**TRƯỜNG ĐẠI HỌC TRÀ VINH**

**TRƯỜNG KỸ THUẬT VÀ CÔNG NGHỆ**



**BÀI TẬP BÁO CÁO KẾT THÚC MÔN**

**NHÓM 4:**

**- Thị Sóc Kha**

**- Trần Thị Xuân Mai**

**- Nguyễn Thành Lý**

**Trà Vinh, ngày 26 tháng 10 năm 2025**

1. **Lý do chọn đề tài:**

- Ngày nay với sự phát triển của công nghệ hiện đại, dữ liệu trở thành nguồn tài nguyên vô cùng quan trọng, đóng vai trò cho xã hội hiện nay, song song đó cũng nhiều dữ liệu mang đầy virus và lừa đảo, trong đó tin nhắn của thư rác (spam mail) là một ví dụ điển hình nó xuất hiện ngày càng nhiều, gây nhiều phiền toái. Vì vậy, ta cần có một thuật toán Support Vector Machine ( SVM) dữ liệu để phân loại email thành “spam” và “không spam” là rất cần thiết.

1. **Mục tiêu nghiên cứu:**

- Tìm hiểu về Support Vector Machine (SVM) trong phân loại văn bản và xây dựng công cụ lọc emil spam.

1. **Đối tượng và phạm vi nghiên cứu:**

- Đối tượng nghiên cứu: Support Vector Machine (SVM) trong phân loại văn bản và xây dựng công cụ lọc email spam.

- Phạm vi nghiên cứu:

+ Không gian nghiên cứu: Trường Đại Học Trà Vinh

+ Thời gian: ngày 24/10/2025 .Vì thời gian này sự xuất hiện ngày càng nhiều của các loại tin nhắn rác và các tin nhắn chứa các nội dung lừa đảo đến các bạn học sinh-sinh viên.

1. **Nội dung nghiên cứu:**

- Đề tài tập trung nghiên cứu tìm hiểu về Support Vector Machine (SVM) và ứng dụng của SVM trong phân loại văn bản, xây dựng công cụ lọc emil spam.

1. **Phương pháp nghiên cứu:**

- Phương pháp nghiên cứu Support Vector Machine (SVM) trong phân loại văn bản và ứng dụng lọc email spam.

- Tổng hợp và nghiên cứu tài liệu từ sách, internet và một số trang học lập trình như Scikit-learn, Kaggle và GeeksforGeeks. Thu thập, chọn lọc các tài liệu về học máy (Machine Learning), thuật toán SVM, siêu phẳng phân tách (hyperplane) và cách SVM được dùng để phân loại dữ liệu văn bản, tìm hiểu về quy trình xử lý ngôn ngữ tự nhiên gồm: loại bỏ kí tự đặc biệt, chuẩn hóa chữ, tách từ, bỏ stopwords và nguyên lý của TF-IDF là phương pháp biến đổi văn bản thành dạng số để máy tính hiểu được nội dung.

* Thu thập dữ liệu: gồm các email thật (spam và không spam). Các email spam thường có đặc điểm như chứa từ khóa quảng cáo, limk, hay yêu cầu nhấn vào liên kết.
* Tiền xử lý: Trước khi phân loại, hệ thống sẽ loại bỏ từ dư, làm sạch văn bản, rồi chuyển dữ liệu thành dạng TF-IDF vector
* Phân loại bằng SVM: Hệ thống SVM sẽ phân tích đặc trưng của từng email, tìm đường phân tách tối ưu giữa email spam và không spam.
* So sánh dự đoán với dữ liệu thực tế để đo độ chính xác, trong nhiều nghiên cứu cho thấy SVM có độ chính xác khoảng 90-95%, cao hơn một số mô hình khác như Naive Bayes.
* Nguồn tham khảo: Các trang web học lập trình (<https://aicandy.vn/support-vector-machine-svm-xu-ly-du-lieu-phi-tuyen/>, <https://codelearn.io/sharing/scikit-learn-trong-python-la-gi?srsltid=AfmBOoqcTSoTVXd7wBTP1dsoPZ6PnY-Lm2NZy-r9FylazF6X63KiS-IB> ), video hướng dẫn (<https://youtu.be/SiSpgDoq02g?si=k-4O9alPUQQr6E-V>).
* Công cụ trình bày: Microsoft Word.
* Hình ảnh mô phỏng: Sử dụng sơ đồ như hình Support vector Machine để giải thích trực quan.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Nội dung do AI tạo ra có thể không chính xác.

CHƯƠNG 1: Cơ sở lý thuyết Support Vector Machine (SVM) trong phân loại văn bản.

* 1. Giới thiệu SVM trong phân loại văn bản:

- Phân loại văn bản là quá trình sử dụng thuật toán học máy hoặc xử lý ngôn ngữ tự nhiên (NLP) để tự động gán nhãn cho các văn bản dựa trên nội dung của chúng.

Ví dụ:

+ Email được chia thành hai loại: Spam (thư rác) và Ham (bình thường).

+ Bài viết có thể được phân loại theo chủ đề như “Thể thao”, “Giáo dục”, “Giải trí”, “Kinh tế”…

+ Bình luận trên mạng xã hội có thể được nhận diện là tích cực, tiêu cực hoặc trung lập.

* 1. **Mục tiêu của phân loại văn bản:**

- Mục tiêu chính của phân loại văn bản là giúp máy tính hiểu nội dung và ý nghĩa của văn bản, từ đó:

+ Tự động sắp xếp, lọc, hoặc tìm kiếm thông tin nhanh chóng.

+ Tiết kiệm thời gian xử lý dữ liệu so với việc làm thủ công.

+ Hỗ trợ ra quyết định trong các hệ thống thông minh (như phát hiện thư rác, lọc tin tức, đánh giá ý kiến người dùng.

* 1. **Cách thức hoạt động:**
* Quy trình phân loại văn bản gồm các bước cơ bản:

**-** Thu thập dữ liệu: Tập hợp văn bản cần phân loại (ví dụ: email, bài viết, bình luận).

- Tiền xử lý: Loại bỏ ký tự đặc biệt, chuyển chữ thường, tách từ, xóa từ dừng (stop words).

- Biểu diễn văn bản: Chuyển văn bản sang dạng số bằng các mô hình như

+ Bag of Words (BoW)

+ TF-IDF (Term Frequency – Inverse Document Frequency)

- Huấn luyện mô hình: Sử dụng các thuật toán học máy như

+ Support Vector Machine (SVM)

+ Naive Bayes

+ Decision Tree

+ Deep Learning (CNN, RNN, BERT)

- Phân loại & đánh giá: Dùng mô hình đã huấn luyện để dự đoán nhãn cho văn bản mới và đánh giá độ chính xác.

* 1. **Phân loại tuyến tính và phi tuyến:**

SVM tuyến tính (Linear SVM): dùng khi dữ liệu có thể phân tách bằng một đường thẳng hoặc siêu phẳng.

- SVM phi tuyến (Nonlinear SVM): dùng hàm nhân (Kernel) để ánh xạ dữ liệu sang không gian có nhiều chiều hơn, giúp phân tách được dữ liệu phức tạp.

* **Ví dụ:**
  1. **Ứng dụng SVM trong phân loại văn bản:**
* Biểu diễn dữ liệu văn bản:

- Trước khi đưa vào SVM, văn bản phải được biểu diễn thành dạng số.

Phổ biến nhất là:

+ Bag of Words (BoW): Đếm tần suất từ xuất hiện.

+ TF-IDF (Term Frequency – Inverse Document Frequency): Giảm trọng số của các từ phổ biến và tăng trọng số cho các từ hiếm nhưng có giá trị phân biệt cao.

CHƯƠNG 2