Project Progress Report

Sentiment analysis on Amazon reviews for different product categories

Yasamin Tabatabaee (syt3)
Team name: AmazonSentiment

The main goal of this project is to build a tool for sentiment analysis of Amazon product reviews and perform sentiment analysis on a large collection of reviews from different categories of products, such as clothes, appliances, software, etc. Other goals are to get insight about various factors that correlate with the sentiment of a review, such as the type of words/phrases used, length of the reviews, etc, and also compare different machine learning classifiers for sentiment analysis on this data.

1) Which tasks have been completed?

In the project proposal, I described four main tasks for the project, which includes 1) data cleaning and text preprocessing 2) data analysis 3) training and evaluating machine learning classifiers and 4) designing the final sentiment analysis software. Up to this point, I have completed steps 1 and 2, and parts of steps 3 and 4, but only for a single product category (reviews related to Fashion products). For step 3, I have only used a Random Forest classifier until now. For step 4, I have implemented the base of command-line software that takes a review as input and returns its sentiment (or a probability distribution over its rating), but the code is still preliminary and needs to be extended in several ways.

2) Which tasks are pending?

First of all, the original dataset I was planning to use had more than 233 million Amazon reviews, but until now I have only used a collection of reviews related to Fashion products with about 800K reviews. The set used for training the classifier is even smaller (100K reviews). Therefore, one main extension in the rest of the project would be using data from other product categories, and potentially training classifiers on larger datasets. This of course depends on the capacity of the computational resources, but even on this relatively small dataset, I got an accuracy of about 81% with a Random Forest classifier, that is reasonably good, and perhaps training on larger datasets is not necessary.

For part 3, I plan to study other classifiers (e.g. SVM, Naive Bayes, potentially deep models etc) to find out which is better on the data and then pick the best performing classifier for the software in part 4. I have implemented a basic code base for part 4, but it still needs additional work to be fully functioning.

3) Are you facing any challenges?

One challenge is the unbalanced nature of the review data, that is, for most product categories, the number of products with positive reviews is much larger than the number of products with negative reviews and the number of neutral reviews is even lower. This leads to the trained classifiers not seeing enough data points for neutral/negative reviews and potentially becoming biased. To reduce this bias, I am planning to try resampling techniques, i.e. draw repetitive samples from underrepresented categories to make the training data more balanced and see if that potentially improves the accuracy of the classifier.

Another challenge is the large size of the review datasets (more than 200 million reviews), and the limited computational resources. I am currently using Google colab's free plan for training classifiers and data analysis, which has a limit of 12GB of RAM which allows for handling small subsets of the data at once (less than 1 million reviews). I might explore other resources, such as the UIUC campus cluster, or just limit the data used for training the classifiers to a small subset of reviews and analyze different product categories separately.