Northwestern University

MSiA-490 Fall 2020

Amazon Kindle Store Review Text Classification Project

Siqi Li

Project Topic

For this project, I will conduct the comparison of three multi-class text classification approaches (BERT, LSTM, fastText). Specifically, I will train each of the model on the training set (80% of data) and evaluate their performance on the test set (20% of data) using metrics such as accuracy, precision, recall, and F-1 score. The best model will then be productized to take review text(s) as input and the output corresponding predicted review score (1-5).

Dataset

The dataset chosen for this project is Kindle Store Review data from Amazon Review Data (2018).

This dataset is an updated version of the Amazon review dataset released in 2014. The raw dataset contains information such as reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category information, price, brand, and image features) for 5,722,988 reviews in Amazon Kindle Store.

Due to computational limitation, only the first 1,000,000 records are used in this text classification project. Each record has the overall review score (integer from 1 to 5), the review text (string), and the datetime (yyyy-mm-dd) on which the review was written.

Link to dataset: http://deepyeti.ucsd.edu/jianmo/amazon/index.html

Citation

- Justifying recommendations using distantly-labeled reviews and fined-grained aspects. Jianmo Ni, Jiacheng Li, Julian McAuley. Empirical Methods in Natural Language Processing (EMNLP), 2019.