Helpful Reviews

Milestone Report

Amazon has been under scrutiny lately for false reviews and fraudulent products. I am working to develop machine learning algorithms to determine the reliability of reviews and reviewers.

I have finished stage three of the following four-step process:

1) Wrangle dataset
2) Create new metric
3) Analyze data
4) Apply machine learning techniques

I wrangled an 8.9 million row dataset of Amazon Book Reviews (Julian McCauley, UCSD). The original data was fairly clean. In my 15,000 sample, one review was eliminated due to the number of helpful reviews exceeding the number of total reviews. Two additional reviews were eliminated because there was no text in their reviews. All reviews with no votes were also eliminated.

I developed a new target column, Helpful Rating, weighted primarily by the percentage of helpful reviews and the total number of helpful reviews. The goal of the project is to predict Helpful Ratings based on the column containing the text of the review.

Exploratory data analysis was conducted on all columns to detect patterns within the data. Additional columns were added including Review Length, Average Sentence Length, and Average Word Length.

My initial findings are as follows:

1) The median review is 5.0 and the mean review is 4.25.
2) Users are more likely to click (thumbs up or thumbs down) if they like a given review.
3) Review Length, Average Sentence Length, and Average Word Length were all positively correlated with Helpful Rating near the 20th percentile.
4) There are a disproportionate number of reviews with Helpful Votes and Total Votes close to 0. Theses distributions are strongly right skewed.
5) The new Helpful Rating metric may be applied to any thumbs-up/thumbs-down system.

These preprocessing steps may be utilized for other datasets including other Amazon products, book review sites, and any dataset that includes reviews. Furthermore, the

new Helpful Rating metric work can be of value to any company interested in transforming text or vote counts into meaningful percentile rankings.

Corey J Wade
May 2018