Helpful Reviews

Final Report

Is a given review helpful?

I analyzed 8.9 million book reviews from 1996 to 2014 from Amazon Book Reviewers.

After cleaning the data and performing data analysis as summarized in the Milestone Report, I prepared the data for machine learning.

All reviews were compiled into a corpus using standard normalization techniques that included using lowercase letters and eliminating special characters, stop words and extra white spaces.

Individual reviews were converted into a sparse matrix using the standard bag of words technique with CountVectorizer and TfidfVectorizer. I iterated over various ngrams and consistently obtained the best results with CountVectorizer(ngram_range=(1,2)). (This includes all 1 and 2-word combinations.)

With all reviews converted into numerical matrices, I focused on the target column, the Helpful Rating. I improved the rating by developing a skewed_to_linear piecewise function that results in a ranking system for any dataset that keeps skewness intact.

To create a binary metric, I split the Helpful Rating column into two numbers as follows.

0 : unhelpful , < 50%,
1 : helpful > 85%.

Cutting out the middle is justifiable because it's unclear whether these reviews are helpful or not. On the contrary, scores of under 50% and over 85% were clearly unhelpful and helpful.

I tried Logistic Regression, Naive Bayes, Decision Trees, Random Forests, and Deep Learning techniques with a range of hyperparameters and cross-validation.

Logistic Regression consistently performed best with 90% AUC Means. The Confusion Matrix revealed precision of 80+% 0 classifiers and 90+% 1 classifiers. The hold-out test set scored 88%.

While Naive Bayes, Decision Trees and Random Forests consistently underperformed, Deep Learning is a leading candidate for further research. With more data available, and more features that may be engineered, Deep Learning is expected to produce better results.

In summary, it is possible to predict whether given reviews are helpful with almost 90% accuracy. Further research and development should only improve results.

Corey Wade
September 2018
https://github.com/coreyjwade/Helpful_Reviews