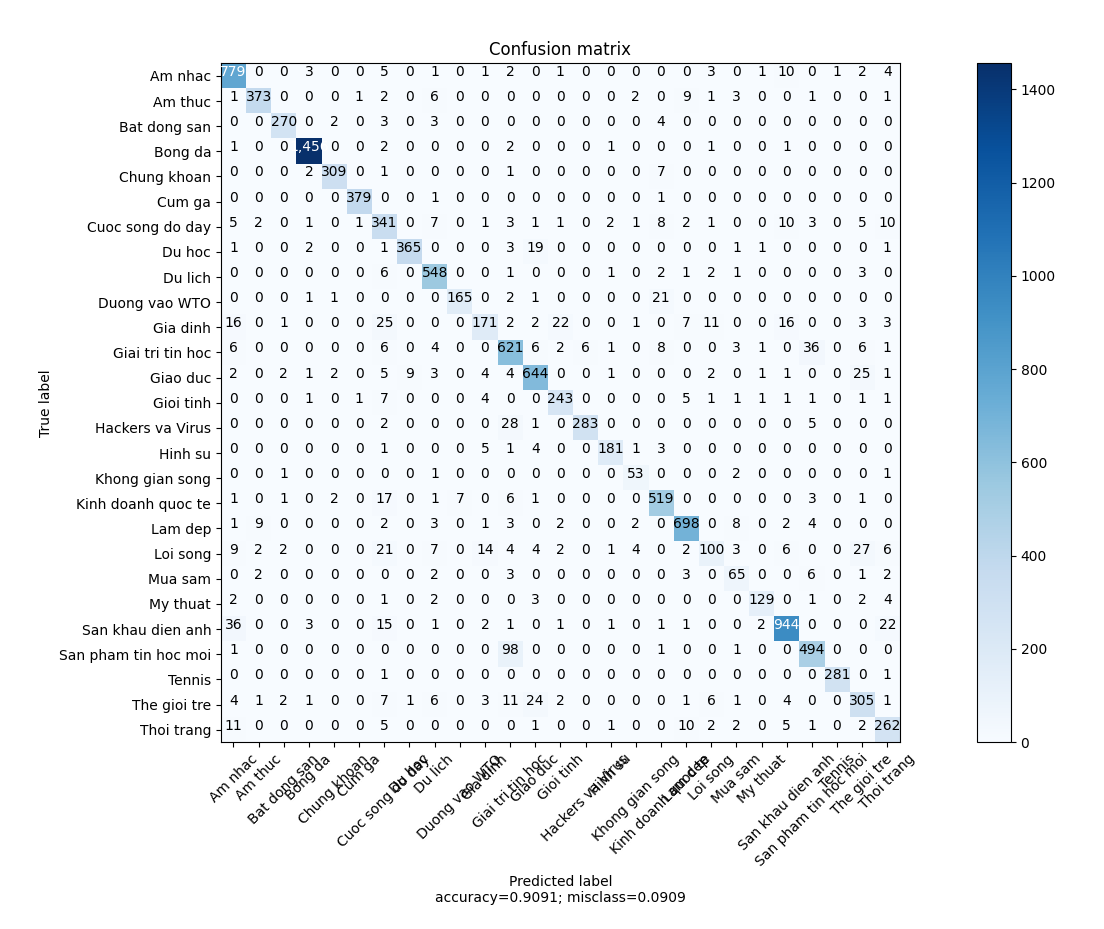
TEST\_THREE - 8/7/2021

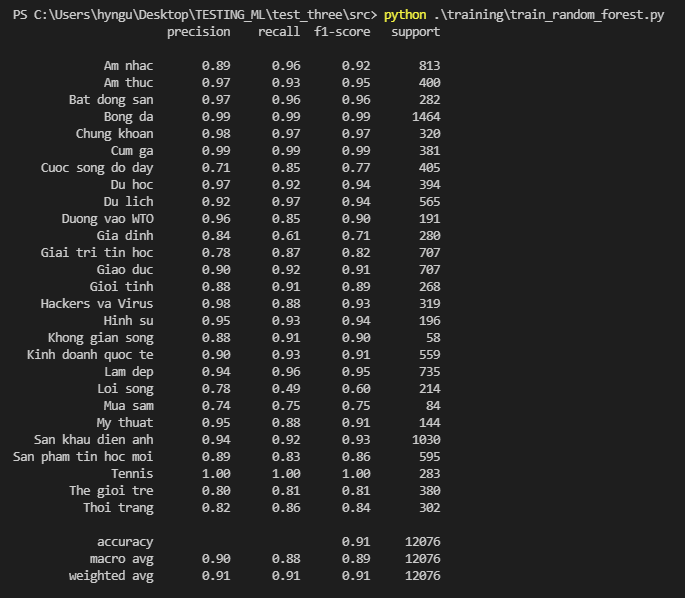
**- Sửa đổi: Kết quả lần trước thấp là do em viết sai hàm load file text <><><><>**

**- Mô hình Random Forest**

**Kết quả ma trận nhầm lẫn giữa các lớp của mô hình:**



**Kết quả bảng phân phối Precision-Recall:**



**Đánh giá ban đầu:**

+ Tập dữ liệu huấn luyện hiện tại còn khá ít dữ liệu huấn luyện: Tập train có khoảng 14000 file, tập test có khoảng 12000 file => không tương xứng theo các tỷ lệ hiện hành

+ Độ chính xác trên tập test khoảng 91%, nhãn “Loi song” có điểm F1-score còn khá thấp

+ Sau khi phân tích thử một số đoạn text nhận thấy nội dung, câu từ của một số chủ đề có liên quan đến nhau, Ví dụ chủ đề Âm nhạc cũng được đề cập trong một số đoạn văn nói về Giới tính, ...

+ Các tham số được thử nghiệm trên mô hình Random Forest:

**n\_estimators=150,**

**criterion="gini",**

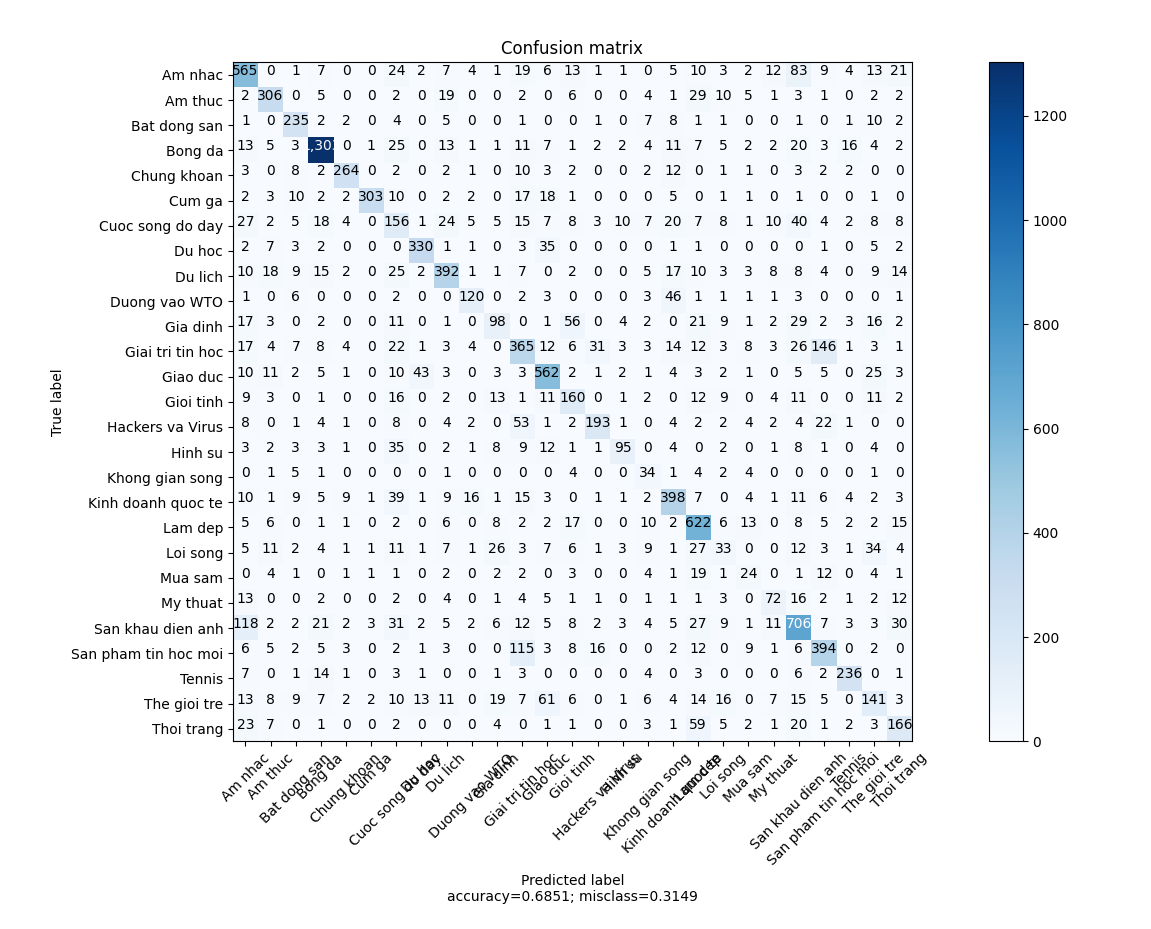
**min\_samples\_leaf=1,**

**min\_samples\_split=2, bootstrap=False**

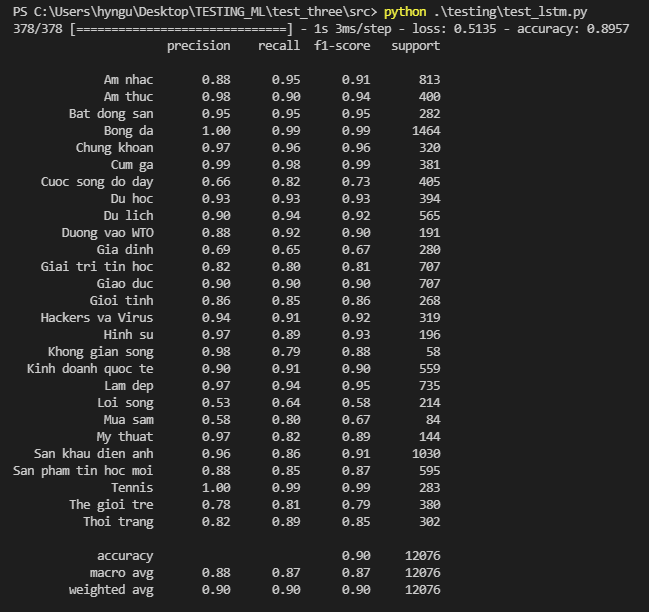
+ Các tham số này được tuning dựa trên phương pháp GridSearch duyệt qua một list các tham số được cho trước.

**- Mô hình LSTM**

**Kết quả ma trận nhầm lẫn giữa các lớp của mô hình:**



**Kết quả bảng phân phối Precision-Recall:**



**Đánh giá ban đầu:**

+ Độ chính xác của mô hình sau khoảng 20 epoch là 89.57%

+ Mô hình hiện tại có độ chính xác thấp hơn phương pháp RandomForest => Đặc trưng TFIDF thích hợp hơn với các phương pháp học máy cơ bản => Đề xuất dùng BERT để mã hoá vector cho từ.

+ Phương pháp chọn các siêu tham số (hyperparmeters):

+ Dựa trên số lượng dữ liệu để chọn batch\_size cho hợp lý: Ở đây, dữ liệu cho mỗi nhãn còn khá ít nên chọn batch\_size = 32

+ Learning\_rate được thử nghiệm dựa trên sự giảm của hàm loss: Khi thử nghiệm giá trị 1e-6 thì thấy hàm loss giảm rất chậm, quyết định lựa chọn giá trị learning\_rate=0.001

**- Hướng dẫn chạy thử nghiệm code**

+ Cài đặt môi trường, các thư viện cần thiết:

**pip install -r requirements.txt**

**pip install -e.**

+ Copy đoạn text cần kiểm tra vào file **test.txt** trong thư mục **Test\_files**

+ Truy cập vào thư mục src:

**cd src/**

+ Tiến hành chạy 1 trong 2 câu lệnh sau để thử nghiệm mô hình:

**python testing/test\_random\_forest.py** hoặc **python testing/test\_lstm.py**