|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP.HCM  **TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

*TP. Hồ Chí Minh, ngày 11 tháng 09 năm 2022*

**THUYẾT MINH**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM 2022**

# THÔNG TIN CHUNG

1. **Tên đề tài:** **THIẾT KẾ HỆ THỐNG RPA TÍCH HỢP NLP ĐỂ TỰ ĐỘNG HÓA QUY TRÌNH CHUYỂN TIẾP EMAIL TRONG DOANH NGHIỆP.**
2. **Loại hình nghiên cứu:**

☐ Nghiên cứu cơ bản

🗶 Nghiên cứu ứng dụng

☐ Nghiên cứu triển khai

1. **Lĩnh vực nghiên cứu:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **TÊN LĨNH VỰC** | **CHUYÊN NGÀNH** | **Chọn** |
| **1** | Lĩnh vực Giáo dục | - Giáo dục học  - Quản lý giáo dục  - Tâm lý giáo dục  - Giáo dục thể chất | **⬜**  **⬜**  **⬜**  **⬜** |
| **2** | Lĩnh vực Kinh tế | - Tài chính - ngân hàng - chứng khoán - kế toán - Kiểm toán, bảo hiểm - tín dụng  - Thương mại - quản trị kinh doanh và du lịch - marketing  - Kinh tế học - kinh tế phát triển - kinh tế chính trị | **⬜**  **⬜**  **⬜** |
| **3** | Lĩnh vực Pháp lý | - Luật Dân sự  - Luật Hình sự  - Luật Kinh tế  - Luật Hành chính  - Luật Quốc tế | **⬜**  **⬜**  **⬜**  **⬜**  **⬜** |
| **4** | Lĩnh vực Công nghệ thông tin | - Toán tin  - Công nghệ thông tin | **⬜**  **🗶** |

1. **Thời gian thực hiện**: 6 tháng
2. **Người hướng dẫn**

Học hàm, học vị, họ và tên: Tiến sĩ Nguyễn Thôn Dã

Khoa/Bộ môn/Trung tâm: Khoa Hệ thống thông tin

Điện thoại: 0906230232 Email: dant@uel.edu.vn

1. **Sinh viên thực hiện đề tài**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Lớp** | **Khóa** | **Khoa** | **Email** | **Nhiệm vụ** |
| **1** | Nguyễn Quang Học | K214060427 | K21406C | K21 | Hệ thống thông tin | hocnq21406c@st.uel.edu.vn | Nhóm trưởng |
| **2** | La Thế Anh | K214060389 | K21406T | K21 | Hệ thống thông tin | anhlt21406@st.uel.edu.vn | Thành viên |
| **3** | Vũ Lưu Hoàng Lan | K214060429 | K21406T | K21 | Hệ thống thông tin | lanvlh21406c@st.uel.edu.vn | Thành viên |
| **4** | Giả Ngọc Thảo Ly | K214060430 | K21406T | K21 | Hệ thống thông tin | lygnt21406c@st.uel.edu.vn | Thành viên |
| **5** | Phan Ngọc Bảo Tâm | K214061268 | K21406C | K21 | Hệ thống thông tin | tampnb21406c@st.uel.edu.vn | Thành viên |

1. **MÔ TẢ NGHIÊN CỨU**

**B1. Tổng quan tình hình nghiên cứu trong, ngoài nước**

Hiện nay, việc phân loại email đã và đang là một trong những vấn đề quan tâm của mọi người từ cá nhân đến các tổ chức, đặc biệt là các doanh nghiệp. Đối mặt với thực trạng đó, trên thế giới cũng như ở Việt Nam đã xuất hiện nhiều bài báo nói về việc phân loại dữ liệu email, loại bỏ các thư rác theo mô hình công nghệ Machine Learning.

# B1.1 Tình hình nghiên cứu tại Việt Nam

**B1.1.1 Nghiên cứu các phương pháp phân loại văn bản và ứng dụng vào phân loại thư điện tử**

- Trong những năm gần đây, spam hay các email không mong muốn đã trở thành một vấn nạn và đe dọa khả năng giao tiếp của con người trên các kênh liên lạc thư điện tử, đó là một trong những thách thức lớn mà khách hàng và các nhà cung cấp dịch vụ cần đối phó. Vì vậy, cần một hệ thống phân loại email chính xác qua nghiên cứu các phương pháp phân loại văn bản và ứng dụng thư điện tử bằng thuật toán Naïve Bayes (Lê Thanh Trà, 2013).[[1]](#endnote-1)

**B.1.1.2 Phân loại thư rác với giải thuật boosting cây quyết định ngẫu nhiên xiên phân đơn giản**

Hằng ngày, mỗi doanh nghiệp đều nhận nhiều thư điện tử không phục vụ cho mục đích thiết thực của họ, mà thường chỉ là những thư quảng cáo, thư phản động, chơi đánh bạc, thậm chí là những đoạn mã độc hại, đồi trụy khác, mà chúng ta gọi đó là thư rác. Việc phân loại thư rác với giải thuật Boosting cây quyết định ngẫu nhiên xiên phân đơn giản (Boosting of Random Oblique Decision Stump) sẽ đạt được hiệu quả cao hơn so với SVM (Support Vector Machine) và Naïve Bayes. (Huỳnh Phụng Toàn, Nguyễn Vũ Lâm, Nguyễn Minh Trung và Đỗ Thanh Nghị, 2011).[[2]](#endnote-2)

**B1.2 Tình hình nghiên cứu trên thế giới**

- Mỗi doanh nghiệp thường nhận và gửi rất nhiều thông tin qua email mỗi ngày. Tuy nhiên, số lượng email khổng lồ đã mang lại một số khó khăn, chẳng hạn như gửi thư rác, đe dọa quyền riêng tư hoặc rò rỉ thông tin. Việc thực hiện phân loại mức độ bảo mật email của dữ liệu mất cân bằng sử dụng mạng nơ-ron nhân tạo là vô cùng cần thiết. (Huang, Jen-Wei; Chiang, Chia-Wen; Chang, Jia-Wei, 2018)[[3]](#endnote-3)

- Những kẻ gửi thư rác có thể bỏ qua các phương pháp máy học có giám sát để phân loại email spam đã được thiết lập từ lâu - đặc biệt là những phương pháp xử lý các tính năng tiêu đề. Đề xuất sử dụng một phương pháp phân loại thư rác mới sử dụng các tính năng dựa trên ngôn ngữ và nội dung email và khả năng đọc được kết hợp với các tính năng nhiệm vụ dựa trên nội dung đã sử dụng trước đây. Mặc dù chỉ áp dụng cho các email bằng tiếng Anh, nhưng kết quả cho thấy rằng phương pháp của chúng em cũng có thể là một phương tiện tuyệt vời để phân loại email spam bằng các ngôn ngữ khác. (Shams, Rushdi; Mercer, Robert E tại Hội nghị quốc tế IEEE, 2013)[[4]](#endnote-4)

**B.1.3 Đánh giá chung**

Trên thực tế còn rất nhiều nghiên cứu liên quan khác. Tuy nhiên, theo như phân tích trên, hầu như việc phân loại email đã và đang là một trong những vấn đề quan tâm của mọi người từ cá nhân đến các tổ chức. Tuy nhiên, việc phân loại email trước đó chỉ dừng lại ở việc phân loại email rác, spam, ... Các công tác còn lại vẫn do con người tiếp tục thực hiện. Điều này vẫn còn tốn khá nhiều thời gian và chi phí lãng phí. Ngoài ra, các nghiên cứu hiện vẫn còn dừng lại ở mặt lý thuyết, chưa có sự ứng dụng nhiều trong thực tiễn. Dựa trên thực trạng đó, sự cấp thiết của đề tài nhóm chúng em mang đến đó chính là một hệ thống phân loại email chính xác và gửi chuyển phát email đó đến đúng từng phòng ban.

# B2. Ý tưởng khoa học, tính cấp thiết và tính mới

# B2.1 Ý tưởng khoa học và tính cấp thiết

**B2.1.1 Giới thiệu chung**

Trong thời đại công nghệ thông tin ngày càng phát triển như hiện nay thì việc liên lạc qua Internet đã dần trở nên phổ biến, chiếm phần lớn trong các phương thức giao tiếp gián tiếp của con người. Email là một phương tiện liên lạc có chi phí thấp, hiệu quả và nhanh chóng trên Internet. Chính vì thế, nó thường được sử dụng như một kênh trao đổi thông tin chính thức giữa các doanh nghiệp. Tuy nhiên, việc xử lý email chưa bao giờ là dễ dàng, theo một thống kê của Phòng nghiên cứu Statista 2022 chỉ rõ, có hơn 300 tỷ email được gửi đi mỗi ngày, điều này có nghĩa là mỗi giây có hơn 3,5 triệu email được gửi đi, bao gồm cả email cá nhân và doanh nghiệp. Với thực tế đó, vấn đề đặt ra cho chúng ta 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.

Trong các cơ quan doanh nghiệp, việc xử lý và phân loại email được diễn ra hằng ngày bởi các nhân viên văn thư, tuy nhiên việc này tốn rất nhiều thời gian, công sức và đôi khi không chính xác. 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. Do đó tính cấp thiết cần phải có một công cụ phân loại tự động giúp cho công việc này nhanh, đơn giản và chính xác. Xuất phát từ nhu cầu ấy, nhóm chúng em quyết định chọn đề tài “Thiết kế hệ thống RPA tích hợp NPL để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp” làm đề tài nghiên cứu khoa học.

**B2.1.2 Ý tưởng nghiên cứu**

Để có thể thực hiện “Thiết kế hệ thống RPA tích hợp NPL để tự động hóa quy trình chuyển tiếp Email trong doanh nghiệp”, nhóm chúng em thực hiện 2 giai đoạn chính. Giai đoạn 1 là tìm mô hình khai phá văn bản (Text Mining) tốt nhất thông qua việc đánh giá các mô hình khai phá văn bản cho khoảng 10.000 email. Mục đích của giai đoạn này là phân loại văn bản. Trong lĩnh vực máy học (Machine Learning) và xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), phân loại văn bản là một bài toán xử lý văn bản cổ điển, nhằm phân một văn bản mới vào nhóm các văn bản cho trước dựa trên sự tương đồng của văn bản đó so với nhóm văn bản.[[5]](#endnote-5) Đặc điểm nổi bật của bài toán này là sự đa dạng của chủ đề văn bản và tính đa chủ đề của văn bản. Thực tế rằng, 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. 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 phân loại tự động. Vấn đề phân loại văn bản được nhiều nhà khoa học quan tâm với các hướng tiếp cận khác nhau. Một cách tiếp cận được nhiều nhà nghiên cứu sử dụng là phương pháp máy học, có thể liệt kê một số thuật toán được nhiều nhà nghiên cứu sử dụng như: Máy học véc-tơ hỗ trợ (Support Vector Machines - SVM), Cây quyết định (Decision Tree), Naïve Bayes.[[6]](#endnote-6) Trong nghiên cứu này, nhóm chúng em sẽ thực hiện so sánh trên một số mô hình khai thác văn bản và tìm ra mô hình tốt nhất cho việc phân loại.

Giai đoạn 2, nhóm chúng em sẽ xây dựng một hệ thống phân loại và chuyển phát email tự động. Khi hệ thống RPA (Robotic Process Automation) nhận email gửi đến, RPA lấy nội dung của email chuyển cho chương trình Python để nhận dạng, email được phân loại là công việc liên quan đến phòng ban nào thì sẽ được RPA chuyển phát đến phòng ban tương ứng.

**B2.1.3 Ý nghĩa đề tài**

Việc xây dựng thành công “Hệ thống tự động hóa quy trình nhận dạng, phân loại và chuyển phát email trong nội bộ doanh nghiệp” sẽ có thể được áp dụng vào nhiều ứng dụng cụ thể trong đời sống, góp phần giảm thiểu sự tiêu tốn về thời gian và công sức con người. Đồng thời, việc xây dựng thành công công cụ mở ra điều kiện thuận lợi cho các bài toán xử lý ngôn ngữ tự nhiên khác.

**B2.2 Tính mới**

Nghiên cứu sử dụng và tích hợp được nhiều lý thuyết, mô hình từ các nghiên cứu đi trước để xây dựng một mô hình tổng quan, hoàn thiện hơn, chính xác hơn trong việc phân tích, nhận dạng và phân loại email.

Bên cạnh đó, nghiên cứu còn áp dụng công nghệ RPA kết hợp NLP tự động hóa một cách hoàn toàn quy trình nhận và phân loại email.

# B3. Kết quả nghiên cứu sơ khởi

**B3.1 Khai phá dữ liệu email - phân loại ngôn ngữ tự nhiên**

- Tổng quan ngắn gọn về các thuật toán phân loại văn bản (Kowsari, ; Jafari Meimandi, ; Heidarysafa, ; Mendu, ; Barnes, ; Brown, 2019).[[7]](#endnote-7) Tổng quan này bao gồm:

* Các phương pháp trích xuất đối tượng văn khác nhau: Làm sạch văn bản và tiền xử lý (Text Cleaning and Pre-processing), Biểu diễn từ cú pháp (Syntactic Word Representation), Nhúng từ (Word Embedding); ...
* Các phương pháp giảm kích thước: Phân tích thành phần (Component Analysis), Phân tích phân biệt tuyến tính (LDA), Hệ số hóa ma trận không âm (NMF), Nhúng Stochastic Neighbor phân tán T (t-SNE); ...
* Các thuật toán và kỹ thuật: Hồi quy Logistic (Logistic Regression), Phân loại Naïve Bayes, K- Hàng xóm gần nhất (K-Nearest Neighbor), Máy vectơ hỗ trợ (SVM), ...
* Các phương pháp đánh giá: Trung bình vĩ mô và Trung bình vi mô (Macro-Averaging and Micro-Averaging), Điểm F (F Score), Hệ số tương quan Matthews (MCC), Đặc điểm hoạt động của máy thu (ROC), Khu vực dưới đường cong ROC (AUC), ...

- Kỹ thuật phân loại email hiệu quả và hiệu quả dựa trên phương pháp lọc dữ liệu. Kỹ thuật lọc sáng tạo sử dụng phương pháp chọn phiên bản (ISM) để giảm các cá thể dữ liệu vô nghĩa từ mô hình đào tạo và sau đó phân loại dữ liệu thử nghiệm. Mục tiêu của ISM là xác định các trường hợp (ví dụ, mẫu) nào trong kho tài liệu email nên được chọn làm đại diện cho toàn bộ tập dữ liệu mà không làm mất thông tin đáng kể.[[8]](#endnote-8) (Islam, Rafiqul; Xiang, Yang, 2010)

- So sánh giữa hai cách tiếp cận khác nhau để phân loại email dựa trên danh mục của chúng. Naive Bayes và Hidden Markov Model (HMM). Naive Bayes Classifier dựa trên xác suất có điều kiện. Nó nhanh và hoạt động tốt với tập dữ liệu nhỏ. Nó coi các từ độc lập như một đặc điểm. HMM là một mô hình xác suất, tổng quát, cung cấp cho chúng ta sự phân phối qua các chuỗi quan sát. HMM có thể xử lý các đầu vào có độ dài thay đổi và giúp các chương trình đi đến quyết định khả dĩ nhất, dựa trên cả các quyết định trước đó và dữ liệu hiện tại.[[9]](#endnote-9) (Gomes, Sebastian Romy; Saroar, Sk Golam; Mosfaiul, Md; Telot, Alam; Khan, Behroz Newaz; Chakrabarty, Amitabha; Mostakim, Moin, 2017)

- Một biến thể đa nhãn của phân loại email. ML-EC2 là một thuật toán kết hợp dựa trên phân nhóm văn bản, phân loại văn bản, tính toán kỳ hạn thường xuyên (dựa trên phân bổ dirichlet tiềm ẩn) và kỹ thuật ánh xạ kỳ hạn phân loại. Nó là một ví dụ về phân loại sử dụng kỹ thuật phân cụm văn bản. Nó nghiên cứu vấn đề trong đó mỗi cụm email đại diện cho một nhãn lớp duy nhất trong khi nó được liên kết với tập hợp các nhãn cụm.[[10]](#endnote-10) (Sharaff, Aakanksha; Nagwani, Naresh Kumar, 2020)

**B3.2 Phần mềm UiPath**

- Tìm hiểu về Tự động hóa xử lý rô bốt (RPA) và cách xây dựng bot bằng UiPath, vận dụng những ví dụ thực tế để giải thích những kiến thức cơ bản về UiPath như xây dựng chatbox, các quy trình tự động, ... Tự động hóa quy trình bằng robot sử dụng UiPath giải thích chi tiết các tính năng và chức năng khác nhau của nền tảng RPA bao gồm phát triển, gỡ lỗi và xử lý lỗi.[[11]](#endnote-11) (Adeel Javed, Anum Sundrani, Nadia Malik, Sidney Madison Prescott, 2021)

- UiPath là nhà cung cấp tự động hóa quy trình Robotics hàng đầu vì nó cung cấp một nền tảng phần mềm hoàn chỉnh để giúp các công ty hoặc tổ chức tự động hóa bất kỳ quy trình nào một cách hiệu quả. Tính năng quét dữ liệu cho phép bạn trích xuất dữ liệu có cấu trúc từ bất kỳ trình duyệt hoặc ứng dụng nào và ghi lại dữ liệu đó vào trang web, tệp csv hoặc sang trang tính excel. Tự động hóa email là yêu cầu phổ biến nhất đối với nhiều công ty trên toàn cầu. Sử dụng Uipath, chúng em cũng có thể tự động hóa quá trình này.[[12]](#endnote-12) (V PremaLatha, BM Krishna, BN Kumar, TV Rajesh, 2020)

- Chúng em bắt đầu tìm hiểu về phần mềm UiPath Studio thông qua các khóa học và sách (Alok Mani Tripathi, 2018), từng bước nắm rõ công cụ UiPath để tự động hóa các quy trình lặp đi lặp lại hằng ngày. Thực hiện các dự án mẫu như: tự động thu thập tin tức mới nhất gửi về email, tự động lưu giá cổ phiếu trên web về file excel, tự động nhận dạng chữ viết trên file PDF (Nandan Mullakara, 2019), ... Đặc biệt tìm hiểu sâu về Email Automation để phục vụ cho đề tài.

**B4. Tài liệu tham khảo**

1. Huỳnh Phụng Toàn, Nguyễn Vũ Lâm, Nguyễn Minh Trung và Đỗ Thanh Nghị (2011), *Phân loại thư rác với giải thuật boosting cây quyết định ngẫu nhiên xiên phân đơn giản.* Trường Đại học Cần Thơ: Tạp chí Khoa học 2011 số 19b, pp. 1-9*;*
2. Shams, Rushdi; Mercer, Robert E. (2013), *Classifying Spam Emails Using Text and Readability Features*, 2013 IEEE 13th International Conference on Data Mining, pp. 657 – 666;
3. Sebastiani, F. (2002), *Machine learning in automated text categorization*, ACM Comput. Surv.34, pp. 1-47;
4. Lê Thanh Trà (2013), *Nghiên cứu các phương pháp phân loại văn bản và ứng dụng vào phân loại thư điện tử*, Luận văn Tiến sĩ, Học viện Công nghệ bưu chính viễn thông, truy cập từ <http://dlib.ptit.edu.vn/handle/HVCNBCVT/1380>;
5. Hoàng, N. H., Đông, N. V., Phong, N. V. & Liên, D. T. P. (2020), *Giáo trình: Phương pháp nghiên cứu khoa học*, Trường Đại học Tài chính – Marketing, Nhà xuất bản Thành phố Hồ Chí Minh;
6. Đệ, T. C., & Khang, P. N. (2012), *Phân loại văn bản với máy học vector hỗ trợ và cây quyết định*. Tạp chí Khoa học Trường Đại học Cần Thơ, (21a), 52-63. Truy vấn từ <https://ctujsvn.ctu.edu.vn/index.php/ctujsvn/article/view/1155>;
7. Hữu Đệ - ATP, *Các phương pháp nghiên cứu khoa học*, <https://winerp.vn/tag/so-sanh-cac-phuong-phap-nghien-cuu-khoa-hoc>;
8. Max Bramer (2020), *Principles of Data Mining*, University of Portsmouth, UK;
9. Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, Brown, (2019). *Text Classification Algorithms: A Survey*. Information, 10(4), 150–. doi:10.3390/info10040150;
10. Islam, Rafiqul; Xiang, Yang (2010). *Email classification using data reduction method*. [IEEE 5th International ICST Conference on Communications and Networking in China - Beijing, China] Proceedings of the 5th International ICST Conference on Communications and Networking in China - Proceedings of the 5th International ICST Conference on Communications and Networking in China, doi:10.4108/chinacom.2010.59;
11. Gomes, Sebastian Romy; Saroar, Sk Golam; Mosfaiul, Md; Telot, Alam; Khan, Behroz Newaz; Chakrabarty, Amitabha; Mostakim, Moin (2017). *A comparative approach to email classification using Naive Bayes classifier and hidden Markov model*, [IEEE 2017 4th International Conference on Advances in Electrical Engineering (ICAEE) - Dhaka, Bangladesh (2017.9.28-2017.9.30)] 2017 4th International Conference on Advances in Electrical Engineering (ICAEE) -, 482–487, doi:10.1109/ICAEE.2017.8255404;
12. Sharaff, Aakanksha; Nagwani, Naresh Kumar (2020). ML-EC2. *International Journal of Web-Based Learning and Teaching Technologies*, 15(2), 19–33. doi:10.4018/IJWLTT.2020040102;
13. Huang, Jen-Wei, Chiang, Chia-Wen, Chang, Jia-Wei (2018), *Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise,* Engineering Applications of Artificial Intelligence, 75(): pp. 11-21;
14. Adeel Javed, Anum Sundrani, Nadia Malik, Sidney Madison Prescott (2021). *Robotic Process Automation using UiPath StudioX* – doi:10.1007/978-1-4842-6794-3;
15. Nandan Mullakara and Arin Kumar Asokan (2019), *Robotic Process Automation Projects: Build real-world RPA solutions using UiPath and Automation Anywhere*, nhà xuất bản Packt;
16. Alok Mani Tripathi (2018), *Learning Robotic Process Automation*, nhà xuất bản Packt;
17. V PremaLatha, BM Krishna, BN Kumar, TV Rajesh, (2020). *Automation Using Robotic Process Automation (RPA)*, truy vấn từ <http://xajzkjdx.cn/gallery/472-april2020.pdf>.

# B5. Kế hoạch và phương pháp nghiên cứu

**B5.1 Mục tiêu**

*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), đồng thời, thúc đẩy việc nghiên cứu một nền tảng công nghệ mới về tự động hóa thông minh (Intelligent Automation).

*Về công nghệ*

Thiết kế một hệ thống ứng dụng công nghệ Machine Learning để phân loại chủ đề (Natural Language Processing), gắn nhãn Email và sử dụng 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.

**B5.2 Nội dung**

**B5.2.1 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 thực nghiệm: Tiến hành thực nghiệm các thuật toán, mô hình và thực nghiệm hệ thống để đánh giá mức độ hiệu quả, từ đó đưa ra lựa chọn phù hợp và hoàn thiện 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.

**B5.2.2 Nội dung nghiên cứu**

*Nội dung 1: Nghiên cứu, tìm kiếm và phân tích dữ liệu email về các phòng ban trong doanh nghiệp*

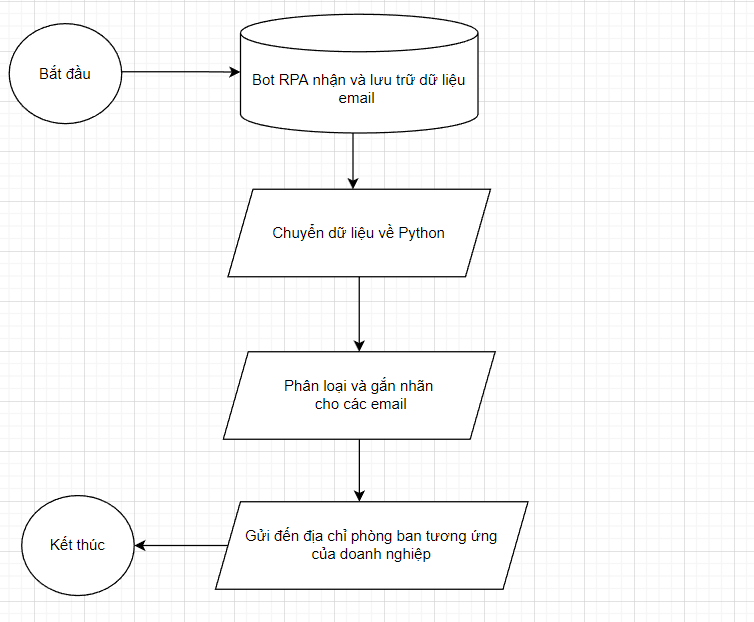
* Dựa trên nội dung các email được gửi về doanh nghiệp, nhóm tiến hành nghiên cứu đặc tính và điểm chung của các email được gửi về trên từng phòng ban nội bộ.
* Phân tích và lựa chọn nhãn phù phòng ban phù hợp với nội dung email.
* Dựa trên dữ liệu về email và nhãn tương ứng, nhóm tiến hành tổng hợp thành một tập dữ liệu chung để sử dụng cho quá trình nghiên cứu.

*Nội dung 2: Nghiên cứu mô hình Text Mining phù hợp*

* Tìm hiểu về các mô hình Text Mining như Logistic Regression, Naive Bayes, Decision Tree, K-Nearest Neighbor, ...
* Từ tập dữ liệu đã được tạo ở nội dung 1, nhóm thực hiện thử nghiệm các mô hình Text Mining để so sánh dựa vào khả năng thu hồi (Recall), độ chính xác (Precision), F1-score, Keppa, MAE, RMSE, … Từ đó tìm ra mô hình tốt nhất.
* Áp dụng mô hình tốt nhất đã thử nghiệm, nhận xét kết quả, thảo luận và đưa ra hướng phát triển tối ưu.

*Nội dung 3: Xây dựng hệ thống[[13]](#endnote-13) RPA*

* Hệ thống RPA bao gồm các quy trình:
* Hệ thống nhận và đọc dữ liệu từ email.
  + Chuyển dữ liệu email vào Python, sử dụng mô hình Text Mining để nhận dạng và phân loại.
* Hệ thống dựa vào kết quả phân loại và tự động gửi đến các phòng ban.



*Hình 7: Lưu đồ thể hiện quy trình hệ thống RPA*

*Nội dung 4: Đánh giá mức độ hiệu quả của hệ thống*

* Triển khai thử nghiệm hệ thống.
* Nghiên cứu, xây dựng khung chuẩn để đánh giá mức độ hiệu quả của hệ thống và tìm ra biện pháp tối ưu hệ thống.

**B5.3 Kế hoạch nghiên cứu**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Công việc** | **Người thực hiện** | **Kết quả mong muốn** | **Thời gian** | **Ghi chú** |
| Tìm hiểu các đặc tính từng loại email cho các phòng ban và chọn các loại nhãn phù hợp | Bảo Tâm  Giả Ly | Hiểu rõ hơn về đặc tính của loại email gửi về phòng ban đó | Tháng 9 - Tháng 10 |  |
| Nghiên cứu các mô hình Text Mining, chọn ra mô hình phù hợp nhất | Hoàng Lan  Thế Anh  Quang Học | Chọn ra mô hình Text  Mining phù hợp | Tháng 9 - Tháng 11 |  |
| Xây dựng hệ thống RPA tự động đọc và gửi email | Quang Học  Bảo Tâm  Giả Ly | Hệ thống hoàn chỉnh và không gặp lỗi khi chạy | Tháng 10 - Tháng 11 |  |
| Xây hệ thống RPA kết hợp với mô hình text mining đã được chọn | Bảo Tâm  Thế Anh  Hoàng Lan | Hệ thống tự động đọc email, gắn nhãn và gửi về phòng ban tương ứng | Tháng 11- Tháng 12 |  |
| Đánh giá kết quả hệ thống khi chạy và tìm cách cải tiến | Cả nhóm | Nâng cao khả năng nhận dạng của hệ thống | Tháng 1- Tháng 2 |  |

# B6. Sản phẩm của đề tài

* 01 báo cáo nghiên cứu khoa học tổng kết hoàn chỉnh;
* 01 chương trình demo;
* 01 bài báo đăng trên hội thảo hoặc tạp chí quốc tế.

1. Lê Thanh Trà (2013), *Nghiên cứu các phương pháp phân loại văn bản và ứng dụng vào phân loại thư điện tử, Luận văn Tiến sĩ*, Học viện Công nghệ bưu chính viễn thông, truy cập từ <http://dlib.ptit.edu.vn/handle/HVCNBCVT/1380>; [↑](#endnote-ref-1)
2. Huỳnh Phụng Toàn, Nguyễn Vũ Lâm, Nguyễn Minh Trung và Đỗ Thanh Nghị (2011), *Phân loại thư rác với giải thuật boosting cây quyết định ngẫu nhiên xiên phân đơn giản*. Trường Đại học Cần Thơ: Tạp chí Khoa học 2011 số 19b, pp. 1-9; [↑](#endnote-ref-2)
3. Huang, Jen-Wei, Chiang, Chia-Wen, Chang, Jia-Wei (2018), *Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise*, Engineering Applications of Artificial Intelligence, 75(): pp. 11-21; [↑](#endnote-ref-3)
4. Shams, Rushdi; Mercer, Robert E. (2013), *Classifying Spam Emails Using Text and Readability Features*, 2013 IEEE 13th International Conference on Data Mining, pp. 657 – 666; [↑](#endnote-ref-4)
5. Sebastiani, F. (2002), *Machine learning in automated text categorization*, ACM Comput. Surv.34, pp. 1-47; [↑](#endnote-ref-5)
6. Đệ, T. C., & Khang, P. N. (2012), *Phân loại văn bản với máy học vector hỗ trợ và cây quyết định*. Tạp chí Khoa học Trường Đại học Cần Thơ, (21a), 52-63. Truy vấn từ <https://ctujsvn.ctu.edu.vn/index.php/ctujsvn/article/view/1155>; [↑](#endnote-ref-6)
7. Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, Brown, (2019). *Text Classification Algorithms: A Survey*. Information, 10(4), 150–. doi:10.3390/info10040150; [↑](#endnote-ref-7)
8. Islam, Rafiqul; Xiang, Yang (2010). *Email classification using data reduction method*. [IEEE 5th International ICST Conference on Communications and Networking in China - Beijing, China] Proceedings of the 5th International ICST Conference on Communications and Networking in China - Proceedings of the 5th International ICST Conference on Communications and Networking in China, doi:10.4108/chinacom.2010.59; [↑](#endnote-ref-8)
9. Gomes, Sebastian Romy; Saroar, Sk Golam; Mosfaiul, Md; Telot, Alam; Khan, Behroz Newaz; Chakrabarty, Amitabha; Mostakim, Moin (2017). *A comparative approach to email classification using Naive Bayes classifier and hidden Markov model*, [IEEE 2017 4th International Conference on Advances in Electrical Engineering (ICAEE) - Dhaka, Bangladesh (2017.9.28-2017.9.30)] 2017 4th International Conference on Advances in Electrical Engineering (ICAEE) -, 482–487, doi:10.1109/ICAEE.2017.8255404; [↑](#endnote-ref-9)
10. Sharaff, Aakanksha; Nagwani, Naresh Kumar (2020). ML-EC2. *International Journal of Web-Based Learning and Teaching Technologies*, 15(2), 19–33. doi:10.4018/IJWLTT.2020040102; [↑](#endnote-ref-10)
11. Adeel Javed, Anum Sundrani, Nadia Malik, Sidney Madison Prescott (2021). *Robotic Process Automation using UiPath StudioX* – doi:10.1007/978-1-4842-6794-3; [↑](#endnote-ref-11)
12. V PremaLatha, BM Krishna, BN Kumar, TV Rajesh, (2020). *Automation Using Robotic Process Automation (RPA)*, truy vấn từ <http://xajzkjdx.cn/gallery/472-april2020.pdf>; [↑](#endnote-ref-12)
13. |  |  |
    | --- | --- |
    | *Ngày 11 tháng 09 năm 2022*  **Nhóm trưởng nhóm nghiên cứu**  *(Họ tên, chữ ký)*    **Nguyễn Quang Học** | *Ngày 11 tháng 09 năm 2022*  **Người hướng dẫn**  *(Họ tên, chữ ký)* |

    *Ngày 11 tháng 09 năm 2022*

    **Lãnh đạo Đơn vị**

    *(Họ tên, chữ ký)* [↑](#endnote-ref-13)