AiPalette
252 North Bridge Rd,
Singapore 179103

Open Table is a restaurant review platform that helps you to discover great restaurants around 8 countries in the world with its database of over 50,000 restaurants. The platform has recently been pushing to expand into more countries in Asia and has set eyes on Thailand. Further, it has partnered up with a local data provider for restaurant data and has obtained a huge corpus of dataset which it wants to leverage in building a classification model that can identify the rating from review. This in turn will give them another revenue line because of the analytics they will provide to the restaurants onboarded in Thailand.

Dataset:

The data attached with the email consists of three files - training data, test data and sample submission.

1. Each row in training data corresponding to a record contains a review along with a rating score that ranges from 1 to 5 stars. It contains close to 40,000 reviews.
2. Test data consists of review ID and a review. We need to predict rating for the review. The test data has close to 6,000 reviews.
3. Final submission data csv file should have 2 columns - ID (corresponding to review in test data) and its predicted rating

For opening up of files in Thai, you may use Google Sheets as otherwise on local machine, it might not show the Thai characters properly.

Key features of the model:

A rating prediction model using only textual information. Pre-trained weights are ok but the prediction model has to be your own. A proper justification of the techniques used is required.

Output:

For measuring the output quality, Mean F1 score will be used. The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives to all actual positives (tp + fn). The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

The output will be a word doc where you describe your approach and a csv file with format same as sample submission.