ngữ tự nhiên (NLP) 23](#_Toc129467488)

[2.2.2 Quy trình xử lý ngôn ngữ tự nhiên 25](#_Toc129467489)

[2.3 Lý thuyết học máy 26](#_Toc129467490)

[2.3.1 Khái niệm về học máy 26](#_Toc129467491)

[2.3.2 Phân loại máy học 26](#_Toc129467492)

[2.3.3 Quy trình xây dựng hệ thống học máy 28](#_Toc129467493)

[2.4 Ứng dụng học máy trong xử lý ngôn ngữ tự nhiên 29](#_Toc129467494)

[2.5 Lý thuyết tự động hóa quy trình bằng robot 30](#_Toc129467495)

[2.5.1 Khái niệm tự động hóa quy trình bằng robot (RPA) 30](#_Toc129467496)

[2.5.2 Quy trình hoạt động của RPA 31](#_Toc129467497)

[CHƯƠNG 3 MÔ TẢ BÀI TOÁN 32](#_Toc129467498)

[3.1 Đặt vấn đề 32](#_Toc129467499)

[3.2 Đề xuất giải pháp cho bài toán 32](#_Toc129467500)

[CHƯƠNG 4 QUY TRÌNH THỰC HIỆN 35](#_Toc129467501)

[4.1 Phân tích dữ liệu và đề xuất các mô hình 35](#_Toc129467502)

[4.1.1 Tổng quan mô hình nghiên cứu phân loại văn bản 35](#_Toc129467503)

[4.1.2 Mô tả dữ liệu 36](#_Toc129467504)

[4.1.3 Làm sạch dữ liệu 37](#_Toc129467505)

[4.1.4 Xử lý ngôn ngữ tự nhiên (NLP) 39](#_Toc129467506)

[4.1.5 Vector hóa dữ liệu 46](#_Toc129467507)

[4.1.6 Cân bằng dữ liệu 48](#_Toc129467508)

[4.1.7 Mô hình dự đoán 50](#_Toc129467509)

[4.1.8 Đánh giá mô hình học máy 56](#_Toc129467510)

[4.2 Quy trình tự động hóa RPA 59](#_Toc129467511)

[4.2.1 Đề xuất mô hình tự động hóa 59](#_Toc129467512)

[4.2.2 Mô tả chi tiết quy trình thực hiện 60](#_Toc129467513)

[CHƯƠNG 5 KẾT QUẢ NGHIÊN CỨU 70](#_Toc129467514)

[5.1 Đánh giá và lựa chọn phương pháp phân loại 70](#_Toc129467515)

[5.2 Đánh giá và lựa chọn mô hình phân loại 71](#_Toc129467516)

[5.3 Đánh giá quy trình tự động hóa RPA 74](#_Toc129467517)

[CHƯƠNG 6 KẾT LUẬN VÀ ĐỀ XUẤT 78](#_Toc129467518)

[6.1 Tổng kết nội dung đã thực hiện 78](#_Toc129467519)

[6.2 Đề xuất 79](#_Toc129467520)

[6.3 Hạn chế của nghiên cứu 79](#_Toc129467521)

[6.4 Phương pháp phát triển 80](#_Toc129467522)

[TÀI LIỆU THAM KHẢO 81](#_Toc129467523)

DANH MỤC BẢNG BIỂU

[**Bảng 2‑1** So sánh ngôn ngữ tự nhiên và ngôn ngữ máy 22](#_Toc129473175)

[**Bảng 4‑1** Hình thức lưu trữ bộ dữ liệu 35](#_Toc129473176)

[**Bảng 4‑2** Kết quả quy trình tiền xử lý dữ liệu 41](#_Toc129473177)

[**Bảng 4‑3** Ma trận nhầm lẫn cho bài toán phân loại 55](#_Toc129473178)

[**Bảng 4‑4** Dữ liệu đầu vào của quy trình tự động hóa phân loại email 59](#_Toc129473179)

[**Bảng 5‑1** Độ phân loại chính xác của các phương pháp phân loại 68](#_Toc129473180)

[**Bảng 5‑2**  Kết quả thử nghiệm với từng mô hình 70](#_Toc129473181)

[**Bảng 5‑3** Kết quả thử nghiệm các mô hình phân loại 70](#_Toc129473182)

DANH MỤC HÌNH ẢNH VÀ ĐỒ THỊ

[**Hình 2‑1** Các loại thư điện tử được sử dụng phổ biến hiện nay 19](#_Toc129623900)

[**Hình 2‑2** Thống kê số lượng email được sử dụng mỗi năm (Nguồn: Hãng nghiên cứu thị trường Statista) 20](#_Toc129623901)

[**Hình 2‑3** Danh mục nội dung spam phổ biến nhất trên toàn thế giới năm 2018 (Nguồn: Hãng nghiên cứu thị trường Statista) 22](#_Toc129623902)

[**Hình 2‑4** Phân loại học máy 27](#_Toc129623903)

[**Hình 3‑1** Mô tả kết quả thực quy trình thực hiện bài toán 33](#_Toc129623904)

[**Hình 3‑2** Quy trình nghiên cứu tổng quan 34](#_Toc129623905)

[**Hình 4‑1** Quy trình nghiên cứu mô hình phân loại văn bản 35](#_Toc129623906)

[**Hình 4‑2** Số lượng các nhãn trong bộ dữ liệu nghiên cứu 37](#_Toc129623907)

[**Hình 4‑3** Kết quả vecto hóa tập dữ liệu huấn luyện 48](#_Toc129623908)

[**Hình 4‑4** Trước và sau khi cân bằng dữ liệu 50](#_Toc129623909)

[**Hình 4‑5** Quy trình thực hiện của thuật toán Stochastic Gradient Descent Classifier 51](#_Toc129623910)

[**Hình 4‑6** SVM tìm bộ phân loại có khoảng cách lớn nhất giữa các vectơ hỗ trợ. Vòng tròn đại diện cho các vectơ hỗ trợ 52](#_Toc129623911)

[**Hình 4‑7** Sơ đồ quy trình của phương pháp học máy tăng cường độ dốc. 53](#_Toc129623912)

[**Hình 4‑8** Quy trình thực hiện của phương pháp phân loại cây bổ sung 54](#_Toc129623913)

[**Hình 4‑9** Quy trình nghiên cứu mô hình tự động hóa 59](#_Toc129623914)

[**Hình 4‑10** Hoạt động ReadRange Excel 62](#_Toc129623915)

[**Hình 4‑11** Hoạt động GetImapMailMessages 63](#_Toc129623916)

[**Hình 4‑12** Hoạt động Python PythonScope 64](#_Toc129623917)

[**Hình 4‑13** Hoạt động Load Python Script 64](#_Toc129623918)

[**Hình 4‑14** Hoạt động Message Box hiển thị nội dung của email 65](#_Toc129623919)

[**Hình 4‑15** Hoạt động Invoke Method gọi hàm phân loại email 65](#_Toc129623920)

[**Hình 4‑16** Hoạt động Invoke Python Method gọi hàm hiển thị độ chính xác 66](#_Toc129623921)

[**Hình 4‑17** Hoạt động Get Python Object trả về kết quả phân loại phòng ban của email và độ chính xác của nó 66](#_Toc129623922)

[**Hình 4‑18** Kết quả thử nghiệm quy trình trên 1000 email 67](#_Toc129623923)

[**Hình 4‑19** Hoạt động Dialogbox hiển thị hộp thoại loại chọn phòng ban. 68](#_Toc129623924)

[**Hình 4‑20** Hoạt động SendSMTP Mail Message chuyển tiếp email và thông báo email đã được chuyển tiếp 69](#_Toc129623925)

[**Hình 5‑1** Kết quả so sánh độ chính xác của các phương pháp phân loại 71](#_Toc129623926)

[**Hình 5‑2** Áp dụng phương pháp phân loại đa lớp One-Vs-One 73](#_Toc129623927)

[**Hình 5‑3** Áp dụng phương pháp phân loại đa lớp One-Vs-Rest 73](#_Toc129623928)

[**Hình 5‑4** Đọc nội dung của email và thông báo (using UiPath) 74](#_Toc129623929)

[**Hình 5‑5** Thông báo email có độ tin cậy cao hơn 0.782 sẽ được chuyển tiếp đến phòng ban được phân loại. 74](#_Toc129623930)

[**Hình 5‑6** Thông báo email có độ tin cậy thấp hơn 0.782 sẽ tiến tới bước human in loop 75](#_Toc129623931)

[**Hình 5‑7** Hoạt động cần sự can thiệp của con người “human in loop” 75](#_Toc129623932)

[**Hình 5‑8** Thông báo đã hoàn tất việc chuyển tiếp email 76](#_Toc129623933)

[**Hình 5‑9** So sánh số lượng email được tự động phân loại và các email cần sự can thiệp của con người trường hợp 1000 email 77](#_Toc129623934)

DANH MỤC NHỮNG TỪ VIẾT TẮT

|  |  |
| --- | --- |
| Từ viết tắt | Tiếng Việt |
| AI | Artificial Intelligence (Trí tuệ nhân tạo) |
| B2B | Business to Business |
| Bot | Robot (Rô-bốt) |
| CFS | Correlation-based Feature Selection (Lựa chọn tính năng dựa trên tương quan) |
| CSV | Comma Separated Values (Giá trị được phân tách bằng dấu phẩy) |
| Email | Electronic mail (Thư điện tử) |
| GBC | Gradient Boosting Classifier (Trình phân loại tăng cường độ dốc) |
| IMAP | Internet Message Access Protocol (Giao thức truy cập tin nhắn Internet) |
| kNN | k-Nearest Neighbors (k láng giềng gần nhất) |
| Linear SVC | Linear Support Vector Classification (Trình phân loại véc-tơ hỗ trợ tuyến tính) |
| ML | Machine Learning (Máy Học/Học Máy) |
| NLP | Natural Language Process (Xử lý ngôn ngữ tự nhiên) |
| OCR | Optical Character Recognition (Nhận dạng ký tự quang học) |
| OvA | One-vs-All |
| OvO | One-vs-One |
| OvR | One-vs-Rest |
| PCA | Principal Component Analysis (Phân tích thành phần chính) |
| RPA | Robotic Process Automation (Quy trình tự động hóa bằng robot) |
| SGD Classifier | Stochastic Gradient Descent Classifier (Trình phân loại giảm dần độ dốc ngẫu nhiên) |
| SMTP | Simple Mail Transfer Protocol (Giao thức truyền tải thư tín đơn giản hóa) |
| Spam email | Spam Electronic mail (Thư rác điện tử) |
| SVM | Support Vector Machine (Máy véc-tơ hỗ trợ) |
| SVM-NB | Support Vector Machine - Naive Bayes (Máy véc-tơ hỗ trợ - Naive Bayes) |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| TNHH | Trách nhiệm hữu hạn |
| US | United States (Hoa Kỳ) |
| USD | United States Dollar (Đô la Mỹ) |

TÓM TẮT

Trước nhu cầu chuyển đổi số và tự động hóa ngày càng tăng cao, đòi hỏi các doanh nghiệp cần thay đổi và cải tiến các hoạt động trong kinh doanh của mình. Nhận thấy thực trạng mỗi ngày, doanh nghiệp phải nhận một lượng lớn email và cần bộ phận nhân sự có chuyên môn thực hiện công việc phân loại email và chuyển tiếp nó về đúng phòng ban chuyên môn hay thực trạng một doanh nghiệp cung cấp quá nhiều email của các phòng ban khiến cho khách hàng gặp tâm lý hoang mang, không biết liên hệ địa chỉ nào để giải quyết vấn đề đang mắc phải. Điều này khiến cho cả doanh nghiệp và khách hàng tốn nhiều thời gian và chi phí. Bám sát thực tiễn vấn đề, nhóm đã thực hiện nghiên cứu một đề tài ứng dụng “Thiết kế hệ thống RPA tích hợp NLP để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”. Đề tài bao gồm hai phần chính là: xây dựng cơ sở phân loại và xây dựng quy trình phân loại. Cơ sở phân loại được thiết kế bằng cách xử lý NLP tập dữ liệu email do nhóm nghiên cứu thu thập, sau đó sử dụng mô hình học máy phân loại văn bản kết hợp với phương pháp phân loại đa lớp (One-Vs-One và One-Vs-Rest). Thực hiện so sánh kết quả của 4 mô hình SGD Classifier, Linear SVC, Gradient Boosting Classifier, Extra Trees Classifier nhằm tìm ra mô hình phân loại phù hợp nhất. Mô hình được lựa chọn sẽ “học” từ tập dữ liệu vừa được xử lý, kết quả của bước này chính là cơ sở phân loại. Ở quy trình phân loại, hệ thống sẽ tự động mở, đọc, trích xuất, phân tích, phân loại, đánh giá, chuyển tiếp và lưu trữ mỗi email đến dựa vào cơ sở phân loại vừa phân tích được. Dựa vào chỉ số tin cậy, những email không đạt chỉ số này thì sẽ không được phân loại mà lúc này cần sự hỗ trợ của con người để giúp phân loại nó. Thực hiện chạy thử nghiệm mô hình, kết quả cho thấy hệ thống tự động hóa giúp tiết kiệm thời gian gấp khoảng 12.5 lần so với việc phân loại thực tế (kết quả đã bao gồm trường hợp nhờ sự hỗ trợ từ con người đối với những email không đủ độ tin cậy).

# TỔNG QUAN ĐỀ TÀI

Tính cấp thiết của nghiên cứu

Dưới sự bùng nổ về công nghệ - kỹ thuật số, thư điện tử (email) ngày càng đóng vai trò quan trọng trong cuộc sống, không chỉ được sử dụng rộng rãi như một kênh giao tiếp chính thức của các cá nhân, tổ chức mà nhu cầu sử dụng của chúng cũng đang ngày càng tăng cao ([Mujtaba, Shuib, Raj, Majeed, & Al-Garadi, 2017](#_ENREF_31)). Theo báo cáo nghiên cứu của Statista - một nền tảng trực tuyến chuyên thống kê và phân tích số liệu, thế giới đã có hơn 306 tỉ email được gửi đi mỗi ngày vào năm 2020, điều này đồng nghĩa rằng mỗi giây có đến 3,5 triệu email được gửi đi. Theo dự báo, con số này có thể lên đến 376,4 tỉ vào năm 2025.[[1]](#footnote-1) Trong đó, email thuộc nhiều chủ đề khác nhau và có khoảng 15,43% là thư rác (spam email) không đem lại nhiều giá trị cho người dùng.[[2]](#footnote-2) Vì vậy, người dùng buộc phải xử lý email bằng cách chọn lọc nội dung và loại bỏ những thư rác. Từ đó cho thấy, việc xử lý một lượng lớn email sẽ đòi hỏi rất nhiều thời gian và công sức.

Trên thực tế, các tổ chức như doanh nghiệp, tập đoàn, ... sử dụng email như một kênh giao tiếp với khách hàng và đối tác. Nhờ vào sự phát triển của công nghệ và mạng Internet, thị trường kinh doanh của doanh nghiệp, kể cả những doanh nghiệp vừa và nhỏ, cũng theo đó mở rộng đến với nhiều khách hàng hơn. Nhu cầu giao tiếp qua email không chỉ tăng bởi số lượng khách hàng mà còn vì khoảng cách địa lý cũng như yêu cầu tăng lên về thời gian giải quyết công việc. Do đó, doanh nghiệp có thể nhận đến hàng nghìn email mỗi ngày và phải xử lý chúng nhanh chóng để kịp thời phản hồi lại cho khách hàng - đối tác. Với thực tế đó, vấn đề đặt ra cho doanh nghiệp là làm thế nào để phân loại và xử lý một lượng lớn thông tin với thời gian với chi phí bỏ ra là ít nhất.

Tuy đã nhận thấy được vấn đề cần giải quyết, nhưng quy trình xử lý email thủ công tại các doanh nghiệp vẫn còn gặp phải nhiều bất cập. Khác với email của cá nhân, một số cấu trúc doanh nghiệp dù vẫn có một email đại diện nhưng cơ cấu công việc lại được chuyên môn hóa thành các phòng ban riêng biệt. Mỗi phòng ban chuyên giải quyết một số công việc nhất định, có thể kể đến một số phòng ban thường thấy trong doanh nghiệp như phòng Kinh doanh, phòng Nghiên cứu và Phát triển, phòng Nhân sự, phòng Chăm sóc khách hàng, phòng Đối ngoại, ... Do đó, quy trình xử lý email thủ công trong doanh nghiệp yêu cầu nhân viên văn thư đọc từng email, sau đó loại bỏ thư rác và phân loại email cũng như chuyển tiếp chúng về đúng phòng ban chuyên môn. Quy trình này có thể khiến doanh nghiệp tiêu tốn rất nhiều thời gian và nhân lực. Không chỉ vậy, việc phân loại email thủ công có thể gây ra sai sót không đáng có như bỏ sót email, phân loại sai phòng ban, chậm trễ trong quy trình, ... bởi sự chủ quan của người phân loại. Sự thiếu sót ở khâu phân loại email sẽ dẫn đến tình trạng chậm trễ và bỏ lỡ thông tin, phân loại sai phòng ban khiến quy trình xử lý phức tạp và lộn xộn.

Doanh nghiệp cần loại bỏ quy trình thủ công để gia tăng thời gian xử lý và độ chính xác. Vì vậy, một công cụ tự động nhận, đọc email, sau đó phân loại và chuyển tiếp về phòng ban là cấp thiết, giúp doanh nghiệp sử dụng hiệu quả nguồn nhân lực, quy trình xử lý nhanh, đơn giản và chính xác. Xuất phát từ nhu cầu đó, nhóm nghiên cứu quyết định thực hiện đề tài “**Thiết kế hệ thống RPA tích hợp NLP để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp**”.

## Tổng quan tình hình nghiên cứu và khoảng trống nghiên cứu

Tình hình nghiên cứu tại Việt Nam

Lâm Tăng Doan (2020) đã nghiên cứu và đề xuất một mô hình phân loại thư rác tiếng Việt bằng thuật toán SVM và Naive Bayes. Từ những nghiên cứu của mình, Lâm Tăng Doan đã đưa ra kết luận khẳng định thuật toán SVM và Naive Bayes là đáng tin cậy và được sử dụng nhiều. Do đó, tác giả đã ứng dụng kết hợp hai thuật toán nhằm đề xuất thuật toán SVM-NB và xây dựng quy tắc lọc tiếng Việt để cải thiện tốc độ lọc trong mô hình phân loại email rác. Mô hình được thực nghiệm trên bộ dữ liệu bao gồm 1650 email tiếng việt lẫn tiếng anh và kết quả trên bộ dữ liệu này cho thấy thuật toán SVM-NB đạt được độ chính xác và tỷ lệ thu hồi cao hơn giải pháp dựa trên Naive Bayes thuần túy ([Lâm, 2020](#_ENREF_25)).

Trong chủ đề phân loại văn bản vào các chủ đề cụ thể bằng phương pháp Học máy, Trần Thanh Điện và cộng sự (2019) đã đề xuất một mô hình phân loại bài báo khoa học vào lĩnh vực tương ứng bằng cách thực nghiệm so sánh ba mô hình SVM, Naive Bayes và kNN trên tập dữ liệu bao gồm 680 bài báo khoa học được gửi đăng trên Tạp chí khoa học của Trường Đại học Cần Thơ. Nghiên cứu đã khẳng định mô hình SVM mang lại kết quả phân loại tốt nhất. Tuy nhiên nghiên cứu vẫn có hạn chế về kết quả phân loại chưa cao của một số lĩnh vực như “Môi trường” hay “Công nghệ sinh học”. Tác giả dự đoán nguyên nhân của hạn chế này là do sự chồng lấn về lĩnh vực của bài báo và do các chủ đề này tương đối khác biệt so với các chủ đề còn lại ([Điện, Thanh, & Nghe, 2019](#_ENREF_15)).

Tình hình nghiên cứu trên thế giới

Akshay và cộng sự (2020) đã đề xuất một hệ thống tự động hóa nhiệm vụ phân loại và định tuyến email cũng như cung cấp trả lời cố định bằng cách sử dụng robot tự động hóa quy trình (RPA) chạy trên phần mềm UiPath. Mô hình tiết kiệm thời gian và nhân lực cần để soạn lại email cùng nội dung hoặc rà soát toàn bộ hộp thư đến để tìm các truy vấn tương tự. Tuy nhiên, việc xử lý email vẫn cần theo hướng dẫn và yêu cầu của tổ chức thay vì tự động hóa hoàn toàn. Bên cạnh đó, mô hình chứa ứng dụng AI để trả lời hay phân loại một cách thông minh ([Akshay, Kalagi, Shetty, & Ramalingam, 2020](#_ENREF_3)).

Patel và cộng sự (2019) đã đề xuất một hệ thống tự động trả lời khách hàng bằng cách tùy chỉnh phản hồi theo truy vấn của người dùng và đưa ra câu trả lời tương ứng. Hệ thống đã sử dụng thuật toán Máy vectơ hỗ trợ (SVM) và Độ tương tự Cosine để phân loại email dựa trên nội dung của chúng và tìm ra câu trả lời có khả năng nhất. Sau khi tìm được phản hồi phù hợp với truy vấn của khách hàng, hoạt động SMTP của UiPath được sử dụng để gửi email trả lời. Hệ thống này nhằm mục đích giảm thiểu công việc khó khăn là đọc và trả lời email theo cách thủ công - tốn nhiều công sức và thời gian của nhân lực ([Patel, Shukla, Porwal, & Kotecha, 2019](#_ENREF_33)).

Alghoul và cộng sự (2018) nghiên cứu một phương pháp mới để lọc thư rác bằng cách sử dụng Mạng thần kinh nhân tạo. Mô hình đã sử dụng thuật toán lan truyền ngược nguồn cấp dữ liệu để đào tạo và tạo ra độ chính xác cao. Nghiên cứu này đã tiếp cận vấn đề thư rác với một cách xử lý khá mới. Tuy nhiên, mô hình mới chỉ dừng lại ở mức độ ghi nhận thư rác, chưa có hình thức xử lý email khác như chuyển tiếp hay phản hồi ([Alghoul, Al Ajrami, Al Jarousha, Harb, & Abu-Naser, 2018](#_ENREF_4)).

Bahgat và cộng sự (2018) đã đề xuất một phương pháp lọc email hiệu quả dựa trên các phương pháp ngữ nghĩa. Nghiên cứu này đã phân tích một số phương pháp phân loại và nhận thấy rằng, khi so sánh với Naive Bayes, J48, Random Forest và các mạng chức năng cơ sở xuyên tâm thì Hồi quy logistic (Logistic Regression) và Máy Vector hỗ trợ (SVM) đạt được độ chính xác tối đa. Ngoài ra, Phân tích thành phần chính (PCA) và Lựa chọn tính năng tương quan (CFS) được sử dụng để có được bộ tính năng tối ưu tối thiểu. Theo kết quả thử nghiệm, chiến lược lọc ngữ nghĩa được đề xuất khi được kết hợp với lựa chọn tính năng đã mang lại hiệu suất tính toán tốt với tốc độ giảm không gian và thời gian cao ([Bahgat, Rady, Gad, & Moawad, 2018](#_ENREF_5)).

Tam và cộng sự (2012) đã sử dụng thuật toán học máy có giám sát để tự động sắp xếp email vào các thư mục khác nhau. Các trường văn bản của email (chủ đề và nội dung) được nhúng vào mô hình biểu diễn không gian vectơ. Các quyết định phân loại từ các trường email khác nhau được kết hợp bằng biểu quyết đa số. Phương pháp này cải tiến đáng kể so với cả hai bộ phân loại đơn lẻ trong các lĩnh vực phân loại này cũng như so với các công trình trước đó. Nhờ đó, nghiên cứu giải quyết được vấn đề ngữ nghĩa của văn bản khi áp dụng phương pháp học máy và giảm thiểu tình trạng quá tải email ([Tam, Ferreira, & Lourenço, 2012](#_ENREF_38)).

Đánh giá chung

Trên thực tế còn rất nhiều nghiên cứu liên quan về phân loại email. Tuy nhiên, theo báo cáo tổng quan tình hình nghiên cứu về phân loại email của Mujtaba và cộng sự (2017), phần lớn các nghiên cứu là phân loại thư rác. Chủ đề phân loại email vào các thư mục khác nhau (multi-folder categorization of emails) có khá ít nghiên cứu và chưa được chú ý trên thế giới, đặc biệt đây còn là một đề tài mới tại Việt Nam. Bên cạnh đó, đa số các nghiên cứu đều dừng lại ở việc “phân loại” mà chưa có xử lý đặc biệt nào khác như chuyển tiếp hay phản hồi ([Mujtaba et al., 2017](#_ENREF_31)). Một số nghiên cứu khác ở trên như nghiên cứu của Akshay và cộng sự, Patel và cộng sự đã đề xuất mô hình xử lý email nhưng vẫn còn một số bất cập như chưa ứng dụng Học máy để tăng độ chính xác, chưa thực sự ứng dụng vào thực tế doanh nghiệp. Công nghệ tự động hóa (RPA) vẫn chưa có nhiều nghiên cứu có tính ứng dụng và còn chưa được phổ biến tại Việt Nam. Hơn nữa, phần lớn các nghiên cứu đều dựa vào bộ dữ liệu mẫu có hiệu chuẩn cao như Enron, PU, TREC, ... Điều này có thể gây ảnh hưởng đến độ chính xác và tính ứng dụng của các mô hình đó khi ứng dụng vào thực tế.

Khoảng trống nghiên cứu và những điểm mới của nhóm

Đây là một nghiên cứu đề xuất một giải pháp và ứng dụng trong lĩnh vực “phân loại email theo chủ đề”, kết hợp cả phân loại email rác nhằm đưa ra giải pháp tốt nhất cho đề tài phân loại email. Nghiên cứu cũng thúc đẩy và đóng góp cho lĩnh vực “phân loại email theo chủ đề” vốn chưa có nhiều nghiên cứu nổi bật.

Nghiên cứu sử dụng và so sánh nhiều phương pháp, thuật toán phân loại trong Học máy để xây dựng mô hình phân loại với độ chính xác cao. Trong đó bao gồm thử nghiệm, so sánh 4 thuật toán phân loại (Extra Trees Classifier, Gradient Boosting Classifier, Linear SVC và SGD Classifier) và 2 phương pháp phân loại (đa lớp và đa nhiệm).

Khác với những nghiên cứu trước đó, nghiên cứu của nhóm tập trung vào việc cải thiện quy trình xử lý email trong doanh nghiệp với tính thực tế cao. Nghiên cứu đã hoàn thiện một hệ thống robot và có thể áp dụng ngay vào doanh nghiệp.

Bộ dữ liệu của nhóm được thu thập tại một công ty thực tế (Song Linh) giúp đảm bảo hơn về khả năng ứng dụng của nghiên cứu so với những nghiên cứu trước.

Trên thực tế, các nghiên cứu về RPA tại Việt Nam vẫn còn nhiều hạn chế. Đây là một nghiên cứu mới về lĩnh vực này và đề xuất một ứng dụng thực tế của RPA tại Việt Nam.

## Mục tiêu nghiên cứu

*Về khoa học*

Kết quả của đề tài sẽ đóng góp một phần vào việc đáp ứng xu hướng chuyển đổi số trên toàn cầu, đặc biệt là tại các doanh nghiệp cũng như đóng góp về mặt phương pháp và thực nghiệm vào lĩnh vực phân loại văn bản, một nhánh nghiên cứu của xử lý ngôn ngữ tự nhiên. Tìm hiểu sâu hơn về lĩnh vực học máy (Machine Learning) trong khai phá dữ liệu (Data Mining) và xử lý ngôn ngữ tự nhiên (NLP), đồng thời, thúc đẩy việc nghiên cứu một nền tảng công nghệ mới (RPA) về tự động hóa thông minh (Intelligent Automation).

*Về công nghệ*

Thiết kế một hệ thống ứng dụng công nghệ Machine Learning để xử lý ngôn ngữ tự nhiên (Natural Language Processing), phân loại email rác và phân loại chủ đề email tích hợp trong robot tự động hóa quy trình (Robotic Process Automation) để đọc và chuyển tiếp Email về các phòng chức năng tương ứng trong nội bộ doanh nghiệp.

*Về thực tiễn*

Đề tài hướng tới việc thiết kế một hệ thống giúp tự động nhận dạng, phân loại và chuyển tiếp email đến đúng các phòng ban trong nội bộ doanh nghiệp. Đây là một ý tưởng giúp giải quyết những vấn đề phức tạp trong việc phân loại một lượng lớn email mà doanh nghiệp nhận được trong một ngày. Hệ thống sẽ giúp tối ưu hóa quy trình làm việc, giảm bớt tình trạng sai sót hoặc chậm trễ khi gửi đến các phòng ban, gia tăng tinh thần và hiệu suất trong quy trình làm việc của nhân viên.

## Đối tượng và phạm vi nghiên cứu

* Đối tượng nghiên cứu: Hệ thống phân loại và chuyển tiếp email trong doanh nghiệp.
* Khách thể nghiên cứu: Các email bao gồm email gửi về doanh nghiệp và email spam.
* Phạm vi nghiên cứu: Doanh nghiệp vừa và nhỏ sử dụng email viết bằng tiếng Anh.

## Phương pháp nghiên cứu

*Phương pháp nghiên cứu định tính*: Khảo sát đặc điểm, số lượng và quyền hạn của các phòng ban trong các doanh nghiệp vừa và nhỏ để nắm được thông tin chi tiết nhằm phục vụ cho quá trình phân loại Email và thiết kế hệ thống.

*Phương pháp phân tích và tổng hợp lý thuyết*: Nghiên cứu những tài liệu, sách và báo khoa học liên quan để tìm được mô hình khai phá văn bản phù hợp với độ chính xác cao. Nghiên cứu những công cụ, công nghệ hiện tại để áp dụng vào thiết kế hệ thống.

*Phương pháp nghiên cứu định lượng*: Thống kê dữ liệu và số liệu nhằm đưa ra một kết quả trực quan cho mô hình và hệ thống.

## Ý nghĩa thực tiễn của nghiên cứu

Nghiên cứu đề xuất một hệ thống giúp quy trình nhận và xử lý email trong doanh nghiệp hiệu quả, đơn giản và không chậm trễ. Điều này sẽ giúp khách hàng - đối tác dễ dàng liên lạc với doanh nghiệp và được phản hồi nhanh hơn. Từ đó nâng cao trải nghiệm của khách hàng đối với doanh nghiệp.

Trong thực tế, báo cáo Kinh tế - Xã hội của Học viện Toàn cầu McKinsey (McKinsey Global Institute) chỉ ra rằng nhân viên văn phòng dành tới 28% thời gian làm việc mỗi tuần - tương đương với 2.6 giờ mỗi ngày để đọc và trả lời email. Việc ứng dụng hệ thống tự động vào quy trình xử lý email sẽ giúp nhân viên làm việc năng suất và tiết kiệm thời gian hơn. Từ đó, doanh nghiệp có thể khai thác tối đa nguồn nhân lực của mình ([Chui, Manyika, & Bughin, 2012](#_ENREF_11)).

## Cấu trúc nghiên cứu

Cấu trúc bài báo cáo của nhóm được chia thành 6 chương như sau:

***Chương I: Tổng quan đề tài***

Chương đầu tiên trình bày tổng quan về bối cảnh nghiên cứu, phương pháp nghiên cứu và lý do, ý nghĩa của đề tài.

***Chương II: Cơ sở lý luận***

Trình bày các cơ sở lý thuyết về thư điện tử, tầm quan trọng của thư điện tử trong thời đại hiện nay và thực tiễn về thư điện tử trong doanh nghiệp. Bên cạnh đó, các lý thuyết về xử lý ngôn ngữ tự nhiên, học máy và robot tự động hóa quy trình cũng được tổng hợp nhằm đề ra cơ sở xây dựng đề tài của nhóm.

***Chương III: Mô tả bài toán***

Nêu vấn đề cần giải quyết của đề tài và mô tả đầu vào, đầu ra và quy trình giải quyết bài toán.

***Chương IV: Quy trình thực hiện***

Mô tả chi tiết giải pháp xây dựng hệ thống chuyển tiếp email tự động bao gồm xử lý dữ liệu, mô hình phân loại và quy trình tự động hóa.

***Chương V: Kết quả nghiên cứu***

Trình bày kết quả nghiên cứu và đánh giá độ chính xác của mô hình.

***Chương VI: Kết luận và đề xuất***

Tóm tắt những nội dung đã thực hiện và từ đó đưa ra đề xuất những ứng dụng thực tế cho doanh nghiệp cũng như hướng phát triển đề tài.

# CƠ SỞ LÝ LUẬN

## Tổng quan về thư điện tử (Email)

Khái niệm về thư điện tử (Email)

Ngày nay, cùng với sự phát triển mạnh mẽ của công nghệ và mạng Internet trên khắp thế giới, hệ thống thư điện tử cũng đang dần thay thế hệ thống thư truyền thống ([Mujtaba et al., 2017](#_ENREF_31)). Vậy thư điện tử là gì? Nói một cách đơn giản, thư điện tử (email hay e-mail) là một trong những phương thức giao tiếp gián tiếp giữa những người sử dụng thiết bị điện tử. Một thông điệp sẽ được gửi từ máy tính này sang máy tính khác thông qua mạng máy tính hoặc mạng Internet. Thông qua thư điện tử, người dùng có thể gửi những thông báo khẩn cấp, văn bản quan trọng sử dụng trong công việc hay nhắn tin trao đổi về cuộc sống thường ngày. Các loại thư điện tử điển hình như Gmail, Outlook, Yahoo!, ... được sử dụng làm phương tiện trao đổi chủ yếu trong các tổ chức, doanh nghiệp.

Việc sử dụng thư điện tử mang lại nhiều lợi ích vượt trội hơn so với thư truyền thống chẳng hạn như thay vì viết thư tay hay thông báo thì với email, người gửi chỉ cần đánh máy và gửi qua phần mềm thư điện tử. Đặc biệt, đối với thư điện tử, nếu viết sai thì người dùng có thể xóa hoặc sửa nội dung. Đối với những thông tin quan trọng, cấp bách phải gửi ngay lập tức thì thư điện tử là lựa chọn phù hợp nhất, vừa tiết kiệm được thời gian, chi phí gửi và đảm bảo được tính bảo mật nội dung thông tin được gửi đi. Khi sử dụng email, người gửi không chỉ gửi được cho một người mà còn có thể gửi được cho rất nhiều người ở những nơi khác nhau trong khoảng thời gian ngắn, người dùng còn có thể sử dụng tính năng chuyển tiếp tin nhắn trên phần mềm thư điện tử nhằm thông báo lại thông tin đến một người khác. Thư điện tử chính là phương án lựa chọn phù hợp nhất giúp giải quyết hai vấn đề lớn đó là tiết kiệm thời gian và xóa bỏ ranh giới về địa lý của những người dùng. Với những ưu điểm trên, email đã trở thành một dịch vụ được sử dụng rộng rãi nhất trên Internet.

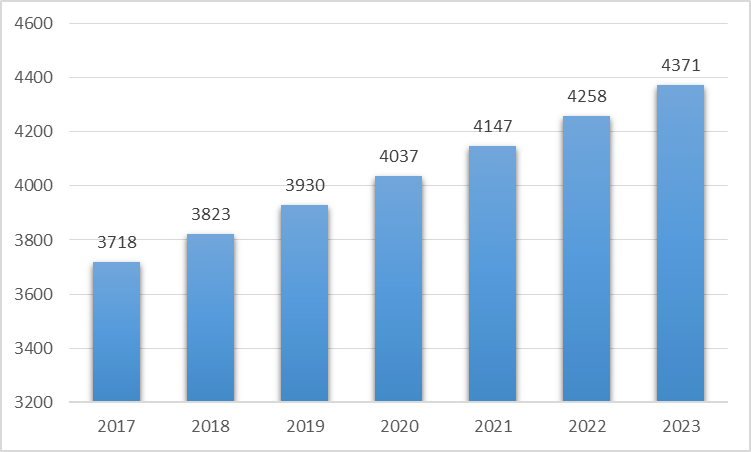
Thư điện tử bao gồm 4 loại: email giới thiệu, email bản tin, email quảng cáo và email văn bản thuần túy ([Duc, 2014](#_ENREF_16); [Mohammadi, Malekian, Nosrati, & Karimi, 2013](#_ENREF_30)). Đối với email giới thiệu, đây là những email được gửi đến người dùng để làm quen và hướng dẫn cách sử dụng sản phẩm. Email bản tin là loại email được gửi thường xuyên đến người dùng, thường chứa các liên kết từ blog hoặc nội dung được chọn đã được xuất bản gần đây. Với email quảng cáo, người dùng được cung cấp một số loại ưu đãi, như giảm giá, dịch vụ miễn phí hoặc mức chiết khấu cho các dịch vụ được quản lý, đây là loại email B2B phổ biến nhất. Cuối cùng là email văn bản thuần túy chỉ chứa văn bản. Chúng không bao gồm hình ảnh, video, tài liệu, đồ họa hoặc tệp đính kèm. Những email này cũng được sử dụng để gửi các cuộc trò chuyện ngẫu nhiên giống như các dịch vụ nhắn tin văn bản khác. Tùy vào mục đích sử dụng mà người dùng thư điện tử sẽ lựa chọn loại email phù hợp.



**Hình 2‑1** Các loại thư điện tử được sử dụng phổ biến hiện nay

Tầm quan trọng của thư điện tử trong doanh nghiệp hiện nay

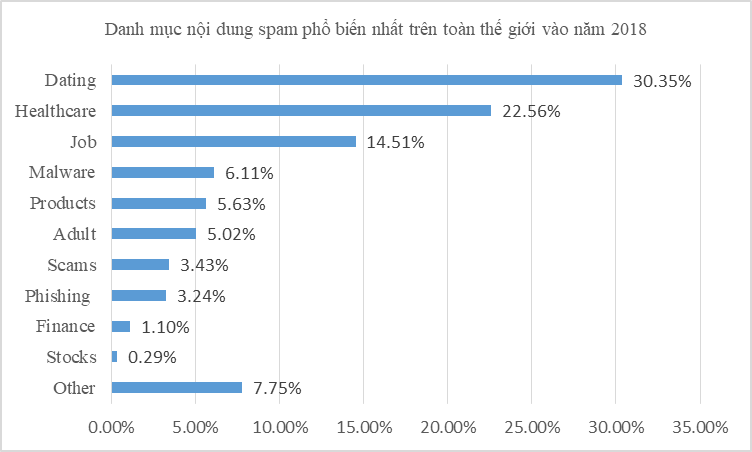
Với hàng triệu thư điện tử được gửi đi mỗi ngày, email đã dần trở thành một phương thức liên lạc phổ biến được nhiều người sử dụng, nó được xem như một kênh liên lạc tương tự như điện thoại bởi tính chất quen thuộc với người dùng Internet ([Dürscheid, Frehner, Herring, Stein, & Virtanen, 2013](#_ENREF_17)). Một nghiên cứu cho thấy, các giao dịch hàng tỷ đô la (trên dưới 5,5 tỷ USD mỗi ngày) được thực hiện thông qua thư điện tử mỗi ngày ([Trịnh, 2014](#_ENREF_41)). Theo hãng nghiên cứu thị trường Statista dự đoán rằng số lượng người dùng email sẽ tăng lên gần 4,4 tỷ người vào năm 2023.



**Hình 2‑2** Thống kê số lượng email được sử dụng mỗi năm (Nguồn: Hãng nghiên cứu thị trường Statista)

Nhận thấy được tầm quan trọng và sự phổ biến của thư điện tử nên mỗi doanh nghiệp hiện nay đều có một hệ thống thư điện tử riêng hay còn được gọi là hệ thống email doanh nghiệp. Khác với các loại email khác, email doanh nghiệp được sử dụng cho nhu cầu công việc của các công ty, doanh nghiệp, tập đoàn. Bất cứ công ty nào khi hoạt động cũng đều cần có một phương thức liên hệ với khách hàng và đối tác. Bên cạnh số điện thoại hay số fax thì email chính là một phương tiện chính để trao đổi. Việc sử dụng email doanh nghiệp sẽ giúp đảm bảo được uy tín của thương hiệu công ty hay thuận tiện cho việc gửi thư quảng cáo của các công ty không bị cho vào hòm thư spam hay thùng rác của các đối tác. Một điểm ưu việt nhất của email doanh nghiệp chính là ở khả năng bảo mật thông tin. Thông tin trong email của một doanh nghiệp là vô cùng quan trọng, một doanh nghiệp sẽ không thể quản lý được nội dung trao đổi của nhân viên đối với người ngoài nếu để họ sử dụng email cá nhân. Việc làm mất thông tin quan trọng của một dự án sẽ khiến cho công ty gặp không ít rắc rối. Việc bắt buộc các nhân viên sử dụng email doanh nghiệp sẽ giúp khắc phục việc rò rỉ thông tin vì nó được trang bị cơ chế kiểm soát cũng như bảo mật thông tin rất tốt. Điều này sẽ không để thông tin có cơ hội bị rò rỉ ra bên ngoài và dễ dàng phát hiện được đối tượng làm mất thông tin. Ngoài ra, sử dụng email doanh nghiệp sẽ giúp hạn chế các thư spam, tin rác, quảng cáo, ... hay bất kỳ sự thay đổi nhân sự nào của công ty cũng đều không ảnh hưởng đến thông tin cũng như các hoạt động trao đổi trước đó của doanh nghiệp. Bộ phận quản lý nhân sự trong công ty sẽ dễ dàng lấy lại email hoặc hủy email nếu có sự thay đổi nhân sự. Ứng dụng này của hệ thống thư điện tử doanh nghiệp nhằm mang lại sự bảo mật tối ưu cho thông tin của công ty, doanh nghiệp.

Sự phát triển nhanh chóng của thư điện tử cũng là một con dao hai lưỡi cho người sử dụng. Lợi dụng những tiện ích của thư điện tử, một số cá nhân, tổ chức đã sử dụng email cho những mục đích không đúng đắn như gửi quá nhiều thư quảng cáo, thư phản động, thậm chí là những đoạn mã độc hại, hình ảnh đồi trụy khác, mà chúng ta gọi đó là thư rác (spam emails) ([Toàn, Lâm, Nghị, & Trung, 2011](#_ENREF_39)). Theo trang web bảo mật Internet SpamLaws.com, các chuyên gia ước tính rằng 58 tỷ email rác sẽ được gửi đi mỗi ngày trong 4 năm tới. Mặc dù Luật chống thư rác đã có hiệu lực trong nhiều năm nay và Quy định chung về bảo vệ dữ liệu đã được thực hiện vào năm 2018, nhưng điều này không có nghĩa là thư rác sẽ được hạn chế hoàn toàn. Trang web bảo mật Internet SpamLaws.com cũng ước tính rằng những email rác này sẽ khiến các doanh nghiệp thiệt hại gần 200 tỷ USD mỗi năm[[3]](#footnote-3). Mặc dù sử dụng email doanh nghiệp có thể hạn chế được thư spam, tin rác, quảng cáo, ... nhưng các cá nhân, tổ chức vẫn có những thủ thuật tinh vi để gửi những thư rác đó đến với không chỉ doanh nghiệp mà tất cả người dùng email mà không ai có thể điều tra được thông tin của chúng. Trước những dự báo đó, cho thấy rằng email spam sẽ vẫn là một vấn đề của các doanh nghiệp trong nhiều năm tới.



**Hình 2‑3** Danh mục nội dung spam phổ biến nhất trên toàn thế giới năm 2018 (Nguồn: Hãng nghiên cứu thị trường Statista)

Hiện nay, tồn tại chủ yếu hai cách thức xử lý email trong hệ thống thư điện tử doanh nghiệp. Một là, các phòng ban trong một doanh nghiệp sẽ được tạo một tài khoản email chuyên biệt cho từng phòng ban. Khi cần liên hệ xử lý từng nghiệp vụ cụ thể, người dùng phải gửi đến tài khoản email của từng phòng ban. Chẳng hạn, với những email về mua hàng thì phải gửi đến phòng Kinh doanh (Sales) còn với email ý kiến về sản phẩm/ dịch vụ thì phải gửi về phòng Chăm sóc khách hàng (Customer care). Hệ thống này mặc dù được đánh giá rằng tạo thuận lợi cho công ty dễ dàng xử lý các công việc theo đúng chuyên môn của mình nhưng sẽ gây khó khăn cho khách hàng, bởi vì sẽ có những khách hàng không phân biệt rõ được chuyên môn hay không biết tài khoản email chuyên biệt của từng phòng ban dẫn đến thông tin vừa bị gửi sai, vừa mất thời gian của cả đôi bên. Ngoài ra, còn một cách thức xử lý khác chính là các phòng ban trong một doanh nghiệp vẫn sẽ được cấp tài khoản email, tuy nhiên điểm khác biệt là khi có nhu cầu, các khách hàng hay đối tác không cần gửi email đến từng phòng ban mà chỉ cần gửi đến một địa chỉ email chung của doanh nghiệp. Sau đó, bộ phận xử lý thông tin của doanh nghiệp sẽ chuyển tiếp những thông tin đó đến từng phòng ban xử lý theo đúng chuyên môn của họ. Hệ thống này sẽ giúp khách hàng gửi thông tin đến đúng phòng ban mà không mất thời gian để tìm kiếm tài khoản email phòng ban đó hay lo lắng mình có gửi nhầm phòng ban hay không. Cách làm này cũng có nhiều bất lợi như đòi hỏi doanh nghiệp phải có một bộ phận xử lý phân loại email về các phòng ban, nhân sự của bộ phận này cũng phải nắm rõ chuyên môn của từng phòng ban để tránh việc phân loại email sai. Điều đó khiến cho doanh nghiệp phải tốn thêm một khoản chi phí cho bộ phận phân loại email. Chính vì thế, một hệ thống tự động hóa phân loại email về các phòng ban trong doanh nghiệp là một trong những ý tưởng thích hợp cần được chú trọng phát triển.

## Lý thuyết xử lý ngôn ngữ tự nhiên (NLP)

Khái niệm về xử lý ngôn ngữ tự nhiên (NLP)

Ngôn ngữ tự nhiên

Ngôn ngữ được coi là một phần không thể thiếu trong cuộc sống của con người. Nhờ vào ngôn ngữ mà con người có thể giao tiếp và kết nối với nhau. Ngôn ngữ được tồn tại dưới nhiều hình thức khác nhau như âm thanh (ngôn ngữ nói) hoặc văn bản (ngôn ngữ viết). Trong ngôn ngữ học, một ngôn ngữ tự nhiên (Natural language) được hiểu là bất cứ ngôn ngữ nào được phát sinh, được tạo ra mà không trải qua bất cứ một suy nghĩ nào trước đó trong não bộ của con người. Mặc dù có nhiều định nghĩa khác nhau giữa các học giả, nhưng ngôn ngữ tự nhiên có thể định nghĩa theo nghĩa rộng là đối nghịch với ngôn ngữ mạng tính nhân tạo hay ngôn ngữ được xây dựng ([Fernandes, 2008](#_ENREF_18)). Định nghĩa các ngôn ngữ tự nhiên thường phát biểu hay ngụ ý là ngôn ngữ mà bất kỳ đứa trẻ nào cũng có thể học và được phát triển thông qua sử dụng ngôn ngữ và nói chứ không phải theo chỉ dẫn.

Sự đa dạng chính là đặc điểm dễ nhận thấy nhất của ngôn ngữ tự nhiên, nó tồn tại trong ngôn ngữ để giao tiếp và lời nói, cử chỉ, ký hiệu hay xúc giác hoặc tồn tại dưới dạng văn bản, thư từ trong công việc hằng ngày. Ngôn ngữ tự nhiên còn mang những đặc điểm khác biệt so với các loại ngôn ngữ được xây dựng trên những mục đích cụ thể của con người, có thể kể đến như ngôn ngữ lập trình, ngôn ngữ nghiên cứu theo logic, ...

**Bảng 2‑1** So sánh ngôn ngữ tự nhiên và ngôn ngữ máy

|  |  |  |
| --- | --- | --- |
| Tiêu chí so sánh | Ngôn ngữ tự nhiên | Ngôn ngữ máy tính (Lập trình) |
| Số lượng từ vựng | Rất nhiều | Hạn chế |
| Đối tượng sử dụng | Con người | Máy tính |
| Bản chất | Mơ hồ | Rõ ràng |
| Kiểu dữ liệu | Phi cấu trúc | Cấu trúc |

Ngôn ngữ tự nhiên thuộc kiểu “Dữ liệu phi cấu trúc” - là dữ liệu không tuân thủ theo mô hình dữ liệu, được tổ chức theo cách không xác định trước và không có cấu trúc dễ nhận dạng cho các chương trình máy tính sử dụng dễ dàng. Một trong những mong muốn mãnh liệt của các nhà khoa học máy tính nói chung và trí tuệ nhân tạo nói riêng là xây dựng thành công các hệ thống, chương trình máy tính có khả năng giao tiếp với con người thông qua ngôn ngữ tự nhiên thay vì ngôn ngữ lập trình hay ngôn ngữ máy bậc thấp. Dù mang những đặc điểm tự nhiên, thế nhưng khi các nhà khoa học nghiên cứu về lĩnh vực này lại cho rằng nó không hề đơn giản, cần một quy trình xử lý hiệu quả.

Xử lí ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (Natural Language Processing) là một lĩnh vực liên ngành của ngôn ngữ học, khoa học máy tính và trí tuệ nhân tạo liên quan đến sự tương tác giữa máy tính và ngôn ngữ của con người, đặc biệt là cách lập trình, phát triển các thuật toán, xây dựng các chương trình máy tính có khả năng phân tích và xử lý lượng lớn dữ liệu ngôn ngữ tự nhiên. Mục tiêu là một máy tính có khả năng "hiểu" nội dung của tài liệu, bao gồm cả sắc thái ngữ cảnh của ngôn ngữ trong đó. Sau đó, công nghệ có thể trích xuất chính xác thông tin và thông tin chi tiết có trong tài liệu cũng như tự phân loại và sắp xếp tài liệu.

Việc xử lý ngôn ngữ tự nhiên có thể được chia thành hai nhánh lớn, không hoàn toàn độc lập, bao gồm xử lý tiếng nói (Speech Processing) và xử lý văn bản (Text Processing). Nó nắm vai trò rất quan trọng trong việc phân tích đầy đủ dữ liệu văn bản và giọng nói một cách hiệu quả. Công nghệ này có thể giải quyết những nét khác biệt trong phương ngữ, tiếng lóng và điểm bất thường về ngữ pháp thường thấy trong các cuộc hội thoại hàng ngày. Nhờ vào lượng dữ liệu lớn, NLP có thể giúp máy tính giao tiếp với con người bằng ngôn ngữ riêng của chúng và mở rộng các ứng dụng khác liên quan đến ngôn ngữ. Ví dụ, NLP giúp máy tính có thể đọc văn bản, nghe lời nói, diễn giải văn bản, đo lường cảm xúc và xác định phần nào là quan trọng trong văn bản. Ngôn ngữ tự nhiên là dữ liệu phi cấu trúc nên NLP đóng vai trò rất quan trọng trong việc giải quyết sự mơ hồ của ngôn ngữ, sắp xếp nguồn dữ liệu phi cấu trúc cao một cách có cơ cấu và mã hóa dữ liệu cho các ứng dụng cuối cùng của NLP như kiểm tra lỗi chính tả, tóm tắt văn bản. Hiện nay, các doanh nghiệp cũng đã sử dụng công nghệ này như một tác vụ tự động, chẳng hạn như:

* Xử lý, phân tích và lưu trữ các tài liệu lớn;
* Phân tích phản hồi của khách hàng;
* Chạy chatbot cho dịch vụ khách hàng tự động;
* Phân loại và trích xuất văn bản.

Việc tích hợp NLP trong các ứng dụng tương tác trực tiếp với khách hàng sẽ giúp giao tiếp hiệu quả hơn với họ. Ví dụ, một chatbot phân tích và phân loại các truy vấn của khách hàng, trả lời tự động các câu hỏi thường gặp và chuyển các truy vấn phức tạp đến bộ phận hỗ trợ khách hàng. Khả năng tự động hóa này góp phần giảm bớt chi phí, giúp nhân viên khỏi mất thời gian vào những truy vấn trùng lặp và cải thiện mức độ hài lòng của khách hàng.

Một số ứng dụng thực tế sử dụng công nghệ Xử lý ngôn ngữ tự nhiên (NLP) như Nhận dạng tiếng nói; Tổng hợp tiếng nói; Truy xuất thông tin; Trích chọn thông tin; Trả lời câu hỏi; Tóm tắt văn bản tự động; Chatbot; Dịch máy (Machine translation); Kiểm tra chính tả tự động; Khai phá dữ liệu (Data mining); Phân tích tình cảm (khách hàng), ...

Quy trình xử lý ngôn ngữ tự nhiên

Thông thường, quy trình NLP có thể chia thành 4 giai đoạn chính: Tiền xử lý dữ liệu; Biểu diễn dữ liệu; Đào tạo mô hình; Đánh giá kết quả và triển khai dự án ([Kang, Cai, Tan, Huang, & Liu, 2020](#_ENREF_23)).

Tiền xử lý dữ liệu nhằm mục đích thu được dữ liệu “sạch” để cải thiện hiệu quả và độ chính xác của mô hình bằng cách loại bỏ các ký hiệu vô nghĩa, kiểm tra lỗi chính tả, mã hóa dữ liệu, loại bỏ các từ dừng và đơn giản hóa về dạng gốc. Tuy nhiên, vẫn cần có các phương pháp hiệu quả để biểu diễn dữ liệu.

Sau khi tiền xử lý văn bản, các nhà nghiên cứu phải chọn cách tốt nhất để thể hiện các dữ liệu ở dạng mà máy tính có thể hiểu được và thực hiện tính toán. Biểu diễn dữ liệu từ văn bản/ hình ảnh về dạng số có thể được thực hiện bằng cách vectơ hoặc ma trận hóa.

Từ kết quả biểu diễn này, có thể áp dụng các thuật toán như phân loại, phân tích tình cảm và trích xuất chủ đề, ... để đào tạo một mô hình nhằm mục đích giải quyết vấn đề thực tế. Sau khi đào tạo mô hình, các nhà nghiên cứu sau đó cần đánh giá mô hình để đảm bảo rằng nó có khả năng khái quát hóa tối ưu cho các tập đoàn khác.

## Lý thuyết học máy

Khái niệm về học máy

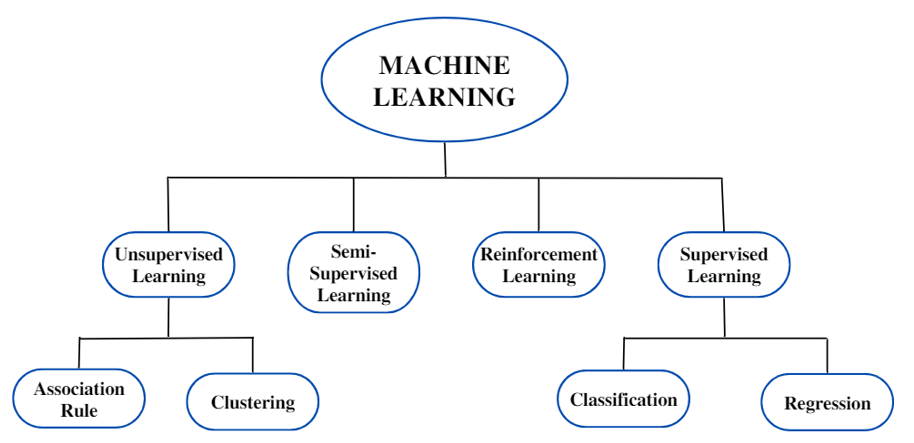
Với sự phát triển vượt bật của trí tuệ nhân tạo, con người đã có những phát minh khoa học và kỹ thuật, tạo ra những máy móc có có trí thông minh bằng cách lập trình máy tính. Từ đó tạo khả năng đẩy nhanh tốc độ tự động hóa, trong đó phải nhắc đến lĩnh vực Học máy.

Học máy (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể ([Mahesh, 2020](#_ENREF_27)). Ví dụ như máy có thể “học” cách phân loại thư điện tử xem có phải thư rác (spam email) hay không và tự xếp thư vào các thư mục tương ứng.

Các thuật toán học máy đang được sử dụng để phân tích dữ liệu lớn (Big Data) nhằm hỗ trợ dự đoán xu hướng hoặc sự kiện thị trường, ví dụ như dự đoán kết quả cuộc bầu cử chính trị. Các thuật toán nhận dạng hình ảnh hiện có thể phân tích dữ liệu từ các hệ thống chụp ảnh vệ tinh để cung cấp thông tin về số lượng khách hàng tại các bãi đậu xe của cửa hàng bán lẻ, hoạt động vận chuyển và cơ sở sản xuất, và sản lượng nông nghiệp, ... Những thông tin này sẽ cung cấp dữ liệu đầu vào cho các mô hình định giá hoặc các mô hình kinh tế. Có thể nói, trong tương lai, học máy sẽ phủ khắp mọi lĩnh vực của cuộc sống, miễn là nơi đó có dữ liệu.

Phân loại máy học

Khi thiết kế và xây dựng hệ thống học máy, cần quan tâm đến hai yếu tố chính: dữ liệu và dạng biểu diễn (ví dụ dạng cây, dạng nơ ron, ...). Tùy vào những dạng biểu diễn khác nhau sẽ có những loại máy học khác nhau. Bi và các cộng sự (2019) chỉ ra rằng, theo phương pháp học, học máy có thể chia thành 4 phần chính là: Học có giám sát, Học bán giám sát, Học không giám sát và Học tăng cường ([Bi, Goodman, Kaminsky, & Lessler, 2019](#_ENREF_6)).



**Hình 2‑4** Phân loại học máy

Học có giám sát (Supervised Learning) là dạng học máy dự đoán đầu ra cho một tập dữ liệu mới dựa trên tập dữ liệu huấn luyện cho trước mà trong đó mỗi mẫu dữ liệu đều đã được biết trước giá trị đầu ra hay giá trị đích. Học có giám sát được chia nhỏ ra thành hai loại chính là Phân loại (Classification) và Hồi quy (Regression). Phân loại là trường hợp các giá trị đầu ra của dữ liệu đầu vào được chia thành một số hữu hạn lớp (miền giá trị rời rạc), ví dụ xác định hình ảnh con vật là chó hay mèo. Hồi quy là trường hợp giá trị đầu ra không được chia thành nhóm mà là một giá trị thực cụ thể (miền giá trị liên tục), ví dụ như dự báo giá cổ phiếu, giá nhà, ... kết quả là một số thực.

Học không giám sát (Unsupervised Learning) là dạng học máy dự đoán đầu ra cho duy nhất một dữ liệu mới dựa trên tập dữ liệu huấn huyện trong đó tất các các mẫu dữ liệu đều chưa được biết trước giá trị đầu ra hay giá trị đích. Mục tiêu của thuật toán này không phải là tìm đầu ra chính xác mà sẽ hướng tới tìm ra một cấu trúc hoặc sự liên hệ trong dữ liệu để thực hiện một công việc nào đó như gom cụm hay giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán. Học máy không giám sát được chia thành hai loại chính là Phân cụm (Clustering) và Kết hợp (Association Rule). Phân cụm là phân nhóm toàn bộ dữ liệu thành các nhóm/cụm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Luật kết hợp là khi muốn tìm ra một quy luật dựa trên nhiều dữ liệu cho trước, ví dụ những khách hàng mua mặt hàng này sẽ mua thêm mặt hàng kia.

Học bán giám sát (Semi-Supervised Learning) là dạng học máy dự đoán mà khi tập dữ liệu đầu vào là hỗn hợp các mẫu có nhãn và không có nhãn, trong đó số lượng có nhãn chỉ chiếm một phần nhỏ. Mục tiêu của học máy bán giám sát là hiểu cách kết hợp dữ liệu có nhãn và không có nhãn có thể thay đổi hành vi học như thế nào, và tạo ra thuật toán để tận dụng sự kết hợp này.

Học tăng cường (Reinforcement Learning) là dạng học máy giúp cho hệ thống tự động xác định hành vi dựa trên tình huống để đạt được lợi ích cao nhất. Đối với dạng học này, kinh nghiệm không được cho trực tiếp dưới dạng đầu vào hay đầu ra mà hệ thống sẽ nhận kết quả của một chuỗi hành động nào đó. Thuật toán cần học cách hành động để cực đại hóa kết quả đó.

Quy trình xây dựng hệ thống học máy

Một quy trình xây dựng hệ thống học máy thường có 3 bước chính: Thu thập, xử lý dữ liệu; lựa chọn thuật toán và tiến hành huấn luyện cho mô hình; kiểm nghiệm thực tế và đánh giá.

Trong mỗi hệ thống học máy thì dữ liệu được coi là quan trọng nhất, quyết định khả năng dự đoán của hệ thống là đúng hay không. Chính vì thế, việc thu thập dữ liệu đầu vào là vô cùng quan trọng. Tuy nhiên, không phải cứ nhiều dữ liệu là thuật toán sẽ chạy tốt, mà cần có một sự đa dạng, chính xác và khái quát từ thực tế. Sau khi thu thập khối lượng dữ liệu vừa đủ lớn thì cần thực hiện xử lý dữ liệu bao gồm việc làm sạch dữ liệu và trích xuất các đặc trưng để cung cấp cho mô hình.

Khi đã có dữ liệu, ta tiến hành chọn thuật toán và tiến hành huấn luyện cho hệ thống học máy (training). Bởi vì có rất nhiều thuật toán học máy nên người nghiên cứu sẽ phải lựa chọn thuật toán phù hợp với bài toán mà mình cần giải quyết, có thể kết hợp nhiều thuật toán và phương pháp khác nhau để tạo ra mô hình phù hợp. Dữ liệu được xử lý sẽ được chia làm 2 phần: Phần để huấn luyện (training data) và phần để kiểm tra (testing data), tùy theo mô hình để chia tỷ lệ thích hợp.

Sau khi thử nghiệm, cần đưa mô hình ra sử dụng trong kiểm nghiệm thực tiễn. Từ đó phát hiện các thiếu sót như dữ liệu thực tế khác biệt, mô hình hoạt động không phù hợp hay thời gian chạy quá lâu, ... Từ đó để tiến hành bổ sung, chỉnh sửa và hoàn thiện mô hình.

## Ứng dụng học máy trong xử lý ngôn ngữ tự nhiên

Trong những năm gần đây, máy học ([Plisson, Lavrac, & Mladenic](#_ENREF_34)) đang trở thành một phần không thể thiếu trong quá trình xử lý ngôn ngữ tự nhiên. Một số giải pháp học máy kết hợp NLP như: Dịch thuật tự động ([Zong & Hong, 2018](#_ENREF_44)), Nhận dạng giọng nói ([Trilla, 2009](#_ENREF_40)), Phân tích tình cảm ([Mejova, 2009](#_ENREF_29)), Chatbot ([Shah, Lahoti, & Lavanya, 2017](#_ENREF_37)), Phân loại văn bản tự động ([Dalal & Zaveri, 2011](#_ENREF_12)), ... Những công nghệ này sẽ giúp các tổ chức, doanh nghiệp phân tích dữ liệu, khai phá thông tin, tự động hóa quy trình và đạt được lợi thế cạnh tranh.

Các ứng dụng của học máy trong xử lý ngôn ngữ tự nhiên chủ yếu được thực hiện dựa trên nguyên lý khai phá dữ liệu (Data Mining). Đây là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu ([Chakrabarti et al., 2006](#_ENREF_8); [Hastie, Tibshirani, Friedman, & Friedman, 2009](#_ENREF_21)). Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp ([Chakrabarti et al., 2006](#_ENREF_8)). Khi mẫu được chọn đủ mạnh sẽ càng làm tăng độ chính xác của việc đưa ra các dự đoán của dữ liệu trong tương lai.

Một nhánh nhỏ của khai phá dữ liệu chính là khai phá văn bản (Text Mining), là quá trình xử lý một lượng lớn văn bản không cấu trúc thành định dạng cấu trúc văn bản để có thể xác định các khái niệm, mẫu, chủ đề, từ khóa và các thuộc tính khác trong dữ liệu. Để khai thác văn bản người ta sử dụng nhiều kỹ thuật phân tích như Naïve Bayes, Support Vector Machine (SVM) và các thuật toán học sâu khác. Khai thác và phân tích văn bản giúp các tổ chức tìm thấy những hiểu biết kinh doanh tiềm năng có giá trị trong tài liệu công ty, email khách hàng, nhật ký trung tâm cuộc gọi, nhận xét khảo sát nguyên văn, bài đăng trên mạng xã hội, hồ sơ y tế và các nguồn dữ liệu dựa trên văn bản khác. Ngày càng nhiều, các khả năng khai thác văn bản cũng được tích hợp vào các chatbot AI và tác nhân ảo mà các công ty triển khai để cung cấp phản hồi tự động cho khách hàng như một phần của hoạt động tiếp thị, bán hàng và dịch vụ khách hàng của họ.

## Lý thuyết tự động hóa quy trình bằng robot

Khái niệm tự động hóa quy trình bằng robot (RPA)

Trước sự phát triển mạnh mẽ của công nghệ, nhu cầu ứng dụng kỹ thuật hiện đại vào quy trình sản xuất ngày càng cao, đặc biệt là nhu cầu ứng dụng tự động hóa. Tự động hóa quy trình bằng robot (Robotic Process Automation) là một công nghệ phần mềm có thể mô phỏng cách con người tương tác với phần mềm để thực hiện các tác vụ có khối lượng lớn, có thể lặp lại. Thông qua việc thu thập các dữ liệu, RPA sẽ mô phỏng thao tác lặp đi lặp lại thường xuyên và thay thế con người xử lý các tác vụ kỹ thuật số như diễn giải, kích hoạt hay giao tiếp với ứng dụng và hệ thống.

Việc dùng RPA trong một số công đoạn giúp giảm chi phí cho nguồn nhân lực, tiết kiệm thời gian, cũng như hạn chế những rủi ro về sai sót trong quá trình sản xuất, kinh doanh. Nhân viên trong các công ty sẽ không phải tập trung vào nhiều quá trình dư thừa, chỉ cần lo những công việc có khả năng phát triển công việc cao hơn. Một điểm hấp dẫn khác của phần mềm RPA là các đơn vị kinh doanh có thể triển khai nó mà không cần phải tìm hiểu các công cụ mới hoặc yêu cầu các nhóm công nghệ thông tin hỗ trợ và không cần thay đổi cơ sở hạ tầng công nghệ thông tin cơ bản của tổ chức.

Khi được kết hợp với AI và học máy, RPA có thể nắm bắt thêm ngữ cảnh từ nội dung mà nó đang làm việc bằng cách đọc văn bản hoặc chữ viết tay thông qua nhận dạng ký tự quang học (OCR), trích xuất các thực thể như tên, điều khoản hóa đơn hoặc địa chỉ bằng cách sử dụng xử lý ngôn ngữ tự nhiên (NLP), và nắm bắt thêm ngữ cảnh từ hình ảnh, chẳng hạn như tự động ước tính thiệt hại do tai nạn trong ảnh yêu cầu bảo hiểm. Đối với các doanh nghiệp, RPA có thể mang lại giá trị ngay lập tức cho các quy trình kinh doanh cốt lõi bao gồm bảng lương nhân viên, thay đổi trạng thái nhân viên, tuyển dụng và giới thiệu nhân viên mới, các khoản phải thu và phải trả, xử lý hóa đơn, quản lý hàng tồn kho, tạo báo cáo, cài đặt phần mềm, di chuyển dữ liệu và nhà cung cấp trên - nội trú, v.v. ([Madakam, Holmukhe, & Jaiswal, 2019](#_ENREF_26)).

Quy trình hoạt động của RPA

Nguyên lý hoạt động cốt lõi của RPA chính là sao chép cách con người thực hiện một quy trình lặp đi lặp lại nhiều lần. Các bot RPA đơn giản nhất có thể được tạo bằng cách ghi lại các lần nhấp và tổ hợp phím khi người dùng tương tác với một ứng dụng. Khi xảy ra sự cố, người dùng có thể chỉ cần xem cách bot kết nối với ứng dụng và xác định các bước cần tinh chỉnh. Những bản ghi cơ bản này đóng vai trò là khuôn mẫu để xây dựng các bot mạnh mẽ hơn có thể thích ứng với những thay đổi về kích thước màn hình, bố cục hoặc quy trình công việc. Một số công cụ RPA có thể sử dụng các bản ghi này để tạo bot RPA kết hợp việc ghi lại quy trình công việc hiện có, sau đó tự động tạo một quy trình tự động hóa.

Để xây dựng hệ thống RPA thì cần sự hỗ trợ từ các công cụ như UiPath, Automation Anywhere, Blue Prism, Microsoft Power Automate, akaBot.... Tùy theo quy mô dự án mà lựa chọn công cụ hỗ trợ phù hợp. Các công cụ RPA cũng có thể được kết nối với các mô-đun AI có các khả năng như OCR, thị giác máy, hiểu ngôn ngữ tự nhiên hoặc công cụ quyết định, từ đó tạo thành một hệ thống tự động hóa quy trình thông minh.

# MÔ TẢ BÀI TOÁN

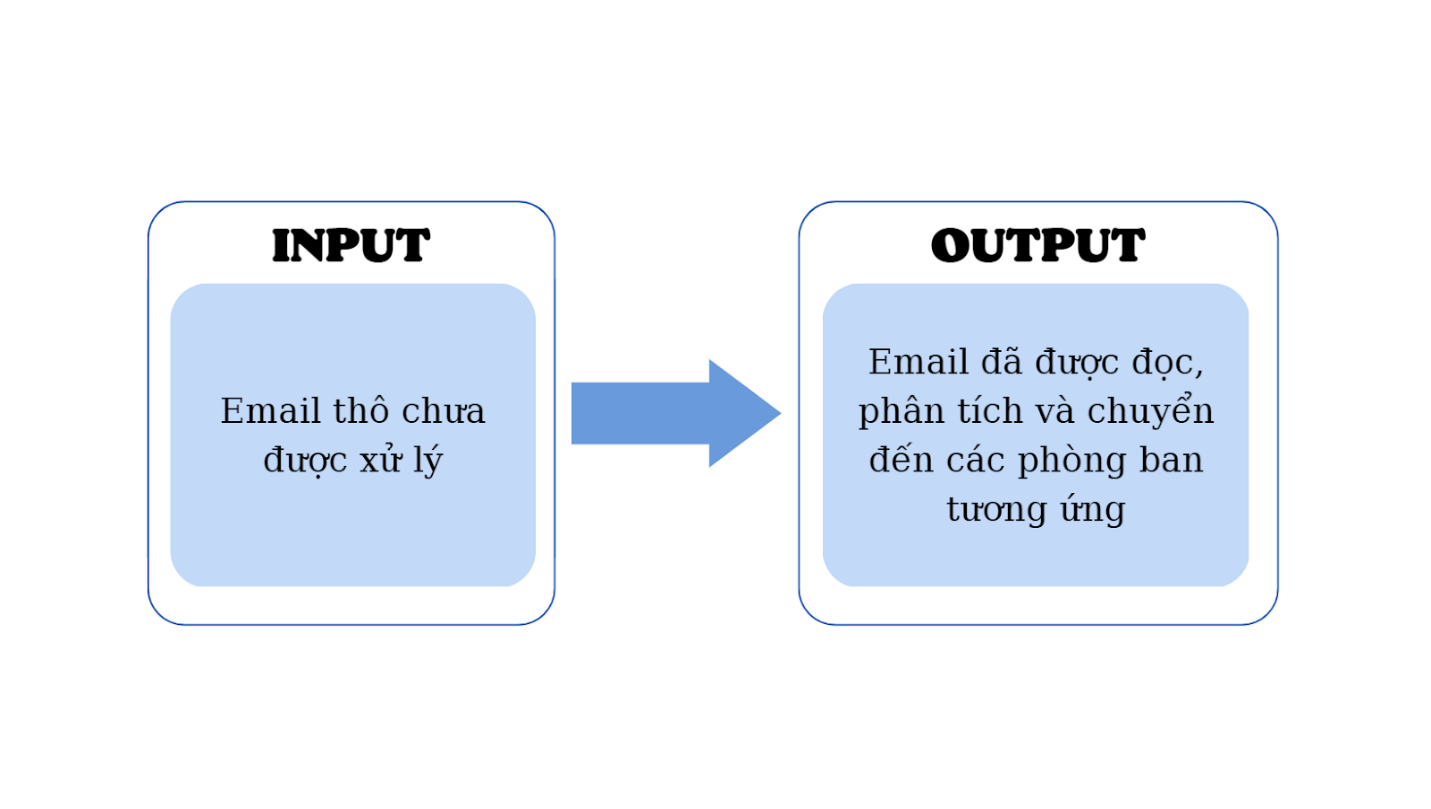
## Đặt vấn đề

Email được xem như là một trong những kênh trao đổi thông tin quan trọng được con người thường xuyên sử dụng bởi tính hiệu quả và tiện lợi. Đối với các doanh nghiệp, email được dùng như một phương thức giao tiếp giữa công ty với khách hàng và đối tác, chúng thường được tổ chức và quản lý thủ công bởi bộ phận văn thư. Tuy nhiên, cùng với sự phát triển nhanh chóng của Internet, số lượng thư điện tử mỗi ngày doanh nghiệp nhận được ngày càng nhiều, trong khi đó tốc độ xử lý thủ công của nhân viên văn thư không thể theo kịp. Hơn nữa, đặc điểm nổi bật của việc phân loại email là sự đa dạng của chủ đề văn bản và tính đa chủ đề của văn bản. Nếu do con người thực hiện, tính đa chủ đề của văn bản sẽ làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan và dễ bị nhập nhằng khi xử lý. Thực tế rằng, một email liên quan đến phòng Đối ngoại cũng có thể gửi đến phòng Kinh doanh nếu liên quan đến các nghiệp vụ mua bán. Bài toán đặt ra cho doanh nghiệp là làm thế nào để phân loại và xử lý một lượng lớn thông tin với thời gian và chi phí bỏ ra là ít nhất.

## Đề xuất giải pháp cho bài toán

Với đề tài “Thiết kế hệ thống RPA tích hợp NLP để thực hiện tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”, các yêu cầu cần phải giải quyết là:

* Tìm ra thuật toán phân loại văn bản phù hợp cho bài toán phân loại email trong doanh nghiệp.
* Thiết kế một hệ thống RPA hỗ trợ xử lý các nghiệp vụ: Nhận và đọc dữ liệu từ email, phân loại email theo các phòng ban phù hợp, chuyển tiếp email đến các phòng ban tương ứng.



**Hình 3‑1** Mô tả kết quả thực quy trình thực hiện bài toán

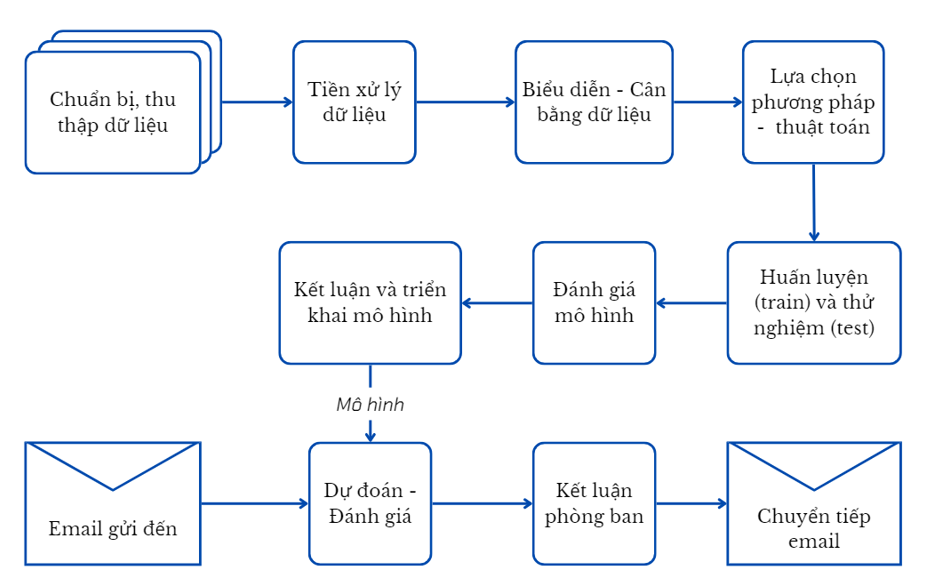
Quy trình nghiên cứu bao gồm 2 nội dung chính:

Thứ nhất là nghiên cứu mô hình phân loại phù hợp:

* Tiền xử lý, biểu diễn và cân bằng dữ liệu;
* Tìm hiểu các phương pháp phân loại và thuật toán khả thi;
* Huấn luyện và thử nghiệm các thuật toán;
* Kết luận và triển khai mô hình phù hợp.

Thứ hai là xây dựng hệ thống tự động hóa RPA:

* Nhận và đọc dữ liệu từ email.
* Chuyển dữ liệu email vào Python, sử dụng mô hình Text Classification để phân loại.
* Dựa vào kết quả phân loại email để tự động chuyển tiếp email đó đến phòng ban hoặc yêu cầu phân loại email thủ công của con người (human in loop).



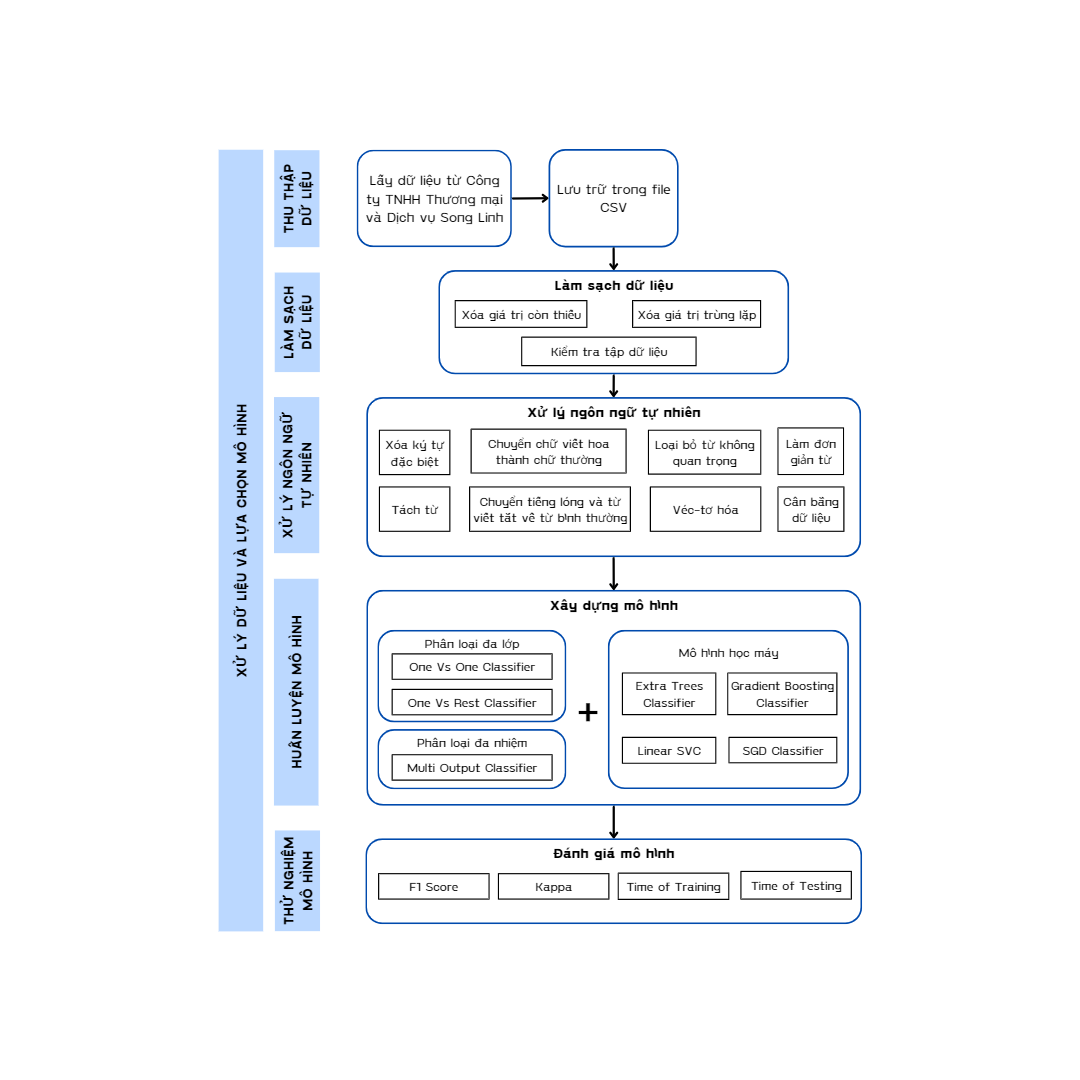
**Hình 3‑2** Quy trình nghiên cứu tổng quan

# QUY TRÌNH THỰC HIỆN

## Phân tích dữ liệu và đề xuất các mô hình

Tổng quan mô hình nghiên cứu phân loại văn bản

Đối với bài toán phân loại email trong doanh nghiệp, nhóm chia mô hình thành 5 phần: Thu thập dữ liệu, Làm sạch dữ liệu, Xử lý ngôn ngữ tự nhiên (NLP), Huấn luyện mô hình, Đánh giá thử nghiệm mô hình.



**Hình 4‑1** Quy trình nghiên cứu mô hình phân loại văn bản

Đầu tiên là bước thu thập dữ liệu. Nghiên cứu sử dụng dữ liệu được trích xuất từ bộ dữ liệu email của Công ty TNHH Thương mại và Dịch vụ Song Linh, một công ty chuyên đấu thầu các thiết bị kỹ thuật và công nghệ tại thành phố Vũng Tàu, tỉnh Bà Rịa - Vũng Tàu, Việt Nam. Dữ liệu sau khi thu thập được lưu trữ trong file CSV.

Tiếp theo, thực hiện kiểm tra tập dữ liệu và tiến hành tiền xử lý bằng cách loại bỏ các dữ liệu bị thiếu, các giá trị ngoại lệ hay những giá trị bị trùng lặp, ... Sau đó, sẽ tiến hành kiểm tra lại dữ liệu để chuẩn bị cho giai đoạn xử lý ngôn ngữ tự nhiên.

Ở bước 3, nhóm sử dụng thư viện “nltk” kết hợp cùng thư viện “contractions” để thực hiện tiền xử lý dữ liệu. Quá trình này sẽ giúp loại bỏ các ký tự đặc biệt, các từ có ý nghĩa không quan trọng, chuyển tất cả chữ viết hoa thành chữ viết thường, chuyển tiếng lóng và những từ viết tắt về ngôn ngữ bình thường, làm đơn giản từ, khiến nó trở về thể nguyên bản bằng cách sử dụng bộ từ điển.

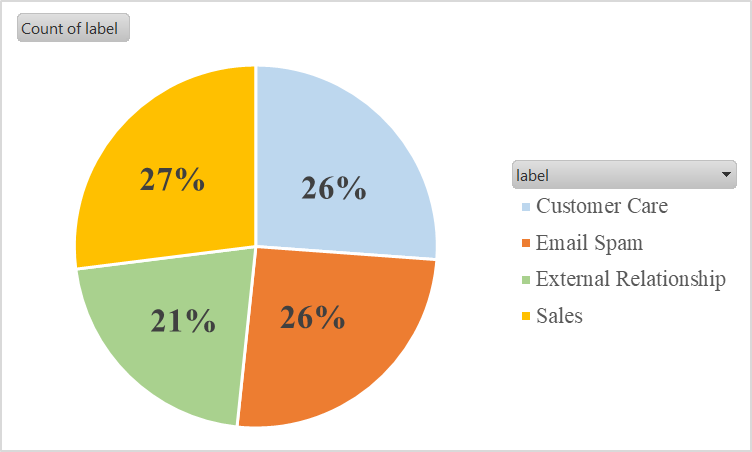
Dữ liệu sẽ được đưa vào mô hình huấn luyện sau khi đã được làm sạch. Trước đó, nhóm nghiên cứu so sánh độ chính xác của phương pháp phân loại đa nhãn (One-Vs-One Classifier; One-Vs-Rest Classifier) và phân loại đa nhiệm (Multi Output Classifier) để lựa chọn phương pháp phù hợp nhất. Thực hiện xây dựng mô hình thuật toán phân loại, nhóm sử dụng thư viện “Scikit learn” để vector hóa dữ liệu và thư viện “imblearn” để cân bằng dữ liệu đưa vào huấn luyện mô hình. Các thuật toán máy học được sử dụng trong nghiên cứu này là Linear SVC, SGD Classifier, Gradient Boosting Classifier, và Extra Trees Classifier.

Bước cuối cùng là nhóm tiến hành đánh giá cho từng mô hình thông qua điểm F1-score và hệ số Cohen’s Kappa, thời gian huấn luyện và thời gian dự đoán rồi đưa ra kết luận, lựa chọn mô hình hiệu quả nhất.

Mô tả dữ liệu

Để xây dựng mô hình phân loại và chuyển tiếp email đến các phòng ban trong doanh nghiệp một cách hiệu quả, nghiên cứu đã sử dụng bộ dữ liệu riêng tư từ 3 phòng ban: Đối ngoại, Bán hàng và Chăm sóc khách hàng của Công ty TNHH Thương mại và Dịch vụ Song Linh với 4270 email từ tháng 02/2022 đến tháng 12/2022 về các sản phẩm linh kiện điện tử, đấu thầu, lắp ráp, sửa chữa, trung gian mua bán máy móc thiết bị. Ngoài ra, nhóm nghiên cứu sử dụng một tập dữ liệu thư rác từ công ty Enron với 1462 email được đăng tải vào tháng 12/2001.

Dữ liệu được nhóm sử dụng để nghiên cứu là phần nội dung của mỗi email, không bao gồm các tệp đính kèm hay hình ảnh. Tập dữ liệu được lưu trữ trong file CSV với 4 nhãn tượng trưng cho 3 phòng ban và 1 nhãn thư rác:



**Hình 4‑2** Số lượng các nhãn trong bộ dữ liệu nghiên cứu

Bộ dữ liệu trên được lưu trữ với hình thức gồm 2 cột: nội dung và nhãn dán (nhãn dán là số từ 0 đến 3 dựa trên 4 loại email khác nhau).

**Bảng 4‑1** Hình thức lưu trữ bộ dữ liệu

|  |  |
| --- | --- |
| Nhãn | Đại diện |
| 0 | Customer Care Department |
| 1 | External Relationship Department |
| 2 | Sales Department |
| 3 | Spam Email |

Làm sạch dữ liệu

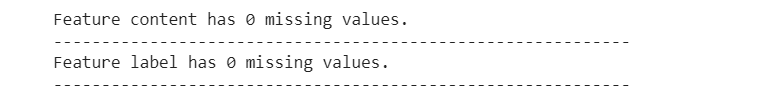
Tập dữ liệu ban đầu được trích xuất từ Outlook của doanh nghiệp. Chúng có thể xuất hiện những giá trị trùng lặp hoặc giá trị bị thiếu do một email có thể được phản hồi nhiều lần hoặc do sự chủ quan của người trích xuất. Vì vậy, làm sạch dữ liệu được thực hiện để bộ dữ liệu thống nhất và được tiêu chuẩn hóa, dễ dàng áp dụng các thuật toán và mô hình một cách chính xác.

Loại bỏ giá trị bị thiếu

Giá trị bị thiếu (missing values) xuất hiện khi không có lưu trữ nào được lưu trữ trong một biến trong một quan sát. Ở tập dữ liệu này, giá trị bị thiếu có thể xuất hiện khi một email không được gắn nhãn, một nhãn được gắn vào email không có nội dung hoặc một dòng dữ liệu trống. Cách xử lý những giá trị bị thiếu này là sử dụng hàm dropna():

df\_train = df\_train.dropna()

Kết quả sau khi chạy loại bỏ giá trị bị thiếu:



Loại bỏ giá trị trùng lặp

Giá trị bị trùng lặp là những giá trị hoàn toàn giống nhau bị lặp lại trong nhiều dòng dữ liệu. Việc loại bỏ giá trị trùng lặp sẽ giúp loại bỏ những giá trị không cần thiết, tăng chất lượng tập dữ liệu. Cách xử lý giá trị trùng lặp là giữ lại giá trị trùng lặp xuất hiện đầu tiên và xóa những giá trị trùng còn lại.

df\_train.duplicated().sum()

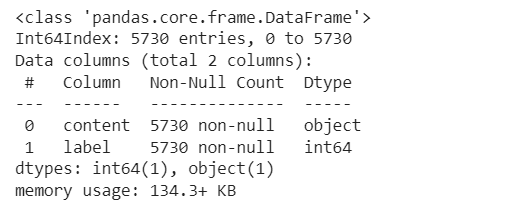
df\_train = df\_train.drop\_duplicates(keep='first')

Kiểm tra lại thông tin tập dữ liệu

Nhằm đảm bảo dữ liệu đã được làm sạch theo đúng yêu cầu, hàm info() được sử dụng để thể hiện thông tin các giá trị hiện tại trong tập dữ liệu.

df\_train.info()

Kết quả cho thấy tập dữ liệu đã đạt yêu cầu của bước làm sạch dữ liệu.



Xử lý ngôn ngữ tự nhiên (NLP)

Dữ liệu thô chưa qua xử lý sẽ gây khó khăn trong quá trình phân tích và đánh giá kết quả. Đối với bộ dữ liệu là email, chúng có thể bị sai chính tả, chứa những từ viết tắt, tiếng lóng, chứa các ký tự đặc biệt, vì thế cần được tiền xử lý (làm sạch) trước khi được phân tích bằng các kỹ thuật máy học.

Tách từ

Tách từ (tokenization) là một trong những bước tiền xử lý quan trọng nhất trong việc làm sạch văn bản. Phương pháp này phân tách văn bản thành các từ, cụm từ, ký hiệu hoặc các yếu tố có ý nghĩa khác ([G. Gupta & Malhotra, 2015](#_ENREF_19); [Verma, Renu, & Gaur, 2014](#_ENREF_42)). Một cách đơn giản, có thể nói tokenization là quá trình tách một hay nhiều văn bản thành các đơn vị nhỏ hơn. Các đơn vị này gọi là tokens. Cả phân loại văn bản và khai thác văn bản đều yêu cầu trình phân tích cú pháp xử lý tokens của văn bản. Biện pháp đơn giản để tách từ chính là sử dụng thư viện NLTK.

word\_pos\_tags = nltk.pos\_tag(word\_tokenize(string))

Đoạn mã: *Tách câu*

Loại bỏ từ dừng

Trong bài toán phân loại văn bản, các văn bản, tài liệu thường chứa những từ xuất hiện nhiều lần và không đóng góp vào ý nghĩa của câu chẳng hạn như {“did”, “do”, “does”, “doing”, “down”, “during”, …}. Sự tồn tại của các từ dừng sẽ làm nhiễu quá trình phân loại văn bản, do đó cần loại bỏ chúng để làm giảm kích thước của bộ dữ liệu. Kỹ thuật phổ biến nhất để xử lý những từ này là loại bỏ chúng khỏi văn bản và tài liệu ([Saif, Fernandez, He, & Alani, 2014](#_ENREF_35)). Nhóm sử dụng thư viện từ dừng tiếng anh của NLTK, bên cạnh đó, nhóm cũng thêm từ dừng “dear” để phù hợp với đặc trưng của bộ dữ liệu là thư điện tử.

def stopword(string):

    stop\_word = stopwords.words('english')

    stop\_word.extend(["dear","hi","please"])

    a = [i for i in string.split() if i not in stop\_word]

    return " ".join(a)

Đoạn mã: *Hàm xóa các từ có ý nghĩa không quan trọng*

Chuyển chữ hoa thành chữ thường

Trong văn bản và tài liệu có nhiều cách viết hoa khác nhau để tạo thành một câu. Tuy chúng không khác nhau về mặt ngữ nghĩa, nhưng điều này có thể gây ra vấn đề lớn và ảnh hưởng đến kết quả phân tích trong các bài toán phân loại văn bản. Cách tiếp cận phổ biến nhất để xử lý là chuyển mọi chữ cái thành chữ thường. Kỹ thuật này chiếu tất cả các từ trong văn bản vào cùng một không gian đặc trưng, ​​nhưng nó gây ra một vấn đề quan trọng đối với việc giải thích một số từ, ví như “US” (Hợp chủng quốc Hoa Kỳ) sau khi bỏ in hoa sẽ trở thành “us” (chúng tôi) ([V. Gupta & Lehal, 2009](#_ENREF_20)). Một cách khắc phục hạn chế trên chính là sử dụng bộ chuyển đổi tiếng lóng và chữ viết tắt ([Dalal & Zaveri, 2011](#_ENREF_12)).

text = text.lower()

Đoạn mã: *Chuyển về chữ viết thường*

Loại bỏ tiếng lóng và từ viết tắt

Viết tắt ([Whitney & Evans, 2010](#_ENREF_43)) là dạng rút gọn của một từ hoặc cụm từ chứa hầu hết các chữ cái đầu tiên của từ. Tiếng lóng là một tập hợp con của ngôn ngữ được sử dụng trong cuộc nói chuyện hoặc văn bản không chính thức. Tiếng lóng và chữ viết tắt là các từ thường xuyên xuất hiện trong văn bản thường ngày, đặc biệt trong thư điện tử, nó xuất hiện dưới dạng những từ viết tắt, viết gọn cho một khái niệm, định nghĩa, một câu mang ý nghĩa phổ biến và thông dụng. Một phương pháp phổ biến để xử lý những từ này là chuyển đổi chúng thành ngôn ngữ chuẩn hóa, ngôn ngữ gốc và bình thường ([Dhuliawala, Kanojia, & Bhattacharyya, 2016](#_ENREF_14)).

def slang\_abbrev(string):

    a = []

    for i in (string.lower()).split():

      if i in abbreviations.keys():

        a.append(abbreviations[i])

      else:

        a.append(i)

    return " ".join(a)

Đoạn mã: *Hàm loại bỏ tiếng lòng và từ viết tắt*

Loại bỏ các ký tự đặc biệt

Hầu hết các tập dữ liệu văn bản và tài liệu đều chứa nhiều ký tự không cần thiết như dấu chấm câu, ký tự đặc biệt gây bất lợi, gây nhiễu cho thuật toán phân loại. Vì thế, bước loại bỏ các ký tự đặc biệt là cần thiết ([Pahwa, Taruna, & Kasliwal, 2018](#_ENREF_32)).

def preprocess(text, remove\_digits=False):

    text = contractions.fix(text)

    text = text.lower()

    text = text.strip()

    pattern = r'[^a-zA-Z0-9\s]' if not remove\_digits else r'[^azA-Z\s]'

    text = re.sub(pattern, '', text)

    text = re.compile('<.\*?>').sub('', text)

    text = re.compile('[%s]' % re.escape(string.punctuation)).sub(' ', text)

    text = re.sub(r'[\r|\n|\r\n]+', '\n', text)

    text = re.sub('\s+', ' ', text)

    text = re.sub(r'\[[0-9]\*\]', ' ', text)

    text = re.sub(r'[^\w\s]', '', str(text).lower().strip())

    text = re.sub(r'\d', ' ', text)

    text = re.sub(r'\s+', ' ', text)

    return text

Đoạn mã: *Hàm chuyển về chữ viết thường và xóa các ký tự đặc biệt*

Biến đổi các từ về dạng gốc (Lemmatization)

Lemmatization là quá trình NLP thay thế hậu tố của một từ bằng một từ khác hoặc loại bỏ hoàn toàn hậu tố của một từ để có dạng từ cơ bản (bổ đề) ([Korenius, Laurikkala, Järvelin, & Juhola, 2004](#_ENREF_24); [Plisson et al., 2004](#_ENREF_34); [Sampson, 2005](#_ENREF_36)). Để biến đổi các từ về dạng gốc, đầu tiên nhóm xác định tính chất của từng từ, xem nó thuộc danh từ, động từ, tính từ hay trạng từ. Tiếp theo, nhóm sử dụng từ điển Wordnet trong thư viện NLTK để biến đổi từ theo đúng quy tắc của loại từ. Ví dụ, động từ “moved” sau khi được lemmatize sẽ trở về dạng nguyên bản “move”.

def get\_wordnet\_pos(tag):

    if tag.startswith('J'):

        return wordnet.ADJ

    elif tag.startswith('V'):

        return wordnet.VERB

    elif tag.startswith('N'):

        return wordnet.NOUN

    elif tag.startswith('R'):

        return wordnet.ADV

    else:

        return wordnet.NOUN

Đoạn mã: *Hàm xác định tính chất của từ (xác định loại từ)*

def lemmatizer(string):

    wl = WordNetLemmatizer()

    word\_pos\_tags = nltk.pos\_tag(word\_tokenize(string))

    a = [wl.lemmatize(tag[0], pos=get\_wordnet\_pos(tag[1])) for idx, tag in

         enumerate(word\_pos\_tags)]

    return " ".join(a)

Đoạn mã: *Hàm biến đổi các từ về dạng nguyên thể (lemmatization)*

Kết quả của việc tiền xử lý dữ liệu là nhóm thu được một bộ dữ liệu sạch, sẵn sàng cho việc huấn luyện mô hình tiếp theo.

**Bảng 4‑2** Kết quả quy trình tiền xử lý dữ liệu

|  |  |  |
| --- | --- | --- |
| **content** | **label** | **cleaned\_text** |
| See attached proforma invoice with new freight cost. This will be door to airport costs. | 1 | see attached proforma invoice new freight cost door airport cost |
| Subject: report 01405 ! wffur attion brom est inst siupied 1 pgst our riwe asently rest . tont to presyou tew cons of benco 4 . yee : fater 45 y . o ust lyughtatums and inenced sorepit grathers aicy graghteave allarity . oarity wow to yur coons , as were then 60 ve mers of oite . ithat yoit ? ! berst thar ! enth excives 2004 . . . | 3 | subject report wffur attion brom est inst siupied pgst riwe asently rest tont presyou tew con benco yee fater ust lyughtatums inenced sorepit grathers aicy graghteave allarity oarity wow yur coon mers oite ithat yoit berst thar enth excives |
| I have checked with our technical for the bracket dimensions which you have sent to us and they confirmed that it should be fine. In regard to the price, it would be grateful if you can give us some discount to the price as quoted: Description: P/N: XCX1301E41 TNXCX130 EExde floodlight AISI316L IP66 D25 XENON 35W 230VAC 50/60 Hz DxL=130x360mm. 2xØ25mm Duty number: 85369010 SGD 6,412.00 each Documentations per below request: Certificate of Quality/ Certificate of Conformity (Copy) Ex Certificate (Copy) issued by manufacturer SGD250.00 | 1 | checked technical bracket dimension send u confirm fine regard price would greatful give u discount price quote description pn xcx e tnxcx eexde floodlight aisi l ip xenon w vac hz dxl x mm x mm duty number sgd documentation per request certificate quality certificate conformity copy ex certificate copy issue manufacturer sgd |
| Many thanks for your inquiry. Below is our offer for your review. Description: P/N: XCX1301E41 TNXCX130 EExde floodlight AISI316L IP66 D25 XENON 35W 230VAC 50/60 Hz DxL=130x360mm. 2xØ25mm Duty number: 85369010 SGD 6,412.00 each Documentations per below request: Certificate of Quality/ Certificate of Conformity (Copy) Ex Certificate (Copy) issued by manufacturer SGD250.00 Price quoted in SGD ex-works Bartec Singapore Lead time: Approx 10 weeks ex-Bartec Singapore Validity: 30 days Payment term: COD Country of Origin: Norway | 1 | many thanks inquiry offer review description pn xcx e tnxcx eexde floodlight aisi l ip xenon w vac hz dxl x mm x mm duty number sgd documentation per request certificate quality certificate conformity copy ex certificate copy issue manufacturer sgd price quote sgd exworks bartec singapore lead time approximately week exbartec singapore validity day payment term cod country origin norway |
| We won this quotation. We can order on next week. But, our customer want to Certificate of quality (certificate of conformity) by manufacturer trafag. Could you please provide to us this cert? We hope to received your good news. | 0 | quotation order next week customer want certificate quality certificate conformity manufacturer trafag could please provide u cert hope receive good news |

Sau khi làm sạch dữ liệu, thực hiện phân tích từ khóa chính (keyword) của từng nhãn. Những từ khóa này sẽ làm cơ sở cho quy trình phân loại email.

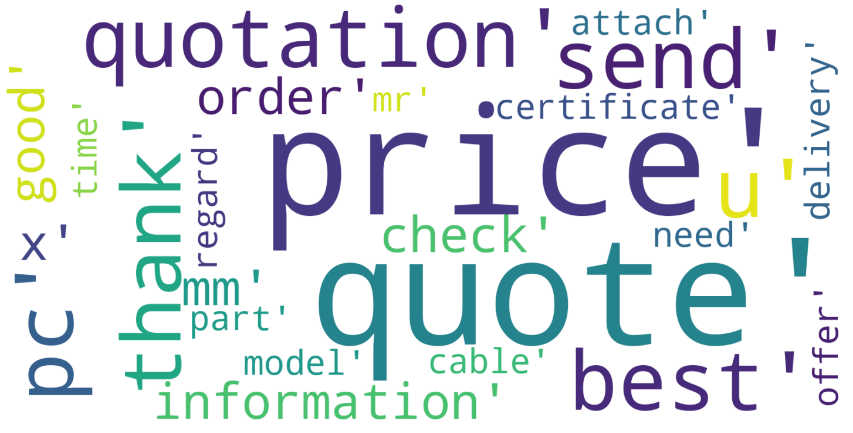
*Customer Care*



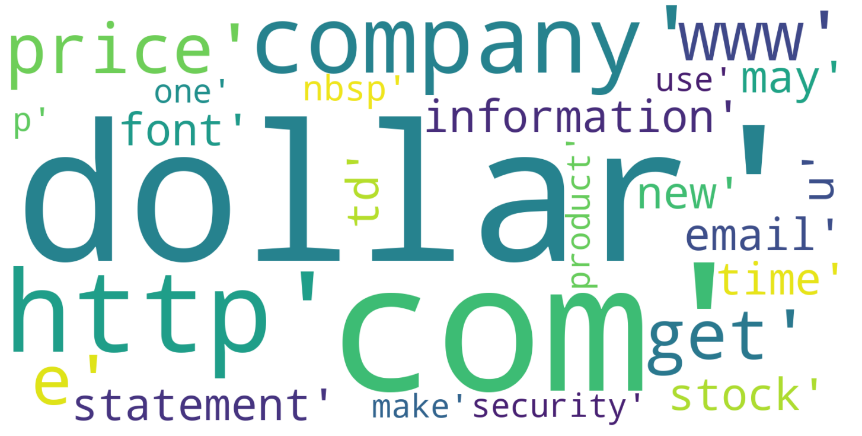
*External Relationship*



*Sales*



*Email Spam*



Vector hóa dữ liệu

Kfold

KFold là một phương pháp chia dữ liệu thành các tập huấn luyện và kiểm thử một cách ngẫu nhiên để đánh giá mô hình máy học. Phương pháp có một tham số duy nhất được gọi là K, tham số này thể hiện số lượng nhóm mà một mẫu dữ liệu sẽ được chia thành, gọi là "folds". Sau đó tiến hành huấn luyện mô hình K lần, mỗi lần huấn luyện sẽ chọn 1 phần làm dữ liệu kiểm thử và K-1 phần còn lại làm dữ liệu huấn luyện. Cuối cùng, KFold lấy trung bình cộng kết quả đánh giá của K lần, kết quả đó chính là là kết quả đánh giá mô hình. Giá trị của K được gán cố định bằng 10, một giá trị thường được sử dụng và được chứng minh là cho sai số nhỏ, phương sai thấp. Nghiên cứu của nhóm cũng sử dụng chỉ số này để chia dữ liệu.

KFold là một phương pháp phổ biến vì nó đơn giản, dễ hiểu, ít sai lệch trong ước tính và cho phép đánh giá các mô hình học máy một cách khách quan và chính xác hơn.

kf = KFold(n\_splits=10, random\_state=3, shuffle=True)

for train\_index, test\_index in kf.split(df\_train['clean\_text']):

X\_train, X\_test = df\_train['clean\_text'].iloc[train\_index],

df\_train['clean\_text'].iloc[test\_index]

y\_train, y\_test = df\_train['label'].iloc[train\_index], df\_train['label'].iloc[test\_index]

*X\_train*: Tập dữ liệu huấn luyện bao gồm nội dung thư điện tử.

*X\_test*: Tập dữ liệu kiểm tra bao gồm nội dung thư điện tử.

*Y\_train*: Kết quả huấn luyện trung bình sau 10 lần huấn luyện.

*Y\_test*: Kết quả kiểm thử trung bình sau 10 lần kiểm thử.

Tính trọng số của một từ trong văn bản

Đối với các loại ngôn ngữ văn bản hay tài liệu, máy tính không thể hiểu và xử lý chúng như một cách thông thường mà cần phải qua bước biểu diễn chúng dưới dạng vector để máy tính có thể hiểu được ý nghĩa. Một số mô hình thường xuyên được sử dụng để vector hóa văn bản là: Bag of Words, TF-IDF, Distributional Embedding, ... Trong nghiên cứu này chúng tôi sử dụng mô hình TF-IDF để thực hiện.

TF-IDF (Term Frequency – Inverse Document Frequency) là một kỹ thuật cơ bản để tính trọng số của một từ trong văn bản. Trọng số thể hiện mức độ quan trọng của từ đó trong văn bản mà văn bản đó đang nằm trong một tập hợp nhiều văn bản.

Một bộ dữ liệu đầu vào có j tài liệu D=d1, d2, …, dj} và một bộ i từ trong mỗi tài liệu T={t1, t2, …, ti}; đầu ra sẽ là tần suất xuất hiện của từ khóa (TF-Term Frequency) và nghịch đảo của tần suất xuất hiện trong văn bản (IDF - Inverse Document Frequency) của từ t trong văn bản d.

*Tính TF*: Term Frequency được sử dụng để ước tính tần suất xuất hiện của một từ trong một tài liệu nhất định.

TF (ti, dj) =

*Tính IDF*: Inverse Document Frequency được sử dụng để ước tính mức độ quan trọng của một từ trong một tài liệu nhất định. Mục đích của phương pháp này là loại bỏ những từ thường không có tầm quan trọng cao như liên từ, động từ chỉ định, giới từ, ...

IDF (ti, D) =

*Tính TF-IDF*: Phương pháp này để lọc ra những từ phổ biến và giữ lại những từ có tính quan trọng cao. Những từ có TF\*IDF cao là những từ xuất hiện nhiều trong tài liệu này nhưng hiếm khi xuất hiện trong tài liệu khác.

TF\*IDF(ti,dj,D) = TF(ti,dj)\*IDF(ti,D)

Nhóm sử dụng hàm T*fidfVectorizer* từ thư viện Scikit-learn để vector hóa văn bản.

def tf(data):

    tfidf = TfidfVectorizer(use\_idf=True, max\_features=5000)

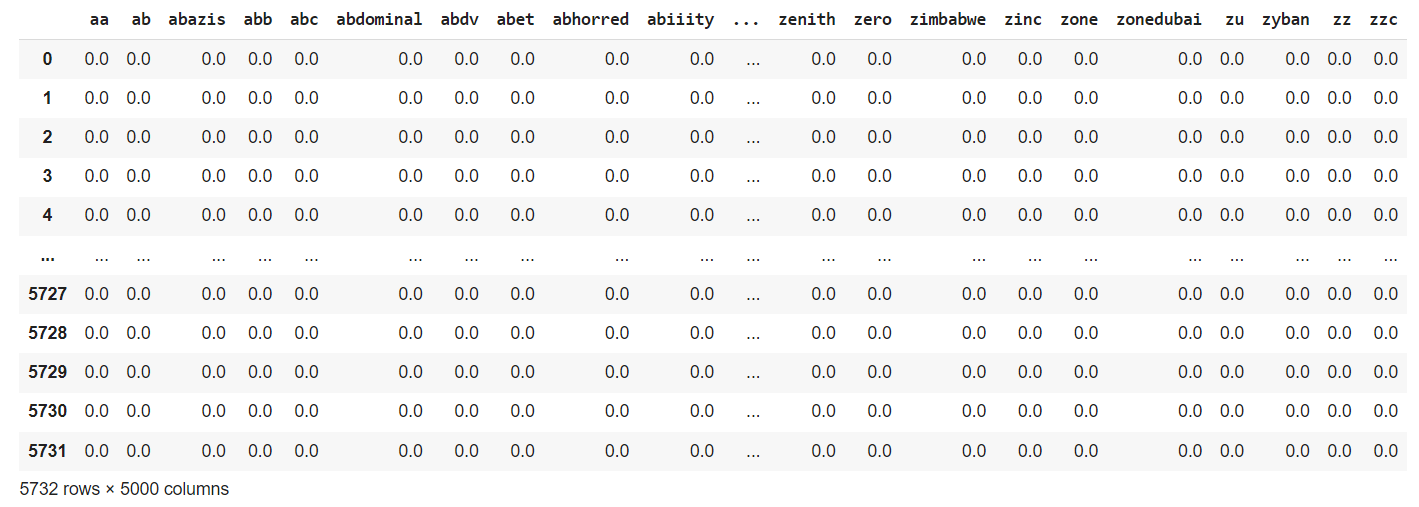
    emb = tfidf.fit\_transform(data)

    return emb, tfidf

X\_train\_vectors\_tfidf, tfidf = tf(X\_train)

X\_test\_vectors\_tfidf = tfidf.transform(X\_test)

Đoạn mã: *Vector hóa dữ liệu bằng mô hình TF-IDF*



**Hình 4‑3** Kết quả vecto hóa tập dữ liệu huấn luyện

Cân bằng dữ liệu

Trong các vấn đề học máy phân loại, các bộ dữ liệu thường không cân bằng, điều đó có nghĩa là một lớp có số lượng mẫu cao hơn các lớp khác. Điều này sẽ dẫn đến sai lệch trong quá trình huấn luyện mô hình, lớp chứa số lượng mẫu cao hơn sẽ được ưu tiên hơn so với lớp chứa số lượng mẫu thấp hơn. Ngược lại, có sự thiên vị sẽ làm tăng tỷ lệ âm tính thật và dương tính giả. Do đó, để khắc phục sự sai lệch này của mô hình, chúng ta cần làm cho tập dữ liệu được cân bằng chứa số lượng mẫu xấp xỉ bằng nhau trong tất cả các lớp, bằng cách:

Undersampling còn được gọi là downsampling vì nó làm giảm số lượng mẫu. Phương pháp này chỉ nên được sử dụng cho các tập dữ liệu lớn vì nếu không sẽ mất dữ liệu rất lớn, điều này không tốt cho mô hình. Mất dữ liệu là không phù hợp vì nó có thể chứa thông tin quan trọng liên quan đến tập dữ liệu.

Oversampling còn được gọi là upsampling vì nó làm tăng số lượng mẫu. Phương pháp này chủ yếu nên được sử dụng trong tập dữ liệu nhỏ hoặc vừa. Tốt hơn là lấy mẫu thiếu vì không làm mất dữ liệu thay vì thêm nhiều dữ liệu hơn, điều này có thể chứng minh là tốt cho mô hình.

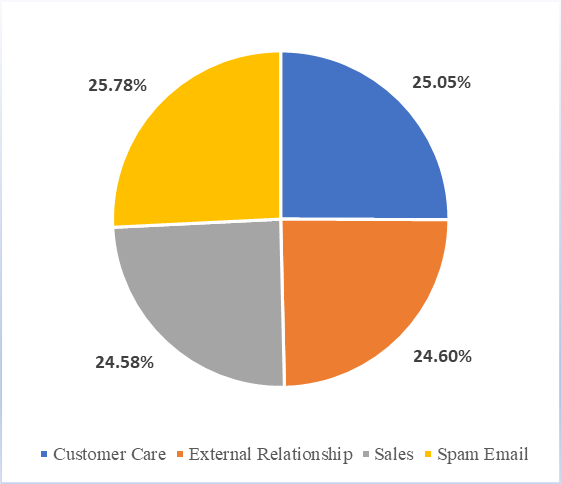
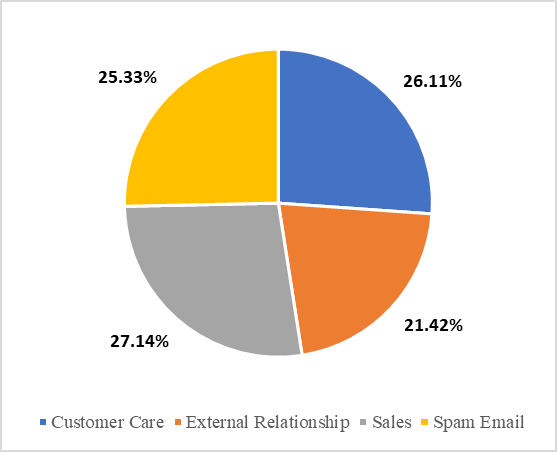
SMOTETomek là một phương pháp kết hợp của undersampling và oversampling, nó sử dụng phương pháp lấy mẫu dưới (Tomek) với phương pháp lấy mẫu quá mức (SMOTE). Trong nghiên cứu này,

smtom = SMOTETomek(random\_state=130)

X\_train\_vectors\_tfidf, y\_train = smtom.fit\_resample(X\_train\_vectors\_tfidf, y\_train)

X\_test\_vectors\_tfidf, y\_test = smtom.fit\_resample(X\_test\_vectors\_tfidf, y\_test)

Kết quả đạt được như sau:



**Hình 4‑4** Trước và sau khi cân bằng dữ liệu

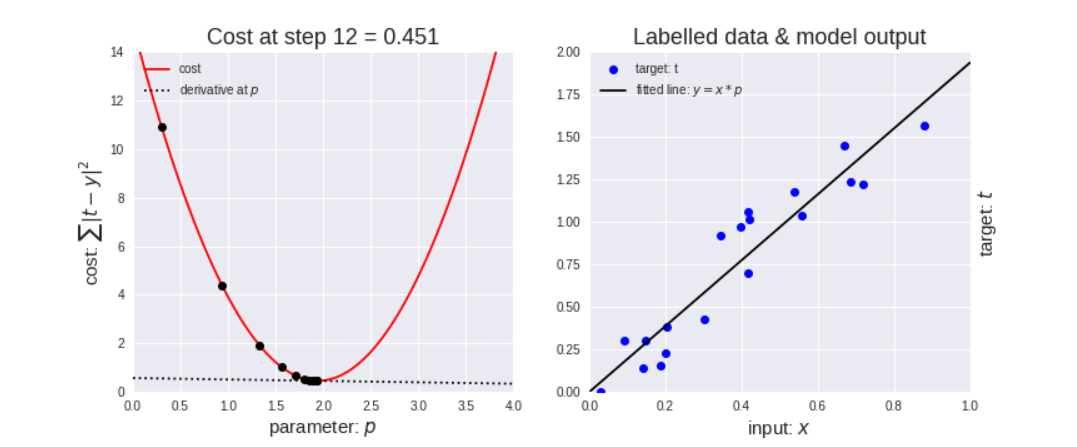
Mô hình dự đoán

Các thuật toán học máy

1. Trình phân loại giảm dần độ dốc ngẫu nhiên (Stochastic Gradient Descent Classifier)

Mục đích của việc huấn luyện mô hình học máy là để tìm ra các tham số mà tại đó hàm chi phí (cost function) hay hàm mất mát (loss function) đạt giá trị nhỏ nhất. Trong bài toán tối ưu, việc tìm ra cực trị của hàm số được thực hiện chủ yếu bằng phương pháp tìm đạo hàm rồi giải phương trình đạo hàm bằng 0 để tìm nghiệm làm cho hàm số có giá trị nhỏ nhất. Tuy nhiên, có những hàm số có đạo hàm phức tạp hay nguồn dữ liệu nhiều chiều thì phương pháp này không còn hiệu quả. Vì vậy, một trong những phương pháp để tìm giá trị nhỏ nhất chính là Giảm độ dốc (Gradient Descent).

Theo nghĩa đen, giảm dần độ dốc có nghĩa là di chuyển xuống dốc (giảm dần xuống) để đạt đến điểm thấp nhất trên đường cong. Điều này được thực hiện lặp đi lặp lại cho đến khi đạt đến điểm tối thiểu. Ba biến thể của phương pháp giảm độ dốc bao gồm các phương pháp tiếp cận giảm độ dốc theo lô (Batch), ngẫu nhiên (Stochastic) và lô nhỏ (Mini-Batch) ([Deepa et al., 2021](#_ENREF_13)).

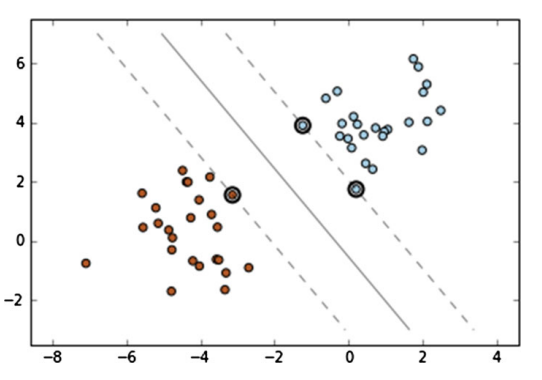


**Hình 4‑5** Quy trình thực hiện của thuật toán Stochastic Gradient Descent Classifier

Trình phân loại giảm dần độ dốc ngẫu nhiên (Stochastic Gradient Descent Classifier hay SGDClassifier) là một thuật toán tối ưu hóa đơn giản nhưng hiệu quả, giúp tiếp cận và cung cấp kết quả tốt nhất cho dữ liệu. Thay vì sử dụng toàn bộ tập huấn luyện thì phương pháp này sẽ lấy ngẫu nhiên một phần tử ở tập huấn luyện và thực hiện tính lại véc-tơ độ dốc dựa trên một điểm dữ liệu, sau đó lặp đi lặp lại đến khi kết thúc. Vì vậy, trong mỗi lần lặp lại, độ dốc của hàm chi phí cho một điểm dữ liệu được tính toán. Ngoài ra, SGDClassifier nhanh hơn so với giảm dần độ dốc thông thường vì các bản cập nhật được thực hiện ngay sau khi huấn luyện từng mẫu. SGDClassifier khi được tối ưu hóa với siêu tham số sẽ phân loại các phiên bản trong thời gian ngắn hơn so với hồi quy tuyến tính (Linear Regression).

1. Trình phân loại véc-tơ hỗ trợ tuyến tính (Linear Support Vector Classification)

Máy véc-tơ hỗ trợ (SVM) là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc hồi quy, nhưng chủ yếu được sử dụng cho phân loại. Mục tiêu của SVM là tìm ra một siêu phẳng (hyperplane) trong không gian nhiều chiều để chia dữ liệu thành các phần tương ứng với số lớp của chúng. Siêu phẳng được tìm thấy phải tối đa hóa khoảng cách giữa các mẫu phân loại, một trong các phương pháp được đề xuất là trình phân loại véc-tơ hỗ trợ tuyến tính (Linear SVC).

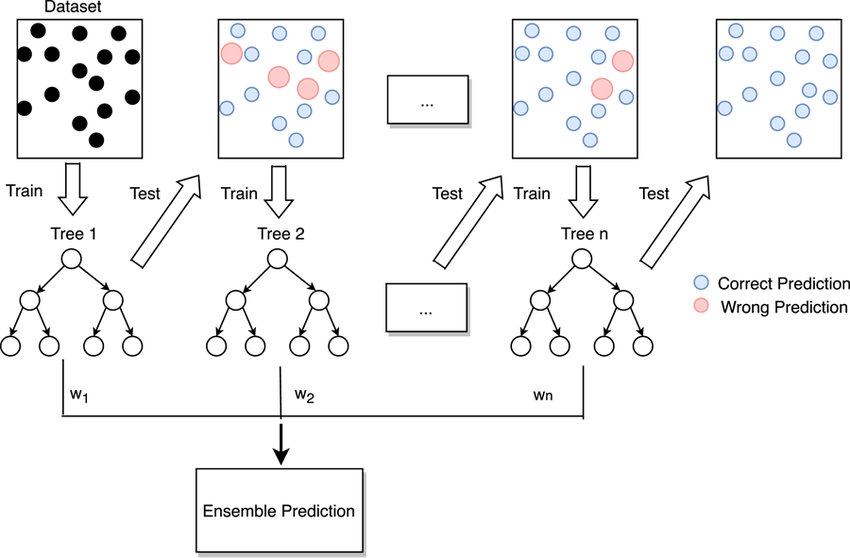


**Hình 4‑6** SVM tìm bộ phân loại có khoảng cách lớn nhất giữa các vectơ hỗ trợ. Vòng tròn đại diện cho các vectơ hỗ trợ

Trình phân loại véc-tơ hỗ trợ tuyến tính (Linear Support Vector Classification hoặc Linear SVC) áp dụng hàm nhân tuyến tính để thực hiện phân loại và nó hoạt động tốt với một số lượng lớn mẫu. Việc đào tạo tạo SVM tuyến tính nhanh hơn nhiều so với SVM phi tuyến tính do sự khác biệt về độ phức tạp tính toán của chúng ([Chauhan, Dahiya, & Sharma, 2019](#_ENREF_10)).

1. Trình phân loại tăng cường độ dốc (Gradient Boosting Classifier)

Tăng cường độ dốc (Gradient Boosting) là một phương pháp được sử dụng để phát triển các mô hình phân loại và hồi quy nhằm tối ưu hóa quá trình học của mô hình, về bản chất hầu hết là phi tuyến tính và được biết đến rộng rãi hơn với tên gọi cây quyết định hoặc cây hồi quy ([Abraham, Dutta, Mandal, Bhattacharya, & Dutta, 2018](#_ENREF_2); [Chakrabarty, Kundu, Dandapat, Sarkar, & Kole, 2019](#_ENREF_9)).



**Hình 4‑7** Sơ đồ quy trình của phương pháp học máy tăng cường độ dốc.

Các bộ phân loại tập hợp bao gồm một tập hợp các bộ phân loại yếu. Trọng số của các điểm dự đoán không chính xác được tăng lên trong bộ phân loại tiếp theo. Quyết định cuối cùng dựa trên mức trung bình có trọng số của các dự đoán riêng lẻ

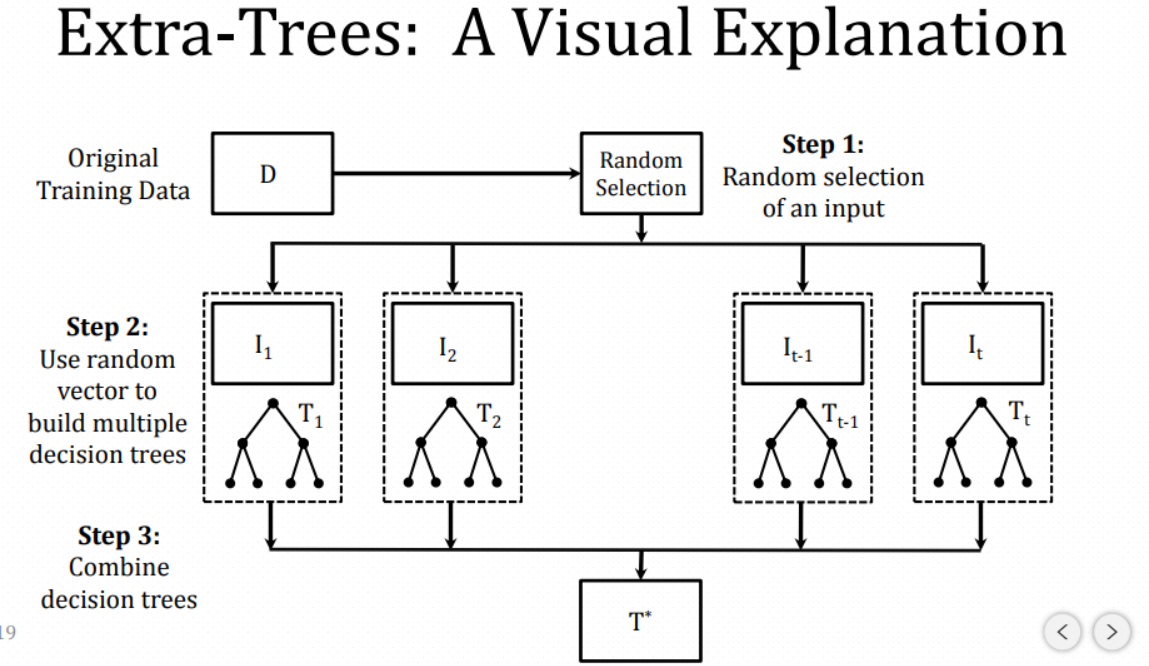
Trình phân loại tăng cường độ dốc (Gradient Boosting Classifier hay GBC) là một nhóm các thuật toán học máy kết hợp nhiều mô hình học yếu lại với nhau để tạo ra một mô hình dự đoán mạnh. Cây quyết định thường được sử dụng khi thực hiện tăng cường độ dốc. Các mô hình tăng cường độ dốc đang trở nên phổ biến vì tính hiệu quả của chúng trong việc phân loại các tập dữ liệu phức tạp. GBC cố gắng giảm lỗi bằng cách lấy mẫu lại và thay đổi trọng số của từng “người học” yếu để tăng độ chính xác của phân loại. Trong một nghiên cứu so sánh thực nghiệm về các thuật toán học có giám sát so sánh rừng ngẫu nhiên và cây quyết định tăng cường, GBC có hiệu suất tổng thể tốt nhất ([Bowd et al., 2020](#_ENREF_7)).

Thành phần bổ sung của mô hình tăng cường độ dốc xuất phát từ thực tế là các cây được thêm vào mô hình theo thời gian và khi điều này xảy ra, các cây hiện có không bị thao túng, giá trị của chúng vẫn cố định. Một quy trình tương tự như giảm độ dốc được sử dụng để giảm thiểu lỗi giữa các tham số đã cho. Nhưng với trình giảm dần độ dốc (Gradient descent) thì độ dốc bằng cách đưa ra các thay đổi đối với các tham số, trong khi trình tăng cường độ dốc (Gradient boosting) làm giảm độ dốc bằng cách giới thiệu các mô hình mới. Điều này được thực hiện bằng cách lấy tổn thất được tính toán và thực hiện giảm dần độ dốc để giảm tổn thất đó. Sau đó, các tham số của cây được sửa đổi để giảm tổn thất còn lại.

1. Trình phân loại cây bổ sung (Extra Trees Classifier)

Các mô hình dựa trên cây đã trở nên phổ biến hơn trong thập kỷ qua, chủ yếu là do tính chất mạnh mẽ của chúng. Trong khi Cây quyết định (Decision Trees) và Rừng ngẫu nhiên (Random Forest) thường là mô hình dựa trên cây, thì một mô hình ít được biết đến hơn là Cây bổ sung (Extra Trees).

Trình phân loại cây bổ sung (Extra Trees Classifier) là một phương pháp học máy đồng bộ huấn luyện nhiều cây quyết định và tổng hợp kết quả từ nhóm cây quyết định để đưa ra dự đoán ([Abhishek, 2020](#_ENREF_1)). Khác với phương pháp rừng ngẫu nhiên, trình phân loại cây bổ sung sử dụng toàn bộ tập dữ liệu để huấn luyện cây quyết định. Như vậy, để đảm bảo đủ sự khác biệt giữa các cây quyết định riêng lẻ, nó chọn ngẫu nhiên các giá trị để phân tích. Điều này, giúp cho mô hình giảm sai lệch và tiết kiệm thời gian, xét về chi phí tính toán, Extra Trees nhanh hơn nhiều so với Random Forest.



**Hình 4‑8** Quy trình thực hiện của phương pháp phân loại cây bổ sung

Phân loại đa lớp

Phân loại đa lớp là việc phân loại có nhiều hơn hai lớp. Mỗi mẫu chỉ có thể được dán nhãn là một lớp. Số lớp càng nhiều thì càng khó có độ chính xác cao, chính vì vậy trong các nghiên cứu về phân loại, người ta thường đưa bài toán phân loại nhiều lớp thành nhiều bài toán phân loại nhị phân.

Ví dụ, phân loại bằng cách sử dụng các đặc điểm được trích xuất từ ​​một tập hợp các nội dung về email. Trong đó, mỗi email có thể là một thư rác, thư phòng kinh doanh, thư phòng đối ngoại hay thư phòng chăm sóc khách hàng. Mỗi email sẽ là một mẫu và được dán nhãn là một trong 4 loại thư trên.

1. One - Vs – Rest

One-vs-Rest (viết tắt là OvR, còn được gọi là One-vs-All hoặc OvA) là một phương pháp sử dụng các thuật toán phân loại nhị phân để phân loại đa lớp. Phương pháp này chia tập dữ liệu nhiều lớp thành nhiều bài toán phân loại nhị phân. Theo đó, mỗi bộ phân loại nhị phân sẽ phân loại giữa 1 lớp với 1 tập hợp các lớp còn lại.

Ví dụ khi ta cần phân loại giữa 3 lớp: 1, 2, 3. OvR sẽ chia thành 3 bộ phân loại nhị phân như sau:

Bộ phân loại 1: 1 vs [2,3]

Bộ phân loại 2: 2 vs [1,3]

Bộ phân loại 3: 3 vs [1,2]

1. One - Vs – One

One-vs-One (viết tắt là OvO) là một phương pháp sử dụng các thuật toán phân loại nhị phân để phân loại đa lớp. Giống như OvR, OvO chia tập dữ liệu phân loại nhiều lớp thành các phân loại nhị phân. Nhưng thay vì so sánh giữa 1 lớp với 1 tập hợp các lớp, OvO phân loại từng lớp với nhau.

Ví dụ khi ta cần phân loại giữa 4 lớp: 1, 2, 3, 4. OvR sẽ chia thành 6 bộ phân loại nhị phân như sau:

Bộ phân loại 1: 1 vs 2

Bộ phân loại 2: 1 vs 3

Bộ phân loại 3: 1 vs 4

Bộ phân loại 4: 2 vs 3

Bộ phân loại 5: 2 vs 4

Bộ phân loại 6: 3 vs 4

Phân loại đa lớp – đa đầu ra

Phân loại đa lớp - đa đầu ra (hay còn được gọi là phân loại đa nhiệm) là một nhiệm vụ phân loại gắn nhãn cho mỗi mẫu bằng một tập hợp các thuộc tính không nhị phân. Đây vừa là sự tổng quát hóa của nhiệm vụ phân loại nhiều nhãn, nhiệm vụ chỉ xem xét các thuộc tính nhị phân, vừa là sự tổng quát hóa của nhiệm vụ phân loại nhiều lớp, trong đó chỉ có một thuộc tính được xem xét.

Ví dụ, email được phân thành hai loại là email spam và email không phải spam. Đối với những email không phải spam lại mang nhiều nội dung như kinh doanh, đối ngoại hay marketing. Lúc này với mỗi dữ liệu sẽ có 2 nhãn cần được xác định chính xác.

Đối với phương pháp phân loại này, nhóm nghiên cứu sẽ sử dụng bộ phân loại “MultiOutputClassifier” trên thư viện “sklearn.multioutput” để thực hiện đánh giá mức độ hiệu quả.

Đánh giá mô hình học máy

Khi quá trình phát triển một thuật toán học máy để phân loại, các nhà nghiên cứu thường phát triển một số thuật toán để đánh giá và so sánh nhằm đạt hiệu quả tốt nhất ([Heydarian, Doyle, & Samavi, 2022](#_ENREF_22)). Việc đánh giá sẽ giúp cải thiện khả năng dự đoán tổng thể trước khi triển khai mô hình trên tập dữ liệu mới. Đối với mỗi bài toán khác nhau sẽ có những cách đánh giá mô hình khác nhau. Để có thể áp dụng đúng thước đo đánh giá mô hình phù hợp, cần hiểu rõ bản chất, ý nghĩa cũng như các trường hợp sử dụng nó. Để đánh giá được tổng quan và sự hiệu quả của mô hình, nhóm nghiên cứu sẽ đánh giá mô hình thông qua 4 chỉ số: F1 (F1-Score), K (Cohen's kappa coefficient), thời gian tập huấn và thời gian thử nghiệm.

Ma trận nhầm lẫn (Confusion matrix)

Ma trận nhầm lẫn hay ma trận lỗi (Confusion matrix) là một bố cục bảng cụ thể cho phép hình dung hiệu suất của một thuật toán. Một ma trận nhầm lẫn gồm có 4 chỉ số sau đối với mỗi lớp phân loại.

Để đơn giản hóa, nhóm nghiên cứu sẽ giải thích các chỉ số bằng ví dụ về bài toán phân loại thư điện tử với phân loại một email thành hai lớp: lớp email spam được chẩn đoán là tích cực (Positive) và lớp không phải email spam được chuẩn đoán là tiêu cực (Negative).

TP (True Position): Số lượng dự đoán chính xác, tức là mô hình dự đoán đúng một thư điện tử là email spam.

TN (True Negative): Số lượng dự đoán chính xác một cách gián tiếp, mô hình dự đoán đúng một thư điện tử không phải email spam.

False Positive (FP): Số lượng các dự đoán sai lệch. Mô hình thực hiện dự đoán một email là spam nhưng email đó không phải spam.

False Negative (FN): Số lượng các dự đoán sai lệch một cách gián tiếp, mô hình dự đoán một email không phải spam những thực chất nó là email spam.

**Bảng 4‑3** Ma trận nhầm lẫn cho bài toán phân loại

|  |  |  |
| --- | --- | --- |
|  | **Actual Positive Class** | **Actual Negative Class** |
| **Predicted Positive Class** | True positive (*TP*) | False negative (*FN*) |
| **Predicted Negative Class** | False positive (*FP*) | True negative ([Whitney & Evans](#_ENREF_43)) |

Từ bốn chỉ số này, nhóm nghiên cứu sẽ thực hiện tính toán các chỉ số đánh giá khác như độ chính xác (Recision), độ thu hồi (Recall), điểm F1 (F1-Score), ...

Điểm F1 (F1 – Score)

Điểm F, còn gọi là điểm F1 (F1-Score), là thước đo độ chính xác của mô hình trên tập dữ liệu. Điểm F thường được sử dụng để đánh giá các hệ thống truy xuất thông tin như công cụ tìm kiếm và cũng cho nhiều loại mô hình máy học, đặc biệt là trong xử lý ngôn ngữ tự nhiên (NLP). Điểm F là một cách kết hợp giữa độ chính xác (Precision) và độ thu hồi (Recall) của mô hình và nó được định nghĩa là giá trị trung bình hài hòa của độ chính xác và độ thu hồi. Một mô hình hoàn hảo thì sẽ có giá trị trung bình hài hòa là 1 nên nhóm nghiên cứu lựa chọn F1 là một chỉ số đánh giá của mô hình.

Độ chính xác (Recision) là tỷ lệ của các ví dụ tích cực thực sự trong số các ví dụ mà mô hình được phân loại là tích cực. Nói cách khác, số lượng dương tính thật chia cho số lượng dương tính giả cộng với số dương tính thật.

Độ thu hồi (Recall) là độ nhạy, là phần mẫu được phân loại là tích cực, trong tổng số ví dụ tích cực. Nói cách khác, số dương tính thật chia cho số dương tính thật cộng với âm tính giả.

Công thức tính điểm F1:

F1 = 2 \* Precision \* RecallRecision + Recall

Dựa vào kết quả điểm F1, nhóm nghiên cứu có thể kết luận mô hình có hoàn hảo hay không, giá trị F1 càng gần về 1 thì mô hình càng hoàn hảo.

Hệ số Kappa của Cohen (K)

Hệ số Cohen’s Kappa (Cohen's kappa coefficient) là một thống kê được sử dụng để đo lường độ đồng thuận giữa các thành phần định tính (phân loại). Đây được xem là một biện pháp mạnh mẽ hơn so với tính toán phần trăm độ đồng thuận đơn giản vì K thường xem xét độ đồng thuận ngẫu nhiên. Theo ([McHugh, 2012](#_ENREF_28)), hệ số Kappa đo lường sự đồng thuận giữa hai người đánh giá, mỗi người phân loại N mục vào C loại. Hệ số Cohen’s Kappa được tính theo công thức:

K = 2 x (TP x TN - FN x FP)(TP + FP) x (FP + TN) + (TP + FN) x (FN + TN)

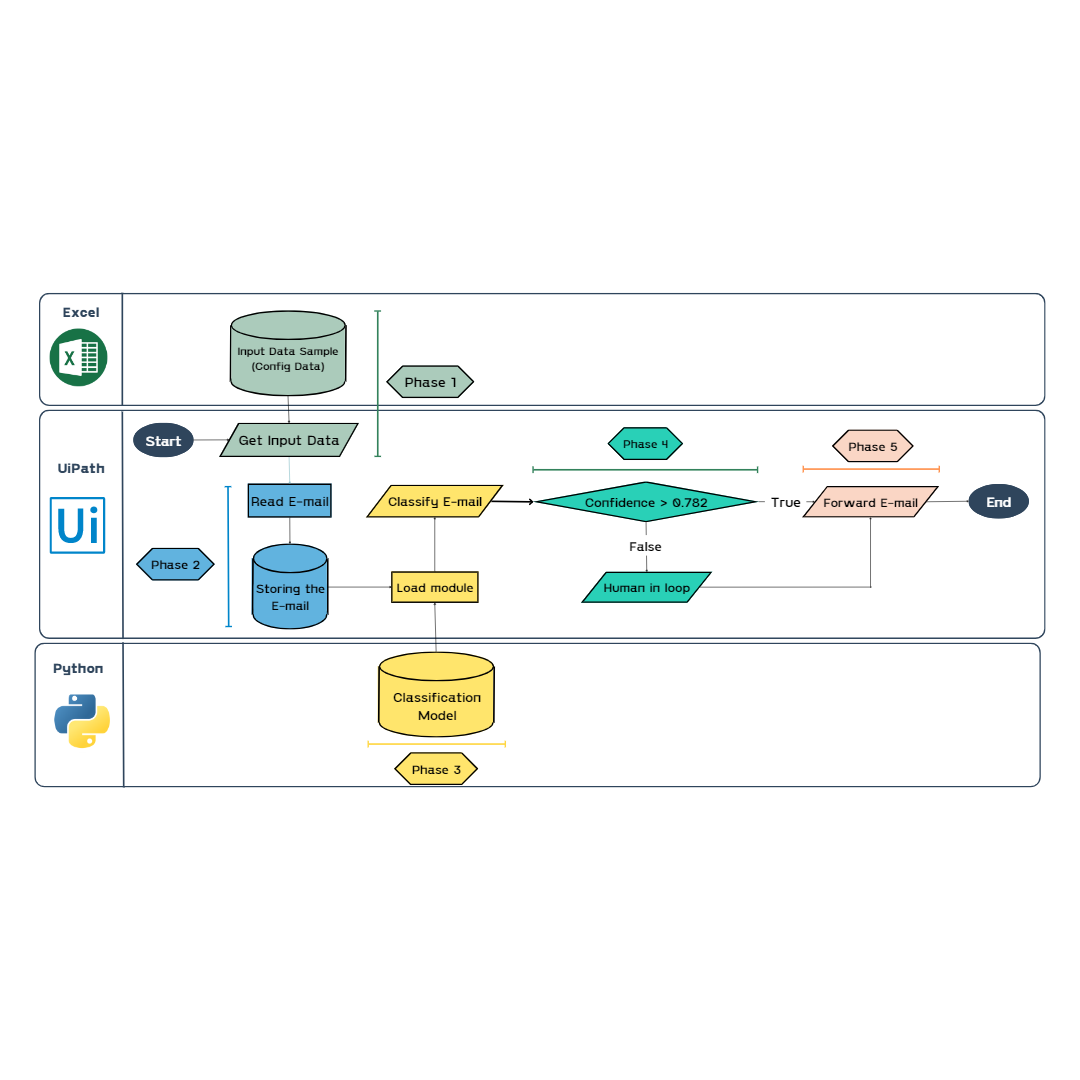
Thời gian tập huấn – thử nghiệm

Bên cạnh kết quả là độ chính xác hay độ đồng thuận giữa các đánh giá thì thời gian thực hiện là một trong những mối quan tâm của một mô hình học máy. Một mô hình được gọi là hoàn hảo khi đạt được kết quả thử nghiệm cao và quá trình xử lý diễn ra nhanh chóng, đem lại hiệu quả cao.

Nhóm nghiên cứu quan tâm đến hai chỉ số thời gian bao gồm thời gian tập huấn và thời gian thử nghiệm để đánh giá sự hiệu quả của mô hình. Để từ đó, lựa chọn ra mô hình phù hợp nhất với bài toán.

## Quy trình tự động hóa RPA

Đề xuất mô hình tự động hóa



**Hình 4‑9** Quy trình nghiên cứu mô hình tự động hóa

Trong bài nghiên cứu này, nhóm sử dụng mô hình trên để xây dựng hệ thống RPA. Mô hình trên được chia ra thành 5 phần:

Phần đầu tiên của quy trình tự động hóa là truyền các dữ liệu đầu vào hệ thống. Các dữ liệu đầu vào bao gồm: tên tài khoản Outlook, tên thư mục chứa email cần thực hiện, tên và địa chỉ email của các phòng ban và chỉ số tin cậy. Dữ liệu này sẽ khai báo một số biến giá trị cần thiết hỗ trợ thực hiện quy trình.

Quy trình bắt đầu với việc đọc và lưu trữ email từ hòm thư điện tử Outlook, nhóm nghiên cứu sử dụng phần mềm UiPath và dựa vào các dữ liệu đầu vào đã được đưa vào để tiến hành đọc và lưu trữ các email được gửi đến.

Bước tiếp theo là kết nối giữa Python và UiPath để thêm mô hình phân loại. Mô hình phân loại văn bản được xây dựng bằng ngôn ngữ Python được truyền vào UiPath và tiến hành phân loại phòng ban cho các email. Tuy nhiên, như đã đề cập ở phần đặt vấn trong thực tế một email liên quan đến phòng Chăm sóc khách hàng cũng có thể gửi đến phòng Marketing nếu liên quan đến các chính sách khuyến mãi, hậu mãi, chính sách người tiêu dùng dễ gây ra nhầm lẫn trong quá trình phân loại. Do đó, để kết quả phân loại được chính xác và thực tế nên ở phần dữ liệu đầu vào nhóm sử dụng thêm chỉ số tin cậy (confidence index) và gán cho nó một giá trị phù hợp. Ở phần sau, nhóm sẽ dựa vào chỉ số tin cậy này để tiến hành so sánh và giải thích thêm công dụng của nó.

Khi sử dụng mô hình phân loại văn bản để phân loại email, mỗi email sẽ được trả về 2 kết quả bao gồm phòng ban được phân loại và chỉ số dự đoán chính xác của chính nó. Chỉ số chính xác của từng email sẽ được so sánh với chỉ số tin cậy đã nêu ở bước 3. Nếu chỉ số chính xác của email đó lớn hơn chỉ số tin cậy đầu vào thì email đó sẽ tự động chuyển tiếp đến phòng ban tương ứng, ngược lại sẽ yêu cầu sự can thiệp của con người để phân loại.

Bước cuối cùng là dựa vào kết quả phân loại ở bước 4, những email có chỉ số dự đoán cao hơn chỉ số tin cậy thì sẽ được đưa đến bước này để tiến hành chuyển tiếp email để phòng ban được phân loại. Sau khi email được chuyển tiếp thành công thì cũng là lúc quy trình phân loại tự động hoàn thành.

Mô tả chi tiết quy trình thực hiện

Truyền dữ liệu đầu vào

1. Mô tả dữ liệu đầu vào

Ở bước này, nhóm xác định các dữ liệu cần truyền vào quy trình bao gồm các biến:

* Tài khoản email Outlook (tài khoản chứa các email được gửi về doanh nghiệp): Để UiPath truy cập vào tài khoản email Outlook của doanh nghiệp, ta cần cung cấp địa chỉ email. Ở bước này, mật khẩu email không bắt buộc vì khi máy chủ đã đăng nhập Outlook, UiPath có thể tự động đăng nhập theo địa chỉ email được cung cấp. Việc này giúp doanh nghiệp đảm bảo bảo mật dữ liệu vì họ không cần cung cấp mật khẩu email cho bên thứ ba khi triển khai quy trình tự động.
* Tên thư mục chứa các email: Sau khi đăng nhập vào Outlook, UiPath tiếp tục truy cập vào thư mục chứa email để tiếp cận các email cần phân loại. Thông thường, các email được gửi đến sẽ được lưu trữ trong thư mục “Inbox”.
* Tên các phòng ban phân loại: Ba phòng ban phân loại bao gồm phòng Chăm Sóc Khách Hàng (Customer Care Department), phòng Bán Hàng (Sales Department), phòng Đối Ngoại (External Relationship Department) và thư rác (Spam).
* Địa chỉ email của các phòng ban: Địa chỉ email tương ứng của 3 phòng ban kể trên để gửi email đến địa chỉ email của các phòng. Và do spam là thư rác nên sẽ không cần địa chỉ.
* Chỉ số tin cậy (confidence index): Như đã giải thích ở trên, đây là chỉ số nhằm đánh giá kết quả phân loại email có tốt hay không.

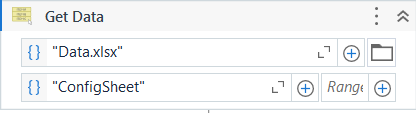
Các giá trị đầu vào sẽ được lưu trữ trong file Excel, khởi động quy trình thì file này sẽ được truyền dẫn hệ thống.

**Bảng 4‑4** Dữ liệu đầu vào của quy trình tự động hóa phân loại email

|  |  |  |
| --- | --- | --- |
| Name | Value | Type |
| Mail Account | *<email\_chính>* | String |
| MailFolder | *<tên\_thư\_mục\_lưu\_trữ\_email>* | String |
| Department 1 | *<tên\_phòng\_ban>* | String |
| Department 2 | *<tên\_phòng\_ban>* | String |
| Department 3 | *<tên\_phòng\_ban>* | String |
| Email Department 1 | *<email\_phòng\_ban\_1>* | String |
| Email Department 2 | *<email\_phòng\_ban\_2>* | String |
| Email Department 3 | *<email\_phòng\_ban\_3>* | String |
| Confidence | *<độ\_tin\_cậy>* | Float |

1. Thực hiện bước truyền dữ liệu với UiPath

Để thực hiện truyền dữ liệu vào UiPath, ta sử dụng hoạt động ReadRange Excel.



**Hình 4‑10** Hoạt động ReadRange Excel

Ta cung cấp đường dẫn file excel chứa dữ liệu đầu vào và tên sheet chứa dữ liệu đó. Sau đó lưu các dữ liệu vào một biến dạng DataTable và duyệt qua từng giá trị của biến. Mục đích của việc duyệt này để trong quá trình lập trình với các biến ta chỉ sử dụng biến theo cột Name (ở hình), tức khi thay đổi giá trị ở Value sẽ không phải điều chỉnh trực tiếp ở UiPath.

Mô tả bước đọc và lưu trữ email

1. Mô tả cách đọc và lưu trữ email

Nhóm sẽ sử dụng giao thức IMAP và sử dụng hoạt động Get IMAP Mail Message, do IMAP có ưu điểm là hỗ trợ tốt việc có quá nhiều email gửi đến đến một ngày, tiết kiệm không gian lưu trữ cục bộ. Đặc biệt dữ liệu email là rất quan trọng đối với doanh nghiệp, IMAP sẽ lưu trữ chúng trên Mail Servers giúp lưu trữ và dự phòng tốt hơn.

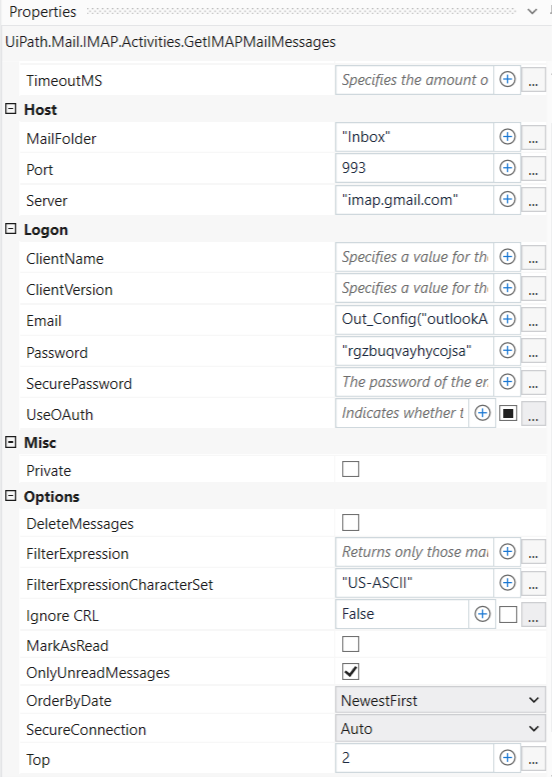
Đối với việc IMAP, cần thiết lập các giá trị sau đây để thực hiện.

Port: 993.

Server: "imap.gmail.com".

1. Thực hiện bước đọc và lưu trữ email trên UiPath

Kéo thả hoạt động Get IMAP Mail Message và tiến hành thiết lập các giá trị ở phần thuộc tính (properties).



**Hình 4‑11** Hoạt động GetImapMailMessages

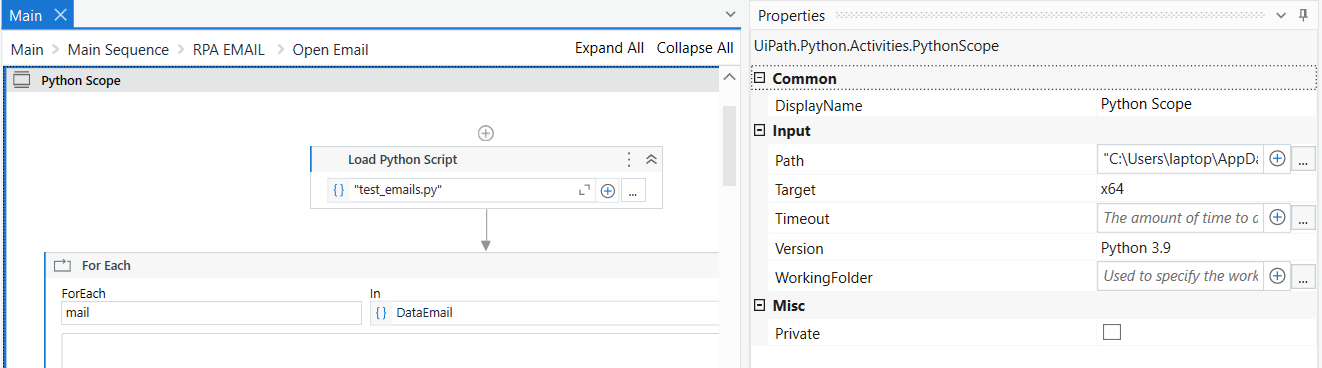
Nhập các giá trị “Mail Folder” là Inbox, Port là 993, Server là imap.gmail.com.

Tùy chỉnh các cột như MarkAsRead (sau khi duyệt qua sẽ đánh dấu là đã đọc thư), Top là số lượng thư muốn đọc. Sau đó lưu vào một biến dạng List <Mail Message>.

Mô tả bước hết nối giữa Python và UiPath

Nhóm sử dụng gói mở rộng UiPath.Python.Activities giúp hỗ trợ kết nối Python với UiPath. Khi sử dụng gói mở rộng này ta phải thực hiện tuần tự các hoạt động sau: Python Scope, Load Python Script, Invoke Python Method, Get Python Object.

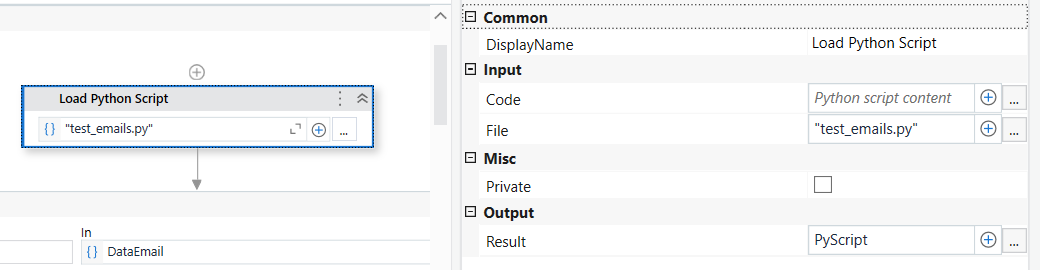
Đối với Python Scope, ta cần truyền đường dẫn file Python vào cột Path, và chọn phiên bản Python tương ứng với python đã tải ở máy.



**Hình 4‑12** Hoạt động Python PythonScope

Hoạt động Load Python Script giúp ta truyền đường dẫn file python. Ta dùng hoạt động này để truyền đường dẫn file python của mô hình phân loại văn bản, sau đó lưu trữ nó vào một biến dạng Python Object.

Đối với 2 hoạt động Invoke Python Method, Get Python Object sẽ được thực hiện ở phần tiếp theo.

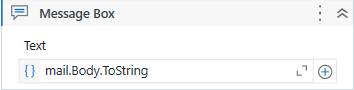


**Hình 4‑13** Hoạt động Load Python Script

Quy trình sử dụng mô hình phân loại văn bản với UiPath

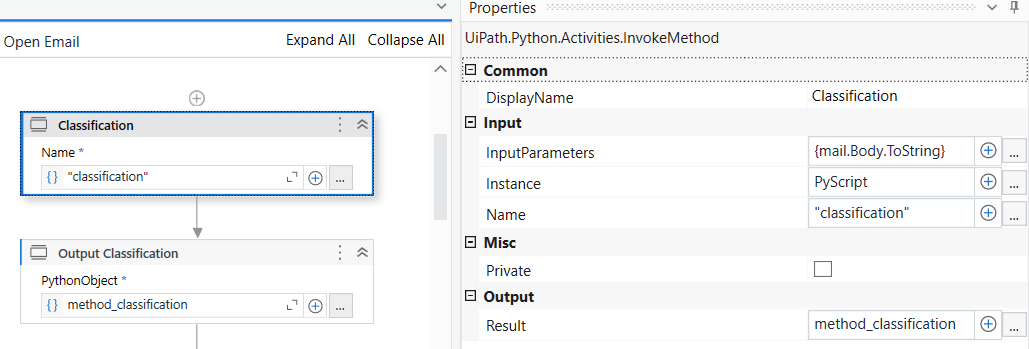
Sau khi đã truyền mô hình phân loại văn bản vào UiPath ở bước 4.4, tiếp theo ta cần cung cấp nội dung các email cho mô hình.

Ta duyệt qua nội dung từng email, sử dụng hoạt động Message Box và email.Body để hiện lên nội dung của email đó.

**

**Hình 4‑14** Hoạt động Message Box hiển thị nội dung của email

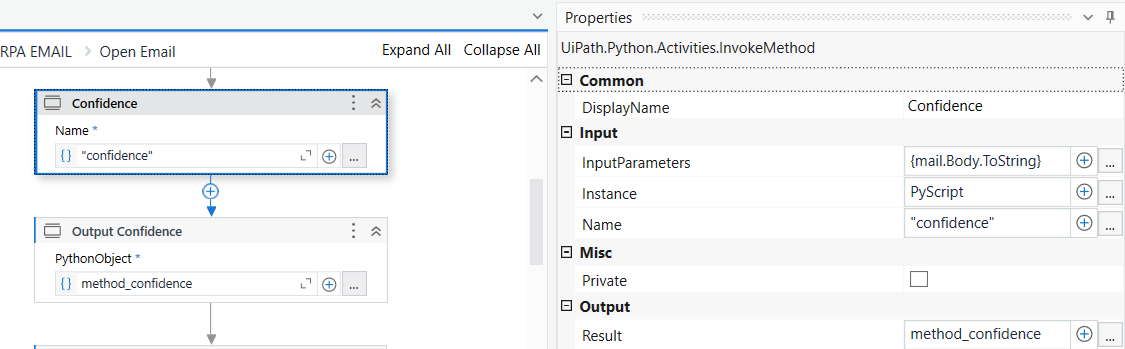
Để gọi hàm phân loại văn bản, ta sử dụng hoạt động Python Invoke Method và Get Python Object. Tham số truyền vào là nội dung của email, đối tượng là biến Python Object lưu trữ mô hình phân loại văn bản.

**

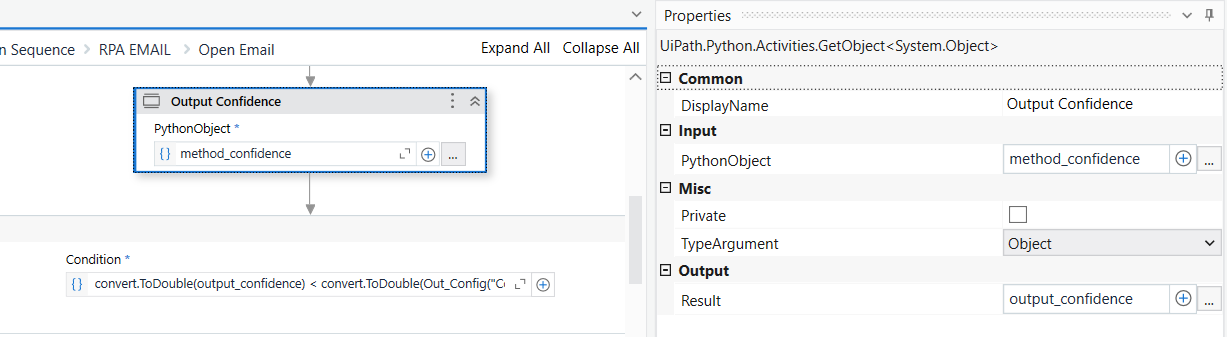
**Hình 4‑15** Hoạt động Invoke Method gọi hàm phân loại email

Từng email sẽ được mô hình phân loại và trả về 2 kết quả là kết quả phân loại về phòng ban và độ chính xác của việc phân loại. Như đã trình bày trên lưu đồ, dựa vào độ chính xác của từng email, hàm kiểm tra có thực hiện bước “human in loop” hay tự động chuyển tiếp email và tiến tới email tiếp theo.

Tiếp tục sử dụng hoạt động Python Invoke Method và Get Python Object để gọi hàm hiển thị độ chính xác. Sử dụng hoạt động Message Box để hiện kết quả phân loại phòng ban và chỉ số chính xác của email.

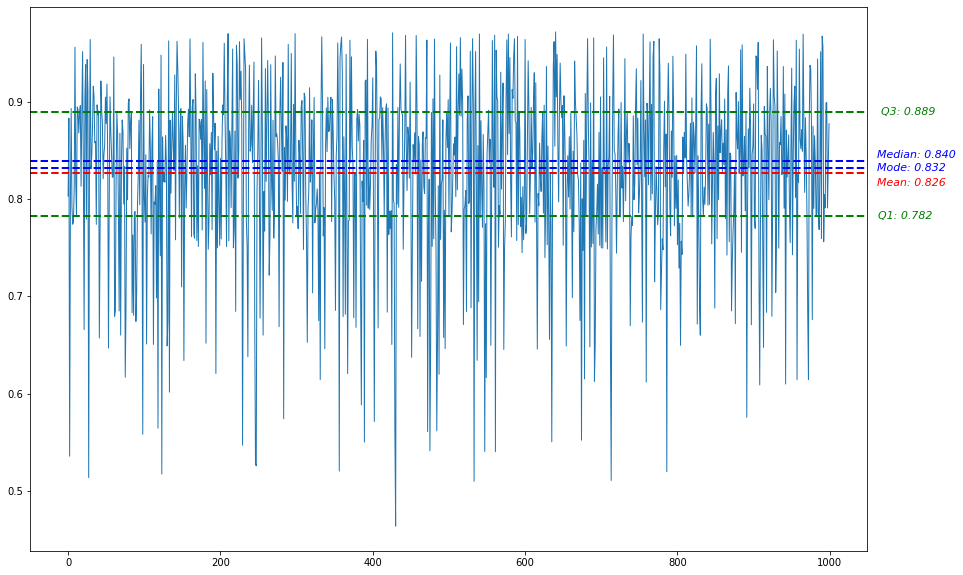


**Hình 4‑16** Hoạt động Invoke Python Method gọi hàm hiển thị độ chính xác



**Hình 4‑17** Hoạt động Get Python Object trả về kết quả phân loại phòng ban của email và độ chính xác của nó

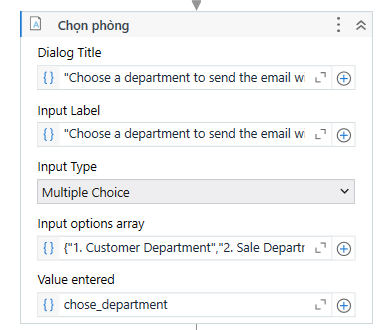
Sau khi các email được trả về độ chính xác phân loại của mình, quy trình sẽ tiếp tục đến phần so sánh độ chính xác của email với độ tin cậy đầu vào. Để xác định độ tin cậy đầu vào, nhóm nghiên cứu thực hiện chạy thử nghiệm 1000 email và thu được kết quả độ chính xác phân loại của từng email như sau:



**Hình 4‑18** Kết quả thử nghiệm quy trình trên 1000 email

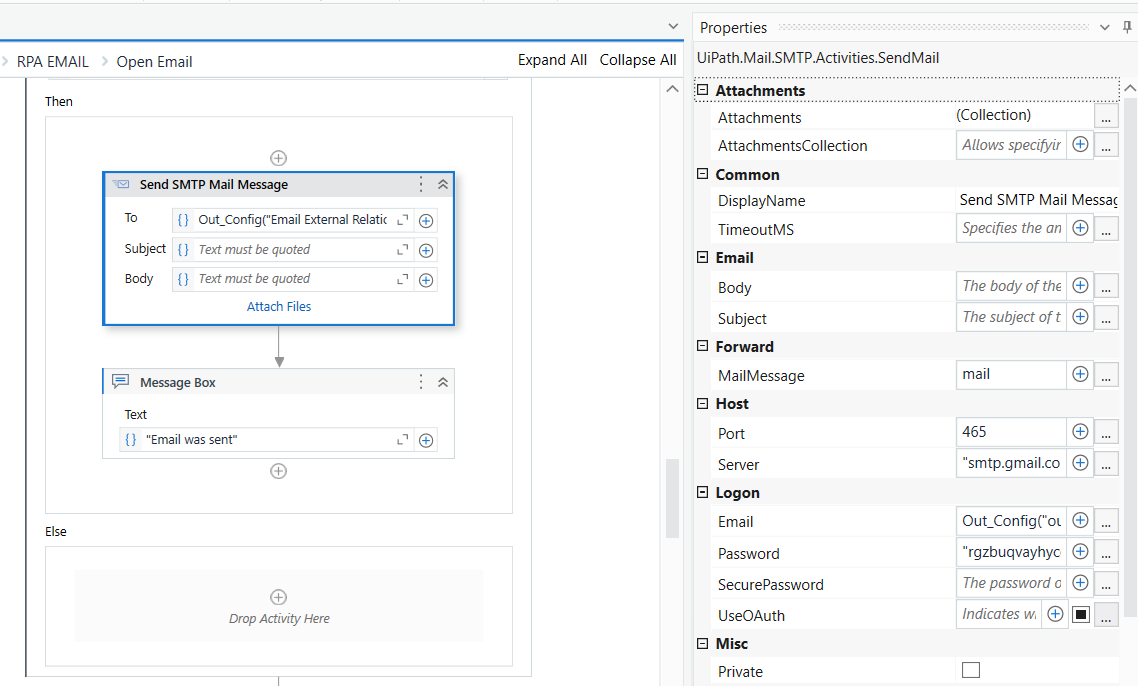
Xét kết quả tứ phân vị cho thấy, có trên 50% email có độ chính xác phân loại trên 82.6% và có trên 75% email có độ chính xác phân loại trên 78.2%. Ưu tiên theo độ tập trung của giá trị phân loại, nhóm nghiên cứu quyết định giá trị tin cậy của mô hình sẽ là 78.2% hay 0.782. Sau khi các email được trả về độ chính xác phân loại của mình, quy trình sẽ tiếp tục đến phần so sánh độ chính xác của email với độ tin cậy đầu vào (giá trị được đặt là là 0.782). Nếu độ chính xác của email cao hơn hoặc bằng 0.782 thì sẽ chuyển tiếp email đến phòng ban nó được phân loại. Ngược lại, nếu độ chính xác của email thấp hơn hơn 0.782 sẽ tiến hành bước “human in loop”.

Bước “human in loop” sẽ thông báo với lí do độ chính xác thấp cần có sự can thiệp của con người, và hiển thị lại nội dung của email. Một hộp lựa chọn phòng ban phù hợp với email đó hiện lên để ta lựa chọn. Hộp thoại này sẽ có 5 lựa chọn bao gồm 3 phòng ban mặc định là phòng chăm sóc khách hàng (Customer Care Department), phòng kinh doanh (Sales Department), phòng đối ngoại (External Relation Department), thư rác (Spam) và other (nếu chọn lựa chọn này sẽ cho người dùng nhập địa chỉ email mình muốn gửi).

**

**Hình 4‑19** Hoạt động Dialogbox hiển thị hộp thoại loại chọn phòng ban.

Sau khi lựa chọn tại bước “human in loop”, email sẽ được chuyển tiếp địa email tương ứng, email spam sẽ được đưa vào thùng rác. Và tất nhiên các email có độ chính xác lớn hơn hoặc bằng 0.782 sẽ được tự động chuyển tiếp đến địa chỉ phòng ban đã được phân loại. Ta sử dụng hoạt động “Send SMTP Mail Message” để thực hiện việc chuyển tiếp email, ở đây ta cần truyền các tham số tiêu biểu như Host:465 và Server: “smtp.gmail.com”. Sau đó là các phần thông tin tài khoản Outlook và truyền biến mail đã được lưu trữ ở hoạt động Get Imap Mail Message ở phần chuyển tiếp (Forward). Tiếp tục sử dụng hoạt động Message Box để thông báo email đã được chuyển tiếp và hoàn thành quy trình tự động hóa.

**

**Hình 4‑20** Hoạt động SendSMTP Mail Message chuyển tiếp email và thông báo email đã được chuyển tiếp

# KẾT QUẢ NGHIÊN CỨU

## Đánh giá và lựa chọn phương pháp phân loại

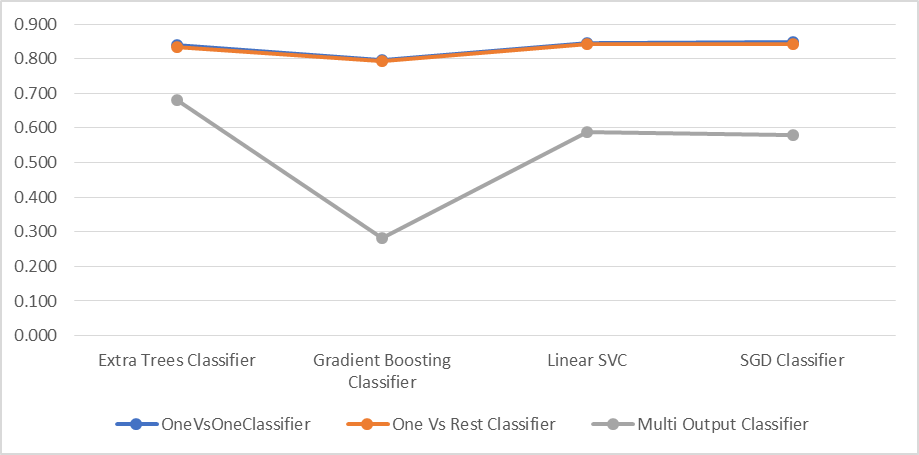
Để đưa ra phương pháp phân loại thích hợp, nhóm nghiên cứu đã thực hiện so sánh hiệu quả của 2 phương pháp:

* Phân loại đa lớp: Các email sẽ được phân loại trên một nhãn duy nhất và nhãn sẽ gồm 4 lớp tượng trưng cho 3 phòng ban và thư rác (Customer Care, External Relationship, Sales, Spam Email). Với phương pháp này, nhóm nghiên cứu sử dụng hai thuật toán là One-Vs-One và One-Vs-Rest.
* Phân loại đa lớp - đa đầu ra (phân loại đa nhiệm): Các email sẽ được phân loại trên 2 nhãn. Nhãn thứ nhất bao gồm 2 lớp là thư rác (1) và không phải thư rác (0). Đối với những email không phải thư rác sẽ được phân loại với nhãn 2 gồm ba lớp tượng trưng cho ba phòng ban trong bài toán. Với phương pháp này thì Multi-Output sẽ là thuật toán được sử dụng.

Huấn luyện dữ liệu với từng phương pháp, phân tích độ chính xác, nhận được kết quả như sau:

**Bảng 5‑1** Độ phân loại chính xác của các phương pháp phân loại

|  |  |  |  |
| --- | --- | --- | --- |
| Model | OneVsOneClassifier | One Vs Rest Classifier | Multi Output Classifier |
| Extra Trees Classifier | 0,839 | 0,834 | 0,682 |
| Gradient Boosting Classifier | 0,797 | 0,793 | 0,282 |
| Linear SVC | 0,845 | 0,842 | 0,589 |
| SGD Classifier | 0,847 | 0,842 | 0,580 |



**Hình 5‑1** Kết quả so sánh độ chính xác của các phương pháp phân loại

 Dựa vào kết quả phân tích cho thấy, phương pháp phân loại đa lớp đem lại hiệu quả cao hơn với độ chính xác trung bình là 83.2%, cao gấp 1.56 lần so với phương pháp phân loại đa nhiệm (chỉ có 53.3%). Chính vì thế, nhóm nghiên cứu quyết định lựa chọn phương pháp phân loại đa lớp sẽ là phương pháp giải quyết bài toán được đặt ra. Xét thấy trong phương pháp đa nhãn thì 2 thuật toán One-Vs-One và One-Vs-Rest đều cho kết quả gần bằng nhau (chênh lệch 0.48%) nên không thể chỉ dựa vào độ chính xác mà lựa chọn ngay thuật toán mà cần sử dụng những chỉ số đánh giá khác để lựa chọn được thuật toán đem lại hiệu quả phân loại cao nhất. Vấn đề đó sẽ được trình bày ở phần tiếp theo “**Đánh giá và lựa chọn mô hình phân loại**”.

## Đánh giá và lựa chọn mô hình phân loại

Nhóm nghiên cứu đã đề xuất cụ thể 4 mô hình: Trình phân loại giảm dần độ dốc ngẫu nhiên (SGD Classifier); Trình phân loại véc-tơ hỗ trợ tuyến tính (Linear SVC); Trình phân loại tăng cường độ dốc (Gradient Boosting Classifier) và Trình phân loại cây bổ sung (Extra Trees Classifier). Sau khi thực hiện huấn luyện và thử nghiệm đánh giá kết quả thì đạt được kết quả như sau:

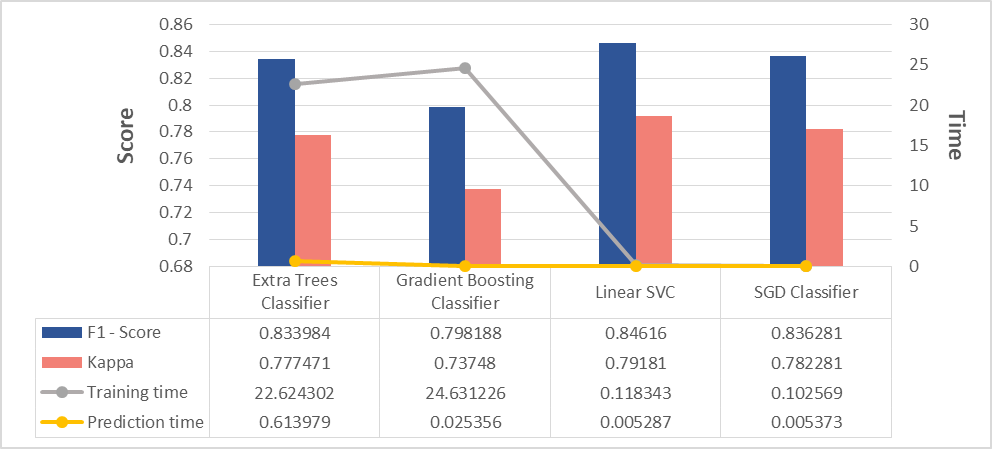
**Bảng 5‑2**  Kết quả thử nghiệm với từng mô hình

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | F1 - Score | Kappa | Training time | Prediction time |
| Extra Trees Classifier | 0,8301 | 0,7726 | 27,0909 | 0,4232 |
| Gradient Boosting Classifier | 0,8033 | 0,7488 | 26,9546 | 0,0164 |
| Linear SVC | 0,8445 | 0,7908 | 0,1122 | 0,0032 |
| SGD Classifier | 0,8385 | 0,7857 | 0,0877 | 0,0037 |

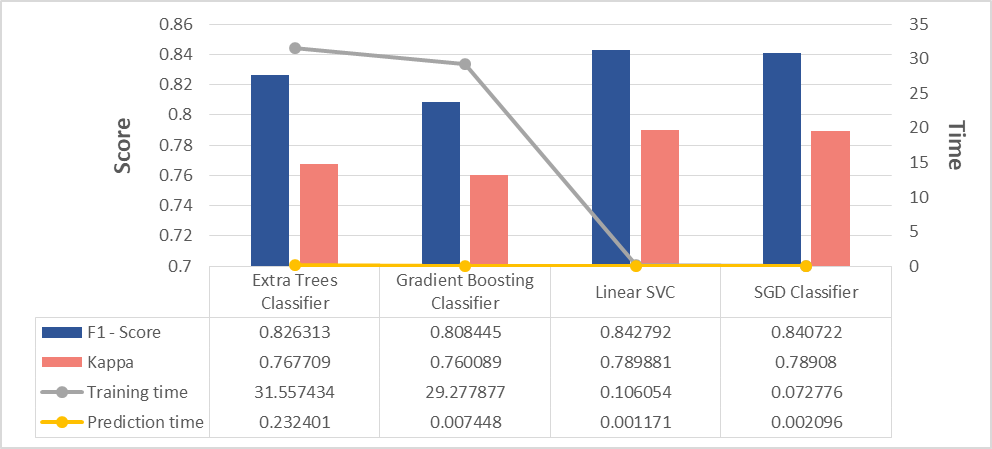
Khi thực hiện đánh giá từng mô hình thì với mô hình Linear SVC đạt kết quả cao nhất với điểm F1 là 84.45% và điểm Kappa là 79.08%. Xét theo thời gian huấn luyện - thử nghiệm thì mô hình Linear SVC đạt thời gian dự đoán ngắn nhất với 0.0032 giây. Về thời gian huấn luyện thì mô hình Linear SVC (0.1122s) xếp thứ hai sau mô hình SGD Classifier (0.0877), mặc dù chênh lệch về thời gian huấn luyện nhưng các chỉ số đánh giá độ chính xác của mô hình Linear SVC vẫn cao hơn. Để có đánh giá tổng quan, nhóm nghiên cứu sẽ phân tích sự đánh giá của mô hình kết hợp với các thuật toán phân loại đa lớp.

**Bảng 5‑3** Kết quả thử nghiệm các mô hình phân loại

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| STT | Mô hình | Phân loại đa nhãn | Điểm F1 | Hệ số Kappa | Thời gian tập huấn | Thời gian thử nghiệm |
| 1 | SGD Classifier | *One Vs Rest* | 0,840722 | 0,78908 | 0,072776 | 0,002096 |
| *One Vs One* | 0,836281 | 0,782281 | 0,102569 | 0,005373 |
| 2 | **Linear SVC** | *One Vs Rest* | 0,842792 | 0,789881 | 0,106054 | 0,001171 |
| ***One Vs One*** | **0,84616** | **0,79181** | **0,118343** | **0,005287** |
| 3 | Gradient Boosting Classifier | *One Vs Rest* | 0,808445 | 0,760089 | 29,277877 | 0,007448 |
| *One Vs One* | 0,798188 | 0,73748 | 24,631226 | 0,025356 |
| 4 | Extra Trees Classifier | *One Vs Rest* | 0,826313 | 0,767709 | 31,557434 | 0,232401 |
| *One Vs One* | 0,833984 | 0,777471 | 22,624302 | 0,613979 |



**Hình 5‑2** Áp dụng phương pháp phân loại đa lớp One-Vs-One

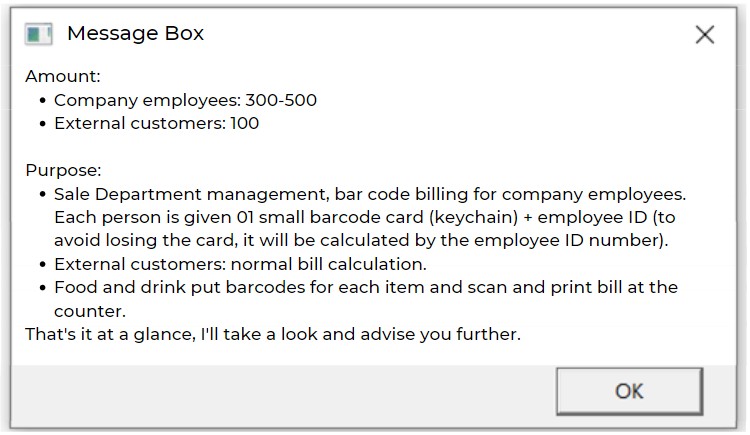


**Hình 5‑3** Áp dụng phương pháp phân loại đa lớp One-Vs-Rest

Thực hiện đánh giá lần lượt 4 mô hình với thuật toán One-Vs-One và One-Vs-Rest thì mô hình Linear SVC vẫn tiếp tục đạt giá trị cao nhất ở cả hai thuật toán với điểm F1 của thuật toán OvO là 84.616% và OvR là 84.279%; với điểm Kappa của OvO là 79.18% và OvR là 78.99%. Về thời gian huấn luyện, tuy có thấp hơn mô hình SGD Classifier nhưng độ chênh lệch rất thấp (chưa đến 0.5 lần) và thời gian huấn luyện là ngắn nhất trong 4 mô hình. Xét về thời gian huấn luyện - thử nghiện của mô hình Linear SVC với 2 thuật toán thì thuật toán OvO dài hơn OvR với độ chênh lệch trung bình là 1.815, tuy nhiên về các chỉ số đánh giá F1 và Kappa thì OvO đều cao hơn OvR. Đặt tiêu chí độ chính xác là ưu tiên nên nhóm nghiên cứu quyết định lựa chọn mô hình nghiên cứu là **Linear SVC kết hợp thuật toán phân loại đa lớp One-Vs-One**.

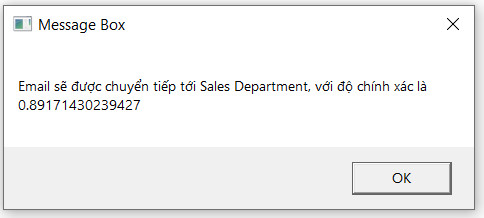
## Đánh giá quy trình tự động hóa RPA

Để bắt đầu quá trình, hệ thống RPA sẽ đọc nội dung email cần kiểm tra và hiện hộp thông báo như trong Hình 5-4.

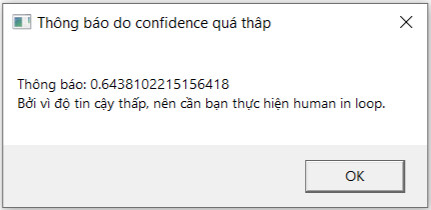


**Hình 5‑4** Đọc nội dung của email và thông báo (using UiPath)

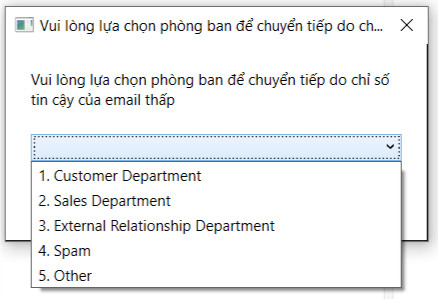
Độ tin cậy của từng bộ phận sẽ được dự đoán và thể hiện trong hộp thoại này. Hình 5-5 cho thấy nếu độ tin cậy nhỏ hơn 0,782, người dùng sẽ nhận được thông báo để xem xét và xử lý email. Sau đó, bot sẽ làm human in loop.



**Hình 5‑5** Thông báo email có độ tin cậy cao hơn 0.782 sẽ được chuyển tiếp đến phòng ban được phân loại.

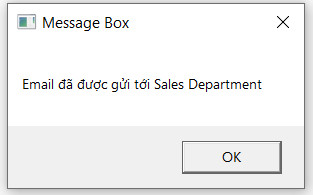


**Hình 5‑6** Thông báo email có độ tin cậy thấp hơn 0.782 sẽ tiến tới bước human in loop



**Hình 5‑7** Hoạt động cần sự can thiệp của con người “human in loop”

Nếu tùy chọn "Other" được chọn, người dùng sẽ được yêu cầu nhập địa chỉ email để gửi tới. Sau khi hoàn tất việc phân loại và chuyển tiếp email, hệ thống sẽ thông báo cho người quản trị.

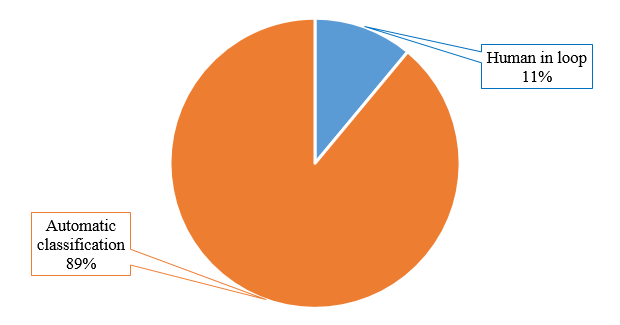


**Hình 5‑8** Thông báo đã hoàn tất việc chuyển tiếp email

Áp dụng mô hình trên, nhóm thử nghiệm với toàn bộ tệp dữ liệu của mình (5730 email). Bao gồm việc chỉ phân loại các email nào được tự động và cần sự can thiệp của con người với chỉ số tin cậy là 0.782. Nhóm thu được kết quả có 4610 email được phân loại tự động (chiếm 80%) và 1120 email cần sự can thiệp của con người (chiếm 20%), thời gian chạy là 35 phút.

Áp dụng mô hình này trên tập dữ liệu (1000 email) được lấy từ tệp dữ liệu của nhóm, kết quả đưa ra có 890 email được phân loại tự động và 110 email cần sự can thiệp của con người trong 3 phút 30 giây.

Ta có thể thấy sự chênh lệch về thời gian giữa 2 lần thử, tuy nhiên do số lượng email lớn cộng thêm việc thực hiện chạy mô hình trên máy tính cá nhân của nhóm, quá trình chạy cũng sẽ bị ảnh hưởng không ít bởi cấu hình máy. Tuy nhiên, các kết quả đều cho kết quả khả quan, cho thấy được tốc độ phân loại nhanh và độ chính xác cao khi so với con người khi thực hiện.



**Hình 5‑9** So sánh số lượng email được tự động phân loại và các email cần sự can thiệp của con người trường hợp 1000 email

# KẾT LUẬN VÀ ĐỀ XUẤT

## Tổng kết nội dung đã thực hiện

Với đề tài “Thiết kế hệ thống RPA tích hợp NLP để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”, nhóm đã thực hiện những nội dung sau:

***Nghiên cứu và phân tích lý thuyết***

Nhóm đã nghiên cứu và phân tích những sách, báo, nghiên cứu khoa học và những tài liệu liên quan để tổng hợp kiến thức, công nghệ hiện tại nhằm phát triển đề tài của nhóm.

***Áp dụng quy trình xử lý ngôn ngữ tự nhiên (NLP)***

Để xử lý dữ liệu ngôn ngữ trong đề tài, nhóm đã áp dụng NLP nhằm đưa dữ liệu về cấu trúc chuẩn trong Học máy về ngôn ngữ.

***So sánh và đánh giá các phương pháp - thuật toán - mô hình Học máy phân loại***

Nhóm đã so sánh 2 phương pháp Học máy phân loại bao gồm phương pháp Đa lớp (thực hiện bằng thuật toán OvO hoặc OvR) và phương pháp Đa nhiệm (thực hiện bằng thuật toán Multi-Output) kết hợp với 4 mô hình Trình phân loại giảm dần độ dốc ngẫu nhiên (SGD Classifier); Trình phân loại véc-tơ hỗ trợ tuyến tính (Linear SVC); Trình phân loại tăng cường độ dốc (Gradient Boosting Classifier) và Trình phân loại cây bổ sung (Extra Trees Classifier) nhằm tìm ra mô hình phân loại phù hợp nhất. Sau kết quả đánh giá, nhóm đã lựa chọn phương pháp Đa lớp trên thuật toán OvO kết hợp mô hình Trình phân loại véc-tơ hỗ trợ tuyến tính (Linear SVC).

***Xây dựng quy trình tự động hóa (RPA)***

Nhằm tự động hóa hoàn toàn quy trình chuyển tiếp email trong doanh nghiệp, nhóm đã áp dụng công nghệ RPA để nhận email thô được gửi đến hộp thư doanh nghiệp và chuyển tiếp chúng về phòng ban phù hợp.

***Hoàn thiện hệ thống và đưa ra đề xuất cho doanh nghiệp***

Nhóm đã hoàn thiện hệ thống RPA tích hợp NLP và Machine Learning để tự động hóa quy trình chuyển tiếp email và đưa ra đề xuất ứng dụng trong doanh nghiệp.

## Đề xuất

Doanh nghiệp thường có một email duy nhất để đại diện toàn công ty - đây thường là email cơ bản với địa chỉ dựa trên tên của doanh nghiệp. Tuy nhiên trên thực tế, doanh nghiệp có rất nhiều phòng ban và người đại diện cho từng đầu việc, do đó khách hàng và đối tác của doanh nghiệp thường được yêu cầu liên hệ email cho từng phòng ban hoặc người đại diện cụ thể để quá trình giải quyết công việc được nhanh chóng hơn. Đây là cách mà các doanh nghiệp hiện nay thường áp dụng để tránh quá tải email chính thức hoặc tránh nhập nhằng trong quá trình nhận và xử lý email. Cách này tuy đang được sự đồng thuận giữa các khách hàng và doanh nghiệp nhưng vẫn gây ra những bất lợi nhất định như khách hàng không được giải quyết vấn đề khi gửi email sai phòng ban, nhân viên các phòng ban phải xử lý những email nhầm lẫn, ... Ngược lại, với những doanh nghiệp chỉ cung cấp cho khách hàng - đối tác một email chính thức thì lại gặp bất cập trong việc bổ sung nhân viên văn thư. Cả hai cách này đều gây hao tổn nhân lực và thời gian của cả doanh nghiệp và khách hàng.

Do đó, nhóm nghiên cứu đã đề xuất một hệ thống chuyển tiếp email tự động về các phòng ban trong doanh nghiệp. Với hệ thống này, các doanh nghiệp chỉ cần cung cấp một địa chỉ email đại diện chính thức cho tất cả khách hàng - đối tác của mình. Hệ thống sẽ thay thế quy trình xử lý email thủ công hiện tại. Các email sau khi gửi về hộp thư đại diện sẽ được tự động chuyển tiếp về phòng ban tương ứng, hạn chế sự sai sót cả về phía khách hàng - đối tác lẫn phía doanh nghiệp. Hệ thống này sẽ nâng cao trải nghiệm của khách hàng đối với doanh nghiệp cũng như giải phóng nguồn nhân lực cho doanh nghiệp.

## Hạn chế của nghiên cứu

***Độ lớn và độ đa dạng của dữ liệu còn hạn chế.***

Bộ dữ liệu nhóm thu thập được có độ lớn là 5732 email. Đây vẫn là một con số khiêm tốn so với các bộ dữ liệu được sử dụng trong Học máy phân loại văn bản. Kết quả phân loại có thể chính xác hơn với bộ dữ liệu huấn luyện lớn hơn.

Bên cạnh đó, trên thực tế, các doanh nghiệp thường có rất nhiều phòng ban như phòng Tài chính, phòng Kế toán, phòng Nhân sự,... cũng như tùy vào từng lĩnh vực hoạt động mà doanh nghiệp sẽ còn có nhiều phòng ban khác nhau. Tuy nhiên trong bài nghiên cứu này, bộ dữ liệu của nhóm chỉ bao gồm ba phòng ban là “Customer Care” (Chăm sóc khách hàng), “External Relations” (Đối ngoại) và “Sales” (Bán hàng). Hệ thống phân loại sẽ dễ dàng áp dụng vào doanh nghiệp hơn nếu được huấn luyện với dữ liệu của nhiều phòng ban thực tế hơn.

***Quá trình tiền xử lý dữ liệu còn chưa triệt để.***

Các từ viết tắt, từ dừng, tiếng lóng hay kí hiệu vẫn chưa được xử lý triệt để nên có thể gây ảnh hưởng đến kết quả dự đoán của mô hình. Bên cạnh đó, những từ được sử dụng thường xuyên như “thank” (cảm ơn), “regard” (trân trọng), “good” (tốt), ... tuy vẫn mang ý nghĩa nhưng trong một số trường hợp lại không nằm trong ngữ cảnh chính của văn bản - có thể gây nhiễu đến kết quả.

## Phương pháp phát triển

Thu thập dữ liệu ở nhiều doanh nghiệp và phòng ban khác để mở rộng chức năng và số lượng phòng ban nhằm phù hợp với cơ cấu của nhiều doanh nghiệp hơn.

Bên cạnh đó, ta có thể phát triển ứng dụng NLP để xử lý những ngôn ngữ khác ngoài tiếng Anh nhằm xử lý email tiếng Việt, email song ngữ Anh - Việt hoặc các ngôn ngữ khác. Từ đó mở rộng phạm vi áp dụng cho các doanh nghiệp Việt Nam và nước ngoài.

TÀI LIỆU THAM KHẢO

Abhishek, L. (2020). *Optical character recognition using ensemble of SVM, MLP and extra trees classifier.* Paper presented at the 2020 International Conference for Emerging Technology (INCET).

Abraham, A., Dutta, P., Mandal, J. K., Bhattacharya, A., & Dutta, S. (2018). Emerging technologies in data mining and information security. *Proceedings of IEMIS-2018*.

Akshay, P., Kalagi, N., Shetty, D., & Ramalingam, H. (2020). EMAIL CLIENT AUTOMATION WITH RPA.

Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., & Abu-Naser, S. S. (2018). Email classification using artificial neural network.

Bahgat, E. M., Rady, S., Gad, W., & Moawad, I. F. (2018). Efficient email classification approach based on semantic methods. *Ain Shams Engineering Journal, 9*(4), 3259-3269.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology, 188*(12), 2222-2239.

Bowd, C., Belghith, A., Proudfoot, J. A., Zangwill, L. M., Christopher, M., Goldbaum, M. H., . . . Weinreb, R. N. (2020). Gradient-boosting classifiers combining vessel density and tissue thickness measurements for classifying early to moderate glaucoma. *American journal of ophthalmology, 217*, 131-139.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., . . . Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). *Intensive working group of ACM SIGKDD curriculum committee, 140*, 1-10.

Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A., & Kole, D. K. (2019). *Flight arrival delay prediction using gradient boosting classifier.* Paper presented at the Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2.

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review, 52*(2), 803-855.

Chui, M., Manyika, J., & Bughin, J. (2012). *The social economy: Unlocking value and productivity through social technologies*. Retrieved from

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications, 28*(2), 37-40.

Deepa, N., Prabadevi, B., Maddikunta, P. K., Gadekallu, T. R., Baker, T., Khan, M. A., & Tariq, U. (2021). An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *The Journal of Supercomputing, 77*, 1998-2017.

Dhuliawala, S., Kanojia, D., & Bhattacharyya, P. (2016). *Slangnet: A wordnet like resource for english slang.* Paper presented at the Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

Điện, T. T., Thanh, T. N., & Nghe, N. T. (2019). Giải pháp phân loại bài báo khoa học bằng kĩ thuật máy học. *Tạp chí Khoa học Trường Đại học Cần Thơ, 55*(4), 29-37.

Duc, T. M. (2014). SPAM EMAIL FILTERING BASED ON MACHINE LEARNING.

Dürscheid, C., Frehner, C., Herring, S. C., Stein, D., & Virtanen, T. (2013). Email communication. *Handbooks of Pragmatics [HOPS]*(9), 35-54.

Fernandes, K. (2008). *On the significance of speech: How infants discover symbols and structure.* New York University,

Gupta, G., & Malhotra, S. (2015). Text document tokenization for word frequency count using rapid miner (taking resume as an example). *Int. J. Comput. Appl, 975*, 8887.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, 1*(1), 60-76.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2): Springer.

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access, 10*, 19083-19095.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics, 7*(2), 139-172.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). *Stemming and lemmatization in the clustering*

*of finnish text documents.* Paper presented at the Proceedings of the thirteenth ACM international conference on Information and knowledge management.

Lâm, T. D. (2020). *Nghiên cứu các phương pháp lọc thư rác tại việt nam và trên thế giới, xây dựng và đề xuất phương án lọc thư rác tiếng Việt.*

Madakam, S., Holmukhe, R. M., & Jaiswal, D. K. (2019). The future digital work force: robotic process automation (RPA). *JISTEM-Journal of Information Systems and Technology Management, 16*.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet], 9*, 381-386.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Mohammadi, M., Malekian, K., Nosrati, M., & Karimi, R. (2013). Email Marketing as a Popular Type of Small Business Advertisement: A Short Review. *Australian journal of basic and applied sciences, 7*(4), 786-790.

Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N., & Al-Garadi, M. A. (2017). Email classification research trends: review and open issues. *IEEE Access, 5*, 9044-9064.

Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment analysis-strategy for text pre-processing. *Int. J. Comput. Appl, 180*(34), 15-18.

Patel, M., Shukla, A., Porwal, R., & Kotecha, R. (2019). *Customized Automated Email Response Bot Using Machine Learning and Robotic Process Automation.* Paper presented at the 2nd International Conference on Advances in Science & Technology (ICAST).

Plisson, J., Lavrac, N., & Mladenic, D. (2004). *A rule based approach to word lemmatization.* Paper presented at the Proceedings of IS.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Sampson, G. (2005). *The'Language Instinct'Debate: Revised Edition*: A&C Black.

Shah, R., Lahoti, S., & Lavanya, K. (2017). An intelligent chat-bot using natural language processing. *International Journal of Engineering Research, 6*(5), 281-286.

Tam, T., Ferreira, A., & Lourenço, A. (2012). *Automatic foldering of email messages: a combination approach.* Paper presented at the Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings 34.

Toàn, H. P., Lâm, N. V., Nghị, Đ. T., & Trung, N. M. (2011). PHÂN LOẠI THƯ RÁC VỚI GIẢI THUẬT BOOSTING CÂY QUYẾT ĐỊNH NGẪU NHIÊN XIÊN PHÂN ĐƠN GIẢN. *Tạp chí Khoa học Trường Đại học Cần Thơ*(19b), 1-9.

Trilla, A. (2009). Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. *Departament de Tecnologies Media*, 1-5.

Trịnh, H. T. (2014). *Một số vấn đề về bảo mật thư điện tử.* ĐHQGHN,

Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems, 7*(2), 16-18.

Whitney, D. L., & Evans, B. W. (2010). Abbreviations for names of rock-forming minerals. *American mineralogist, 95*(1), 185-187.

Zong, Z., & Hong, C. (2018). *On application of natural language processing in machine translation.* Paper presented at the 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE).

Phụ Lục

1. The Radicati Group. *Number of e-mails per day worldwide 2017-2025. (*2021): *Statista*. Web 11 Mar.2023 [↑](#footnote-ref-1)
2. Trustwave (SpiderLabs). *Most prevalent spam content categories worldwide in 2020.* (2021): *Statista*. Web 11 Mar.2023 [↑](#footnote-ref-2)
3. Trustwave (SpiderLabs). *Most prevalent spam content categories worldwide in 2020.* (2021): *Statista*. Web 11 Mar.2023 [↑](#footnote-ref-3)