Sentiment Analysis on Amazon Baby Products Reviews – Milestone Report

Sentiment is a feeling or an emotion triggered by certain action or it could be an opinion/perception about some object/person/event/activity preempted by these feelings.
Internet is a resourceful place with respect to sentiment information. From a user's perspective people are able to post their own content through various social media sites, express their opinion on forums, blogs or online social networking sites and publish their views about certain product/services which can be made available to general public as well as to the companies.

With the advent of online marketing, companies like Amazon, Pandora, Apple, and Yelp have created a massive online user base. These companies rely heavily on the online content generated by the consumers of their product and services via a medium of "reviews".
Amazon is no new or small name. With the online retail industry on a rise, Amazon has established itself as a key player in both domestic and international markets. And to contribute to its continued success, "User reviews" have become proven sales drivers for the company for sure.

Through this project, we would like to study sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis.

Being an Amazon fan and a frequent Amazon user, I got interested in exploring the structure of a large database of Amazon reviews via this project and analyzing this information through effective visualization so as to be a smarter consumer as well as reviewer. Further to my analysis, I would also like to apply some machine learning concepts and methods to find the predictive strength in the helpfulness of each review based on previously collected data. For the sake of simplicity I am going to limit my analysis to only Amazon baby dataset.

We would try to find answers for the following questions.

Is there a relationship between the text and the sentiment of a review?

What is the distribution of the ratings in "baby category?"

What is a common theme in this dataset "more positive" or "more negative"?

What exactly makes a review helpful?

Is there any relationship between the length of the review and its helpfulness?

What important fields and information does the data set have?

This database contains 19 different features:

reviewer ID, asin, reviewerName , helpful, helpful_num, helpful_den, reviewText, overall, summary, unixReviewTime, reviewTime, exclamationcount, questioncount, charcount, wordcount, capcount, avgrating, diffrating, ishelpful.

What kind of cleaning and wrangling did you need to do?

I used a data of over 59,000 reviews of Amazon Baby Products. This dataset was fairly clean. For this project I was interested in only positive and negative reviews, hence eliminated all the records pertaining to "neutral ratings" (Overall = 3). I clubbed Overall ratings = 1 and 2 together and labeled them as negative review in "Sentiment" column. I combined Overall rating = 4 and 5 as positive sentiment.

For building predictive models, here are the other cleaning steps I performed.

Null Removal

We first checked the overall ratings with the presence of NaN in their review columns. This dataset was fairly clean and we dint have many Nan values. So we dropped a few observation with NaN reviews/ Nan overall ratings.

Word Normalization

Word Normalization is the process of reduction of each word to its stem form (by chopping of the affixes). While doing this, we converted the text to lower case. We also got rid of apostrophes (') and (-). We also removed stopwords ("The", "this" etc.)

Bag of words/Tokenization

We converted each review into a vector that machine learning models can understand. This comprised of the following steps:-

1. Generating term frequency: - counting how many times does a word occur in each document.
2. Generating Inverse Document Frequency: - weighting the counts, so that frequent tokens get lower weight.
3. Normalizing the vectors to unit length.

Are there other datasets you can find, use and combine with, to answer the questions that matter?

If time permits I plan to incorporate product details dataset and combine with the reviews dataset so as to explore relationship between product and reviewers. Adding geographical details would also help to understand the trend and identify the sentiment theme prevalent in certain region or certain state. Time series analysis would be a good addition as well to see the graph of reviews for products over time.

Any preliminary exploration you've performed and your initial findings.

- In general positive reviews are common in this dataset.
- We have 50 % of the total reviews assigned as 5 -star.
- Best reviews (5-star) are relatively shorter.
- Longer reviews are more helpful.
- Frequent reviewers write longer and helpful reviews.