Abdelalim Yakoub
Ashit Patel
Dr. Mandy Korpusik
Natural Language Processing
December 15, 2020

Amazing Reviews Rating (Milestone)

Dataset

The dataset used in this project is an amazon product data provided by a professor in UCSD, Dr. Julian McAuley. It contains millions of reviews spanning from 1996 to 2014; however, we only have access to subsets of this data. The subsets are classified into product genres, and the genre we used is the Clothing, Shoes, and Jewelry products. It contains around 300,000 reviews. and each review contains multiple data points, such as the reviewer-ID, reviewer-Name, review-Text, review-Date, overall, …etc, but we will only be using 2 of these data points for each review: the review-Text and overall (rating), where the review-Text is the X-data and overall (rating) is the Y-data.

Model

We implemented a logistic regression model, and started to work on the BERT model, and we will use the bert_demo.ipynb as a reference..

Preliminary Results

Rating	Precision	Recall	F1-Score
1	0.53	0.49	0.51
2	0.36	0.15	0.21
3	0.41	0.28	0.33
4	0.50	0.26	0.34
5	0.76	0.95	0.84

The results show that the model is not doing very well, and that it is biased towards the 5-STAR rating.