SENTIMENT ANALYSIS ON AMAZON BABY PRODUCTS

Springboard Capstone Project

SEPTEMBER 6, 2017

PREPARED BY PRERNA SAXENA

Index

Introduction

NLP stands for Neuro-Linguistic Programming. Neuro refers to neurology; Linguistic is language; programming is knowing about how neural language functions or operates. In other words, learning NLP is like learning the language of our own mind!

NLP is a vast area of study and known to be one of the hardest problem in the field of computer science. It's a name given to the continuous effort and ongoing research for finding different ways to tune computers so that they can analyze, understand, derive meaning from human language in a smart and useful way.

Within NLP, there are many classification tasks. In classification tasks we construct a classification function which can give the correlation between a certain 'feature' D and a class C. These functions can be useful in many ways like *topic classification*, *spam filtering* and *sentiment analysis.* Through this project we would like to study sentiment analysis and how it can be employed in identifying and extracting contextual polarity of text documents.

Sentiment Analysis

What is sentiment? It's a feeling or emotion triggered by certain action or it could be an opinion/perception about some object/person/event/activity preempted by these feelings. That is why sentiment analysis is also frequently referred as opinion mining.

Dictionary.com defines sentiment analysis as:

"A process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral."

Internet is a resourceful place with respect to gathering sentiment information. From a user's perspective people are able to post their own content through various social media channels, express their opinion on forums, blogs, and on social networking sites. They can freely share their opinion and publish their views about certain product/service via various online channels which can be made available to general public. These reviews in turn can then influence

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".

Through this project, we would like to study sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. We would also like to examine the effectiveness of different machine learning techniques for classification of online reviews using models devised from a review corpus using various supervised learning methods.

Amazon Review

With the advent of internet, technological advancement and quick and easy access to savvy devices, buying and selling have become very easy. Online retail marketplaces have become a new norm and in fact have become one of the fastest growing sales channels in the US.

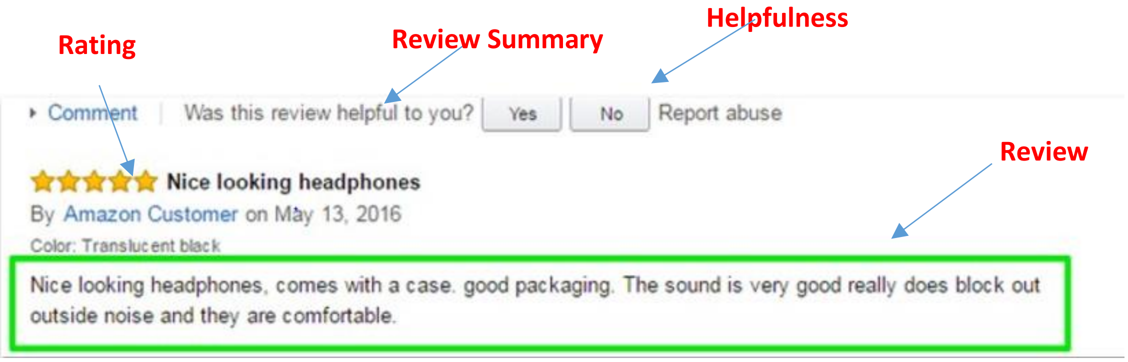
Amazon is no new or small name. With the online retail industry on a rise, Amazon has established itself as a key player in both domestic and international markets. And to contribute to its continued success, “User reviews” have become proven sales drivers for the company for sure. Just like me majority of customers often rely on them to make a purchase decision. *Research shows that online reviews not only help users make their purchasing decisions, but they also help ecommerce businesses obtain more customers*.

But there was one little flaw in the native review system until a little while ago. As much as these reviews were turning helpful for making purchasing decisions, yet, as the number of reviews kept on growing, users were faced with a pile of reviews to scan from for the most useful information they were hunting for. Users had to scroll down to find the most relevant one.

In order to combat this issue, Amazon team improved the display of the review by implementing a voting system along with review. This enables users to up-vote (like if it helped) a review when they believe that it is helpful and down-vote (unlike) a review when they feel like it is not helpful. Furthermore, they also introduced a new ranking system for the reviews based on helpfulness, meaning that the reviews with the high number up-votes will appear at the top, while those with the least up-votes will be displayed on the bottom. By adding this feature, Amazon made it easier for their customers to locate the most helpful reviews on each product whether or not the reviewer gave high ratings for the product.

According to an article published on the Business Insider in 2009, this voting feature contributes to more than 2.7 billion dollars of revenue for Amazon every year.

Properties of Amazon review.



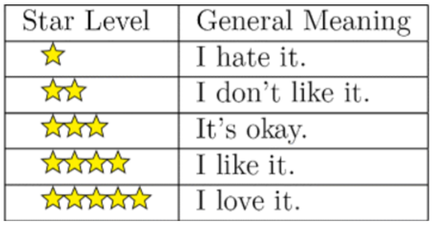
Was this review helpful to you?

★★★★★ **Nice looking headphones**

By Amazon Customer on May 13, 2016

Color: Translucent black

Nice looking headphones, comes with a case. good packaging. The sound is very good really does block out outside noise and they are comfortable.



Star Level	General Meaning
☆	I hate it.
☆☆	I don’t like it.
☆☆☆	It’s okay.
☆☆☆☆	I like it.
☆☆☆☆☆	I love it.

Problem formulation

Encouraged with the increased revenue just by adding this “voting feature” Amazon constantly look for improvements on their review system. Company often releases its historical data to the common public so that newer/better methods and techniques can be developed to map the helpfulness of the new reviews based on the data collected from past reviews and the number of up-votes that those reviews received in the past.

Being an Amazon fan and a frequent Amazon user, I would like to explore the structure of a large database of Amazon reviews via this project and analyze 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?

Who are frequent reviewers? Do they write more helpful reviews?

Amazon Dataset and features

I am going to use a data of over 59,000 reviews of Amazon Baby products that is available via this link here.

http://jmcauley.ucsd.edu/data/amazon/links.html

This database contains 19 different features:

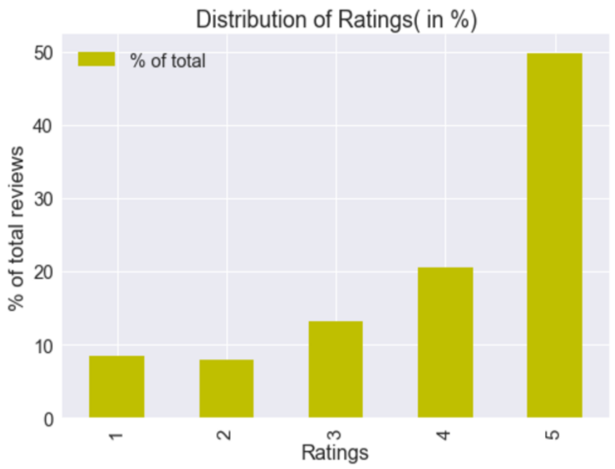
reviewer ID, asin, reviewerName , helpful, helpful_num, helpful_den, reviewText, overall, summary, unixReviewTime, reviewTime, exclamationcount, questioncount, charcount, wordcount, capcount, avgrating, diffrating, ishelpful.

Section -I

Exploratory Data Analysis

Distribution of ratings

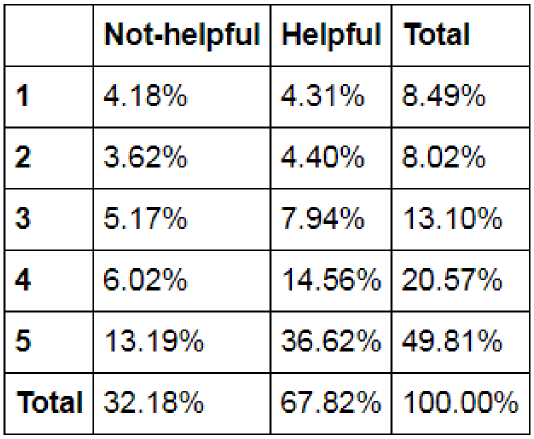
We were able to draw many conclusions from the data that we obtained. Let’s take a look at each of the graphs more closely. First I was interested in finding out the spread of ratings in this dataset.



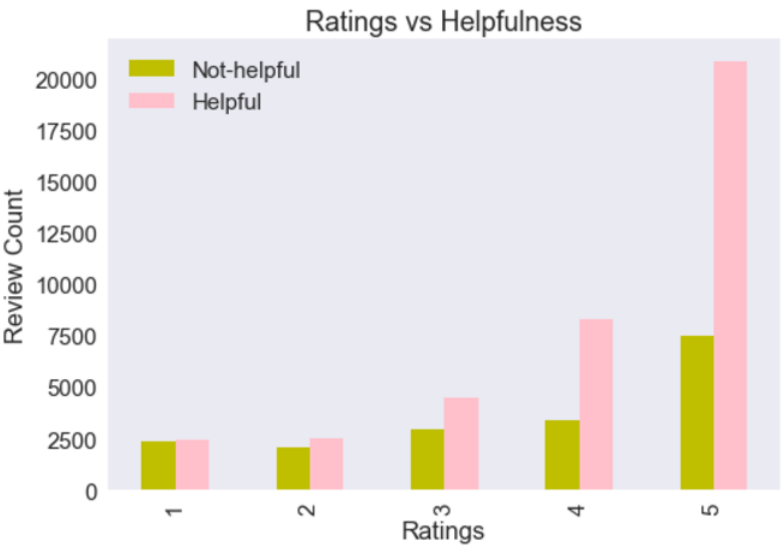
Looking at the distribution of ratings, we see that 5-star reviews constitute a large proportion (50%) of all reviews. The next most prevalent rating is 4-stars (21%), followed by 3-star (13%), 1-star (8.5%), and finally 2-star reviews (8.0%).

Overall Ratings vs Helpfulness

We looked at the percentage of those reviews that users found helpful or not helpful for each Star rating. And we notice that as the ratings increase, the reviews become more helpful. For 5-star reviews, 37% reviews were found helpful and 13% not helpful.



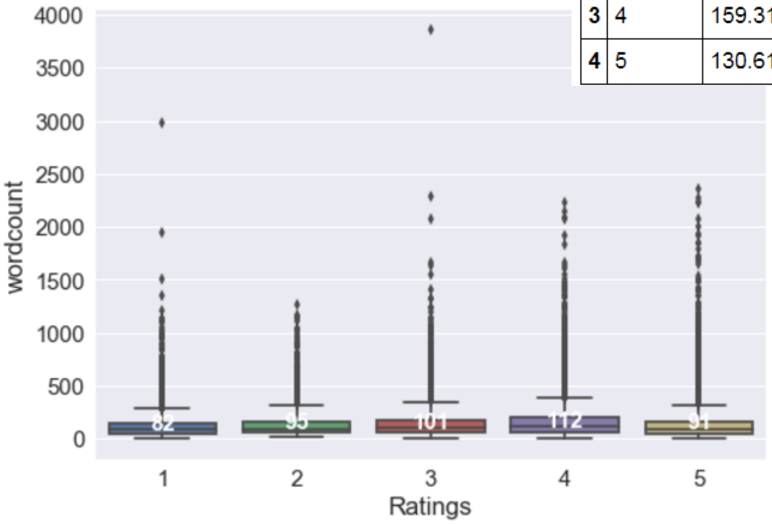
	Not-helpful	Helpful	Total
1	4.18%	4.31%	8.49%
2	3.62%	4.40%	8.02%
3	5.17%	7.94%	13.10%
4	6.02%	14.56%	20.57%
5	13.19%	36.62%	49.81%
Total	32.18%	67.82%	100.00%



How verbose are reviews?

I was interested in knowing the length of review per rating. Below is descriptive statistics I came up with. Top notch ratings that is 5 stars is the second lowest in this dataset.

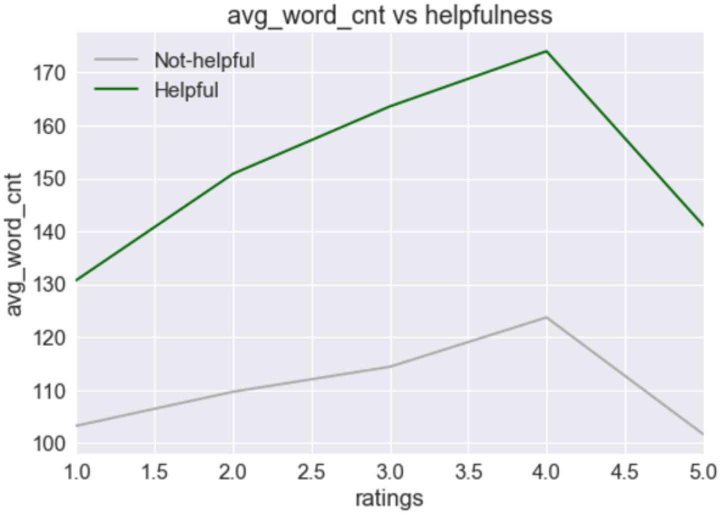
	Ratings	avg(wordcount)	min(wordcount)	max(wordcount)
0	1	117.148677	2	2978
1	2	132.210031	12	1262
2	3	144.172451	4	3855
3	4	159.316293	2	2232
4	5	130.617844	1	2352



Likewise, 5-star reviews have the second lowest median word count (91 words), while 3-star reviews have relatively higher median of about 101 words. So, we see that reviews for the 5 star ratings are relatively shorter in this dataset. 4 stars reviews are lengthy. And not surprising enough, but 1 and 2 stars are shorter too.

Average word count per review on helpfulness index

Word counts for helpful reviews and not helpful reviews have a similar distribution. However, not helpful reviews have a larger concentration of reviews with low word count and helpful reviews have longer reviews. Descriptive reviews are helpful in general.



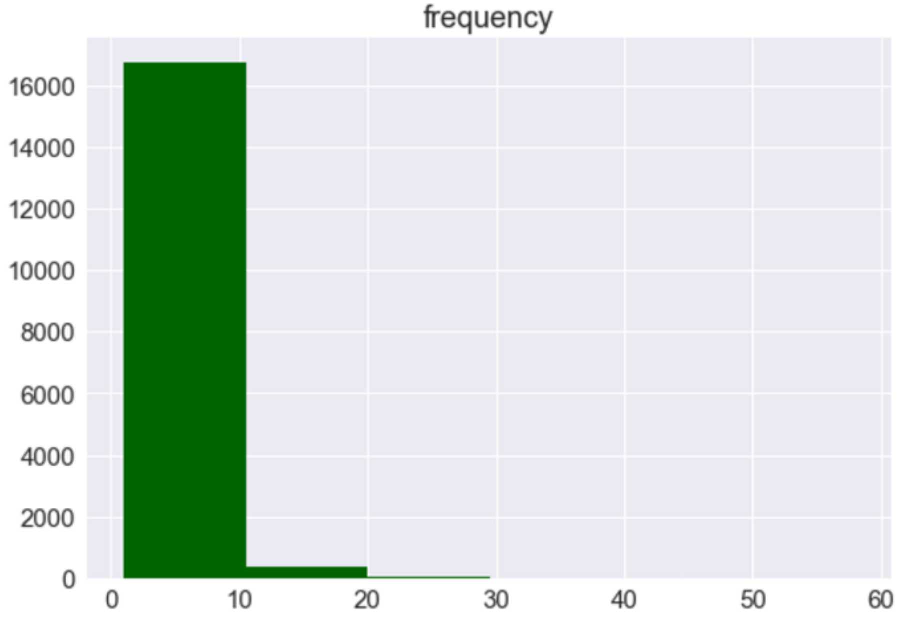
Regular vs Non-Regular Reviewers.

Next I was interested to draw a comparison between the reviews of regular vs not regular reviewers.

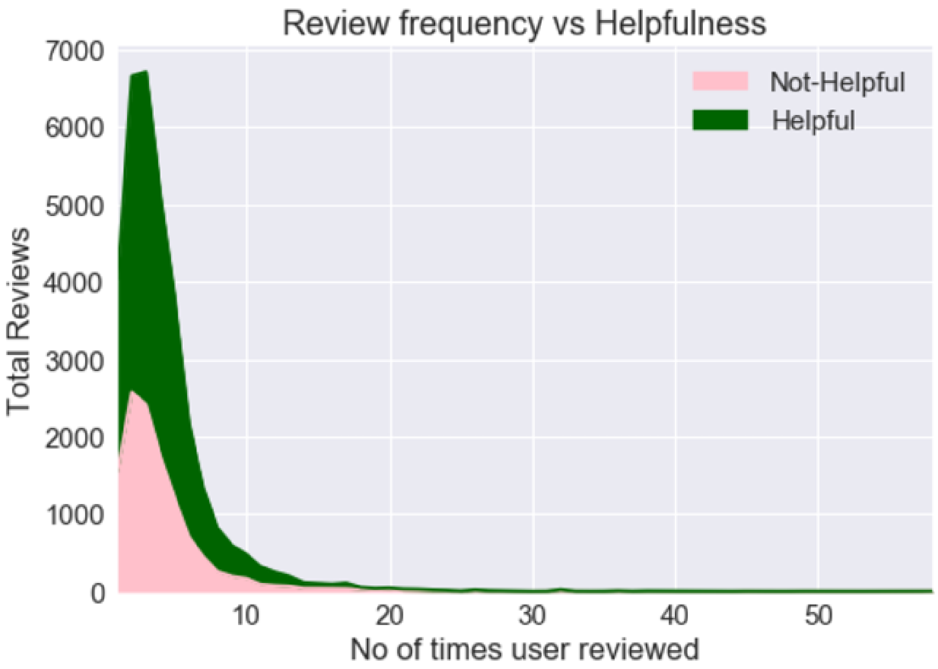
So who are regular reviewers? Our obvious choice would be those customers who have clearly reviewed more than once. The more the better! After analyzing the review counts, we found out that there is a good distinction of review counts between <3 reviews and more than 3 reviews/customer in the dataset. So, we assigned reviewers as frequent reviewers who have more than 3 reviews and vice-versa. The goal here is to identify if there is any behavioral distinction between frequent and not frequent reviewer groups.

In the below histogram, frequency is the number of reviews completed by a given customer on the website. We have quite a good concentration of reviewers in the dataset reviewing in the range of 1-10 times. After analyzing further we found out that:-

- We have 8977 Regular Customers with Review frequency > 2
- We have 8190 customers which are not so regular having frequency <= 2
- Majority of the reviews (7410) are done by customers who have reviewed at least 2 or 3 times.
- There are some outliers too in the 30 – 60 interval.



So, next I wanted to plot a graph in order to determine how these frequent users are impacting helpfulness. In contrast to my assumption I could not detect any striking pattern on the plot generated for Total Frequency of the reviews vs Helpfulness. However, we clearly see more helpful reviews for every range. Also, we notice more helpful reviews than not - useful ones for the users who have reviewed for about 2-8 times as opposed to higher frequency holders. Also, as the review frequency increases that is the number of times user gives review increases, so does the helpful index in general. So we can say more reviews are better!



Inferences from EDA:-

- 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.

Section II

Pre-Processing Data

On the basis of above exploratory analysis of the Amazon Baby product review dataset, we notice some of the characteristic features of the review have a good impact on the overall ratings (*) that they receive. In this section, I plan to employ a dataset of text reviews and corresponding products ratings assigned by each reviewer as labelled training data, and predict product sentiment on a test data set. I would also like to compare the predictive power of different machine learning approaches to text preprocessing employing Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression models.

Here are some of the pre-processing steps I followed.

Null Removal

We first checked the overall ratings with the presence of NaN in their review columns. This dataset was fairly clean and we din't have to deal with too 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, got rid of punctuations, apostrophes (') and (-). We also removed stopwords ("The", "this" etc.)

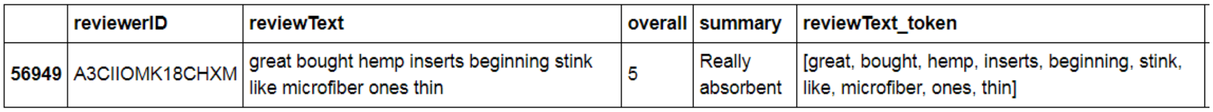
Bag of words/Tokenization

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.

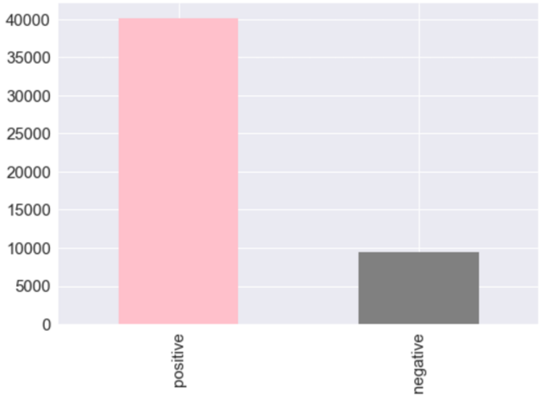
2. Generating Inverse Document Frequency: - weighting the counts, so that frequent tokens get lower weight.
3. normalizing the vectors to unit length

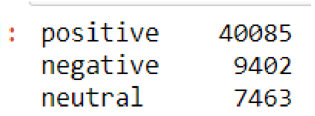
Here is a sample record.



	reviewerID	reviewText	overall	summary	reviewText_token
56949	A3CIIOMK18CHXM	great bought hemp inserts beginning stink like microfiber ones thin	5	Really absorbent	[great, bought, hemp, inserts, beginning, stink, like, microfiber, ones, thin]

Removing Neutral ratings:-



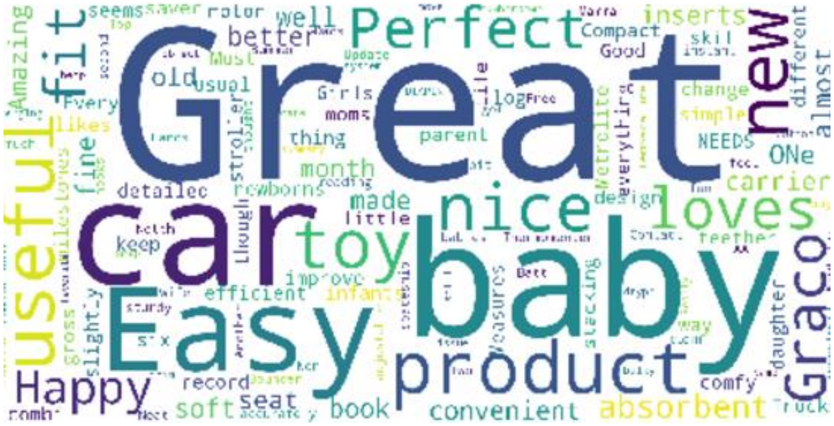


```
: positive    40085
  negative     9402
  neutral      7463
```

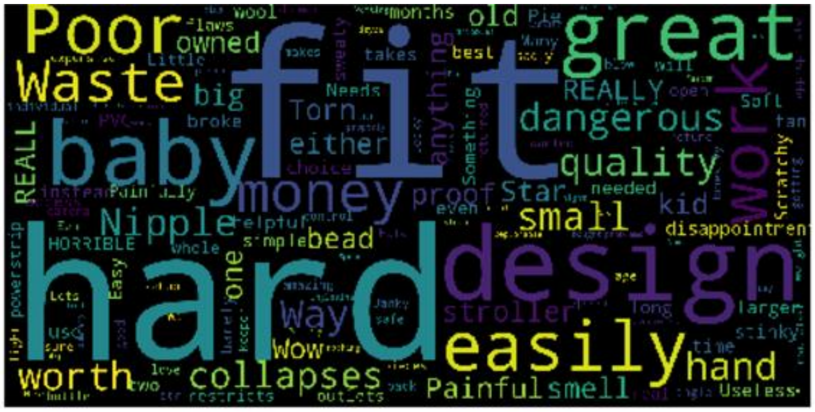
Further I also removed observations with neutral ratings to limit our analysis to binary classification. We reduced our dataset to 49,487 observations. We mapped overall rating to a new column ”Sentiment” and labeled 1 and 2 stars as “negative sentiment” and 4 and 5 star ratings as “positive sentiment”. Next, I was interested to see the word cloud for both the classes.

Word Cloud for positive and negative sentiments

Positive Sentiments: perfect, great, Easy, useful, happy, nice, amazing



Negative Sentiments: poor, waste, hard, painful, dangerous



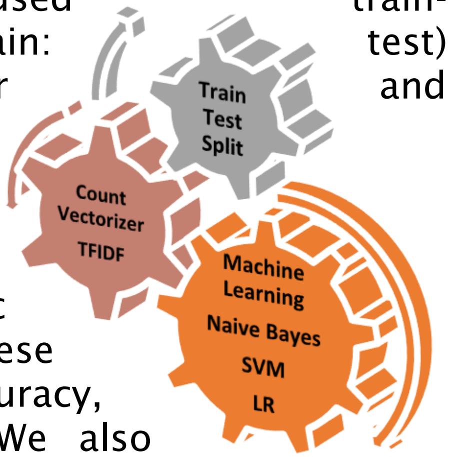
Above is the visual representation of the word frequencies in the positive and negative reviews displaying a direct relationship between the word frequency and the word size in the word cloud? The bigger the word in the cloud, higher is the frequency of the word used in the reviews. We used “Summary” field for the word cloud generation.

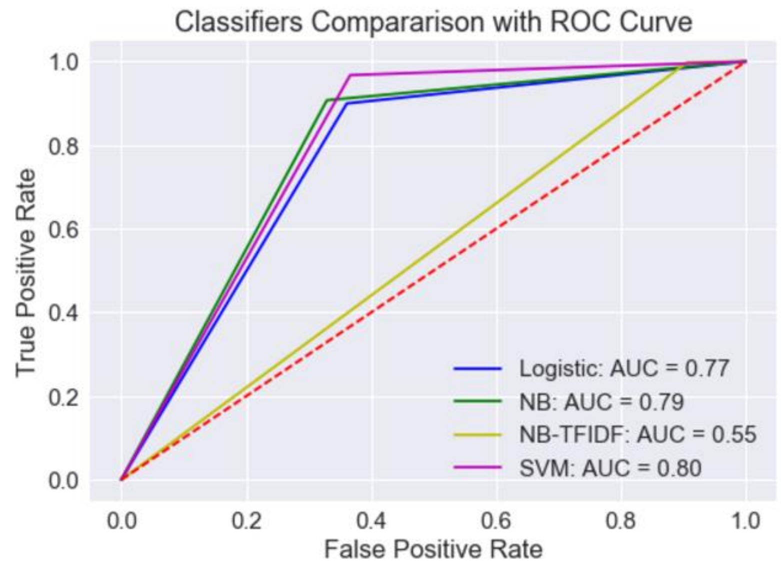
In general above visual helps us in identifying the general theme of the sentiments. However, there are a few words which were found in opposite clouds despite belonging to positive and negative groups. For example: words like “easily” and “quality” have positive elements yet they end up in negative cloud. Likewise, “great” shows up in positive as well as in negative cloud. This is an interesting finding and would need further investigation.

At this point, I had numeric training features created from the Bag of Words and the original sentiment labels for each feature vector as well, so next I was curious to see the predictive strength of these vectors.

Machine Learning

To build a predictive model, we used train-test-split to split the data in 7:3 (train: test) ratio. Then, we used count-vectorizer and tf-IDF transformer to convert the data into numerical features. Further, we employed various algorithms like Naïve Bayes, Support Vector Machine, and Logistic Regression and evaluated each of these models against **metrics** like accuracy, precision, recall, and F-measure. We also generated ROC for the mentioned classifiers and found out that SVM provided us the best results among all the classifiers with an AUC of about .80. Naïve Bayes without tf-IDF was very similar to SVM with linear kernel AUC = .79 which makes up an interesting finding given the small sample size of only 60000 records in our dataset. We got the lowest AUC from Logistic Regression = .77.





Conclusion

With the information we gathered by analyzing the characteristics features of reviews, we see there is a lot of value hidden in the text of the reviews. Amazon can come up with a marketing strategy to encourage users to leave longer reviews as they tend to be more useful in prompting purchasing decisions. Our machine learning model shows that it is possible to predict the rating level of a review with great accuracy just by analyzing the text.

Future Direction

Although, we got a decent score from all our algorithms across the board, I still see definite room for improvement in all. I would begin with investigating all the misclassified labels. I would also like to improve pre-processing techniques, some of which would include decoding HTML entities, eliminating numerals, adding more features etc. I would also like to test with n-grams.

I am also interested to incorporate product details in the 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.

Lastly, I would like to run my models with a larger dataset, ideally, with original McAuley's full dataset and check for accuracy with bigger sample size.

Learnings and Credits

In this project I was able to make use of a range of skills learned through Springboard's Data Science career track. The capstone project gave me a platform to apply the fundamental concepts in the real world scenario. Amazon's publicly available data set was very streamlined and clean to work with. I gained experience in exploratory analysis, data-munging, and data-analysis. I got to study most complex concept of NLP in much detail through this project and was able to touch upon the basics of text classification via several modelling techniques.

Code and the synopsis can be located in the following github link.
GitHub repo: https://github.com/DataGalore/Capstone

Special credit goes to my mentor Kenneth Kihara for the valuable suggestions, feedback and discussions.

References

www.amazon.com

www.sprinboard.com

http://ataspinar.com/2016/02/01/sentiment-analysis-with-bag-of-words-part-2/

https://radimrehurek.com/data_science_python/

https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words

https://snap.stanford.edu/data/amazon-meta.html

https://github.com/AubreyB13/sentiment-analysis/blob/master/Final_NLP_project.ipynb

https://www.kaggle.com/roopalik/amazon-baby-dataset

http://fileadmin.cs.lth.se/cs/education/edan70/LTProjects/2014/Reports/Wallin.pdf

https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

http://ataspinar.com/2015/11/16/text-classification-and-sentiment-analysis/

http://fastml.com/classifying-text-with-bag-of-words-a-tutorial/

https://www.youtube.com/watch?v=c3fnHA6yLeY&list=PL6397E4B26D00A269&index=24