Helpful Reviews

Final Report

Is a given review helpful?

The CJW Independent Review Team analyzed 8.9 million book reviews from 1996 to 2014 from reviewers and/or books with at least 5 reviews.

After cleaning the data and performing data analysis as summarized in our previous report, we prepared the data for machine learning.

All were 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. We iterated over various ngrams and consistently obtained the best results with CountVectorizer(ngram_range=(1,2)). (Note: this includes all 1 and 2-word combinations and is slightly more computationally expensive.)

With all reviews converted into matrices, we focused on the target column, the Helpful Rating. We modified the rating by developing a skewed_to_linear piecewise function that results in a ranking system for any dataset that keeps skewness intact. Development of the skewed_to_linear piecewise function was essential so that it may be repeated elsewhere.

In order to simplify our results, we split the Helpful Rating into two classifiers. 0: unhelpful, <50%, and 1: helpful >80%. The middle data was cut out.

Cutting out the middle data is justifiable because it's unclear whether these reviews are helpful or not. On the contrary, scores of under 50% and over 80% are clearly unhelpful and helpful. T

We tried Logistic Regression, Naive Bayes, Decision Trees, Random Forest, and Deep Learning techniques with a range of parameters.

Logistic Regression consistently performed best with 90% AUC Means that included precision of 80+% 0 classifiers and 90+% 1 classifiers.

Naive Bayes, Decision Trees and Random Forests consistently underperformed, with only Naive Bayes consistently scoring in the 80th percentile. These algorithms may discarded going forward.

Deep Learning, however, 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, we can predict whether reviews given reviews are helpful with approximately 90% accuracy. Further research and development is warranted to improve results.

Corey Wade
September 1, 2018
https://github.com/coreyjwade/Helpful_Reviews