Homework #2 – group based

There are numerous review websites with both ratings/scores and comments, which has resulted in many machine learning and deep learning models to use texts as inputs to predict the ratings/scores. However, regardless of how accurate of these models are, they cannot be directly used by firms. In fact, firms are more interested in knowing what factors cause/ contribute to the high or low ratings of a product or service.
Thus, this homework is an exploratory assignment where each team will:

1. collect data for a product or service reviews with ratings/scores. The data should have at least 10,000 rows, and the more the better.

2. build deep learning models such as LSTM/BERT to get reasonable accuracy of the rating/score prediction

3. More importantly, find factors that can explain what cause a high score and what cause a low score. Factors need to be more specific and can be useful. As an example, sentiment scores have been used to predict ratings. However, sentiment scores do not provide much meaningful values to the firm as it is quite intuitive to say positive sentiment leads to high ratings. As an example of useful score, a comment mentioned sunroof of an auto is specific and useful.

Before the team starts to collect data, please reach to Dr. Xia to get evaluation whether the website/data are fit for the assignment.