NLP - Sentiment Analysis Project
Beauty Product Reviews

Engin Turkmen

March 5, 2019

Contents

1. INTRODUCTION

1.1. General:

Sentiment analysis ,which is a subtopics of Natural Language Processing (NLP), has been gradually becoming more and more popular. It is a contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

Sentiment Analysis has many applications ranging from ecommerce, marketing, to politics and any other research to tackle with text or unstructured text data. Companies, especially in e-commerce, also do sentiment analysis to collect and analyze customer feedback about their products. Besides that, potential customers prefer to review the opinions of existing customers before they purchase a product or use a service of a company. As seen here, there are two parts in e-commerce; one is the online retailer, which wants to maximize e-commerce sales or services, and the other is the consumers, who want to have the best product or service over alternatives.

1.2 Problem:

Multi-class classification on Women e-Trade Comments Problem

The largest internet retailer in the world as measured by revenue and market capitalization in Amazon.com, Inc. which is an American e-commerce and cloud computing company. In this project, Amazon is the client, and asked to predict customers' future inclination on the beauty products based on their data which includes reviews on their products between 2005-2014.

We will use sentiment analysis which is most general use in e-commerce activities. We will build a sentiment analysis model that predicts whether a user like a product or not, based on their review on Amazon. Our dataset consists of customers' reviews and ratings.

1.3 Data Set Description:

Beauty dataset revolving around the reviews written by customers. This is a real commercial data.

This data includes 28798 rows and 9 feature variables. Memory usage is 2.2+ MB.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A6VPK7X53QNAQ	B0000CC64W	AmazonDiva "Keep Calm and Carry On."	[5, 5]	I am a devotee to this serum, it does wonders ...	5.0	If I had to choose only one product to take ca...	1245283200	06 18, 2009
1	A3CHMHGSJSQ02J	B0000CC64W	Anon. A. Non	[2, 2]	As a woman nearing 50, I need all the help I c...	5.0	Makes my skin lovely and smooth	1358467200	01 18, 2013
2	A1V1EP514B5H7Y	B0000CC64W	asiana	[0, 0]	I've used this regenerating serum for more tha...	5.0	Works well at a reasonable price	1322524800	11 29, 2011
3	A1X2LENOF84LCQ	B0000CC64W	D "D"	[62, 75]	I have tried so many products to just be total...	4.0	This does work ladies	1113350400	04 13, 2005
4	A2PATWWZAXHQYA	B0000CC64W	Farnoosh Brock	[1, 1]	I love Oil of Olay. My primary moisturizer is ...	1.0	Did not like the feel/texture of this serum	1387584000	12 21, 2013

Each row corresponds to a customer review, and includes the variables:

reviewerID : ID of the reviewer, e.g. A2SUAM1J3GNN3B - type: object

asin : ID of the product , e.g. 0000013714 – type: object

reviewerName : name of the reviewer – type: object

helpful : helpfulness of the review, e.g. 2/3 – type: object

reviewText : text of the review – type: object

overall : Rating (1,2,3,4,5)– type: float64

summary : summary of the review – type: object

unixReviewTime : time of the review (unix time) – type: int64

reviewTime : time of the review (raw) – type: object

I have downloaded the beauty product review file via link below and opened the file with coding in Jupyter notebook.

Data Source:
http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_10.json.gz

2. DATA WRANGLING

2.1 Inspecting the Data Set:

```
# Basic information on Dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28798 entries, 0 to 28797
Data columns (total 9 columns):
reviewerID        28798 non-null object
asin              28798 non-null object
reviewerName      28576 non-null object
helpful           28798 non-null object
reviewText        28798 non-null object
overall           28798 non-null float64
summary           28798 non-null object
unixReviewTime    28798 non-null int64
reviewTime        28798 non-null object
dtypes: float64(1), int64(1), object(7)
memory usage: 3.4+ MB
```

Amazon beauty products data includes 28798 rows(observations) and 9 columns(feature variables) and its memory usage is 3.4+ MB. In the dataset, we have 7 object, 1 float64 and 1 int64 data types.

222 **'reviewerName'** information is missing in the dataset. Since customers don't give their identity, it may not be reliable to make an analysis on their reviews and ratings. I would prefer to drop the missing values from dataset since we have enough observations to conclude a prediction for sentiment analysis.

We concatenated **'reviewText'** and **'summary'** since both gave the approximately same type of information about product in text format, and later dropped both **'reviewText'** and **'summary'** columns.

'helpful' variable includes positive (thumbs up) and negative (thumbs down) feedback for reviews' information, it was divided into two separate columns: positive feedback was represented as **"pos_feedback"** , and negative feedback was represented as **"neg_feedback"**. Also, **'helpful'** variable was represented as an int64 type in the dataset via that step.

We classified the **'overall'** (ratings) as good and bad in order to make sentiment analysis. For that, we dropped observations whose overall values are equal to 3 since that rating group doesn't give an exact opinion about product whether it is good or bad.

In the dataset, **'reviewerID'** and **'reviwerName'** were used both for identification of customers. We dropped one of them from the dataset. Preferably, I dropped **'reviewerName'** since customer names were not standardized and there were lots of different style to represent them in it.

'unixReviewTime' was dropped since it has already been represented in **'reviewTime'** feature in a more understandable format. Also, **'reviewTime'** was converted to datetime data

type. We created a single **'year'** column to make analysis between other variables in the future work.

We renamed the columns in order to improve practicality/readability of coding :

reviewerID : **"customer"**

asin : **"product"**

reviewerName : column will be dropped.

helpful : positive feedback will be represented as **"pos_feedback"** , neutral feedback will be represented as **"neut_feedback"** and negative feedback will be represented as **"neg_feedback"**.

reviewText : This will be concatenated with "summary" and renamed as **"review_text"**

overall : **"rating"**

summary : it will be dropped after it is concatenated with **"reviewerText"**.

unixReviewTime : column will be dropped.

reviewTime : **"time"**

year : year (a new column will be created from **'time'** column)

2.2 Descriptive Statistics:

	rating	pos_feedback	neg_feedback	year
count	24509.000000	24509.000000	24509.000000	24509.000000
mean	4.364723	0.932392	0.307193	2012.798645
std	0.939641	6.810957	1.082227	1.377252
min	1.000000	0.000000	0.000000	2005.000000
25%	4.000000	0.000000	0.000000	2012.000000
50%	5.000000	0.000000	0.000000	2013.000000
75%	5.000000	1.000000	0.000000	2014.000000
max	5.000000	549.000000	52.000000	2014.000000

- Rating Status:

28576 customer gives ratings along with their reviews for beauty products and mean of the ratings is 4.36, which means that customers prefer to give high ratings for products. Standard deviation and percentiles also show that 1 and 2 ratings for products are rare. To be able to predict the ratings reasonably, we classified them as 'good' and 'bad' above.

According to the statistics on rating stars:

Customers	Rating
576 customers	1 star
1508 customers	2 stars
4067 customers	3 stars
8742 customers	4 stars
13683 customers	5 stars

As a result of our classification, we concluded that 22425 customers gave good ratings and 2084 customers gave bad ratings.

- Feedback Status

7073 customers totally agree with the given reviews and gave positive feedbacks. The mean 0.93 and standard deviation is 6.81 whereas the highest positive feedback number for reviews is 549.

4986 customers don’t agree with given reviews the them. They give negative feedbacks. The mean 0.3 and standard deviation is 1.1 whereas the highest negative feedback number for reviews is 52.

On the other hand, 12450 customers don't give either positive or negative feedbacks for reviews.

- Non-numeric variables statistics:

We have 1340 unique customers and 733 products in this dataset. Each customer averagely gives 18 reviews for products and on the other hand, there is averagely 33 reviews for each product in the website.

2.3 Preprocessing the Text:

Since, text is the most unstructured form of all the available data, various types of noise are present in it and the data is not readily analyzable without any pre-processing. The entire process of cleaning and standardization of text, making it noise-free and ready for analysis is known as text preprocessing. In this section, I apply the following text preprocessing respectively.

Removing HTML tags

We wrote a function to remove the HTML tags which typically does not add much value towards understanding and analyzing text.

Removing accented characters

We wrote a function to convert and standardize accented characters/letters into ASCII characters.

Expanding Contractions

We wrote a function to convert each contraction to its expanded, original form in order to help with text standardization.

Removing Special Characters

We used simple regular expressions(regexes) to remove special characters and symbols which are usually non-alphanumeric characters or even occasional numeric characters.

Lemmatization

We removed word affixes to get to the base form of a word, known as root word.

Removing stopwords

We wrote a function to remove stopwords which have little or no significance in the text.

Building a Text Normalizer

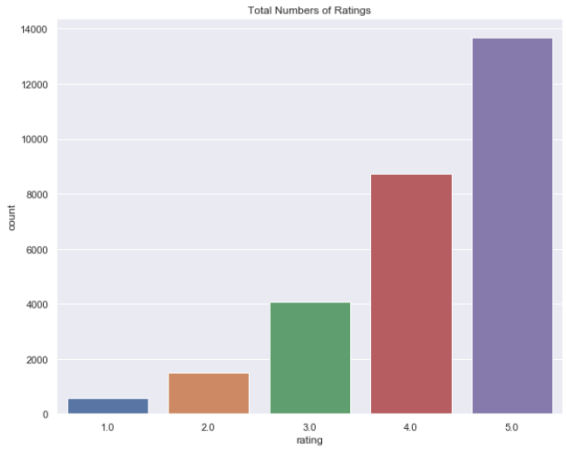
Based on the functions which we have written above and with additional text correction techniques (such as lowercase the text, and remove the extra newlines, white spaces, apostrophes), we built a text normalizer in order to help us to preprocess the new_text document.

After applying text normalizer to 'the review_text' document, we applied tokenizer to create tokens for the clean text. As a result of that, we had 1706537 words in total with a vocabulary size of 25023. Max review length is 1090 whereas min review length is 1 as a word based.

Eventually, after completing all data wrangling phases, we wrote the dataframe to csv file as a 'cleaned_dataset'.

3. DATA STORYTELLING

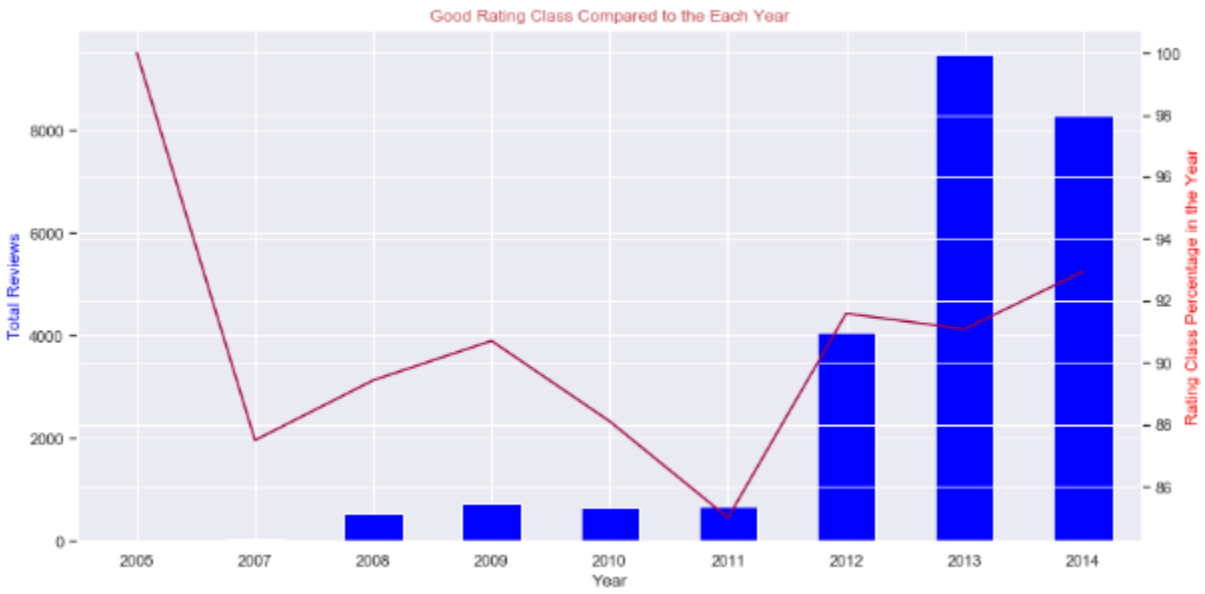
3.1 Target Variable (“rating_class”)

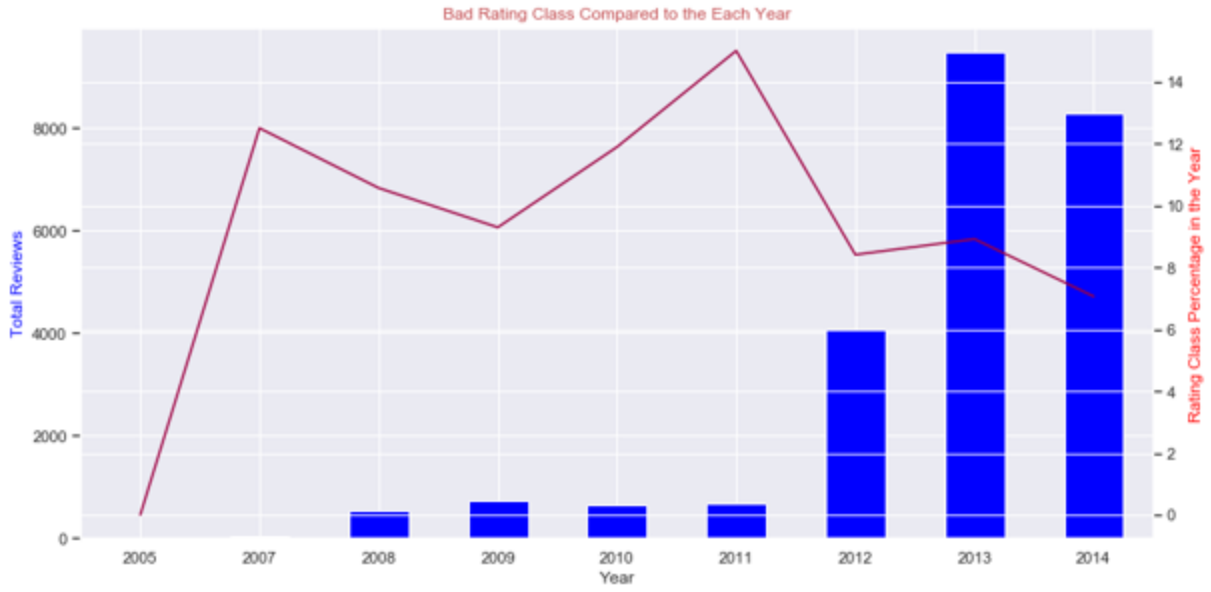


Ratings	Reviews
1 Star	576 Reviews
2 Stars	1508 Reviews
3 Stars	4067 Reviews
4 Stars	8742 Reviews
5 Stars	13683 Reviews

Rating Class	Reviews
Bad	2084 Reviews
Good	22425 Reviews

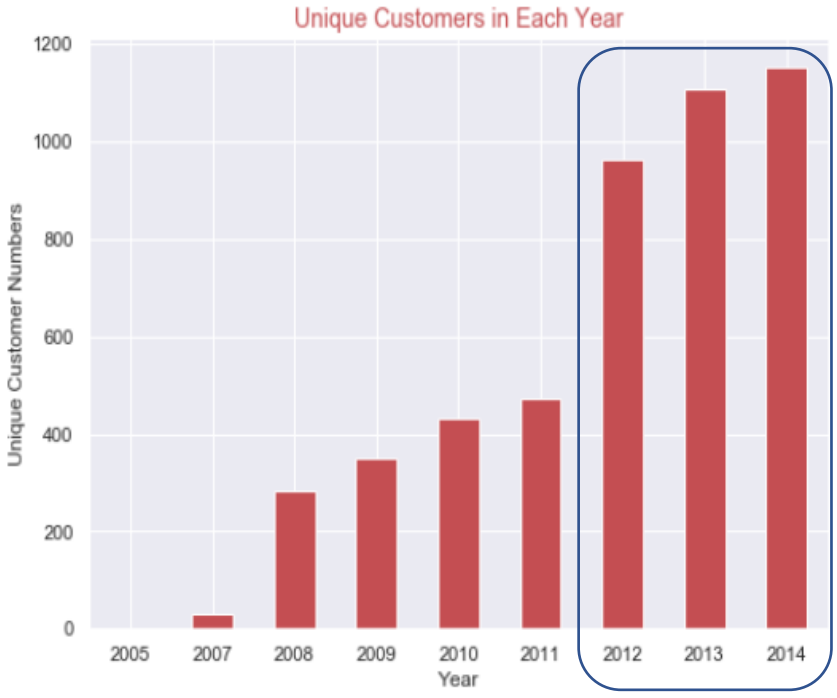
Customers gave rating, which ranged between 1 to 5, for each beauty product they bought in the Amazon online market. According to the statistics between 2005-04-13 and 2014-07-23, ratings shows us that customers were averagely satisfied with the products which they experienced. We diminished those 5 rating categories into two categories such as 'good' and 'bad' in order to implement a sentiment analysis. This is an imbalance between rating classes and especially rating 1 and 2, which is represented as “bad” rating class, are very small portions compared to other ratings.





At above in these two graphs, good and bad ratings progress in terms of each year is shown.

3.2 Customer Feature

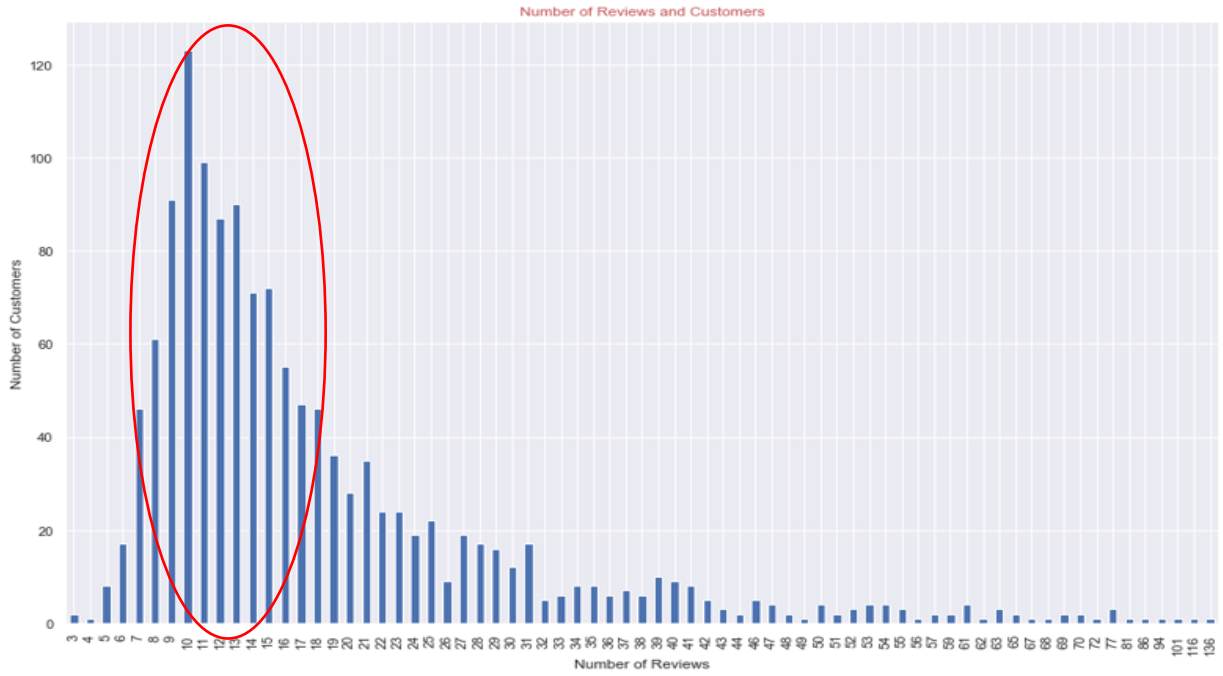


Year	Customers
2005	1 Customer
2007	31 Customers
2008	283 Customers
2009	351 Customers
2010	433 Customers
2011	474 Customers
2012	964 Customers
2013	1108 Customers
2014	1154 Customers

Rating Class	Number of Unique Customers
Bad	747 Customers
Good	1340 Customers

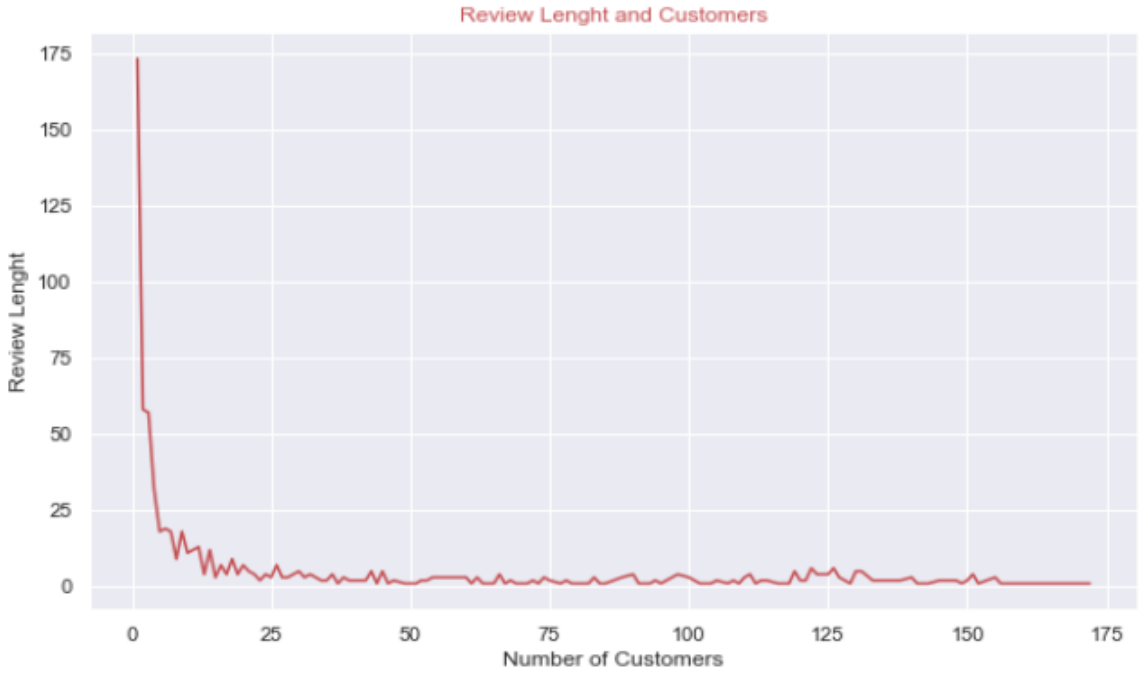
We have total 1340 unique customers who gave good reviews and 747 customers who gave bad reviews in the dataset. As it may be observed in the chart and table, the number of unique customers for each year has increased with the progress of the year.

- **Customers and Number of Reviews**



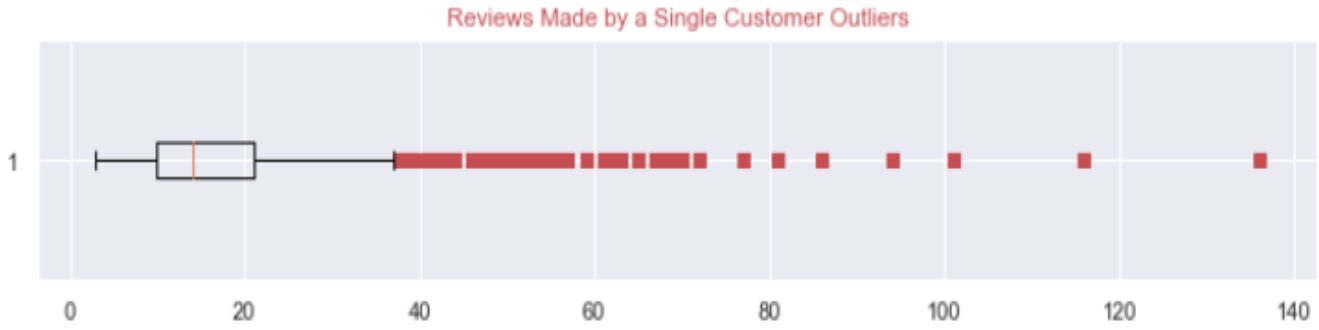
Customers give reviews ranging between 3 and 136 numbers for Amazon beauty products as it may be observed in the chart above. Mostly, customers gave 7-18 reviews.

- **Customers and Review Length**



Most of the customers tend to make short reviews.

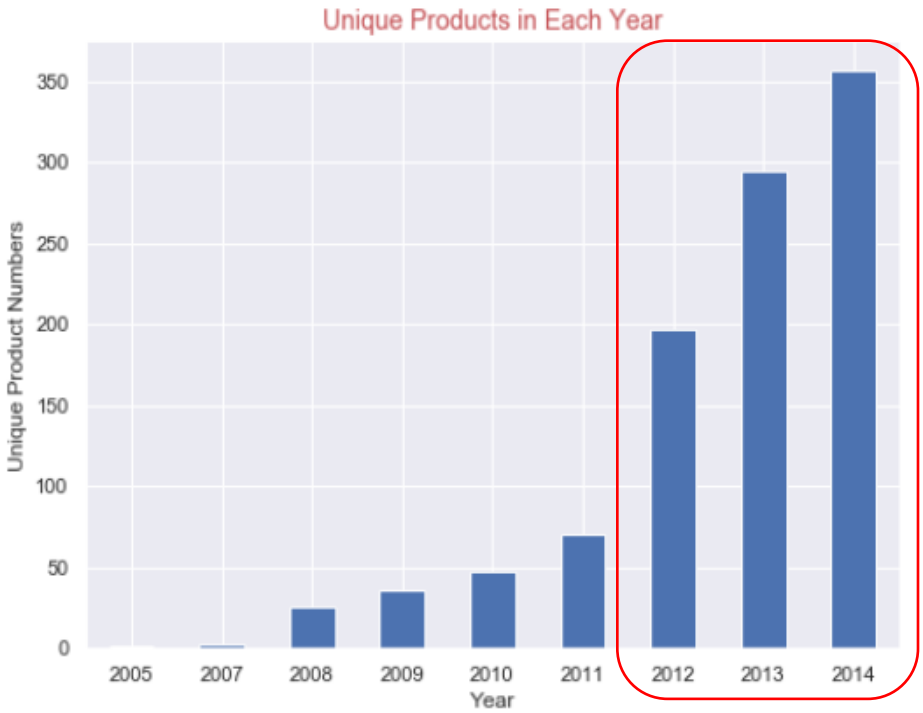
- **Reviews Made by a Singular Customer Outliers**



Most of the customers gave less then 18 reviews. Giving more than 70 reviews is very rare, but we have also customers who made reviews more than 100. Customers who has many reviews may affect the objectiveness of the results and it may affect the score of the model.

3.3 Product Feature

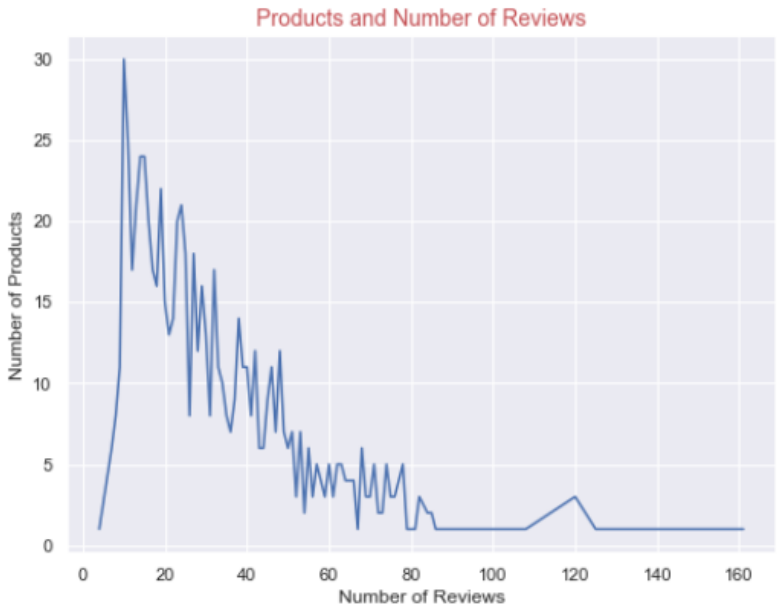
- **Unique Products**



Year	Products
2005	1
2007	3
2008	26
2009	36
2010	47
2011	71
2012	197
2013	294
2014	357

We have total 733 unique products in the dataset which belongs to year between 2005 and 2014. As it may be observed in the chart and table, the number of unique products for each year has increased with the progress of the year.

- **Products and Number of Reviews**



Number of products are compared with the number of reviews that products received in the chart above. There are some products which really received over 100 reviews.

- **Outlier products (in regards of number of reviews)**

There are approximately 33 reviews per product. Some products received more than 100 reviews. There are outlier products. These outliers may be considered in two different aspects:

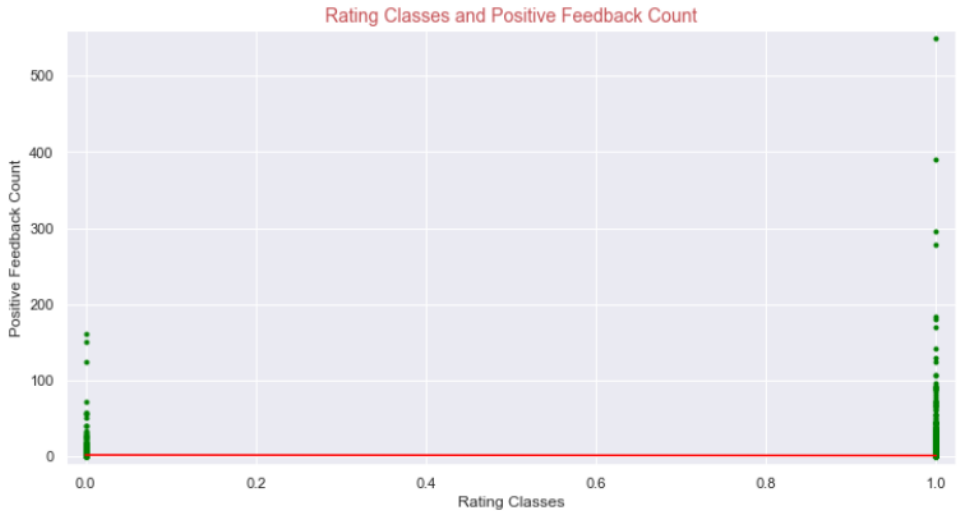
a. The first one, most probably, reviews which were made for a single product share the same or similar words, and this fact may affect the test score.

b. The second one, reviews of outlier products may give clues about the strong and weak points of the related products.

3.4 Feedback Features

- Positive Feedbacks and Rating Classes

There are 6089 good rating class and 984 bad rating class reviews which received positive feedbacks from other customers. The correlation between positive feedback and rating classes (good or bad) is very small and neglectable.



- **Outlier positive feedbacks (in regards of number of feedbacks)**

Company should prioritize to inspect the positive feedback outliers. Because those positive feedback might be for good rating as well as bad ratings for products. That is a very great input for the company usage for further improvements of the related products.

- Negative Feedbacks and Rating Classes

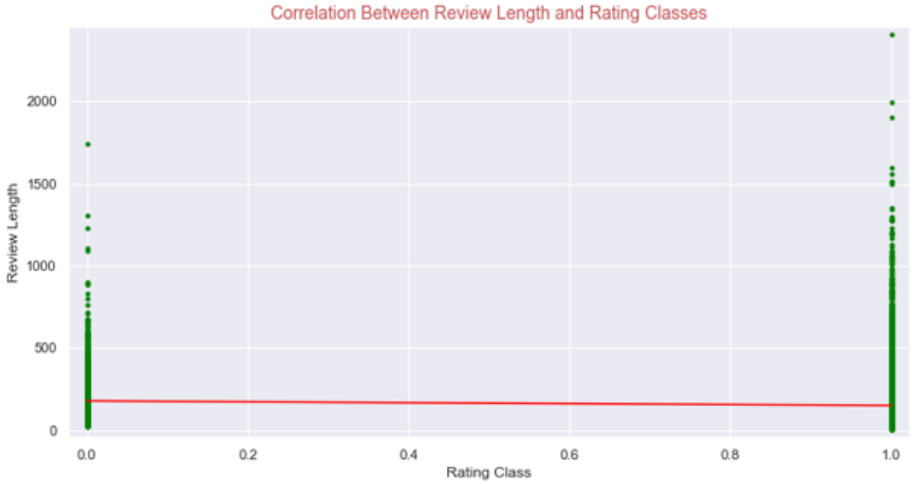
There are 4206 good rating class and 780 bad rating class reviews which received negative feedbacks from other customers. There is a very slight negative correlation between negative feedback and rating classes.

Outlier Negative Feedbacks(in regards of number of feedbacks)

Company should try to figure out the reason why customer intent to give negative feedbacks for good and bad rating class products. Those outliers may help to understand better the improvement areas for company.

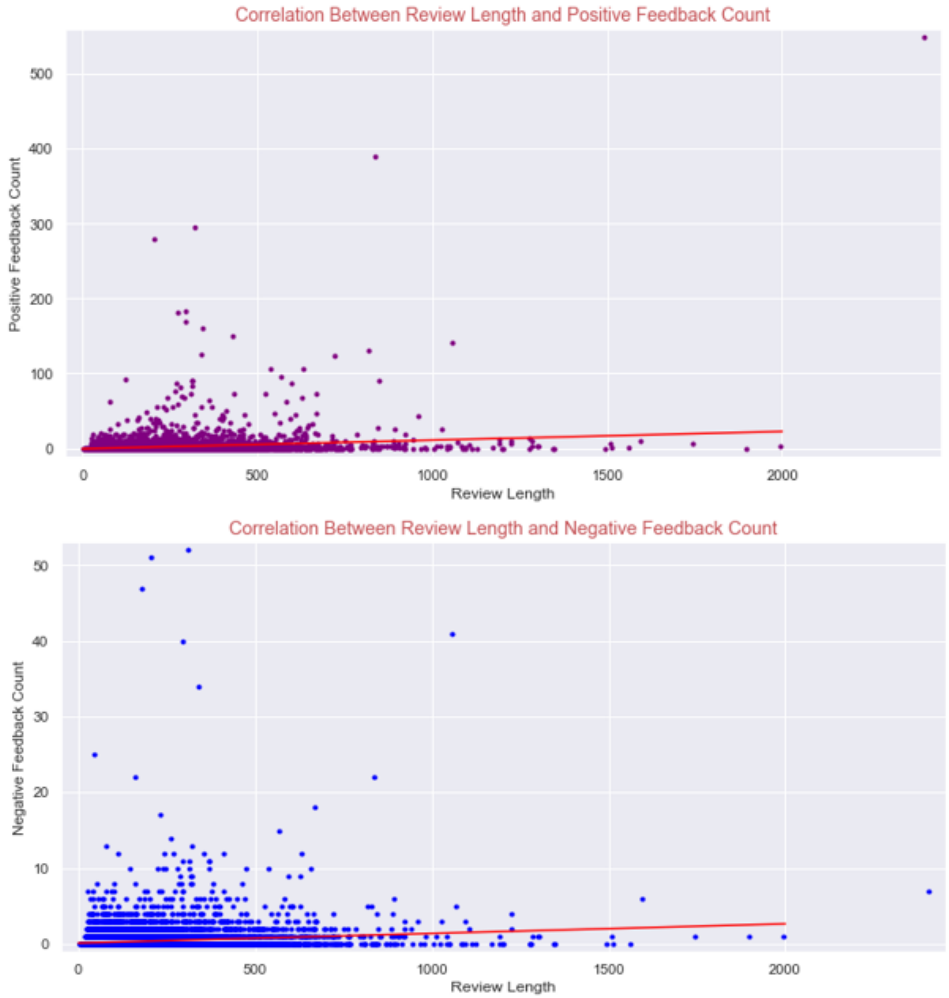
3.5 Review Length Feature

- **Review Length and Rating**



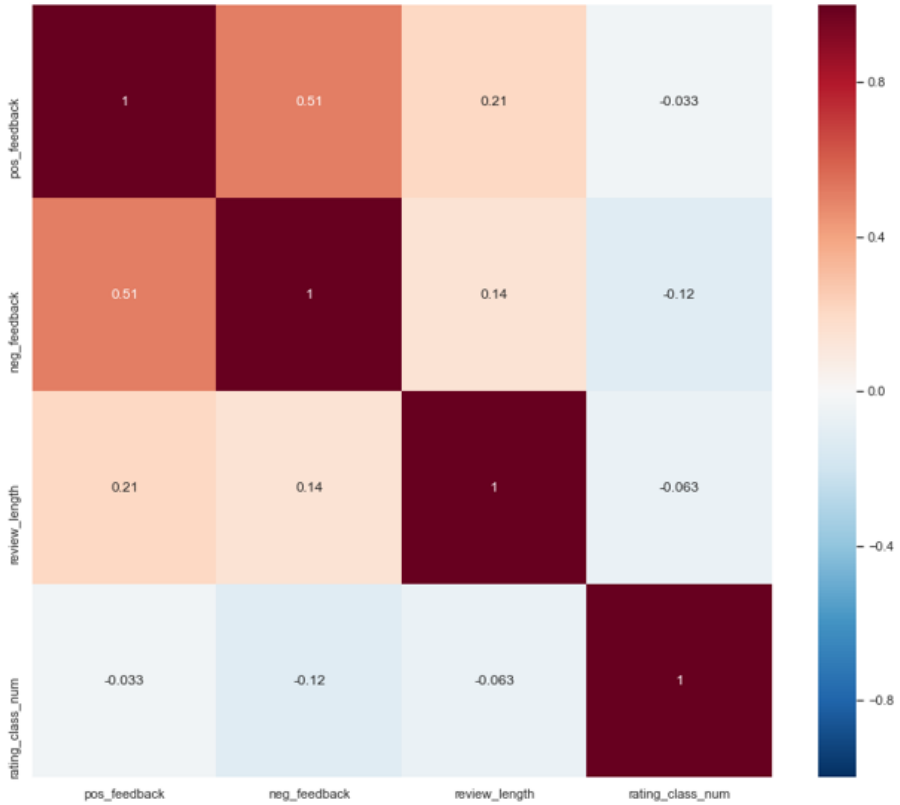
There is a very slight negative correlation between Review Length and Rating Classes. Most of the reviews have less than 200 words.

- **Review Length and Feedbacks**



There is a very slight positive correlation between feedbacks and review length.

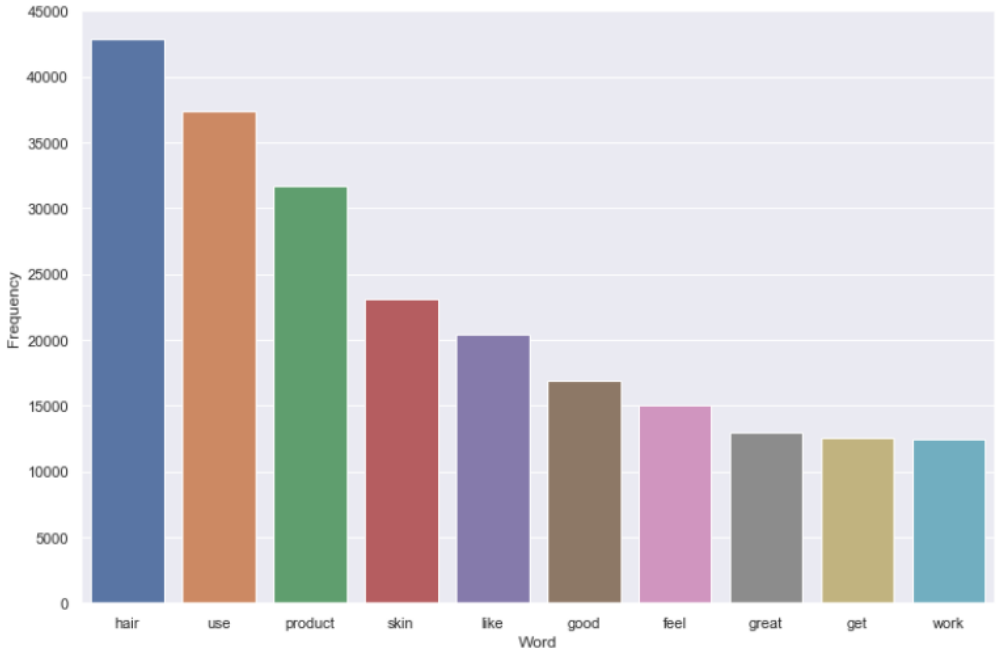
- **Correlation Between Numeric Variables**



There is no strong correlation between any two numeric variables.

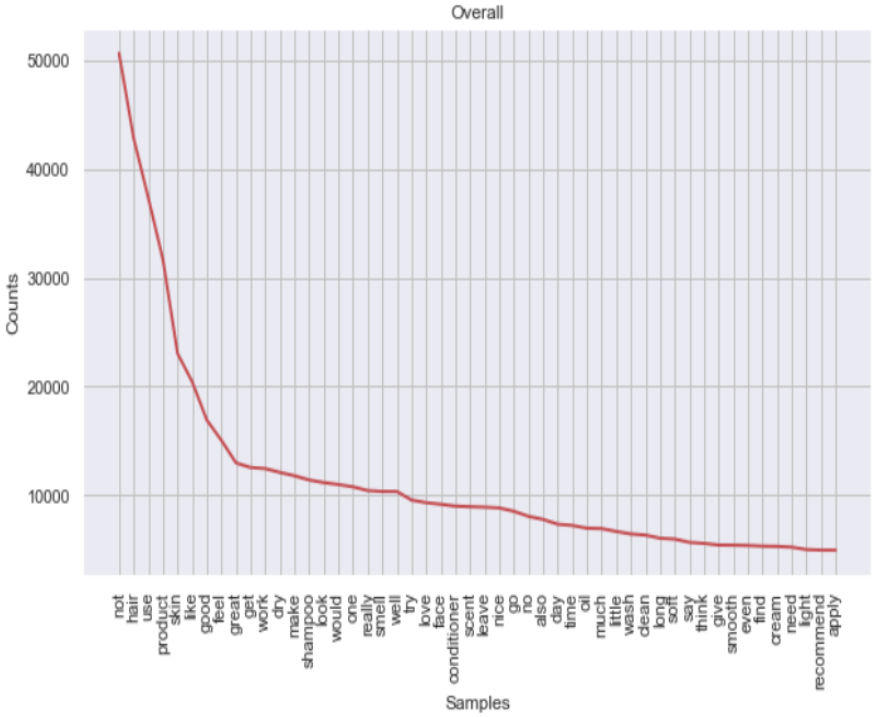
3.6 Review Text Feature

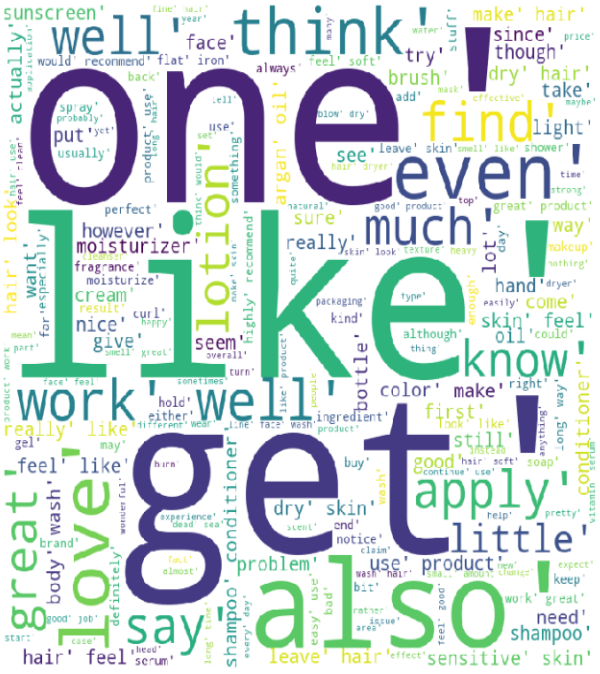
- **Top 10 Words in General**



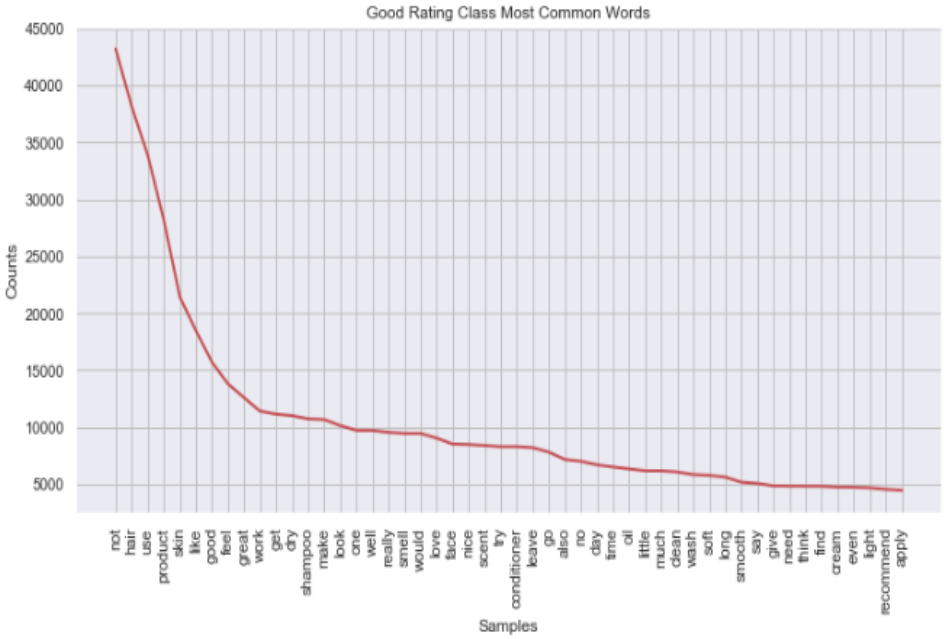
- **Most Common Words with Rating Classes**

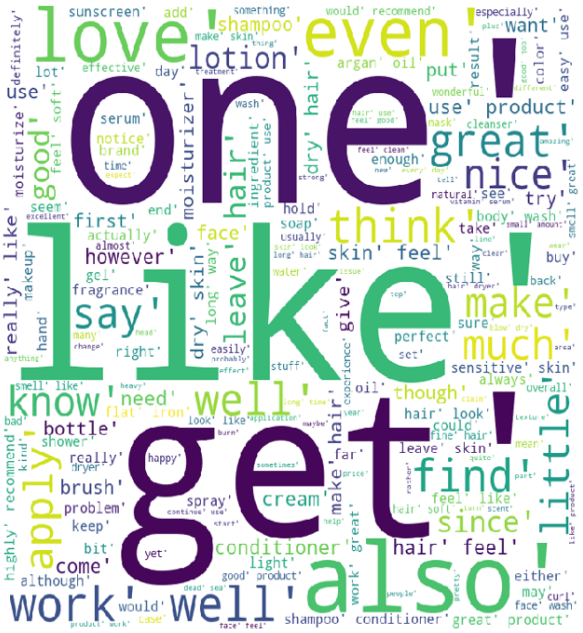
  - **Overall**



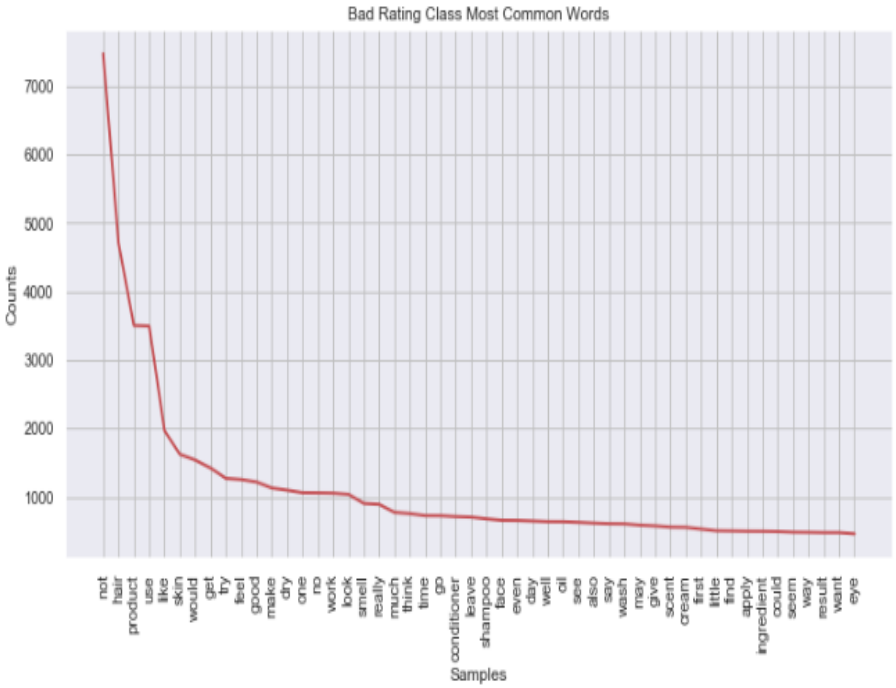


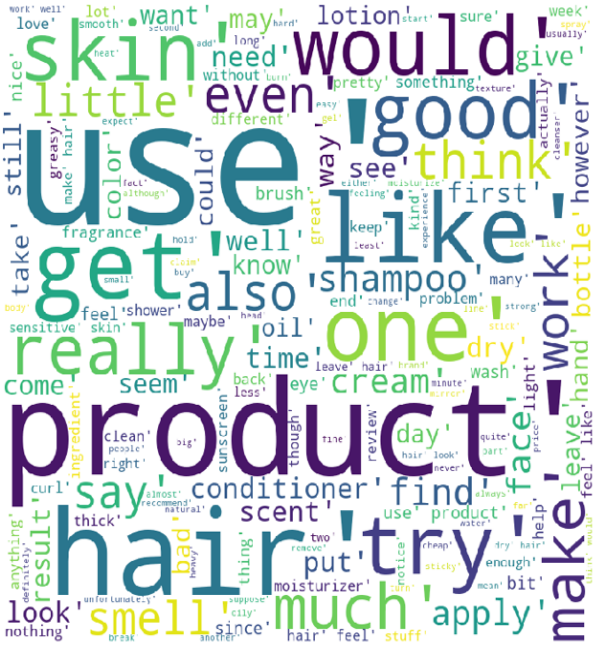
  - **“Good” Rating**





- **“Bad” Rating**





There is a great number of matching words among rating classes.

3.7 Exploratory Data Analysis Summary

* 5 rating categories has been decreased to two categories such as ‘good’ and ‘bad’ and there is an imbalance between rating classes
* There are 1340 unique customers who gave good reviews and 747 customers who gave bad reviews. The number of unique customers for each year has increased with the progress of the year.
* Customers give reviews ranging between 3 and 136 numbers for beauty products. Mostly, customers gave 7-18 reviews for products and tend to make short reviews. Giving more than 70 reviews is very rare, but we have also customers who made reviews more than 100. Customers who has many reviews may affect the objectiveness of the results and it may affect the score of the model.
* We have total 733 unique products and the number of unique products for each year has increased with the progress of the year.
* There are approximately 33 reviews per product. Some products received more than 100 reviews. There are outlier products. These outliers may be considered in two different aspects:
    a. The first one, most probably, reviews which were made for a single product share the same or similar words, and this fact may affect the test score.
    b. The second one, reviews of outlier products may give clues about the strong and weak points of the related products.
* There are 6089 good rating class and 984 bad rating class reviews which received positive feedbacks from other customers. The correlation between positive feedback and rating classes (good or bad) is very small and neglectable.
* Company should prioritize to inspect the positive feedback outliers. Because those positive feedback might be for good rating as well as for bad rating. That is a very great input for the company usage for further improvements of the related products.

***** There are 4206 good rating class and 780 bad rating class reviews which received negative feedbacks from other customers. There is a very slight negative correlation between negative feedback and rating classes.

***** Company should try to figure out the reason why customer intent to give negative feedbacks for good and bad rating class products. Those outliers may help to understand better the improvement areas for their products.

***** There is a very slight negative correlation between review length and rating classes. Most of the reviews have less than 200 words.

* There is a very slight positive correlation between feedbacks and review length.
* There is no strong correlation between any two numeric variables.
* There is a great number of matching words among rating classes.