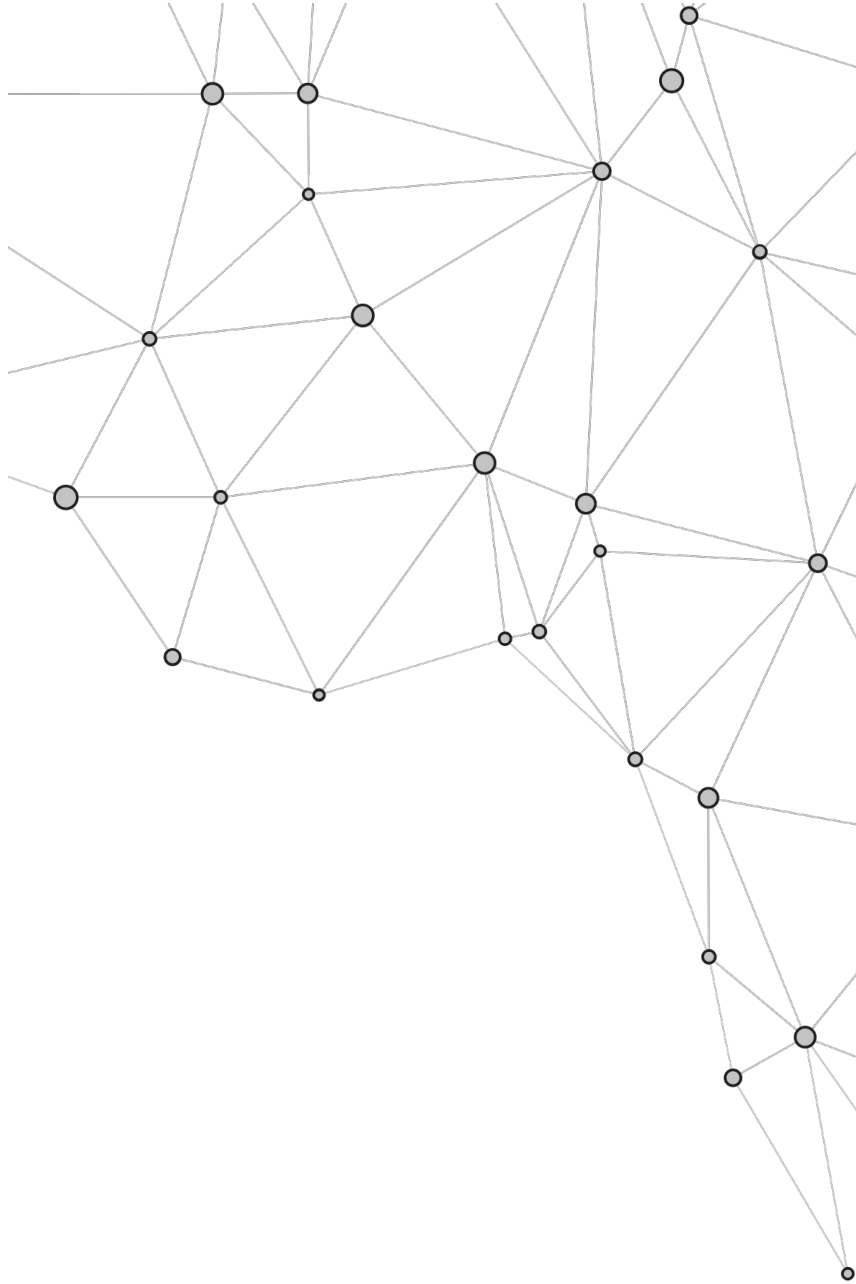
Topic Modelling
Analyze Facebook Post Topic



Danh sách thành viên

Nhóm 11

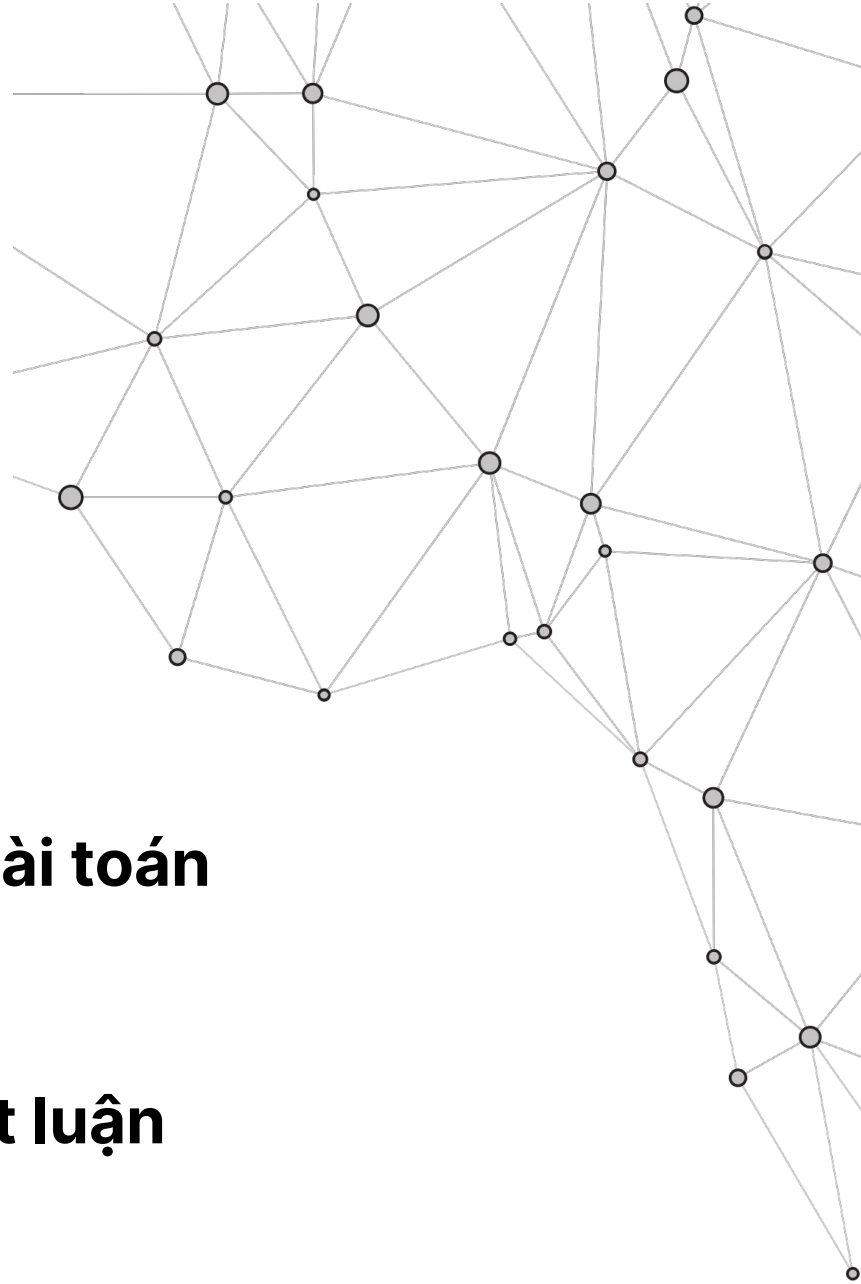
Họ và Tên: Lại Tuấn Anh
MSSV: 18020117
Email: 18020117@vnu.edu.vn

Họ và Tên: Triệu Vũ Hải
MSSV: 18020442
Email: 18020442@vnu.edu.vn

Họ và Tên: Nguyễn Thế Quân
MSSV: 18021030
Email: 18021030@vnu.edu.vn

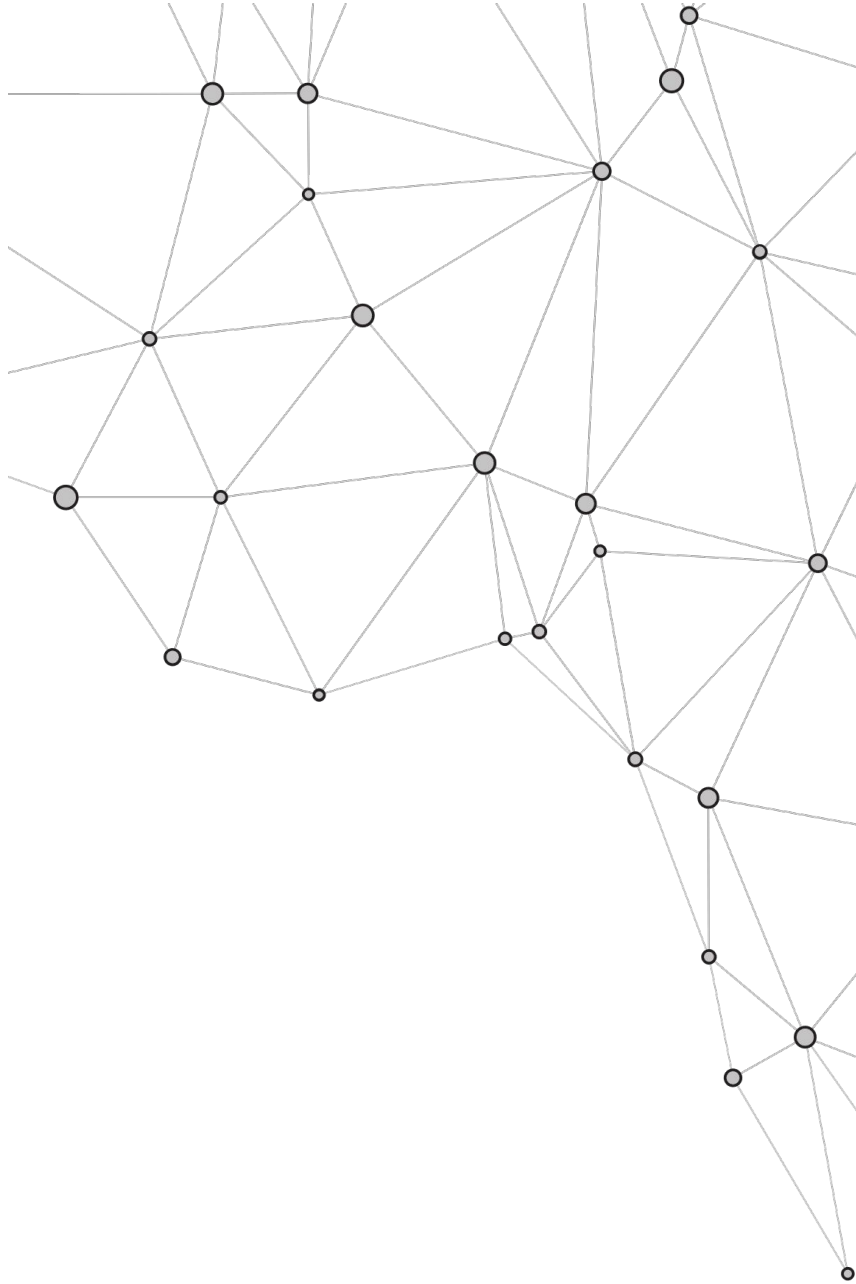
Họ và Tên: Nguyễn Hữu Huy
MSSV: 18020644
Email: 18020644@vnu.edu.vn

Mục lục

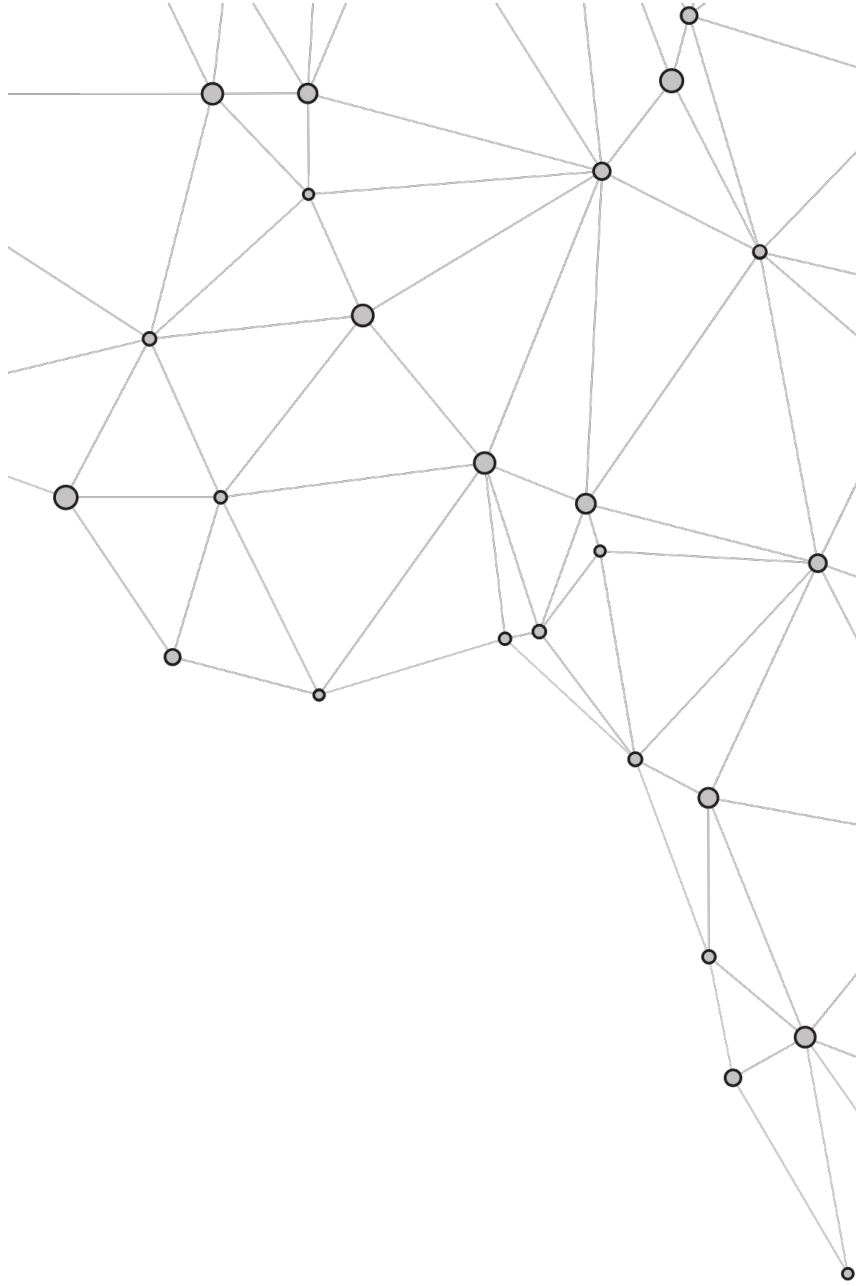


Giới thiệu bài toán

Thực hiện đánh nhãn cho tập dữ liệu văn bản bao gồm 16000 bản ghi và 23 nhãn khác nhau.



Phân tích dữ liệu



Các bước phân tích dữ liệu

- Làm sạch dữ liệu
- Thu thập thông tin dữ liệu
- Visualize dữ liệu

Thu thập thông tin

Kết quả

	label	count	mean_text_len	max_text_len
15	__label__Nha_dat	2542	369.423682	3354
9	__label__Kinh_doanh_va_Cong_nghiep	2355	981.234395	19811
3	__label__Do_an_va_do_uong	2355	612.346072	4475
20	__label__Tai_chinh	1379	1065.898477	19811
13	__label__Mua_sam	1169	507.592814	18166

...

Làm sạch dữ liệu

Vấn đề

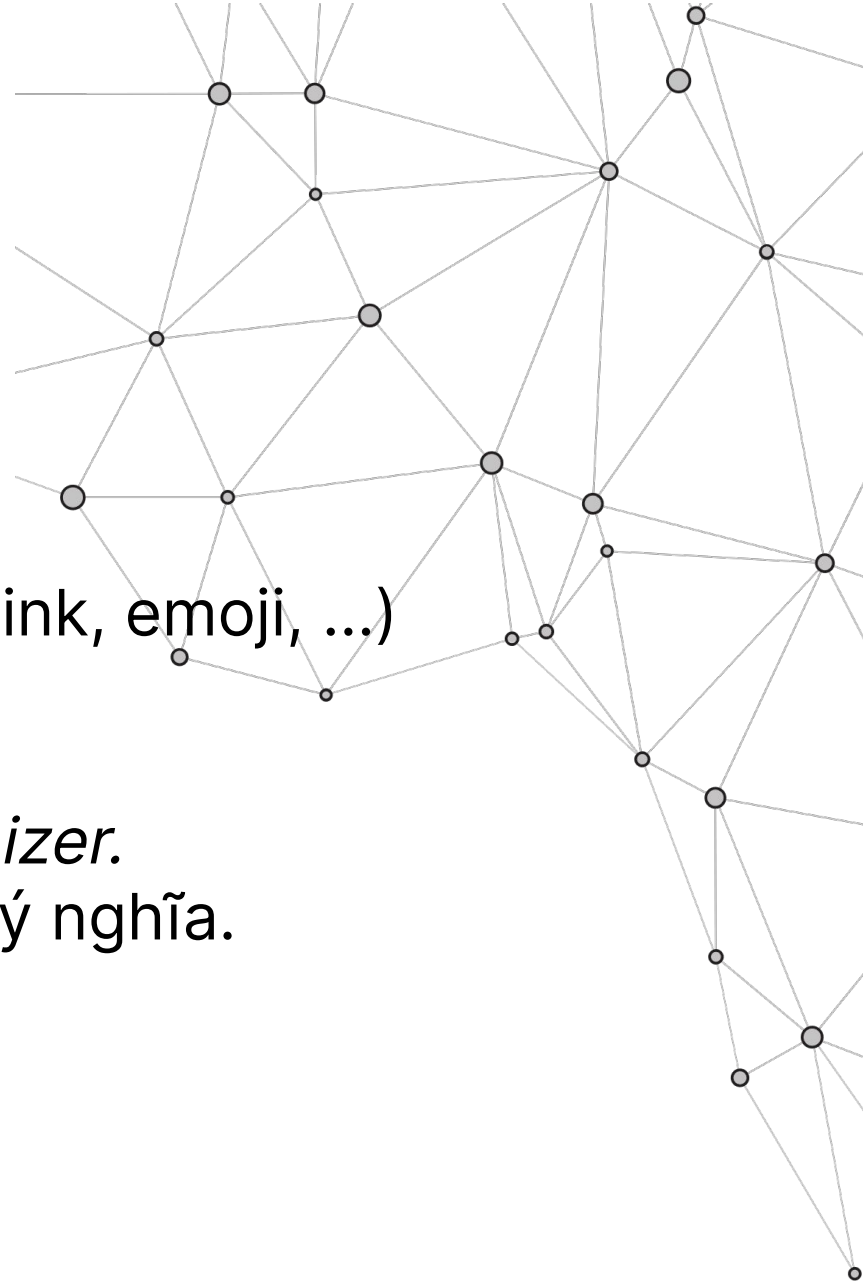
- Dữ liệu là những đoạn văn bản được người dùng tạo ra. Vì vậy, dữ liệu có cấu trúc rời rạc, không tối ưu cho quá trình phân tích dữ liệu.
- Dữ liệu văn bản có nhiều kí tự gây nhiễu (emoji, số điện thoại, đường dẫn), làm ảnh hưởng đến chất lượng mô hình nếu không được xử lý.

Làm sạch dữ liệu

Vấn đề

- Dữ liệu có nhiều từ được sử dụng nhiều lần nhưng không mang nhiều ý nghĩa (*thế, nhưng, tôi, mà,...).*
- Tiếng Việt có nhiều từ ghép. Từ ghép được hợp thành lại từ hai từ đơn có nghĩa hoặc vô nghĩa (*giáo + dục → giáo dục).* Do vậy, ta cần nối các từ ghép lại để phân tích được hiệu quả.

Làm sạch dữ liệu



Hướng giải quyết

- Xóa dữ liệu nhiễu bằng cách sử dụng RegExp. (link, emoji, ...)
- Đưa dữ liệu về chữ in thường.
- Chuẩn hóa dấu trong tiếng Việt.
- Nối hai từ đơn thành một từ ghép bằng *ViTokenizer.*
- Xóa các từ xuất hiện nhiều lần nhưng không có ý nghĩa.

Làm sạch dữ liệu

Kết quả

Cả nhà cho em hỏi với ạ 😍!! Công ty em hay phát sinh một số công việc bắt buộc phải cho nhân viên tăng ca, phát sinh không thường xuyên nhưng công việc mang tính chất dài hạn. Em liệt kê chi phí tăng ca đó vào loại chi phí nào để không cộng gộp đóng BHXH hoặc phải ký hợp đồng nào để không phải tham gia BHXH ạ? <3

công_ty phát_sinh công_việc bắt_buộc nhân_viên ca phát_sinh thường_xuyên công_việc tính_chất dài_hạn liệt_kê chi_phí ca chi_phí cộng gộp đóng bhxh ký hợp_đồng tham_gia bhxh

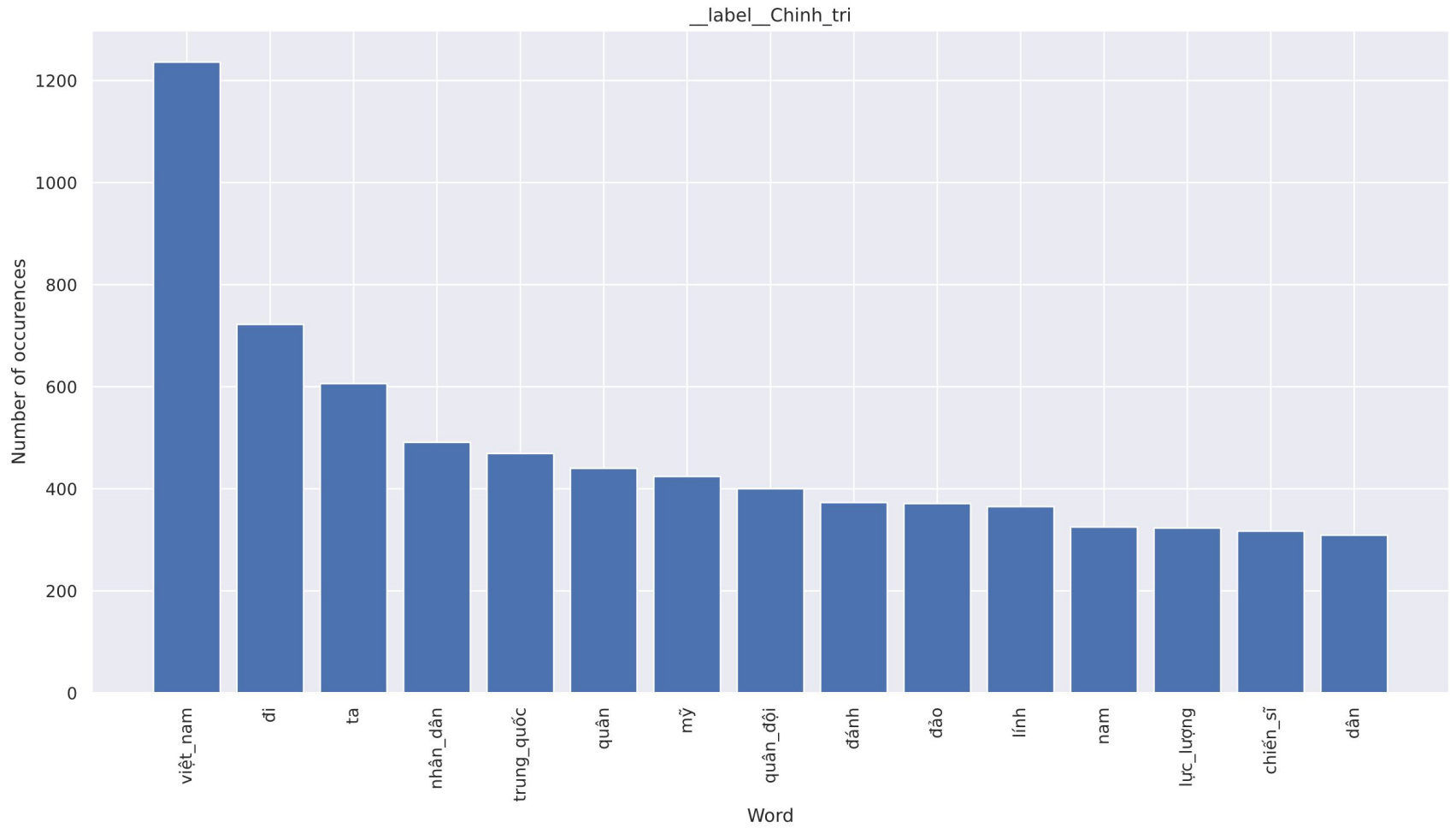
Visualize dữ liệu

Word Cloud

__label__Chinh_tri

Visualize dữ liệu

Bar Chart



Phương pháp giải quyết bài toán

Các bước giải quyết bài toán

- Chia tập dữ liệu đã được làm sạch thành 2 tập: Train và Test theo tỉ lệ 75-25.
- Vector hóa dữ liệu đã được làm sạch bằng công thực TF-IDF.
- Từ những vector được sinh ra, huấn luyện mô hình bằng thuật toán SVM.
- Kiểm tra độ chính xác của thuật toán bằng cách so sánh kết quả dự đoán với kết quả thật.

Phân chia dữ liệu train và test

- Dựa trên dữ liệu được cung cấp, nhóm đã chia tập dữ liệu thành 2 phần:
  - Tập train: sử dụng để xây dựng mô hình giải quyết bài toán dựa trên TFIDF và SVM.
  - Tập test: dữ liệu dùng để đánh giá độ hiệu quả của mô hình, mức độ chính xác của việc phân loại dữ liệu. Dữ liệu test sẽ được bỏ nhãn).

- Cách chia dữ liệu: dữ liệu được chia theo tỉ lệ train:test là 75:25 với mỗi nhãn có trong tập dữ liệu để đảm bảo độ tương đồng tỉ lệ các nhãn giữa tập train và tập test.

Thuật toán TF-IDF

- **Mục đích:** Phản ánh độ quan trọng của mỗi từ hoặc n-gr am đối với văn bản trên toàn bộ tài liệu đầu vào

- **TF:** Term Frequency (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

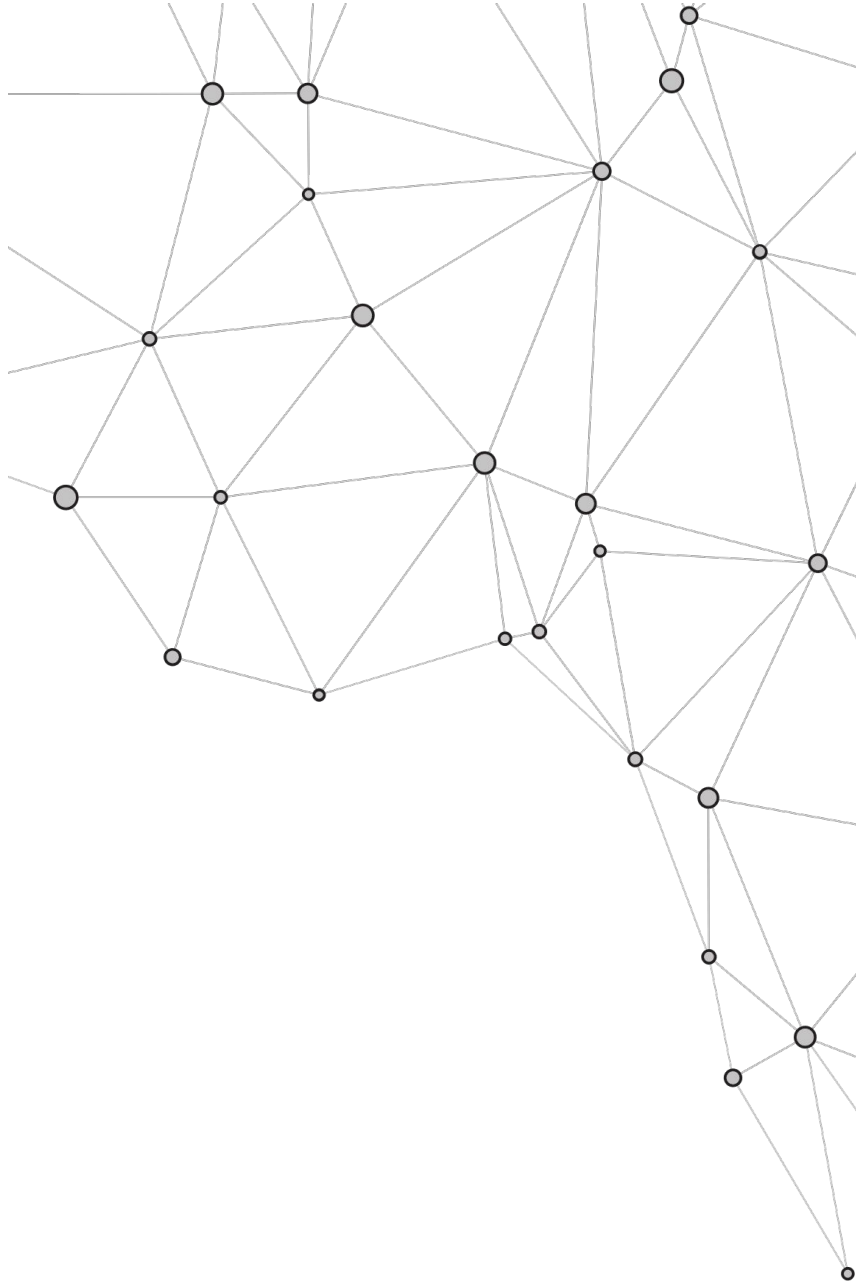
- **IDF:** Inverse Document Frequency (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ

$$\mathrm{tf}(t, d) = \frac{\mathrm{f}(t, d)}{\max\{\mathrm{f}(w, d) : w \in d\}}$$

Thuật toán TF-IDF

Kết quả

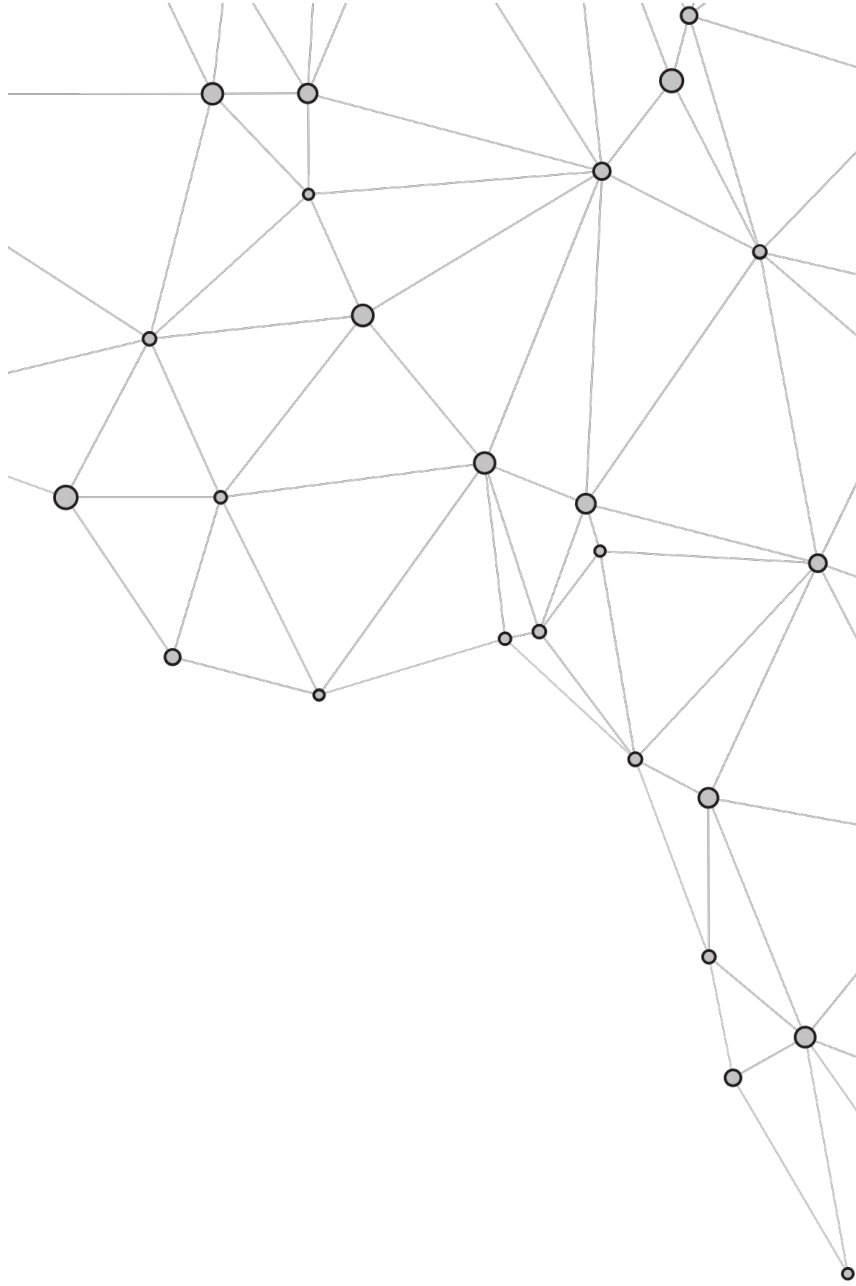
(0, 848) 0.0965859327207001 8
(0, 1344) 0.10323826902420 76
(0, 995) 0.10342093889759825
(0, 1) 0.1119697578238460 9
(0, 927) 0.1196449143209780 5
(0, 1562) 0.13307865987371503
(0, 475) 0.1010960602860630 6
(0, 14) 0.0912390458764926 3
(0, 967) 0.108680967980731
(0, 883) 0.1050673072244670 4



Thuật toán SVM

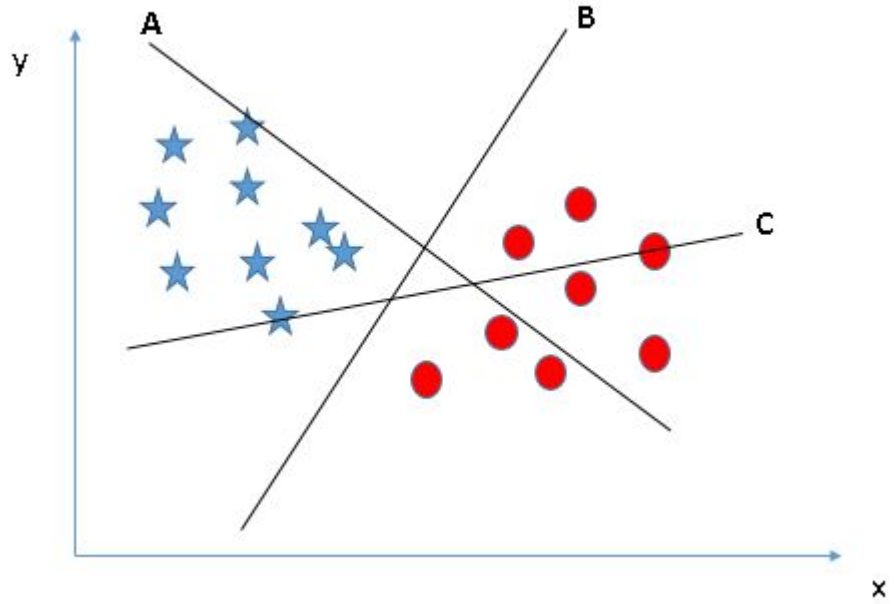
- Mục đích: là thuật toán học có giám sát (supervised learning) được sử dụng để phân lớp dữ liệu

- Ý tưởng: biểu diễn tập training trong không gian vector, mỗi tài liệu là một điểm. Phương pháp sẽ tìm ra mặt siêu phẳng (hyperplane) có thể chia không gian này thành hai lớp riêng biệt.



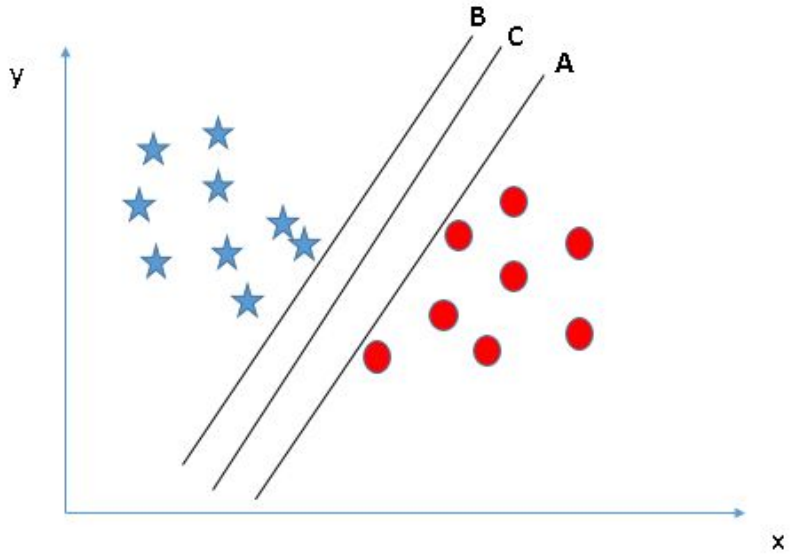
Thuật toán SVM

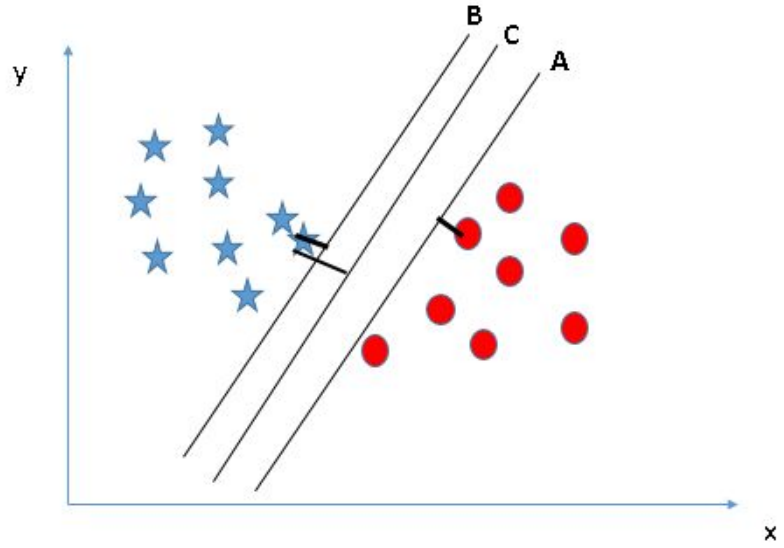
Quy tắc 1: Chọn 1 siêu phẳng để phân chia thành 2 lớp tốt nhất



Thuật toán SVM

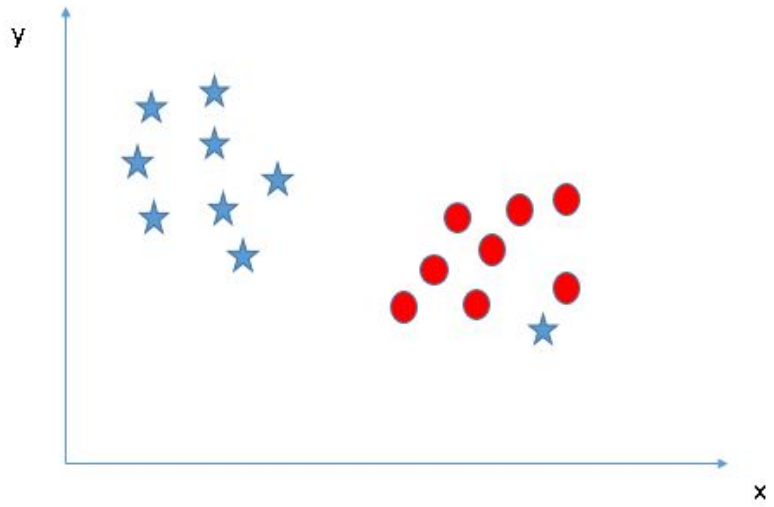
Quy tắc 2: Xác định khoảng cách lớn nhất giữa điểm dữ liệu gần nhất của hai lớp so với siêu phẳng (Margin)

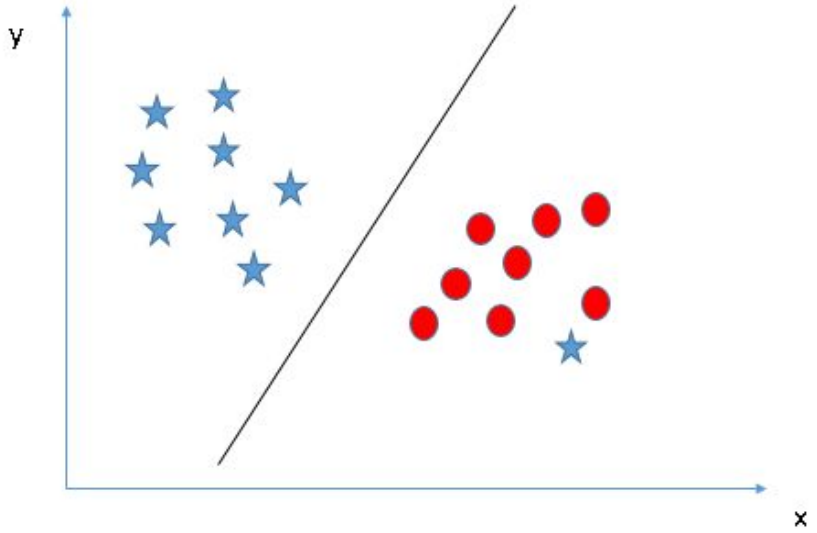




Thuật toán SVM

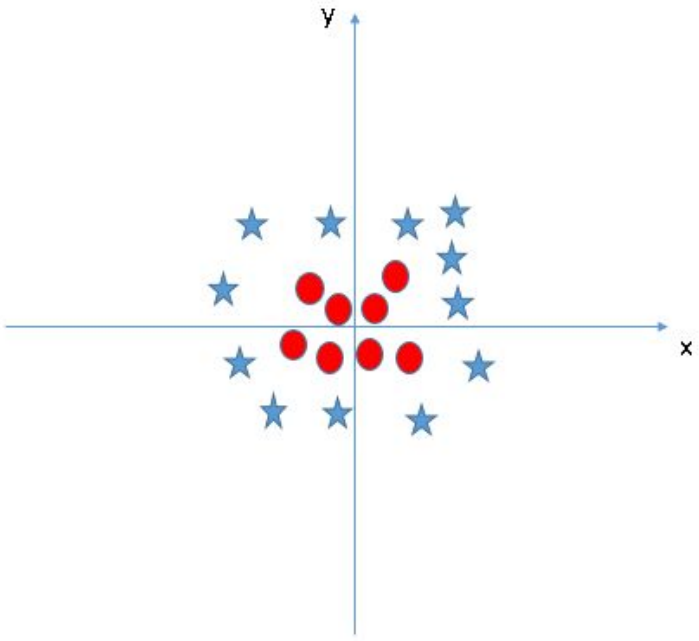
Quy tắc 3: Chấp nhận ngoại lệ cao

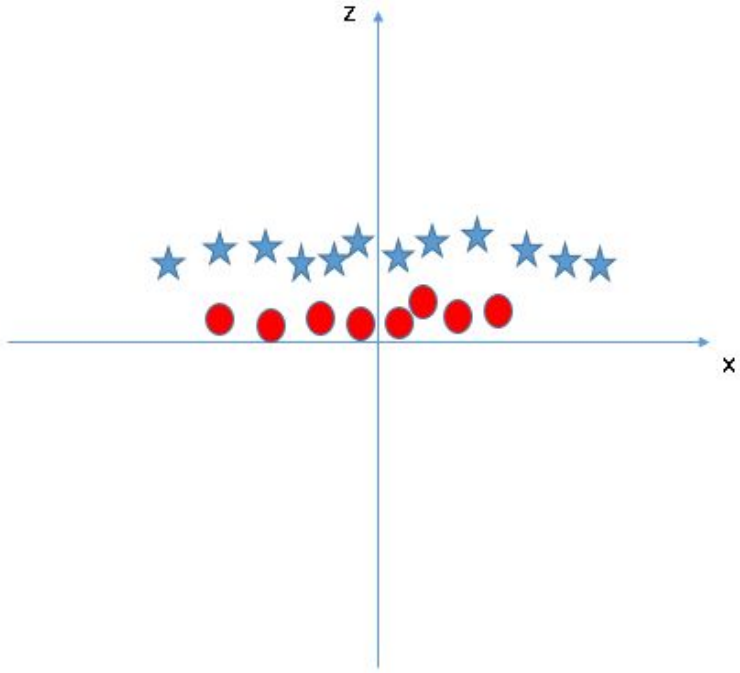




Thuật toán SVM

Quy tắc 4: Thêm tính năng cho SVM





Kết quả thử nghiệm và kết luận

85%

Độ chính xác trung bình

Nhận xét kết quả

Nhóm sử dụng hai phương pháp

- Classification Report: giúp hình dung độ chính xác và độ recall tương ứng với từng label, cũng như kết quả tổng thể với toàn bộ label.
- Confusion Matrix: giúp hình dung được lỗi sai thường nằm ở các label nào để có thể tìm được hướng giải quyết.

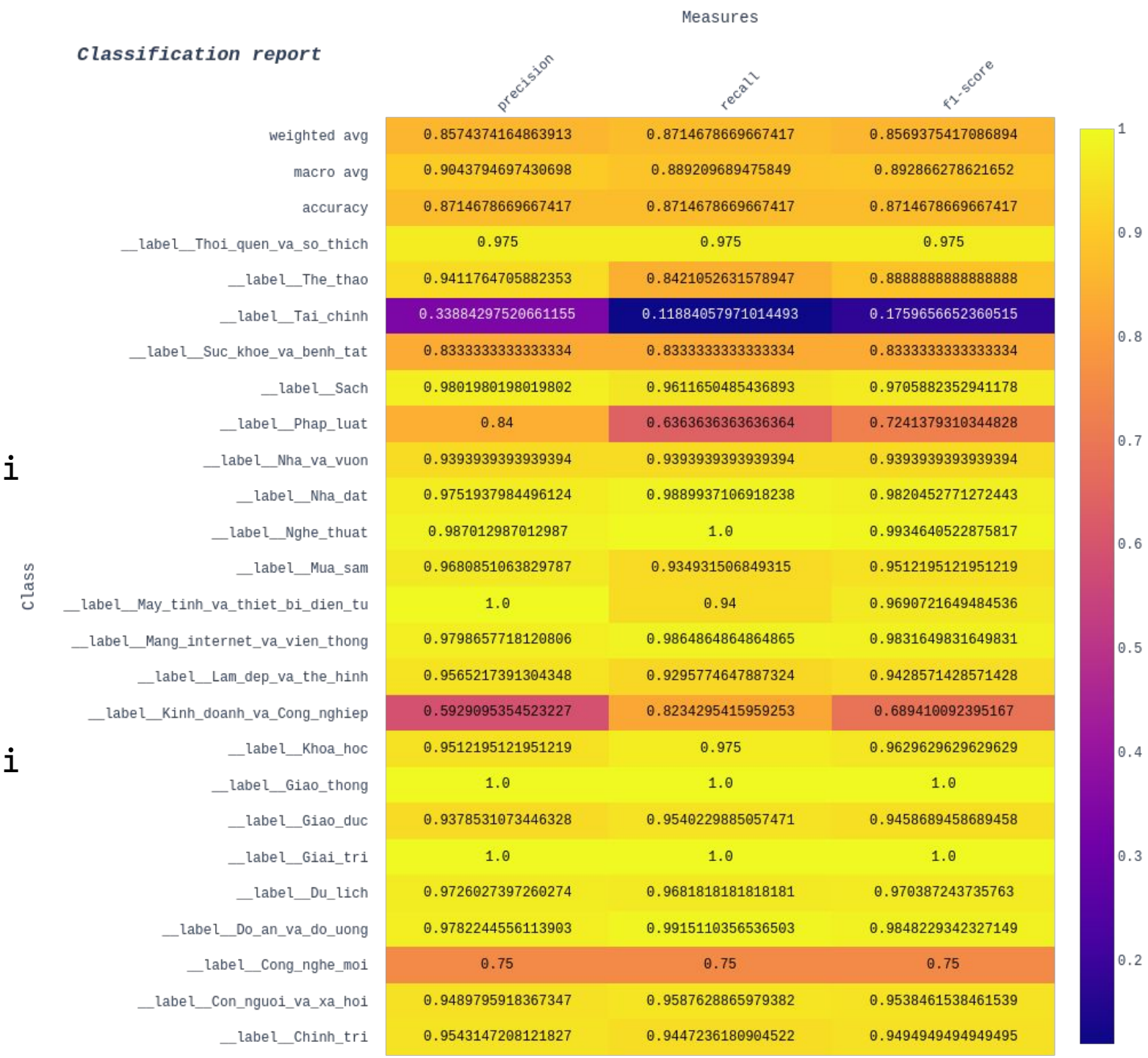
Thống kê

Kết quả tốt

- __label_Giao_thong (100%)
- __label_Giai_tri (100%)
- __label_May_tinh_va_thiet_bi_dien_tu (100%)

Kết quả thấp

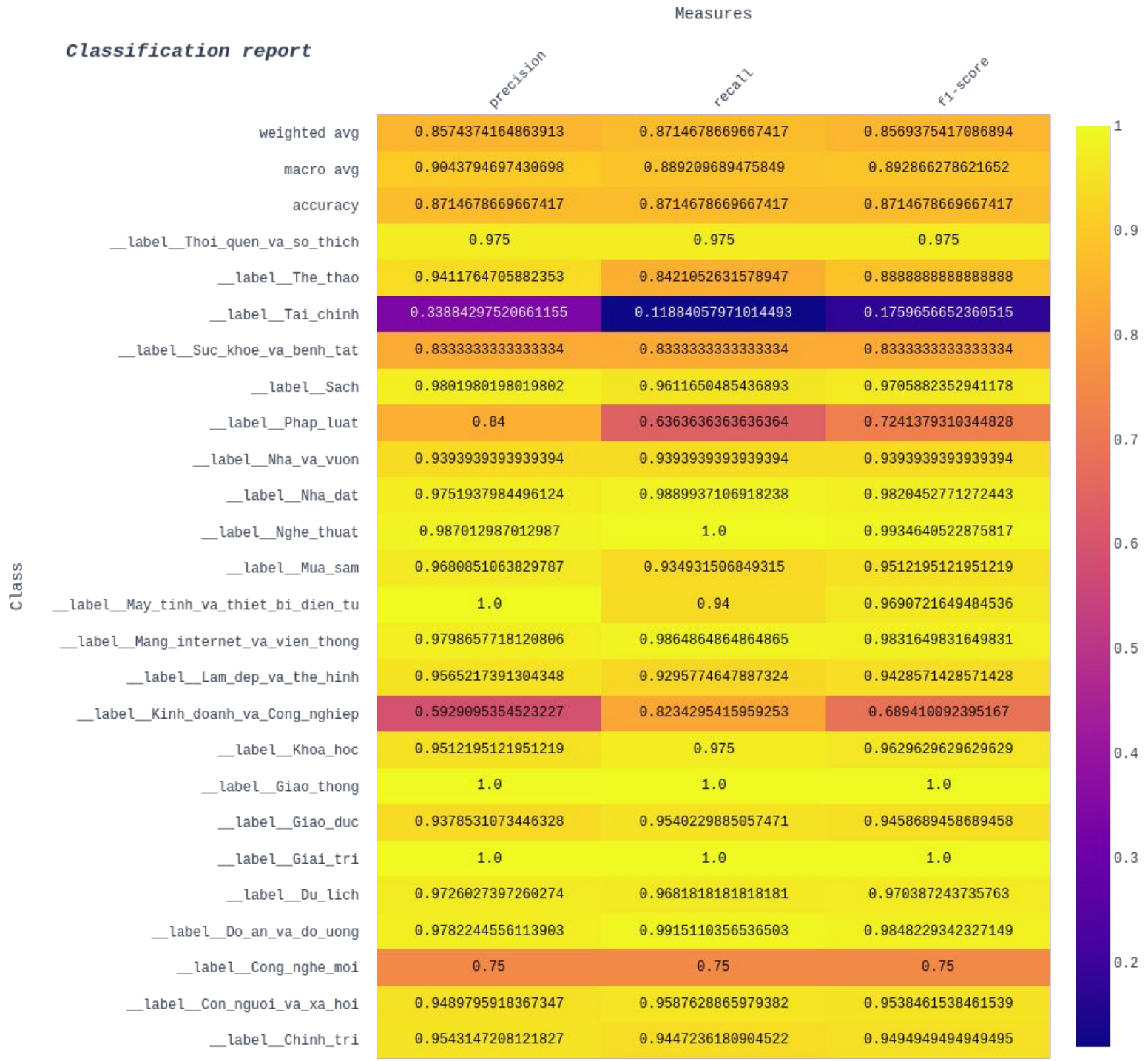
- __label_Tai_chinh (34%)
- __label_Kinh_doanh_va_cong_nghiep (59%)
- __label_Cong_nghe_moi (75%)



Giải thích

Kết quả tốt

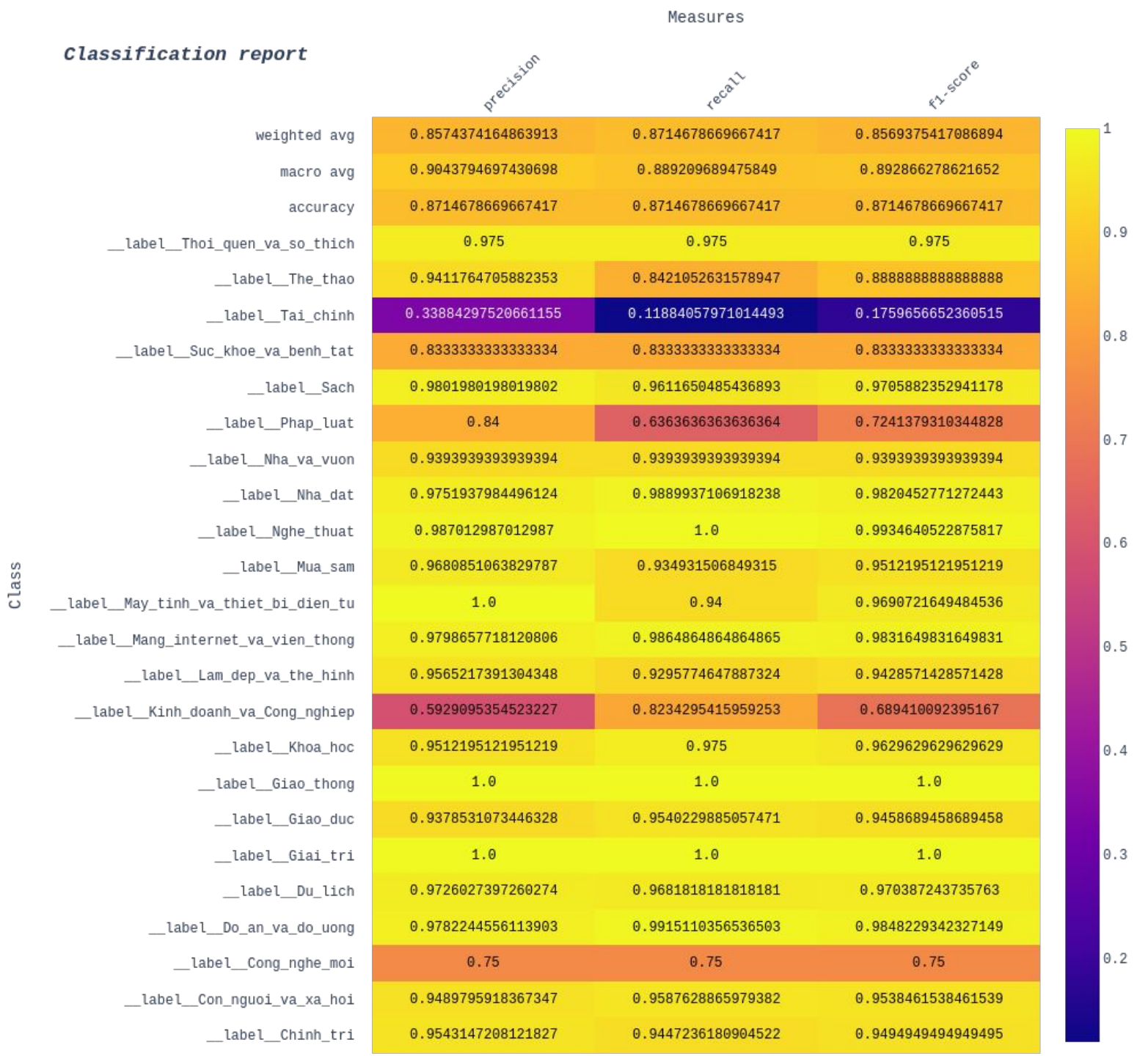
Ba nhãn có kết quả cao nhất đều có các cụm từ chuyên ngành đặc biệt và độc nhất với nhãn của mình.



Giải thích

Kết quả thấp

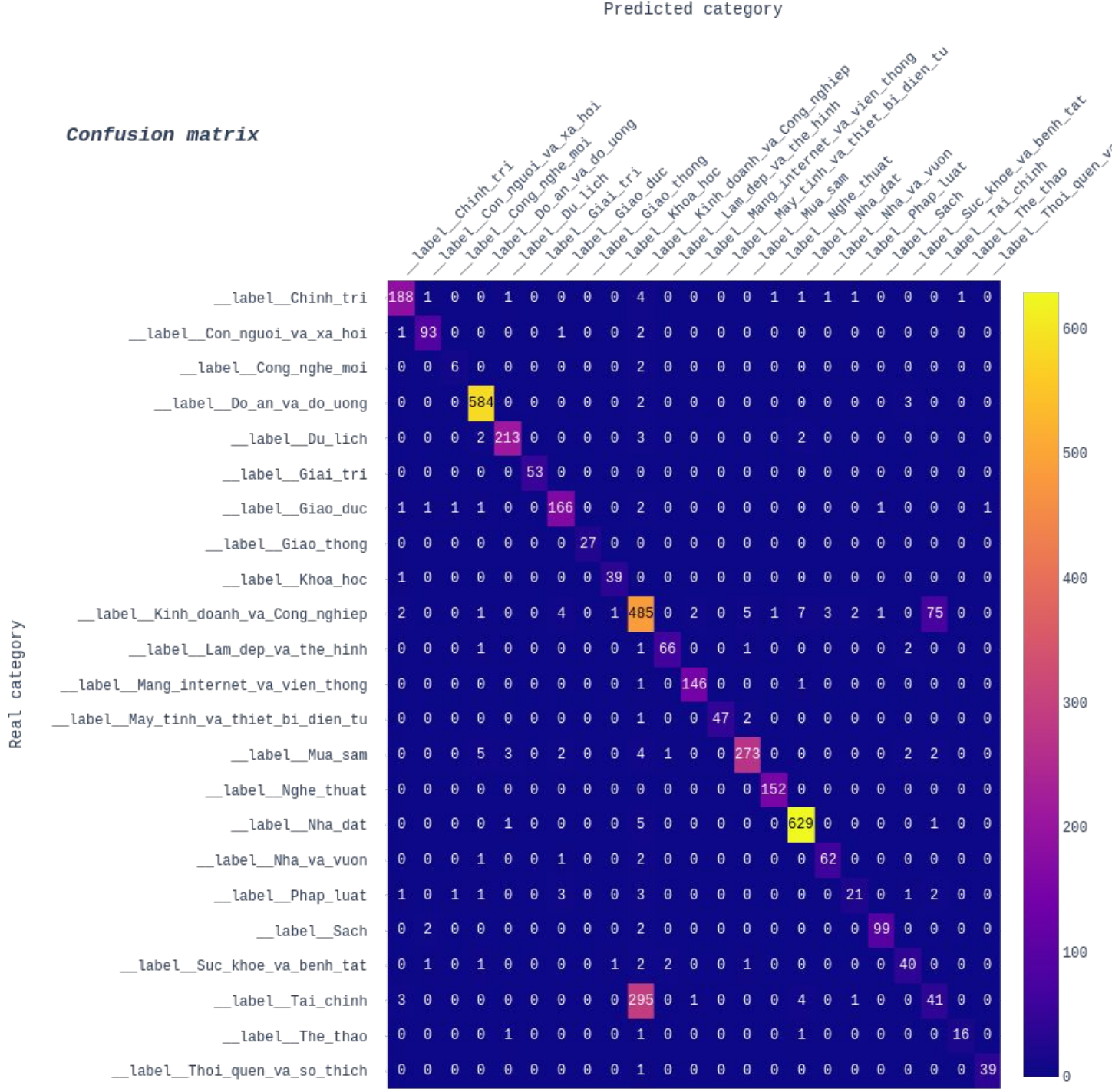
- Với nhãn cong_nghe, kết quả này có thể được giải thích là do nhãn chiếm tỉ lệ rất thấp trong dữ liệu huấn luyện (33 hàng, chiếm 0.2%), nên mô hình không có đủ dữ liệu để đưa ra dự đoán chính xác.
- Với hai nhãn còn lại, độ chính xác thấp là do mô hình thường xuyên dự đoán nhầm tai_chinh thành kinh_doanh và ngược lại. Danh sách từ khóa của hai nhãn có độ tương đồng rất cao, dẫn đến sai sót trong quá trình dự đoán.



Giải thích

Kết quả thấp

Chủ đề Kinh Doanh là một chủ đề rộng, có thể bao hàm nhiều chủ đề nhỏ hơn như mua bán nhà đất, luật kinh doanh, nội thất gia đình, kinh doanh tranh ảnh nghệ thuật,...



Kết luận

Tóm lại, các kết quả nói trên đã phản ánh được bản chất của các chủ đề cũng như hạn chế của phương pháp sử dụng. Các chủ đề thường không độc lập với nhau mà cũng bao hàm và giao nhau, dẫn tới sai sót trong quá trình dự đoán. Mặt khác, phương pháp sử dụng không quan tâm tới cấu trúc câu mà chỉ tập trung vào số lần xuất hiện của từ cũng một phần làm giảm độ chính xác của mô hình.