



Deep Learning

# SmartMail Classifier

Ứng dụng DistilBERT trong tự động  
phân loại và ưu tiên email hỗ trợ  
khách hàng

Nhóm thuyết trình

**Nhóm 07**





# Thành viên nhóm 7

1 Lê Hoàng Giang - K224161809

2 Phạm Nguyễn Gia Huy - K224161817

3 La Nam Khánh - K224161820

4 Trương Thảo Vy - K224161845





# Mục lục

- 1. Giới thiệu
- 2. Bộ dữ liệu
- 3. Database
- 4. Front-End
- 5. Back-end
- 6. Model và Kết quả

# Giới thiệu

## Bối cảnh:

- Doanh nghiệp, đặc biệt trong Customer Support / IT Helpdesk, nhận hàng nghìn email mỗi ngày với nội dung đa dạng.
- Email có thể bao gồm: báo lỗi hệ thống, yêu cầu thay đổi, đề nghị hỗ trợ, hoặc phản nàn.
- Phân loại thủ công email theo loại yêu cầu và mức độ ưu tiên gây tốn thời gian, dễ sai sót, ảnh hưởng tốc độ phản hồi.

**Giải pháp đề xuất:** Ứng dụng Deep Learning để tự động đọc hiểu và phân loại email.

- Tự động phân loại loại yêu cầu: Incident, Problem, Change, Request.
- Tự động xác định mức độ ưu tiên: Low, Medium, High.
- Giúp hệ thống hỗ trợ sắp xếp yêu cầu chính xác, tăng tốc độ xử lý và cải thiện trải nghiệm khách hàng.



# Bộ dữ liệu

Data được lấy từ Kaggle, tên bộ dữ liệu Multilingual Customer Support Tickets (Synthetic) bởi Tobias Bück. Cung cấp các ticket hỗ trợ khách hàng (email yêu cầu, phản hồi...) có nhãn sẵn như loại yêu cầu (type), mức độ ưu tiên (priority), bộ phận (queue), ngôn ngữ (language) để phục vụ cho các bài toán NLP: phân loại, định tuyến ticket, ưu tiên xử lý.

subject	body	answer	type	queue	priority	language	version	tags								
account disruption	dear customer support team! Thank you for reaching out,	· Incident	Technical Support	high	en	51	['Account', 'Disruption', 'Outage', 'IT', 'Tech Support']									
query smart home system int	dear customer support team! Thank you for your inquiry.	· Request	Returns and Exchanges	medium	en	51	['Product', 'Feature', 'Tech Support']									
inquiry regarding invoice det	dear customer support team! We appreciate you reaching	Request	Billing and Payments	low	en	51	['Billing', 'Payment', 'Account', 'Documentation', 'Feedback']									
question marketing agency s	dear support teamnni hope n	Thank you for your inquiry.	Problem	Sales and Pre-Sales	medium	en	51	['Product', 'Feature', 'Feedback', 'Tech Support']								
feature query	dear customer supportnni hc	Thank you for your inquiry.	Request	Technical Support	high	en	51	['Feature', 'Product', 'Documentation', 'Feedback']								
system interruptions	dear customer support team! Thank you for bringing the sy	Incident	Service Outages and Mair	high	en	51	['Outage', 'Disruption', 'Performance', 'IT', 'Tech Support']									
connectivity problems printer	dear support teamnni reporti	Thank you for reaching out n	Incident	Technical Support	medium	en	51	['Network', 'Hardware', 'Performance', 'Bug', 'Compatibility']								
vpn access issue	customer supportnnwe enco	Thank you for reporting this p	Incident	Product Support	medium	en	51	['Network', 'Disruption', 'VPN', 'Tech Support']								
immediate help needed techn	dear customer support teami	Thanks for providing detailed	Problem	IT Support	medium	en	51	['Bug', 'Crash', 'Network', 'Performance', 'Disruption', 'Outage', 'Tech Support']								

# Database

The screenshot shows the MongoDB Compass interface. The left sidebar is titled 'Cluster' and includes links for Overview, Data Explorer (which is selected and highlighted in green), Real Time, Cluster Metrics, Query Insights, Performance Advisor, Online Archive, Command Line Tools, Infrastructure as Code, and SHORTCUTS (Search & Vector Search). The top navigation bar shows 'ORGANIZATION Vy's Org - 2025-10-14', 'PROJECT Project 0', and 'CLUSTER Cluster0'. The main area is titled 'email\_app.predictions' and displays storage details: STORAGE SIZE: 52KB, LOGICAL DATA SIZE: 41.28KB, TOTAL DOCUMENTS: 53, INDEXES TOTAL SIZE: 36KB. It features tabs for Find, Indexes, Schema Anti-Patterns (0), Aggregation, and Search Indexes. A search bar at the top says 'Search Namespaces'. Below the collection title, it says 'Generate queries from natural language in Compass' and has an 'INSERT DOCUMENT' button. A 'Filter' section allows querying with 'Type a query: { field: 'value' }' and includes 'Reset', 'Apply', and 'Options' buttons. Two document snippets are shown:

```

predicted_type : "request"
predicted_priority : "low"
timestamp : 2025-11-13T13:01:25.057+00:00

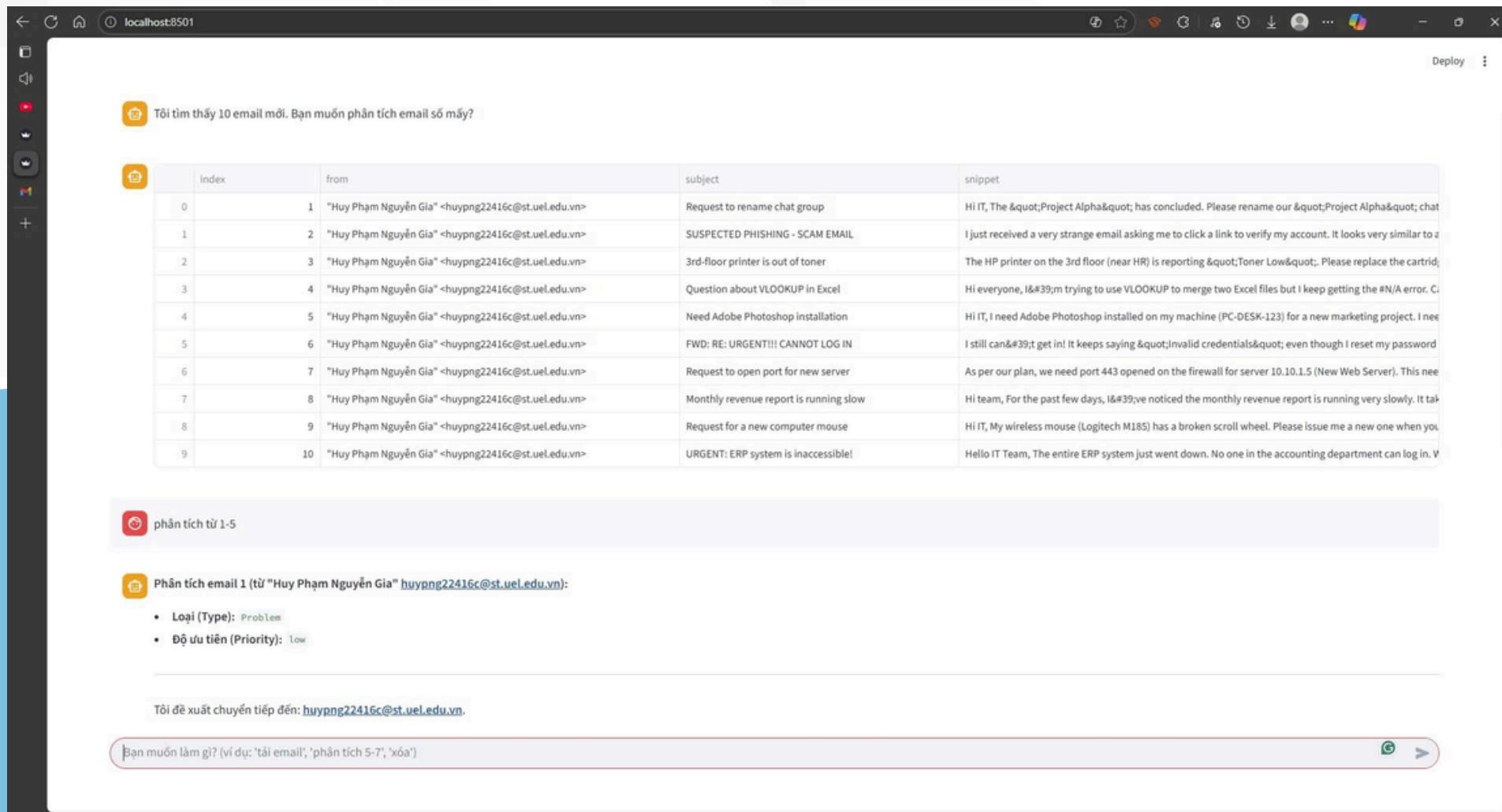
_id: ObjectId('691574e42c3a6c956238ef4f')
cleaned_text : "Thông báo tài khoản Bông Sen Vàng tháng 11.2025 Vietnam Airlines TH Oc..."
predicted_type : "Request"
predicted_priority : "high"
timestamp : 2025-11-13T13:04:20.151+00:00

```

At the bottom, there are 'PREVIOUS' and 'NEXT' buttons, and the text '21-40 of many results'.

Hệ thống backend được xây dựng nhằm quản lý và lưu trữ kết quả dự đoán từ mô hình Deep Learning phân loại email hỗ trợ khách hàng. Sau khi mô hình dự đoán loại yêu cầu (predicted\_type) và mức độ ưu tiên (predicted\_priority) cho mỗi email, kết quả sẽ được gửi đến MongoDB Atlas thông qua API kết nối trực tiếp từ backend.

# Front-end - Streamlit

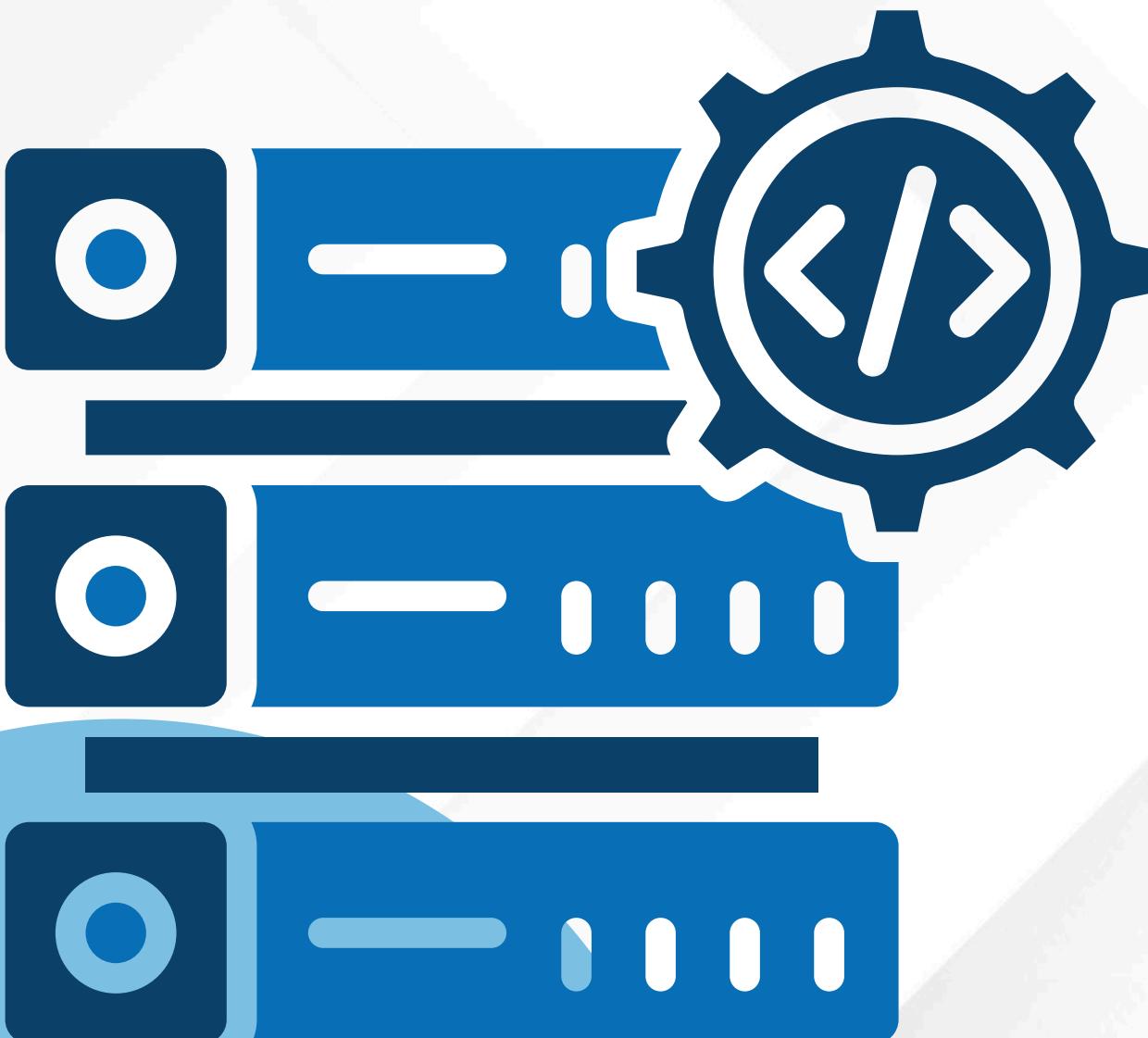


Vai trò: Đây là giao diện chatbox mà người dùng tương tác trực tiếp.

Xử lý chính: Nó xử lý toàn bộ logic phía người dùng, bao gồm:

- Xác thực: Đăng nhập an toàn qua Google OAuth.
- Tương tác Gmail: Tải email mới và thực hiện chuyển tiếp mail.

Luồng hoạt động: Khi người dùng yêu cầu phân tích, Frontend sẽ gọi API của Backend, gửi văn bản email đi. Sau khi nhận kết quả, nó sẽ hiển thị dự đoán cho người dùng và chờ lệnh.



## Back-end - FastAPI

- Vai trò: Đây là "bộ não" của toàn bộ ứng dụng.
- Xử lý chính: Chịu trách nhiệm tải và vận hành mô hình AI (BERT).
- Logic: Nó cung cấp một API endpoint (/predict). Khi nhận được một email, nó sẽ làm sạch văn bản, đưa qua mô hình AI, và trả về dự đoán cho Type và Priority.

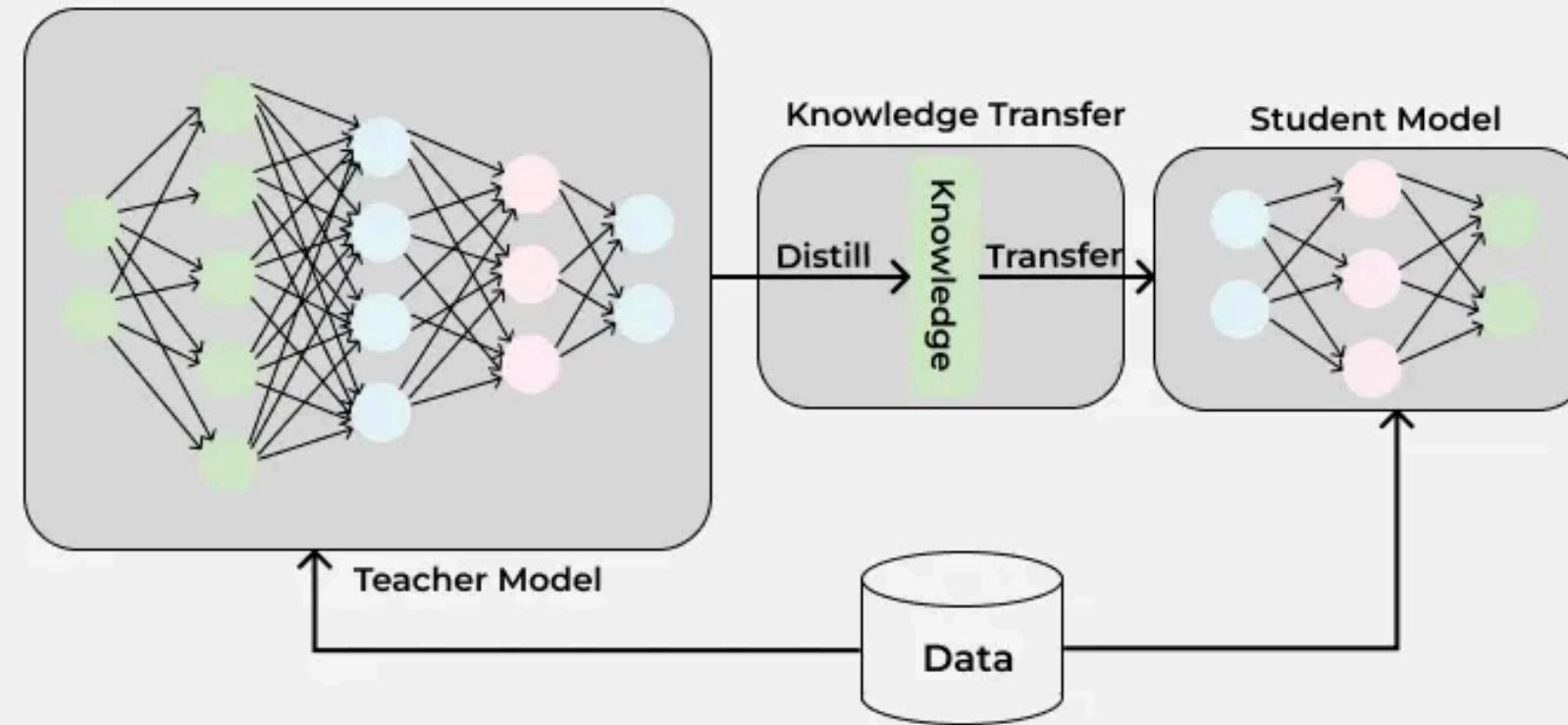
Database: Mọi dự đoán đều được tự động lưu trữ vào MongoDB hỗ trợ cho việc theo dõi và phân tích.

# Model

## Mục tiêu & Dữ liệu

- Mô hình tự động phân loại email hỗ trợ khách hàng theo:
- Type: Change, Incident, Problem, Request
- Priority: Low, Medium, High
- Dữ liệu gồm subject + body, được ghép và làm sạch trước khi huấn luyện.
- Nhãn được mã hóa số, dữ liệu chia theo tỷ lệ 70% train – 15% val – 15% test.

**Teacher-Student model for Knowledge Distillation**



## Phương pháp

- Sử dụng Transformer – DistilBERT nhờ khả năng hiểu ngữ cảnh tốt và phù hợp tiếng Anh.
- Văn bản được tokenize, padding/truncation về độ dài cố định.
- Fine-tune bằng DistilBERTForSequenceClassification, optimizer AdamW, cross-entropy loss.
- Đánh giá bằng precision, recall, F1-score.

# Kết quả

## Type

- Accuracy: 0.80 ; F1-score trung bình: 0.7978.
- Lớp Change và Request đạt hiệu suất cao (F1 > 0.95).
- Lớp Problem thấp hơn do dữ liệu ít và nội dung khó.

## Priority

- Accuracy: 0.61 ; F1-score trung bình: 0.6170.
- Khó phân biệt vì mức Low – Medium – High có sự chồng lấn ngữ nghĩa.

## Kết luận chung

- Mô hình đạt F1-score trung bình 0.7074.
- Hiệu quả tốt trong phân loại Type, và mức khá trong phân loại Priority.
- Khả năng ứng dụng thực tế cao cho hệ thống email Customer Support.

--- Validation Report ---

★ Task 1: TYPE classification:

	precision	recall	f1-score	support
Change	0.94	0.98	0.96	245
Incident	0.77	0.71	0.74	967
Problem	0.53	0.59	0.56	511
Request	0.99	0.99	0.99	715
accuracy			0.80	2438
macro avg	0.81	0.82	0.81	2438
weighted avg	0.80	0.80	0.80	2438

★ Task 2: PRIORITY classification:

	precision	recall	f1-score	support
low	0.80	0.64	0.71	803
medium	0.53	0.54	0.53	840
high	0.57	0.67	0.62	795
accuracy			0.61	2438
macro avg	0.63	0.62	0.62	2438
weighted avg	0.63	0.61	0.62	2438

Final Test Metrics (from best model):

Avg F1: 0.7074

Type F1: 0.7978

Pri F1: 0.6170

# DEMO



Cảm ơn Thầy và  
các bạn đã lắng  
nghe!

 Deep Learning

 Nhóm 07

 SmartMail Classifier