Sentiment Analysis for Amazon Reviews

Huseyin YILMAZ

Summary

Sentiment analysis of product reviews is one of the most popular implementations of NLP (natural language processing). In this analysis, I want to study the correlation between the Amazon product reviews and the rating of the products given by the customers. I used traditional machine learning algorithms and deep neural networks. Seven different traditional machine learning algorithms used with six different bag of words methods (CountVectorizer, TfIdfVectorizer, HashingVectorizer, PCA with SMOTE Combination, Truncated SVD with SMOTE Combination, Word2Vec). Results of methods were compared and visualized. Logistic regression with CouuntVectorizing emerged as the best model.

1. INTRODUCTION

1.a. General

Natural language processing (or NLP) serves numerous use cases when dealing with text or unstructured text data. One of the subtopics of this research is called sentiment analysis or opinion mining, which is, given a bunch of text, we can computationally study peoples opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. Applications of this technique are diverse. For example, businesses always want to find public or consumer opinions and emotions about their products and services. Potential customers also want to know the opinions and emotions of existing users before they use a service or purchase a product. Last but not least, researchers uses these information to do an in-depth analysis of market trends and consumer opinions, which could potentially lead to a better prediction of the stock market. The average human reader will have difficulty identifying relevant sites and accurately summarizing the information and opinions contained in them. Besides, to instruct a computer to recognize sarcasm is indeed a complex and challenging task given that at the moment, computer still cannot think like human beings.

1.b. Problem

Our goal is to build a sentiment analysis model that predicts whether a user liked a product or not, based on their review on Amazon. Our dataset consists of customers' reviews and ratings, which we got from Consumer Reviews of Amazon products. We extracted the features of our dataset and built several supervised model based on that. These models not only include traditional algorithms such as Logistic Regression, Linear SVM, Naive Bayes, Kernel SVM, KNN, Random Forest, Gradient Boosting, XGBoost; but also deep learning with Keras. We compared the accuracy of these models and got a better understanding of the polarized attitudes towards the products.

1.c. Data Set

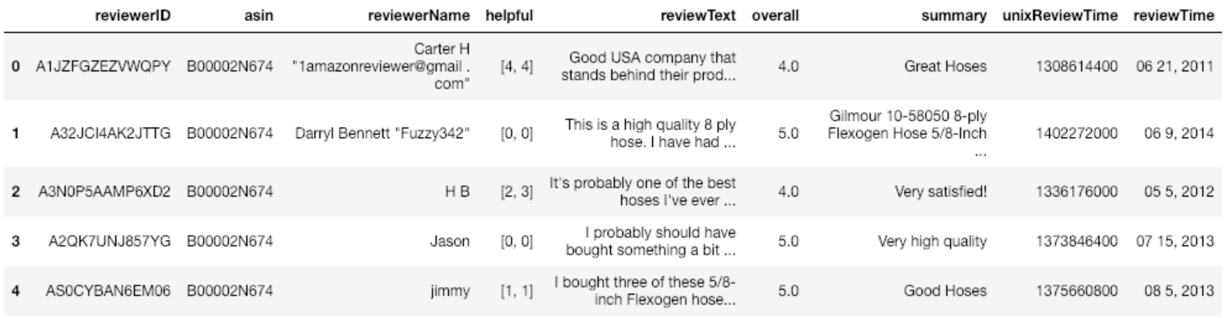
Our dataset comes from consumer reviews of Amazon products which are related with patio, lawn and garden. The data was obtained from Julian McAuley's dataset collection.[1] This dataset has 13,272 data points in total. Each record has below feature:

- Product/productId: asin, e.g. amazon.com/dp/B00006HAXW
- Product/title: title of the product
- Product/price: price of the product
- Review/userId: id of the user, e.g. A1RSDE90N6RSZF
- Review/profileName: name of the user
- Review/helpfulness: fraction of users who found the review helpful
- Review/score: rating of the product
- Review/time: time of the review (unix time)
- Review/summary: review summary
- Review/text: text of the review

2. DATA WRANGLING

2.1. Initial Understanding

The initial look of the data set is as below

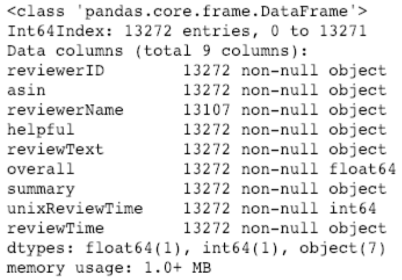


	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A1JZFGZEZVWQPY	B00002N674	Carter H "1amazonreviewer@gmail . com"	[4, 4]	Good USA company that stands behind their prod...	4.0	Great Hoses	1308614400	06 21, 2011
1	A32JCI4AK2JTTG	B00002N674	Darryl Bennett "Fuzzy342"	[0, 0]	This is a high quality 8 ply hose. I have had ...	5.0	Gilmour 10-58050 8-ply Flexogen Hose 5/8-Inch ...	1402272000	06 9, 2014
2	A3N0P5AAMP6XD2	B00002N674	H B	[2, 3]	It's probably one of the best hoses I've ever ...	4.0	Very satisfied!	1336176000	05 5, 2012
3	A2QK7UNJ857YG	B00002N674	Jason	[0, 0]	I probably should have bought something a bit ...	5.0	Very high quality	1373846400	07 15, 2013
4	AS0CYBAN6EM06	B00002N674	jimmy	[1, 1]	I bought three of these 5/8-inch Flexogen hose...	5.0	Good Hoses	1375660800	08 5, 2013

[1] http://seotest.ciberius.info/seo--jmcauley.ucsd.edu/data/amazon/

Information – info( )

One of the basic and common way to understand the date is using “info( )” method. It is simple but tells a lot.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13272 entries, 0 to 13271
Data columns (total 9 columns):
reviewerID        13272 non-null object
asin              13272 non-null object
reviewerName      13107 non-null object
helpful           13272 non-null object
reviewText        13272 non-null object
overall           13272 non-null float64
summary           13272 non-null object
unixReviewTime    13272 non-null int64
reviewTime        13272 non-null object
dtypes: float64(1), int64(1), object(7)
memory usage: 1.0+ MB
```

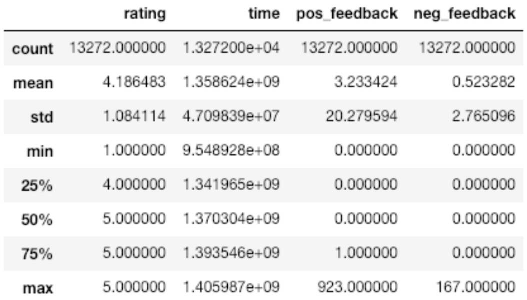
- <u>What we learned from the information:</u>
  - We have the shape, 13272 observations(records or rows) and 9 columns (or variables).
  - There is no missing value.
  - There are two variables related with date but data types are not datetime, one of them is "int64" and the other one is "object". One time related variable will be enough for us, we can drop one of them.
  - We need to figure out that whether the "helpful" variable needs to be converted to numeric type in order to use it.
  - There are two different variables which identify reviewer/user, we can drop one of them.
  - In order to improve practical and readable coding, we need change some of the column names and also we need to convert column names to lowercase.

- <u>Design of reshaping:</u>
  - "reviewerID" --> "customer"
  - "asin" --> "product"
  - "reviewerName" --> column will be droped
  - "reviewText" --> "review_text" (will be merged with “summary”
  - "helpful" --> will be splited in two columns; "pos_feedback" as positive feedback + "neg_feedback" as negative feedback.
  - "overall" --> "rating"
  - "summary" --> as is
  - "unixReviewTime" --> "time"
  - "reviewTime" --> column will be droped

- <u>Issues fixed:</u>
  - 3 new columns created
  - 5 redundant columns dropped
  - Some column names were changed and made lowercase

Statistics summary – describe( )

Numeric features



	rating	time	pos_feedback	neg_feedback
count	13272.000000	1.327200e+04	13272.000000	13272.000000
mean	4.186483	1.358624e+09	3.233424	0.523282
std	1.084114	4.709839e+07	20.279594	2.765096
min	1.000000	9.548928e+08	0.000000	0.000000
25%	4.000000	1.341965e+09	0.000000	0.000000
50%	5.000000	1.370304e+09	0.000000	0.000000
75%	5.000000	1.393546e+09	1.000000	0.000000
max	5.000000	1.405987e+09	923.000000	167.000000

Non-numeric features

- Number of unique customers: 1686
- Number of unique products: 962
- Review per customer: 7.87
- Review per product: 13.79

- Rating:
    - Mean of the ratings is more than 4 out of 5. It means that people are tendentious to giving high ratings. "std" value (1.084) and percentile values show that 1 and 2 star ratings are rare.
    - Small numbers of "ratings under 4" will decrease the predictability of these ratings. To overcome this problem we need to split the ratings in to two groups as "good" and "bad" ratings.

- Total votes (t_votes) and positive votes (p_votes):
    - Their means are more than 3.0 but percentile values shows that more than half of the reviews don't have "helpful"votes.
    - They have outliers and should be cleaned or imputed.

- Non-numeric variables statistics:
    - Some customers have more than one ratings and most probably we have some outliers.
    - All ratings do not belong to diffent different people

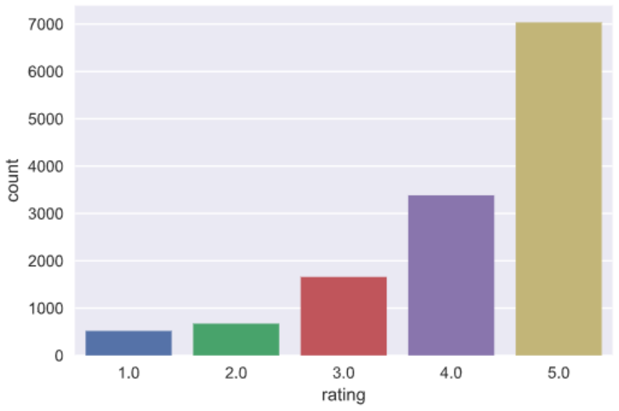
2.2. Text Preprocessing

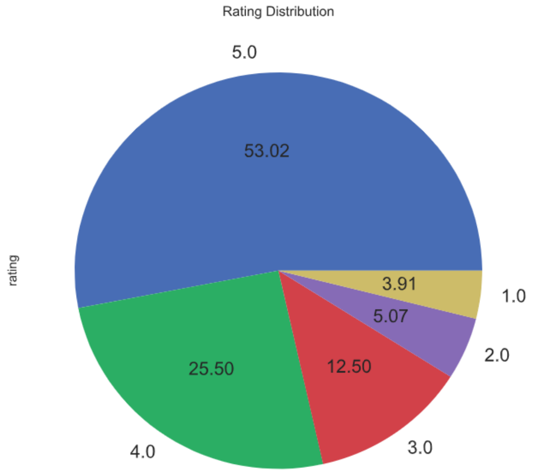
After creating the merged feature, in the context of corpus normalization we applied advanced text cleaning such as:

- Lowercase the text
- Keep only words
- Html removal
- Expanding contractions
- Whitespace and underscores removal
- Special characters removal
- Accent marks removal
- Lemmatization
- Stop words removal

3. EXPLORATORY DATA ANALYSIS

3.1. “Rating” Feature

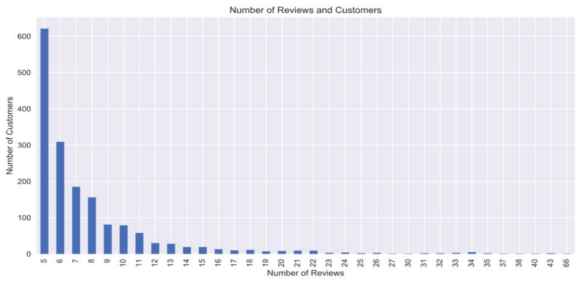




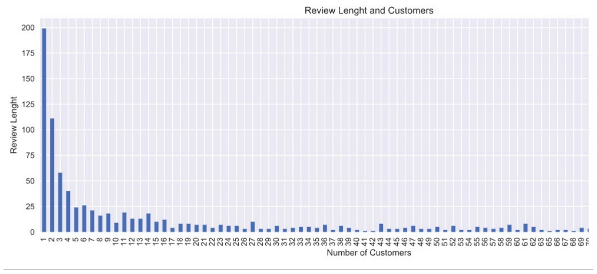
- There is an imbalance between rating classes
- Especially 1 and 2 ratings have small portions according to other classes

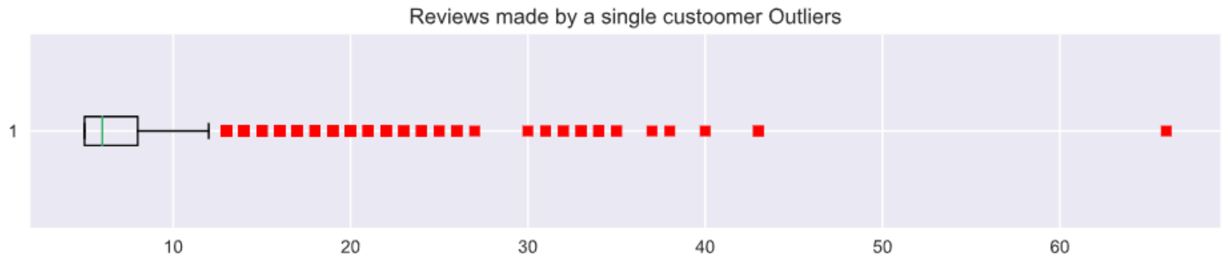
3.2. “Customer” feature

Number of reviews and customers



Review length and customers

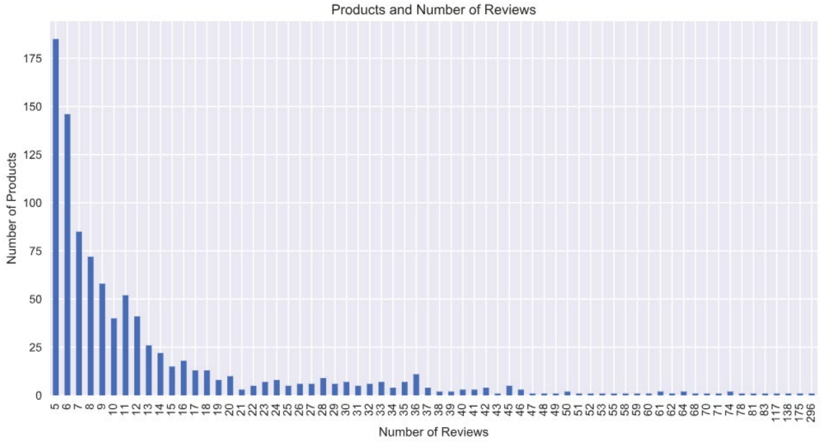


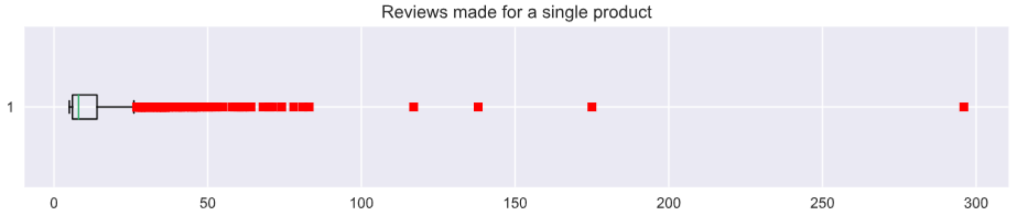


- Most of the customers gave less than 8 reviews (mean = 7.87).
- Giving more than 20 reviews is very rare.
- They are rare but we have customers who made reviews more than 40.
- Customers who has a large number of reviews may affect the objectiveness of the results.
- Here, "customer uniqueness" computed as a metric of "rating class".

- The purpose is understanding "how much are the reviews made by different customers or how much are they populated by same customers.
- For instance, customer uniqueness of "rating class 5" is 0.23, this means 77% of the reviews are given by customers who already made a review before.
- And some customers are populating the review rates which may affect the test scores negatively.
- There are outlier customers in regards of number of reviews.
- This may affect the objectiveness of the rating class.
- So, they may affect the score of the model.
- Most of the customers are tend to make short reviews.

3.3. “Product” feature:





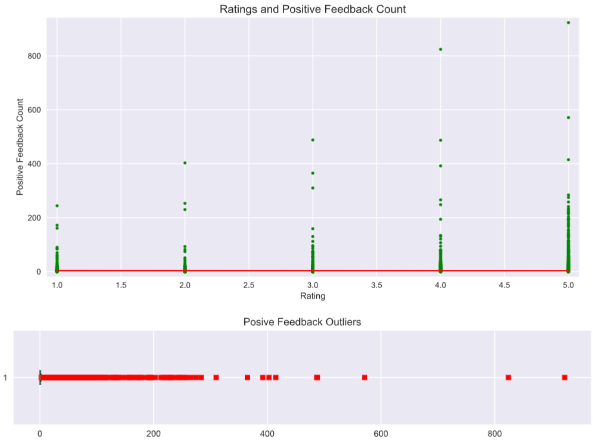
- Most of the products have less than 14 reviews (mean = 13.79)
- Product which have been received more than mean can analysis differently for extracting the strong issues about the products. For instance, cronical problems can be easily detected in this way.
- There are outlier products.
- These outliers can be considered in two different aspects

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

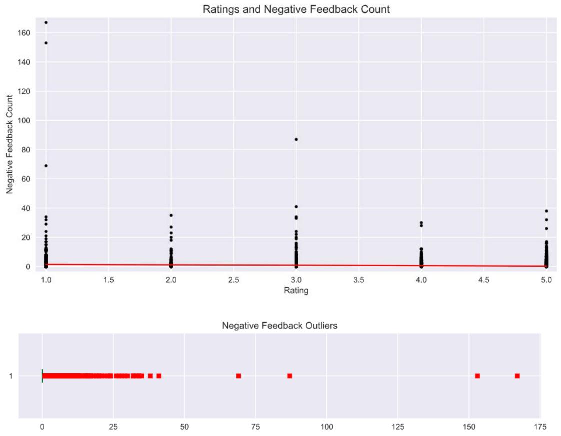
  - The second one, reviews of outlier products may give clues about the strong and weak points of the related products.

3.4. “Feedback” feature:

Positive feedback

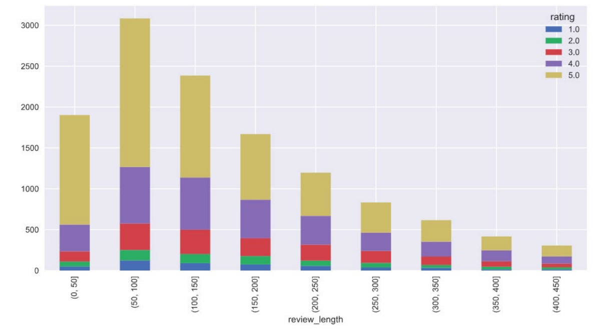


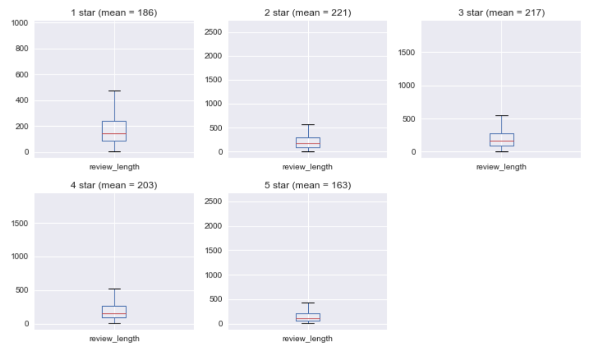
Negative Feedback

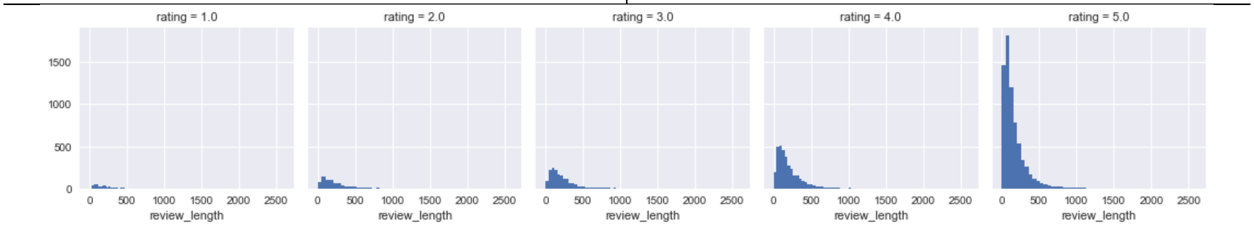


- The correlation (red lines in scatter plots) between feedbacks and ratings are very small and neglectable.
- Feedback outliers can provide the information of what customers like most about a specific product, and company can use this information for further improvements of the related products.

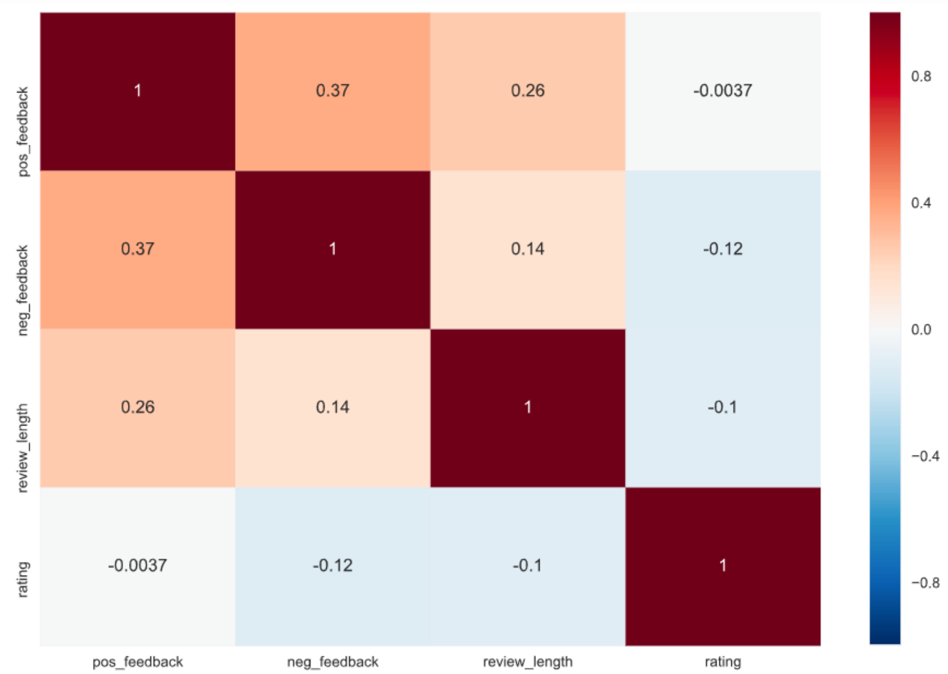
3.5. "Review Length" feature





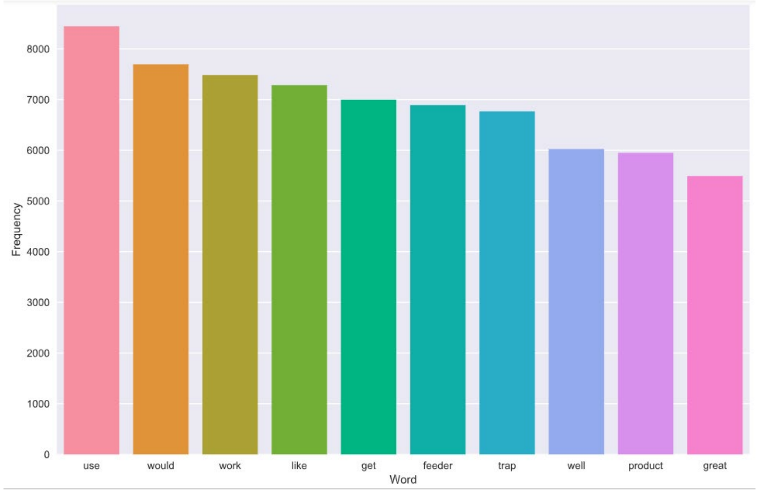


- There is a very slight negative correlation between Review Length and Rating Classes.
- Most of the reviews have less than 200 words.
- Boxplots show us that customers tend to using more words when they give 2 , 3 and 4 ratings.



- As seen above correlation matrix, there is no strong correlation between any two numeric variables.

3.6. “Review text” feature



- Most significant inference of text graphs is that there is a great number of matching words among review texts of different rating classes

3.7. EDA Findings

- There is an imbalance between rating classes, especially 1 and 2 ratings have small portions according to other classes.
- Most of the customers gave less than 8 reviews (mean = 7.87) but there are also outlier customers who have a large number of reviews. A great number of reviews which are made by a single customer may affect the objectiveness and so the test score negatively.
- Customer uniqueness of 4 and 5 ratings are low, it means they are populated by same customers. This may bring the same above drawbacks.
- Most of the products have less than 14 reviews (mean = 13.79). Product which have been received more than mean can be analyzed differently for extracting the strong issues about the products. For instance, chronical problems can be easily detected in this way.
- There is a negative correlation between number of feedbacks and rating classes are not strong. But outlier feedbacks can provide additional information.
- Most of the reviews have less than 200 words. And customers tend to using more words when they give 2, 3 and 4 ratings.
- There is no strong correlation between any two numeric variables.
- Most significant inference of text graphs is that there is a great number of matching words among review texts of different rating classes

- **Recap crucial points:**
  - There is no strong relationship between numeric predictors and target variable
  - Data set is imbalanced in regards of rating classes.
  - There is a great number of matching words among rating classes.

- **Conclusion:**
  - Using numeric variables will not make meaningful contribution to prediction.
  - We can reduce the imbalance with reducing the number of rating classes.