HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF INFORMATION AND TECHNOLOGY

---------------o0o---------------

Seminar II

Detection of rating based on TripAdvisor reviews

Instructor: Dr Phuong Thanh Nguyen

Student: Trung Mai Duc – 20232177M

HANOI, OCTOBER 2024

Table of content

I. Introduction

Hotel prediction is the process of predicting sentiment and the score of reviews based on textual review. This task is a part of both Nature Language Processing (NLP), Deep Learning and has many applications in real life. Hotel review prediction is widely used by hotel management to better understand customer preferences, improve service quality, and respond more effectively to complaints or compliments. It also helps potential guests by summarizing reviews or highlighting key factors influencing a hotel's rating.

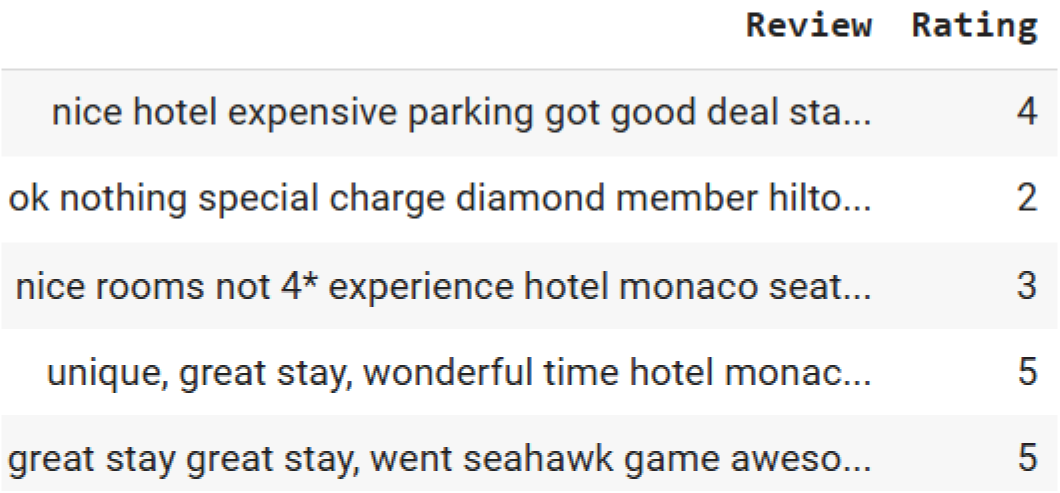
Because of the above benefits, I decided to choose the topic "Detection of rating based on TripAdvisor reviews" as my research topic for the subject.

II. Dataset used

1. Overview

TripAdvisor is a hotel booking application. It provides hotels with corresponding reviews to help users make informed choices. In this project, the score of review is predicted based on textual reviews.

The dataset has 20000 reviews crawled from TripAdvisor. It has two columns: the textual review and score rating.



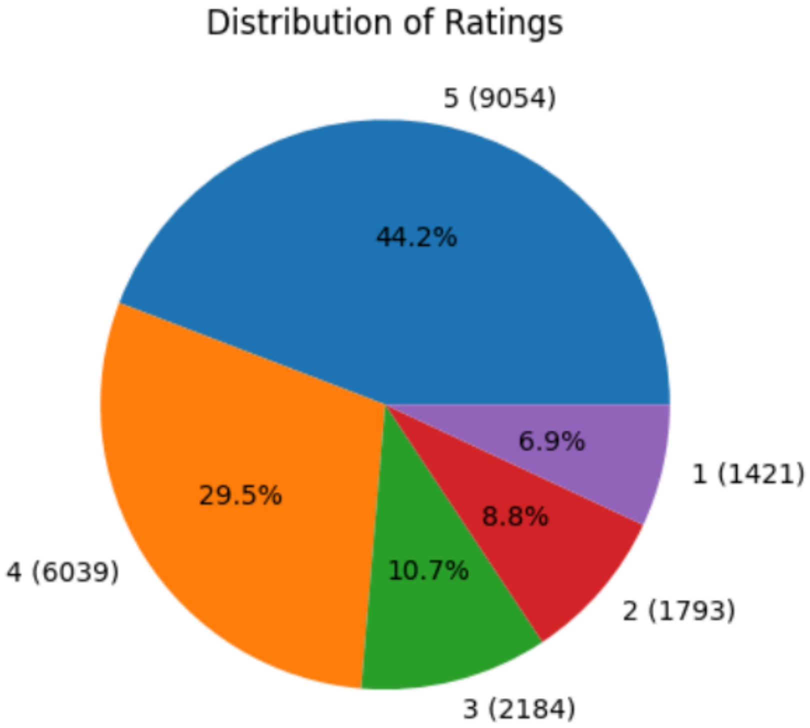
Review	Rating
nice hotel expensive parking got good deal sta...	4
ok nothing special charge diamond member hilto...	2
nice rooms not 4* experience hotel monaco seat...	3
unique, great stay, wonderful time hotel monac...	5
great stay great stay, went seahawk game aweso...	5

2. Detail of dataset

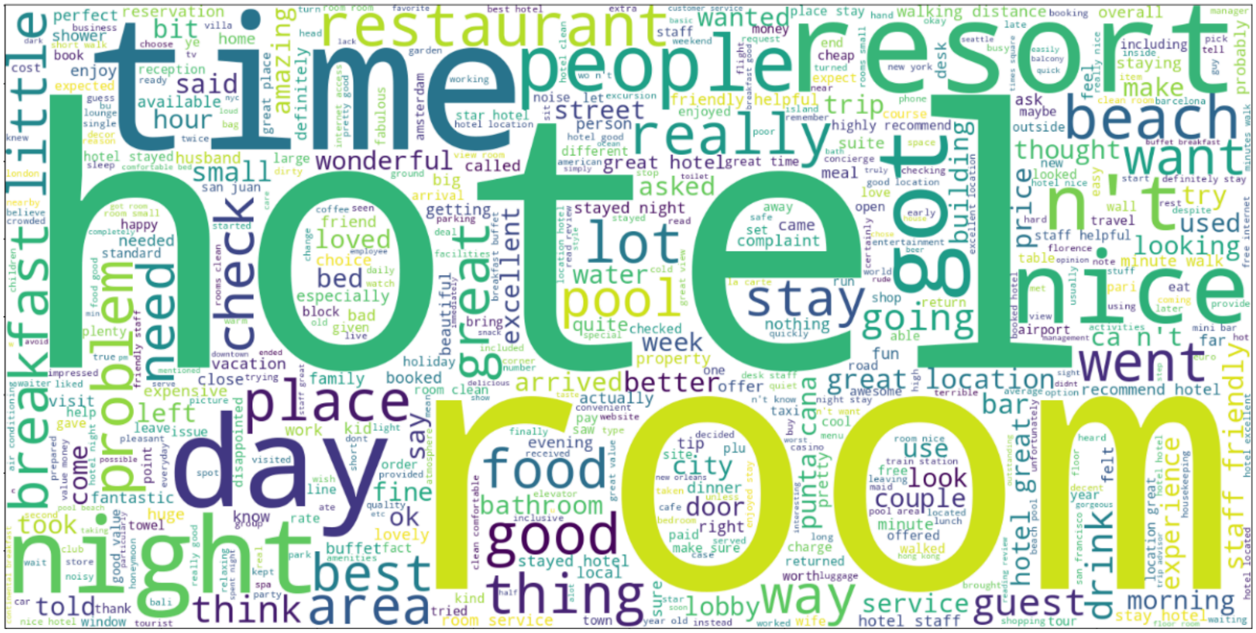
In this section, I will describe the details of the dataset. This is important to have visualization of the dataset. Based on this, we will have the solution for the prediction task.

The reviews appear to be somewhat preprocessed:

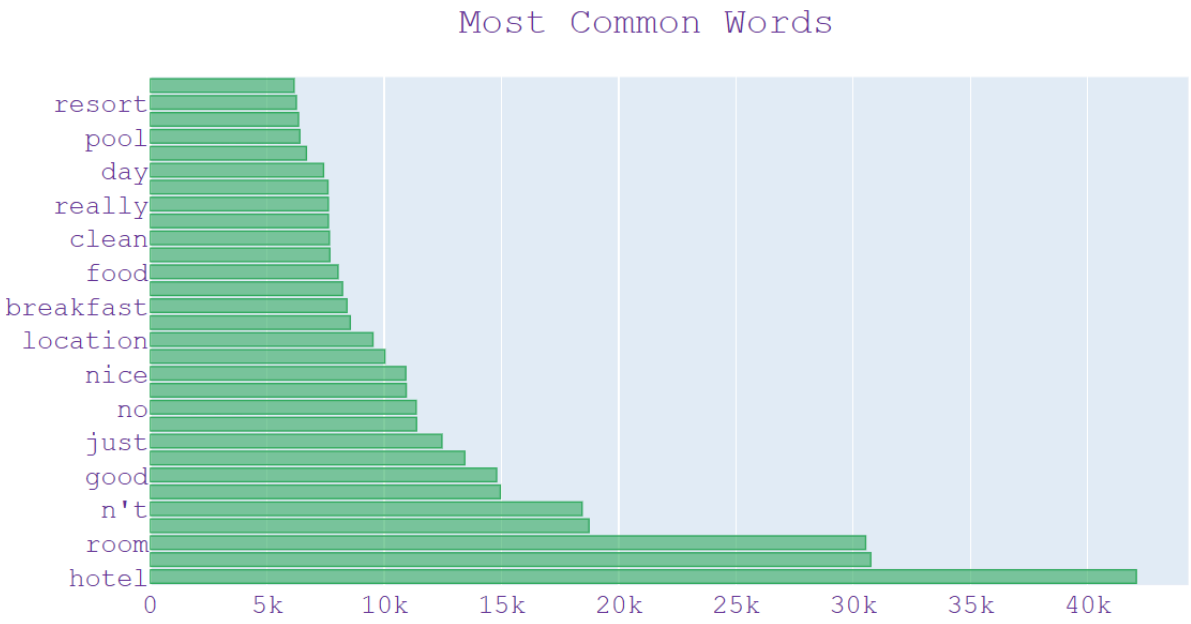
- There are no capital letters or full stops, but there is other punctuation like commas and apostrophes.
- Sentences don't flow very well because stop words have been removed.
- Words have clearly been tokenized, because "n't" appears in isolation a lot.
- There are many plurals, suggesting that words were not converted to singular.
- Some words weren't tokenized correctly due to missing spaces.
- There are some misspellings in the reviews.
- The reviews seem to all end with commas, suggesting that full stops have been converted to commas.
- The ratings are integer from 1 to 5.
- There are 20491 data points, the dataset has no null data.



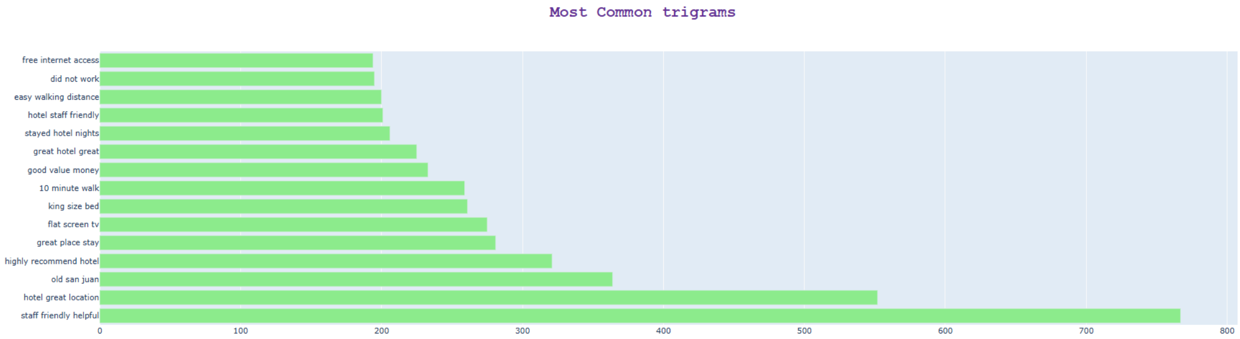
This is the distribution of Ratings in the dataset. Based on the chart, we can clearly see that the ratings are unevenly distributed, making the model harder to learn because of its imbalance. Most of the reviews are positive (4 or 5 score). Although all the reviews are English, there are some odd characters in reviews that makes encoding error.



We also have a word cloud of reviews. Because all the reviews are about hotels, the most common words are about booking like ‘hotel’ or ‘room’. This is described better in the figure below.



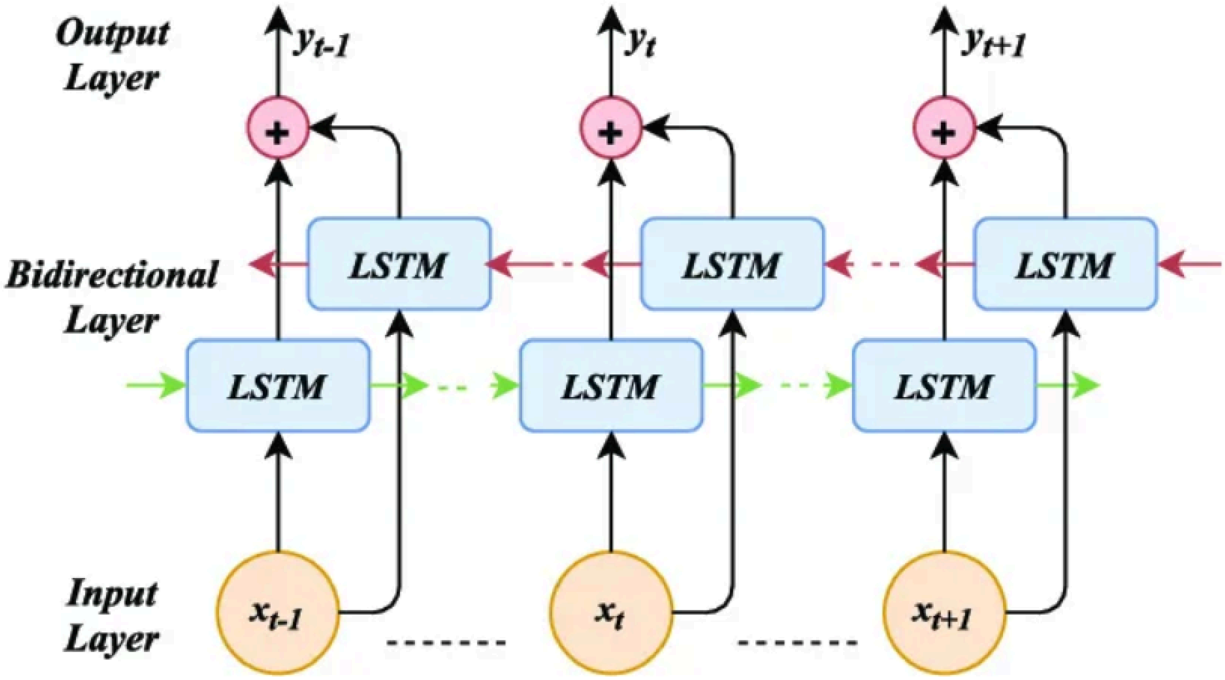
We also have the most common trigrams of the review. Because the majority of reviews are positive, the most common of trigrams are positive.



III. Methodology

1, Classifier technique

There is much previous work to process text for prediction. Most of them focus on summarizing the sequence by the sentiment, or using topic models on text. But for long sequences, Bidirectional LSTM (Bi-LSTM) is a good choice for tasks such as sentiment analysis and detection. The benefit of Bi-LSTM is that it captures not only the context that comes before a specific time step (as in traditional RNNs) but also the context that follows. In this work, the ability of Bi-LSTM help the model understand the entire semantic of the review, thereby predicting the corresponding score.



This is the architecture of Bi-LSTM. By using two LSTM layers: one processing the input sequence in the forward direction and the other in the backward direction, it can process the whole sequence without missing connection between words.

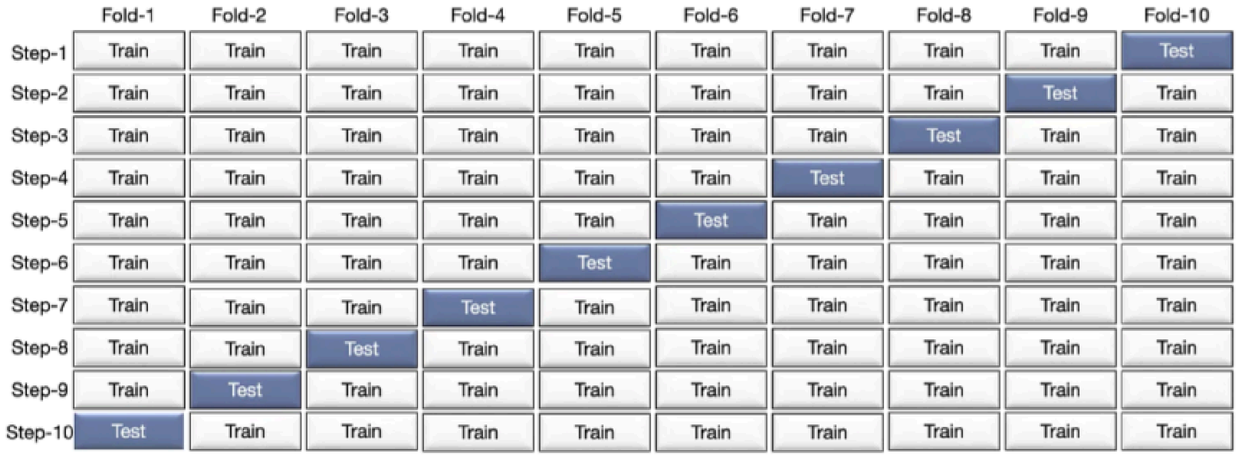
2. Ten-fold cross-validation

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Cross-validation has two main-step: splitting data into fold and rotating the training among them. The splitting technique has the following properties:

- Each fold has approximately the same size.
- Data can be randomly selected in each fold.
- All folds are used to train except one in each step.

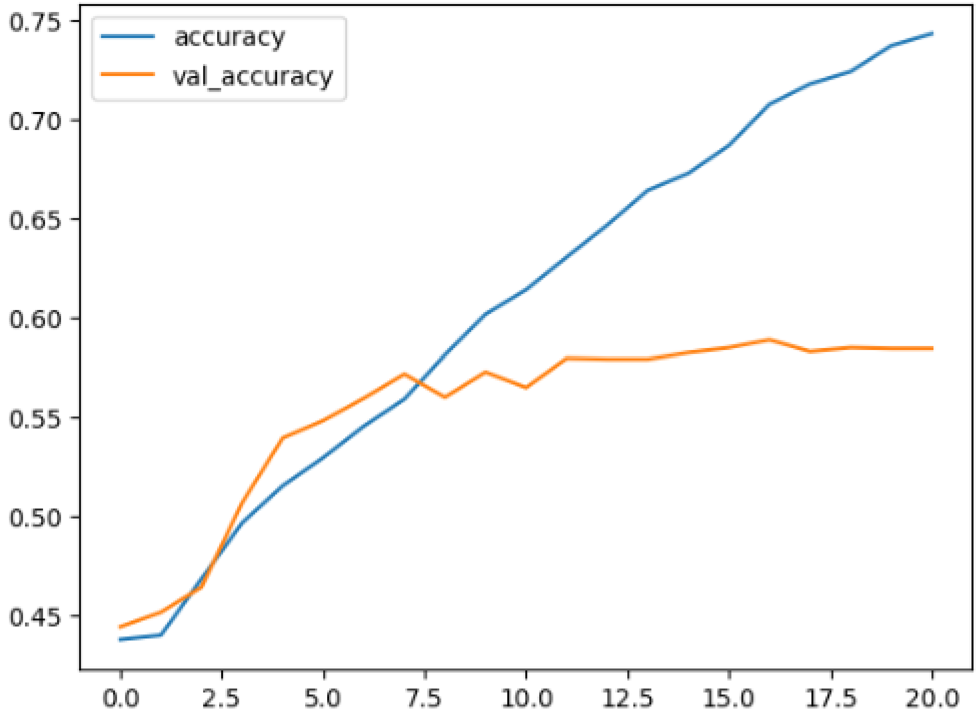
The technique is described below. In this project, I split the dataset into 10 fold and use separate fold in testing for each step. All other folds are used in training.



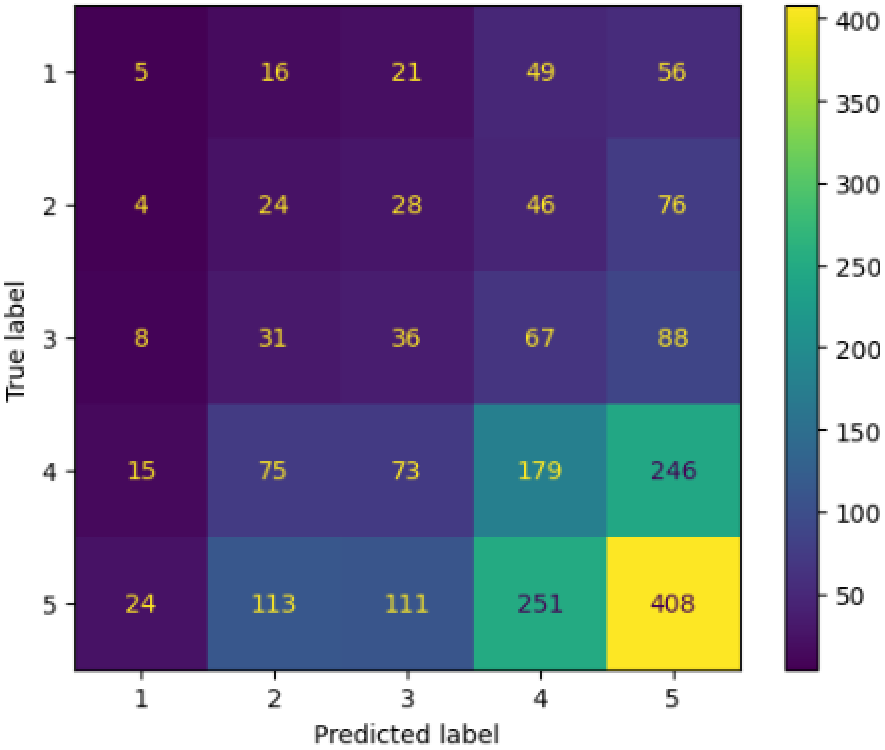
	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Fold-6	Fold-7	Fold-8	Fold-9	Fold-10
Step-1	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test
Step-2	Train	Train	Train	Train	Train	Train	Train	Train	Test	Train
Step-3	Train	Train	Train	Train	Train	Train	Train	Test	Train	Train
Step-4	Train	Train	Train	Train	Train	Train	Test	Train	Train	Train
Step-5	Train	Train	Train	Train	Train	Test	Train	Train	Train	Train
Step-6	Train	Train	Train	Train	Test	Train	Train	Train	Train	Train
Step-7	Train	Train	Train	Test	Train	Train	Train	Train	Train	Train
Step-8	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
Step-9	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
Step-10	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train

IV. Experimental

1, Training with Bi-LSTM



The dataset is splitted into a 9/1 ratio of training/validation. The highest accuracy result is 58.91% at epoch 17. After that, the accuracy slowly reduces, showing that the model is overfit with training data.



The figure above shows the confusion matrix of models after training. Based on the figure, we found that most incorrect predictions were relevant with label '5', which is the majority in the dataset.

2, Ten-fold cross-validation

The table below describe the accuracy of each step

Step	1	2	3	4	5	6	7	8	9	10
Acc(%)	57.8	56.3	55.7	57.7	57.9	58.1	57.9	58.1	56.9	59.3

The average accuracy of all steps is 57.44%.

Although the average accuracy is lower than the result when we train only in 1 step, there are still benefits and drawbacks of this method.

The benefit of method:

- It used all data for training and testing.
- It provides multiple results to make sure that the training is consistent.
- It can help us in fine-tuning parameters for models.

The drawback of method:

- It cost more time and resource than normal training
- Some folds may have too many outliers, making the accuracy of the following step worse.

V, Conclude

In this project, I completed the topic 'Detection of rating based on TripAdvisor reviews'. During the implementation, I understood more about how to process and classify text. Based on that understanding, I built a model based on Bi-LSTM with 58.91% accuracy on the dataset. I also apply ten-fold cross validation to have a more comprehensive view of the data.

In the future, I will develop this project in two different ways: finding another backbone for the model to have better accuracy and finding a way to deal with the imbalance of the dataset for this task.

VI. Reference

Source code: https://github.com/maiductrung99/big-data-project

1. DataMining-ch5 (guidetodatamining.com)

2. https://towardsdatascience.com, 5 Reasons why you should use Cross-Validation in your Data Science Projects

3. Kaggle, Trip Advisor Hotel (https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews)