**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Viện Công Nghệ Thông Tin Và Truyền Thông**

**---🙠**🕮**🙢---**



**Bài tập lớn môn học sâu ứng dụng**

**Đề tài: dự đoán số sao của bình luận trên trang Lazada**

***Giáo viên hướng dẫn: Trịnh Anh Phúc***

***Nhóm: 4***

|  |  |
| --- | --- |
| ***Thành viên nhóm*** | ***MSSV*** |
| ***Đỗ Huy Anh*** | ***20193977*** |
| ***Hoàng Ngọc Lâm*** | ***20194089*** |
| ***Đinh Chí Công*** | ***20193996*** |
| ***Trương Quang Phú*** | ***20194138*** |
| ***Vũ Quang Huy*** | ***20183558*** |
| ***Phùng Bảo Hà*** | ***20190047*** |
| ***Trần Đức Anh Trường*** | ***20183648*** |
| ***Nguyễn Kim Dân*** | ***20194007*** |
| ***Nguyễn Hữu Hải*** | ***20194039*** |

Mục lục

[1. Giới thiệu bài toán 1](#_Toc108517725)

[2. Bộ dữ liệu --Phân bố dữ liệu, tiền xử lý, ... 1](#_Toc108517726)

[2.1. Bộ dữ liệu: 1](#_Toc108517727)

[2.2. Tiền xử lý dữ liệu: 2](#_Toc108517728)

[2.3. Chia tập train, val, test 4](#_Toc108517729)

[3. Thuật toán, model 5](#_Toc108517730)

[4. Thực nghiệm 8](#_Toc108517731)

[5. Tổng kết 9](#_Toc108517732)

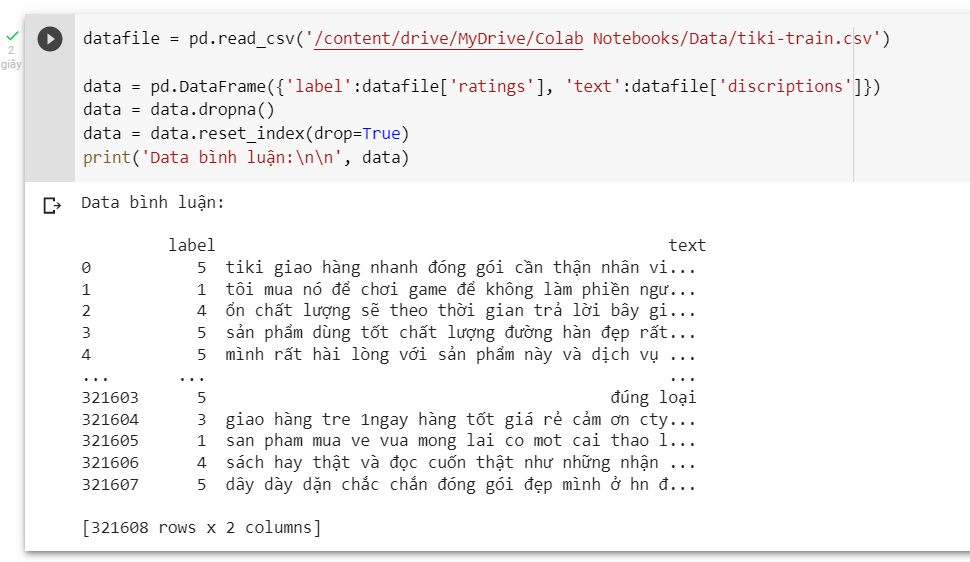
# 1. Giới thiệu bài toán

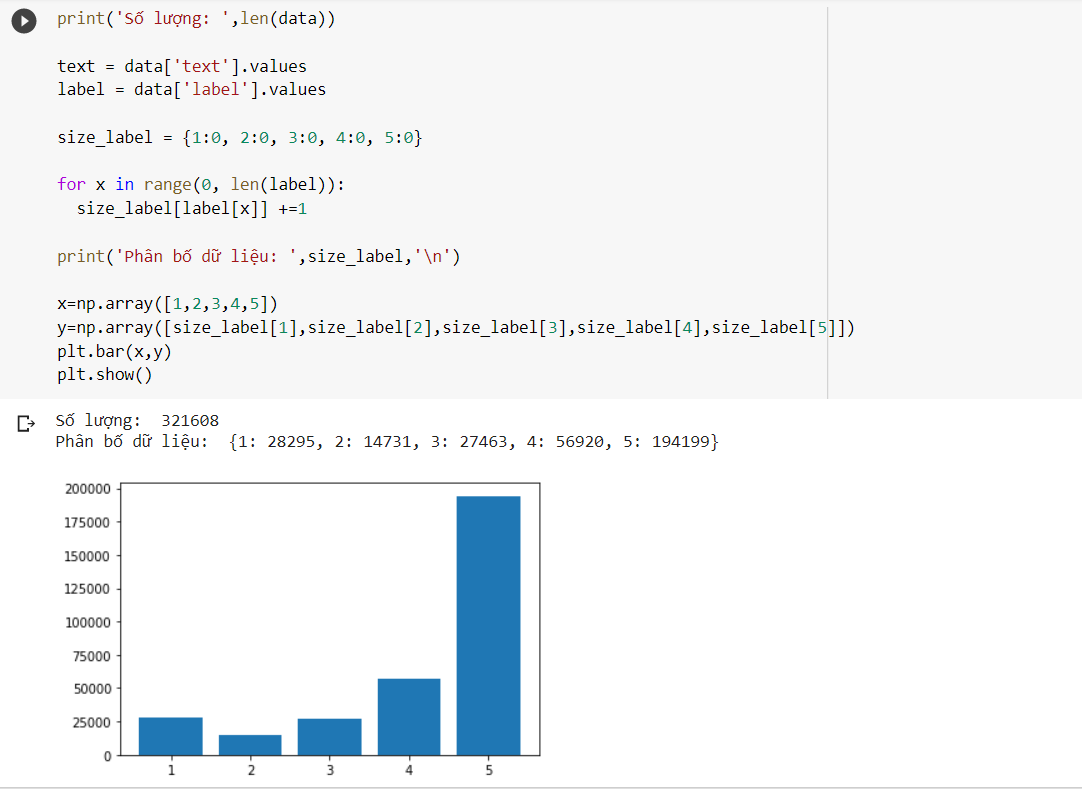
Với sự phát triển của khoa học và công nghệ, dữ liệu không những càng ngày càng lớn hơn mà còn trở lên dễ dàng tiếp cận hơn với tất cả mọi người. Cũng bởi lý do này, trí tuệ nhân tạo, hay cụ thể là học sâu, càng ngày càng mang một vai trò quan trọng để giải quyết cái nguồn dữ liệu khổng lồ đó. Chủ đề mà nhóm bọn em đã làm là sử dụng học sâu, qua một bình luận bất kì trên trang Lazada thì hãy dự đoán số sao của bình luận đó.

# 2. Bộ dữ liệu --Phân bố dữ liệu, tiền xử lý, ...

## 2.1. Bộ dữ liệu:

Dữ liệu là bình luận đánh giá của người mua hàng trên trang Tiki được nhóm tự crawl. Số lượng data examples khoảng 320000. Tập dữ liệu được lưu trong file tiki-train.csv và đã được gán nhãn lần lượt từ 1 tới 5





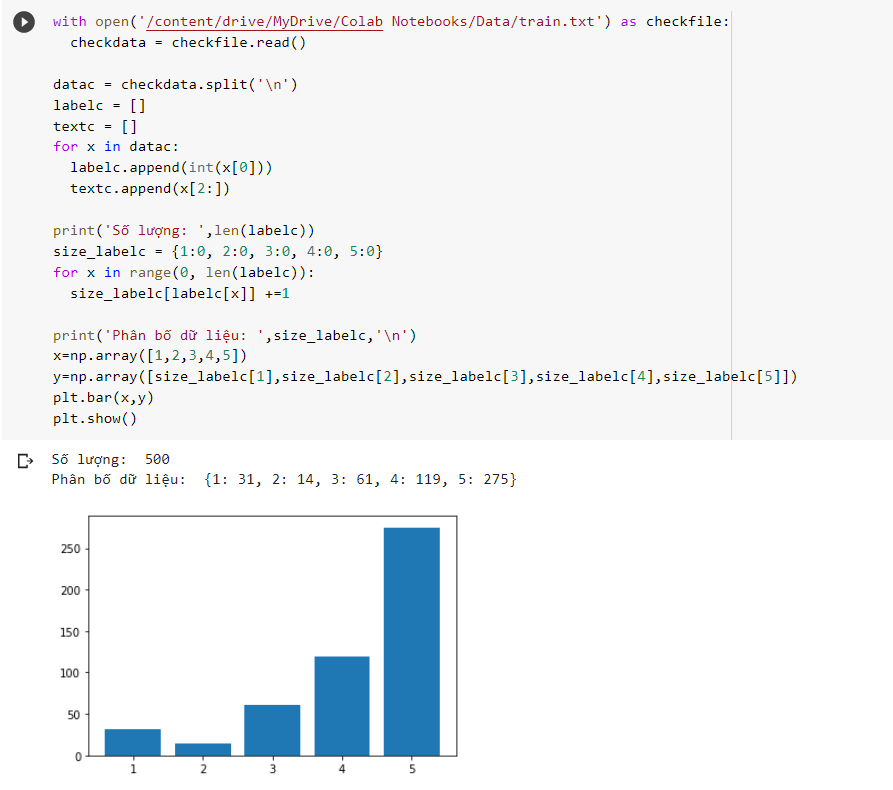
Đánh giá:

* Bình luận 2 sao có ít so với các bình luận khác (14731 lình luận).
* Bình luận 5 sao có số lượng lớn hơn hẳn và có thể gây nhiễu cao.

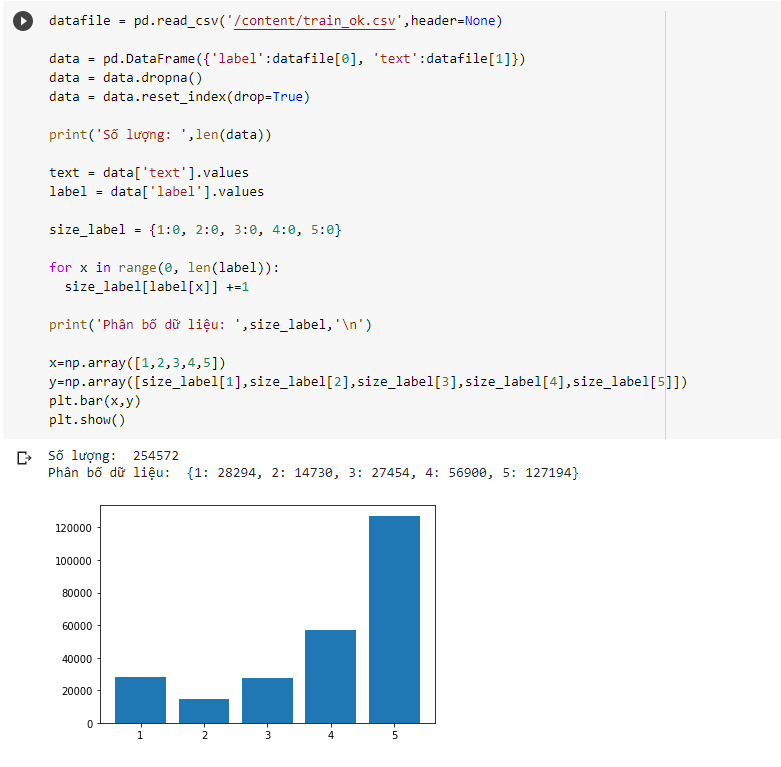
## 2.2. Tiền xử lý dữ liệu:

Nhóm chúng em xử lý dữ liệu đồng thời theo 2 hướng:

* Xử lý từng bình luận đúng quy cách:
  + Chuẩn hóa mã unicode
  + Chuẩn hóa gõ dấu Tiếng Việt
  + Xóa ký tự thừa, khoảng trắng
  + Chuyển đổi các từ viết tắt, teencode
* Xử lý lại phân bố của dữ liệu:
  + Sau khi tham khảo mẫu data train.txt mà thầy cung cấp cùng với sử dụng mô hình để chuẩn đoán tập test.txt, nhóm chúng em đã tiến hành phân bố lại các bình luận.



Cuối cùng nhóm đã tập hợp lại dữ liệu và sử dụng tập huấn luyện cuối cùng để huấn luyện model: train\_ok.csv



## 2.3. Chia tập train, val, test

Vì số lượng bình luận rất nhiều nên nhóm em đã chia tập val chiếm 5% so với dữ liệu ban đầu và tập test chiếm 5% so với dữ liệu còn lại

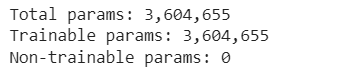


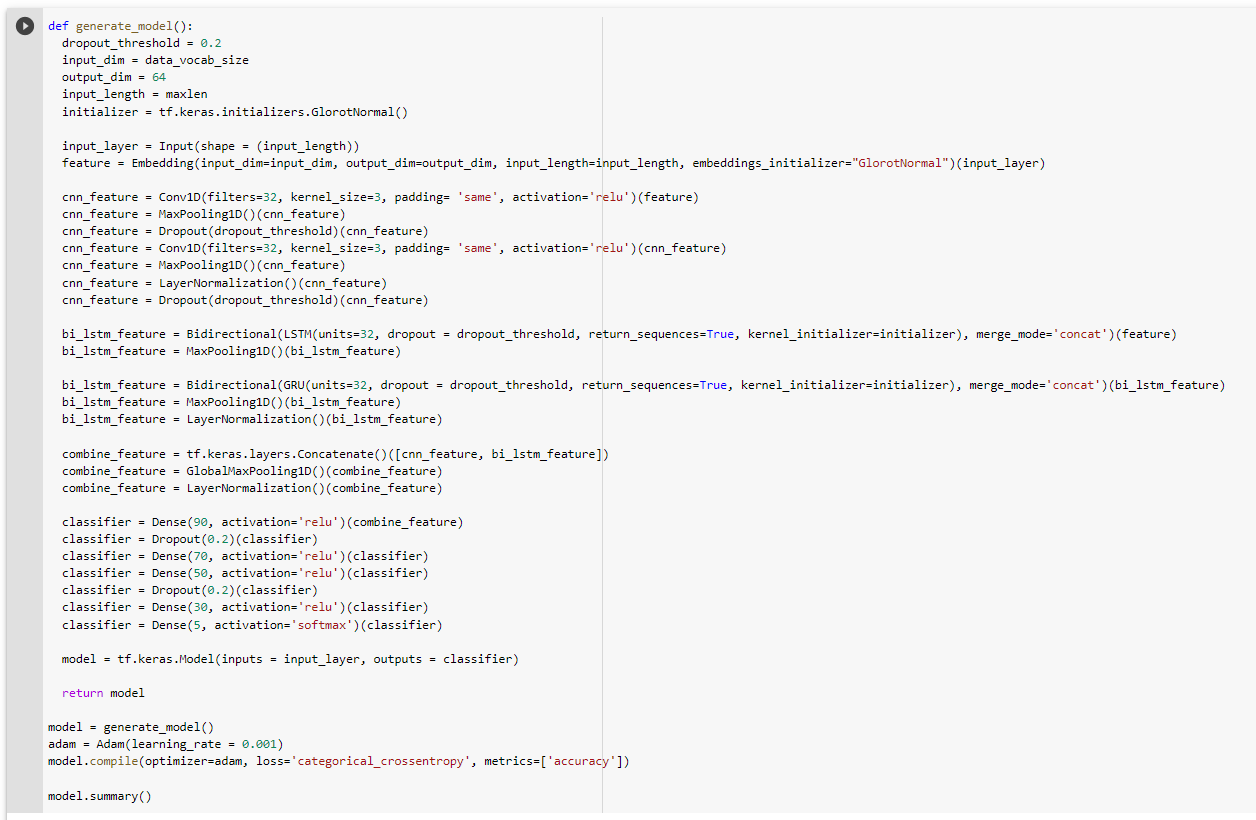
# 3. Thuật toán, model

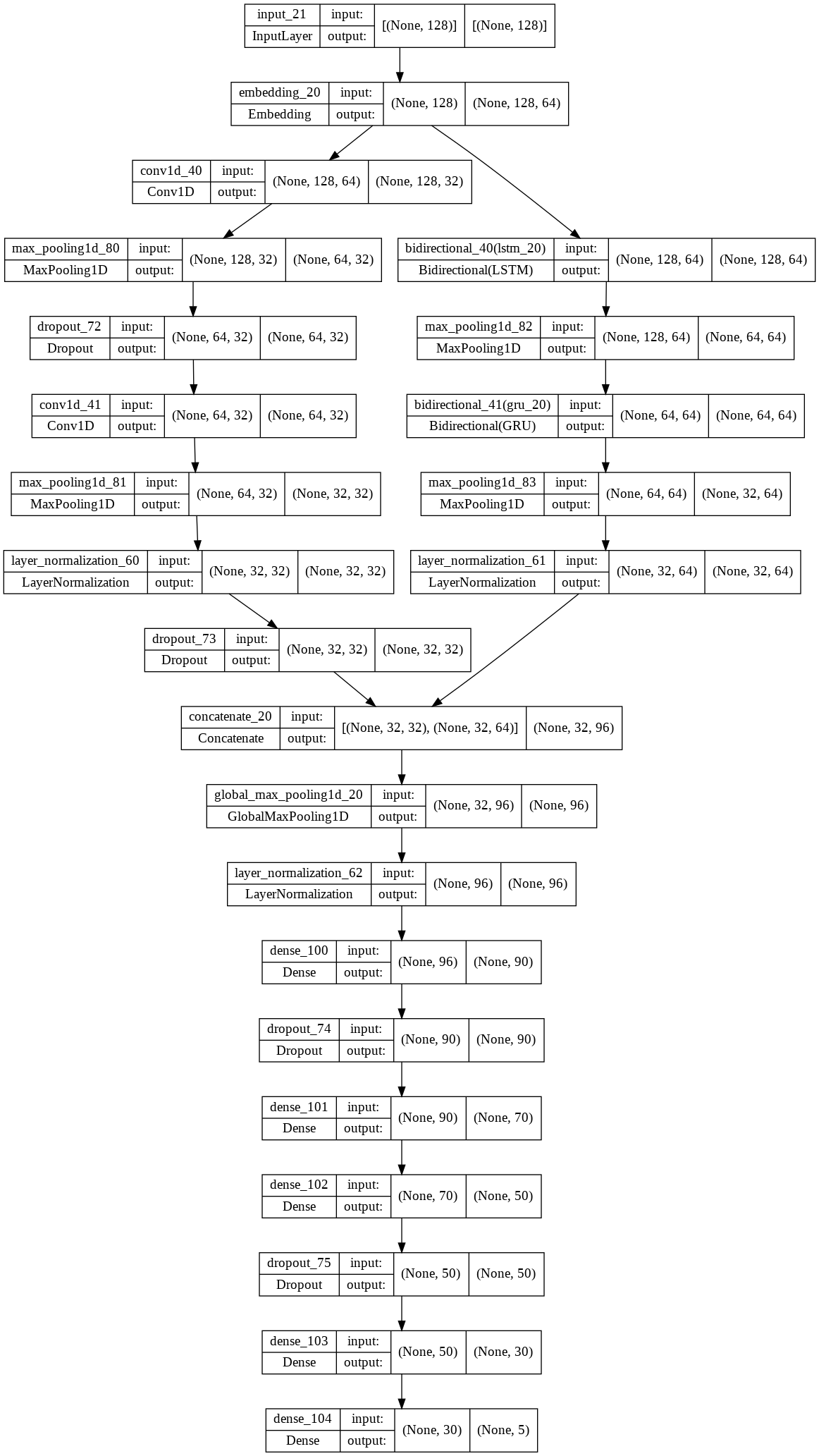
Thông qua nghiên cứu và thử nghiệm các model thì nhóm chúng em đã sử dụng model: CNN & Bidirectional LSTM.

Đây là mô hình kết hợp của mạng CNN và mạng 2 chiều LSTM:

* + Mạng CNN sẽ đánh giá tổng quát bình luận.
  + Mạng 2 chiều LSTM sẽ phân tích cảm xúc theo 2 chiều của câu bình luận.
* Mô hình :
  + Hàm kích hoạt sử dụng: hàm relu và hàm softmax ở cuối khi chia về các nhãn
  + Khởi tạo trọng số: tensorflow.keras.initializers.GlorotNormal()
  + Thuật toán: tensorflow.keras.optimizers.Adam
  + Tốc độ học learning rate: 0.001
  + Hàm loss: categorical\_crossentropy
  + Metrics: accuracy
  + Để tránh overfitting: sử dụng Dropout với tham số 0,2

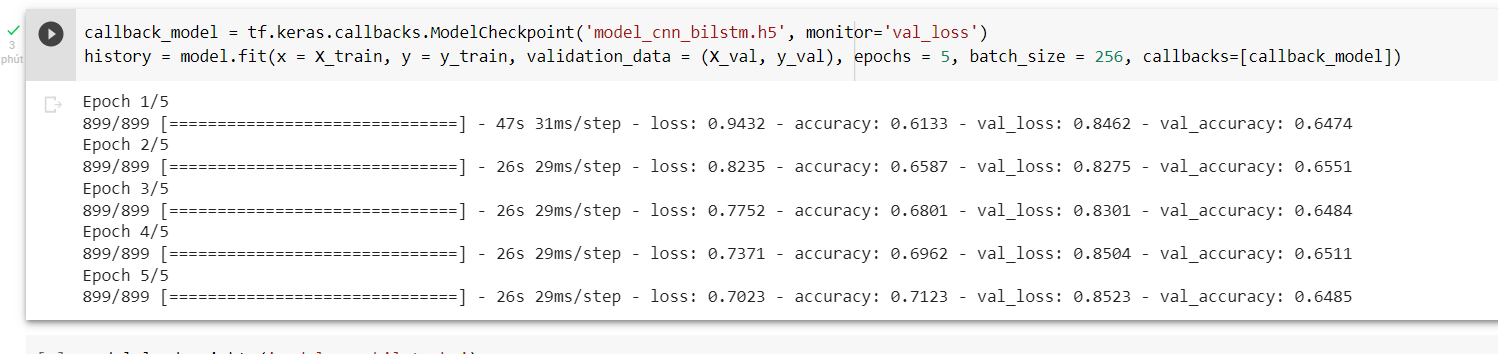




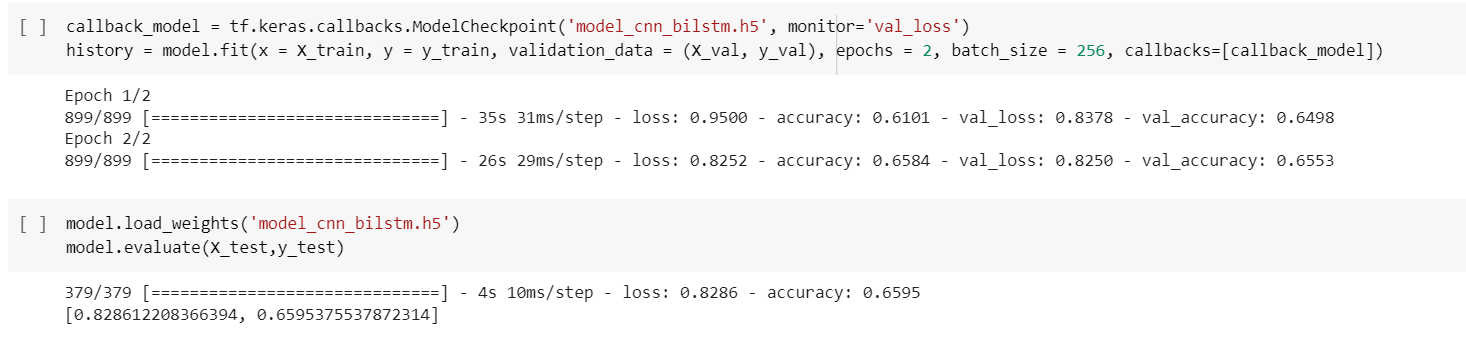


# 4. Thực nghiệm

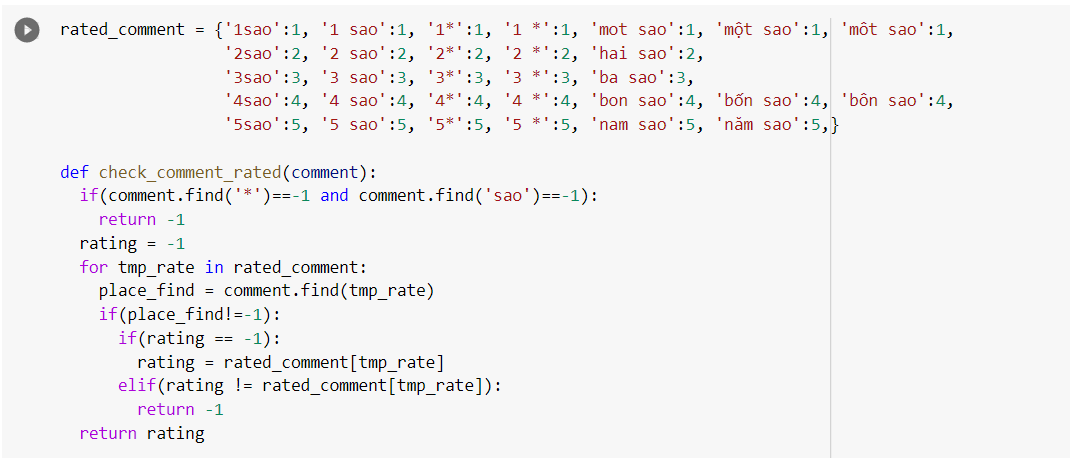
Sau khi thực hiện với nhiều tham số nên nhóm thực nghiệm thử với 5 epochs và batch\_size = 256 và thấy kết quả đã hội tụ ở epoch 2/5:



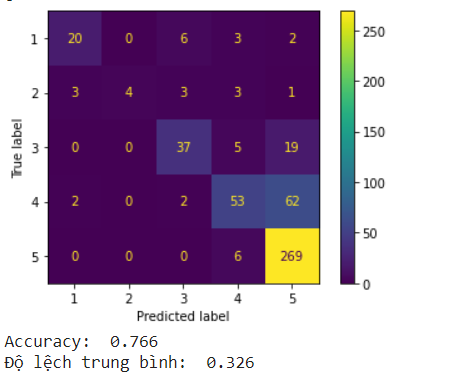
Sau khi chạy thử với kết quả với tập test thu được accuracy = 65,95%



Sau đó nhóm đã áp dụng thêm luật nếu trong bình luận đánh giá có đánh giá sẵn bình luận thì sẽ gán luôn nhãn mà không chạy qua mô hình nữa:



Cuối cùng nhóm qua lại thực hiện kiểm tra trên tập train.txt và nhận được kết quả khá khả quan:



# 5. Tổng kết

Qua bài tập lớn của môn học này, các thành viên trong nhóm đã học được về luồng xử lý một bài toán deep learning cụ thể là bài toán dự đoán số sao của bình luận. Các thành viên đã nắm được các khâu như: lấy thêm dữ lieu, xử lý dữ liệu, tinh chỉnh mô hình,… Trong tương lai có thể nhóm sẽ thử nghiệm những mô hình khác mà chưa có thời gian xem qua để tìm kiếm kết quả dự đoán có độ chinh xác cao nhất.