2024 Data Mining

HW 2

Task introduction

- Product review rating prediction
  - Rating is from 1 to 5 (5 classes classification task)
  - **Model selection: Only DNN, LSTM, or BERT are allowed to be used**
    - Using **models** other than those listed above will result in **0 points**
    - BERT will be introduced in later class
  - **You can use any visualization tool (e.g. matplotlib, Seaborn), along with data/text processing packages(e.g. pandas, csv, regex, NLTK) to do this assignment**
  - **You can also use transformers package to load pretrained BERT models from huggingface**

- Requirement
  - Upload your submission to Kaggle
  - Submit a report and your source code to E3

- Deadine is 5/14 (Tue.) 23:59, no late submission

Dataset Columns

- rating - Consumer ratings of this product (the label that you need to predict)
  - The actual ratings in this dataset are whole numbers without decimal points, and ratings should be outputted directly as integers ranging from 1 to 5
- title - title of the review
- text - content of the review
- verified_purchase - A review is verified_purchase if
  - Bought the item on this site
  - Paid a price available to most shoppers.
- helpful_vote - the number of people that found this review helpful

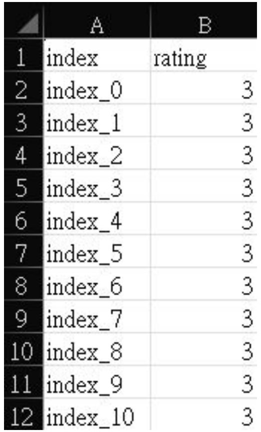
Dataset Download

Reviews dataset

- train.json
  - link

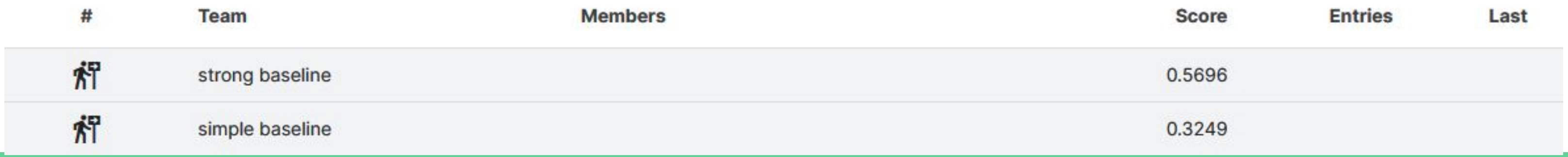
- test.json
  - link

Kaggle Submission



	A	B
1	index	rating
2	index_0	3
3	index_1	3
4	index_2	3
5	index_3	3
6	index_4	3
7	index_5	3
8	index_6	3
9	index_7	3
10	index_8	3
11	index_9	3
12	index_10	3

- Kaggle link
- Display team name : <student ID>
  - There will be a deduction of 5 points for HW 2 if wrong team name
- Submission format
  - A 35001*2 .csv file
    - first row is for the column name and others are your result.
  - Column name must be index and rating
  - Check sample submission
- There is one simple baseline and one strong baseline to beat



#	Team	Members	Score	Entries	Last
	strong baseline		0.5696		
	simple baseline		0.3249		

Kaggle Submission

- The scoring metric is macro F1
- You can submit at most 5 times each day.
- You can choose 3 of the submissions to be considered for the private leaderboard, or will otherwise default to the best public scoring submissions
  You can only view your private leaderboard score after the competition has ended
- Public leaderboard is calculated with 50% of the test data, and private leaderboard is calculated with other 50% of the test data, so the final standings may be different
- Please tune your model parameters using your own validation set instead of adjusting parameters based on the public leaderboard. Otherwise, it's easy to overfit, leading to poor performance on the private leaderboard.

Report Submission

Answer the following 3 questions:

1. How do you select features for your model input, and what preprocessing did you perform to review text?
2. Please describe how you tokenize your data, calculate the distribution of tokenized sequence length of the dataset and explain how you determine the padding size
3. Please compare the impact of using different methods to prepare data for different rating categories

Please answer the questions in detail to receive full points for each question.

Grading policy

- Kaggle (70%)
  - 30% based on the public leaderboard score and 70% based on the private leaderboard score
  - Leaderboard score consists of basic score and ranking score
    - Basic score (55%):

      Over strong baseline : 55

      Over simple bassline : 40

      Under simple baseline : 25
    - Ranking score (15%):

      15-(15/N)*(ranking-1), N=numbers of people in the interval

- Report (30%)
  - 10 for each quesiton

E3 Submission

Submit your source code and report to E3 before 5/14 (Tue.) 23:59.

No late submission !

- Format
  - source code : HW2_<student ID>.py or HW2_<student ID>.ipynb
  - report : HW2_<student ID>.pdf

If you have any question about HW 2, please feel free to contact with TA : Yu-Cheng LIU through email liu2022113.cs11@nycu.edu.tw

Have Fun !

We're using AI instead of biased humans
What did you train the AI on?
What did you train the AI on?
imgflip.com