Amazon Review Rating Prediction - NLP Study

Enes Gokce

May 2020

Table of Contents

Introduction

Online product review networks help to transmit information that customers can use to evaluate products in Internet commerce. These networks frequently include an explicit social component allowing consumers to view both how community members have rated individual product reviews and the social status of individual reviewers (Dhanasobhon, et al., 2007).

Digital networks for product information have redefined traditional "word-of-mouth" social networks by allowing consumers to easily share their opinions and experiences with other members of large-scale online communities (Dellarocas 2003). Many online retailers, such as Amazon.com and BarnesandNoble.com, are augmenting their product markets by building online communities to provide product reviews to other consumers. Likewise, many auction sites, such as Ebay.com, allow consumers to rate product sellers. Such information sharing has the potential to reduce the uncertainty consumers face regarding the quality of a product or a seller.

Several papers in the literature have shown that large-scale information sharing in digital networks may help communicate product/seller quality and build trust between buyers and sellers in online markets (Dhanasobhon, et al., 2007). There are many on-line settings in which users publicly express opinions (Danescu-Niculescu-Mizil et al., 2009). It's easy to assume that while buying a product from an Amazon store, customers first checks the ratings of the products which are displayed as stars. When a more detailed user feedback is desired, the reviews are considered as a resource because reviews carry way more information than the ratings.

The Problem

Resnick (2002) shows that seller reviews in eBay influence the probability of a sale, while Chevalier and Mayzlin (2006) find that product reviews at Amazon.com impact book sales. This raises the question of:

- What makes a review positive and what makes it negative?
- What is the common point of good reviews and bad reviews?

In this study, I will investigate what makes a review a good review and what makes it a bad review. In addition, by using Neural Language Processing (NLP), I developed a prediction model. The prediction model will tell whether the review indicates a positive rating or negative rating. In addition, the study will provide some explanation to misclassification.

The Client

The findings of these study can help sellers on Amazon. By examining what are the coming topics that people write good review and bad review, sellers can make an information-driven investment on product and service improvement.

Amazon itself can use the findings to identify anomalies in reviews. This study itself cannot be used for detecting anomalies but it can provide supplemental information to a study that try to find anomalies in reviews.

The Data

The data used in this project was downloaded from Kaggle. It was uploaded on Kaggle by J. McAuley and J. Leskovec who are Kaggle.com users under the username of Stanford Network Analysis Project. This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all 568.454 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review.

Data includes:

- Reviews from Oct 1999 - Oct 2012
- 568,454 reviews
- 256,059 users
- 74,258 products

Table 1

Column Names and Their Explanation

Feature	Explanation
Id	Row Id
ProductId	Unique identifier for the product
UserId	Unqiue identifier for the user
ProfileName	Profile name of the user
HelpfulnessNumerator	Number of users who found the review helpful
HelpfulnessDenominator	Number of users who indicated whether they found the review helpful or not
Score	Rating between 1 and 5
Time	Timestamp for the review
Summary	Brief summary of the review
Text	Text of the review
Total	10 Rows

The flow of the Study:

1. Feature extraction-1
2. Text Cleaning
3. Feature extraction-2
4. Data Visualization and Data Wrangling
5. Sentiment Analysis
6. Topic Modeling and Latent Dirichlet Allocation (LDA)
7. Predictive Modeling with Machine Learning Algorithms

Feature Extraction-1

For other machine learning studies, I prefer feature extraction after data cleaning. However, for this NLP study, during data cleaning I will lose some of the data. Therefore, the feature that are not possible to obtain after data cleaning were extracted in this process. Here are these features:

1) **Number of stop words:** A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. In Python’s ***nltk*** package, there are 127 English stop words default. These 127 words were ignored.

   In addition, after checking the most frequent words, "br", "also", "im", "ive" words appeared among top 50 frequent words. These words also ignored. At this point, there is a reason why ‘br’ was the most frequent word: in HTML, the
 tag inserts a single line break. After removing punctuations, somehow this script appeared without less-than and bigger than signs.
2) **Number of hashtag characters:** One more interesting feature which we can extract from a review is calculating the number of hashtags or mentions present in it. After data cleaning, it will not be possible obtain this feature.
3) **Number of numerical characters:** Calculating the number of numeric characters which are present in the reviews can be useful. At least, it does not hurt to have such a data.
4) **Number of Uppercase words:** Anger or rage is quite often expressed by writing in UPPERCASE words which makes this a necessary operation to identify those words.

Text cleaning techniques

Before applying NLP techniques on the data, we first need to clean and prepare the data. If this process is not done properly, it can ruin the analysis part totally. Here are the steps that were applied on the data:

1) **Make all text lower case:** The first pre-processing step was transforming the reviews into lower case. This avoids having multiple copies of the same words. For example, while calculating the word count, 'Analytics' and 'analytics' will be taken as different words if we ignore this transformation.
2) **Removing Punctuation:** For now, there is no a meaningful way to analyze punctuations. Thus, they were removed from the text data. With this step, these characters were removed: [!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~]

Data Cleaning

Here are the steps for data cleaning:

Make all text lower case: This is important. Otherwise, 'coffee' and 'Coffee' would be considered as different words.

Removing Punctuation: Punctuations creates noise in the data, should be cleared.

Removal of Stop Words: With this step, I removed all default English stop words in nltk package.

Removing URLs: URLs are another noise in the data that were removed.

Remove html tags: HTML is used extensively on the Internet. But HTML tags themselves are not helpful when processing text.

Removing Emojis: Emojis can be indictor of some emotions that can be related to being customer satisfaction. Unfortunately, we need to remove the emojis in our text analysis because for now, it's not possible to analyze emojis with NLP.

Remove Emoticons: What is the difference between emoji and emoticons?

- :-) is an emoticon
- 😜 → emoji.

Spell Correction: On Amazon reviews, there are plethora of spelling mistakes. Product reviews are sometimes filled with hasty sent reviews that are barely legible at times. In that regard, spelling correction is a useful pre-processing step because this also will help us in reducing multiple copies of words. For example, "Analytics" and "analytcs" will be treated as different words even if they are used in the same sense.

Feature Extraction-2

After text cleaning, more feature extraction was done. These features were extracted after text cleaning because they are more meaningful to obtain at this step. At this point, I tried to extract as many as features. I did not have to worry about whether the features will be useful in the future or not because having extra features do not hurt the text analysis in any way.

Number of Words: This feature tells how many words there are in the review.
Number of characters: How many letters are contained in the review.
Average Word Length: Average number of letters in the words in a review.
Good_Reviews: The data set has Score feature that is between 1 and 5. For this study, Score feature will be converted in Good_Reviews columns. This column has two values:

(0): Reviews that has Score value 1, 2 or 3.

(1) Reviews that has Score value 4 or 5.

In the analysis part, this feature was predicted with Machine Learning models. In this regard, it is a significant variable.

Data Visualization and Data Wrangling

While exploring the data, we will look at the different combinations of features with the help of visuals. This will help us to understand our data better and give us some clue about pattern in data. In addition, an overall looking of the data can help us to understand the shape of the data (Table 1).

Table 1

Descriptive information about features

	Helpfulness Numerator	Helpfulness Denominator	Score	stopwords	punctuation	hastags	numerics	upper	word count	char count	avg word
count	568411	568411	568411	568411	568411.0	568411	568411	568411	568411	568411	568411
mean	2.0	2.0	4.0	32.0	17.0	0.0	0.0	3.0	41.0	269.0	6.0
std	8.0	8.0	1.0	32.0	25.0	0.0	1.0	6.0	43.0	281.0	1.0
min	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	5.0	0.0
25%	0.0	0.0	4.0	13.0	5.0	0.0	0.0	1.0	17.0	110.0	5.0
50%	0.0	1.0	5.0	23.0	10.0	0.0	0.0	2.0	29.0	184.0	6.0
75%	2.0	2.0	5.0	40.0	20.0	0.0	0.0	4.0	50.0	323.0	6.0
max	866.0	923.0	5.0	1295.0	2035.0	34.0	32.0	334.0	1977.0	14782.0	10.0

For the beginning, examining distribution good reviews may help us to understand how this column was shaped. From the Figure 1, it can be seen that this is an unbalanced data. This is an important point to keep in mind because while making prediction, model performance of unbalanced data is different than balanced data.

Figure 1: Distribution of Good Reviews

Examining Helpfulness Numerator

Helpfulness numerator is the number of users who found the review helpful. From the Figure 1, it can be observed that there are extreme values in this column. When I checked reviews that has more than 500 helpful voting. I saw that there are a lot of exactly same reviews (Table 1). This cannot be coincidence. There might be a bot that is writing reviews or some people are copying exactly same reviews for some reason. This anomality was happening even helpfulness is more than 100. As a result, I considered them as a bad data and dropped the reviews that has more than 100 helpfulness voting.

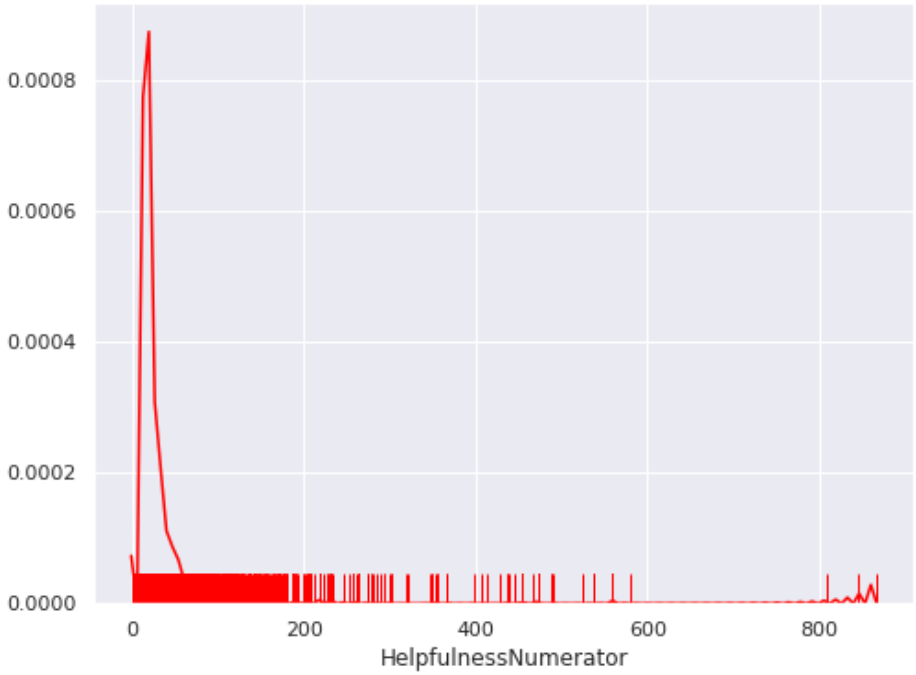

Figure 2: Density plot of Helpfulness Numerator feature

Table 2

Reviews that has more than 500 helpfulness numerators

Reviews
good opportunity sample various timothys kcup ...
ecobrew reusable keurig kcup great brewing cof...
ecobrew reusable keurig kcup great brewing cof...
ecobrew reusable keurig kcup great brewing cof...
see update end reviewbr lamenting frech lettuc...
bought aerogarden wife back may ill start posi...
ordered one fresh whole rabbits arrived head f...
ecobrew reusable keurig kcup great brewing cof…
product called hunmatsuryokucha japanese macch...
ecobrew reusable keurig kcup great brewing cof...
ecobrew reusable keurig kcup great brewing cof...
ecobrew reusable keurig kcup great brewing cof...
huge fan keurig brewing delighted discover sol...
ecobrew reusable keurig kcup great brewing cof...
huge fan keurig brewing delighted discover sol...
ecobrew reusable keurig kcup great brewing cof...
ecobrew reusable keurig kcup great brewing cof...
purchased burrito small shop blocks home unimp...

Examining Number of Stopwords: Number of Stopwords is another feature that could only be obtained before data cleaning. From the Figure 3, we see that most of the reviews have less than 200 stopwords. We can easily assume that more than 200 stopwords are extreme cases. For the further analysis, I checked that reviews that has more than 900 stopwords (Table 3). Again, these appeared exactly same reviews. After checking reviews that has 600 reviews, there appeared still many same reviews. As a result, these reviews were considered as a bad data, and dropped from the data.

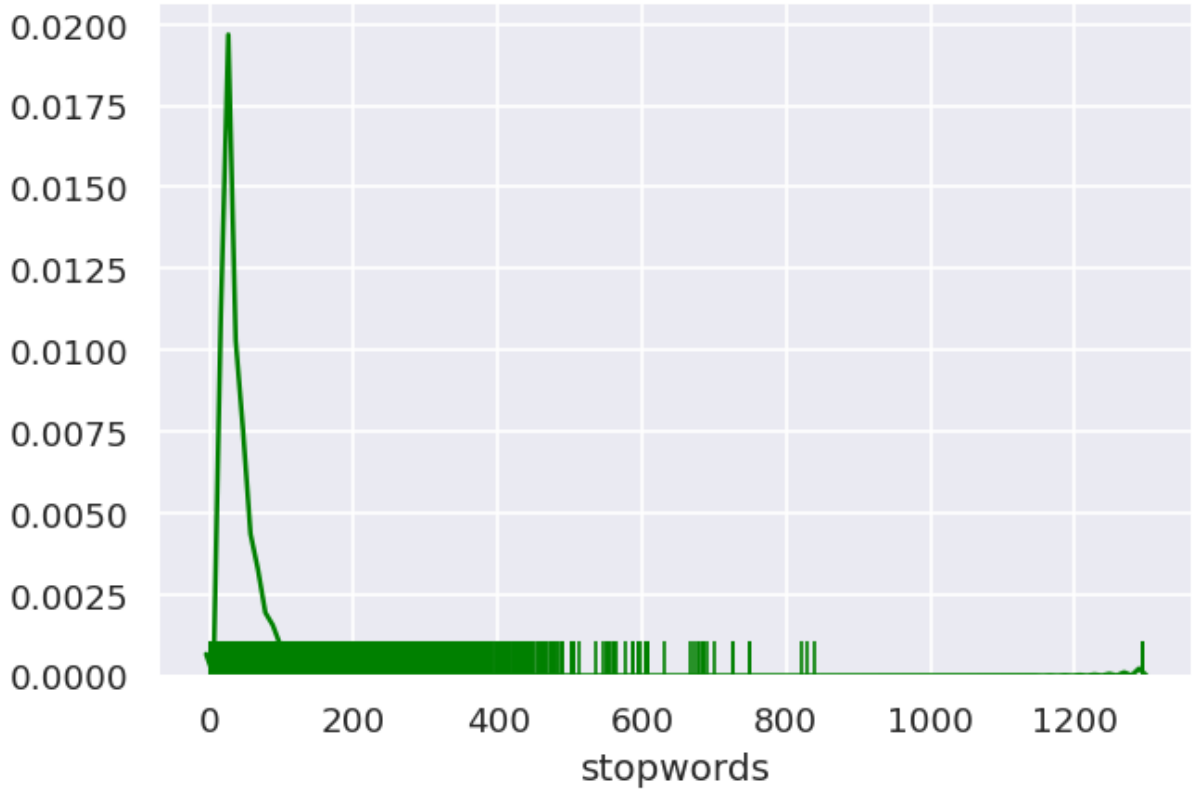

Figure 3: Density plot for stopwords

Table 3

Reviews that has more than 900 stopwords

Reviews
fuzzy wuzzys summarybr somewhat recommended re...
fuzzy wuzzys summarybr somewhat recommended re...
fuzzy wuzzys summarybr somewhat recommended re...

Dropping duplicate reviews

Similar processes were applied on rest of the features. It was observed that same reviews were causing serious problems and appears as extreme values. Therefore, I decided to drop all the duplicate reviews. Before applying this, reviews that have less than 4 words were checked because there is a chance that they can be exactly same reviews coincidentally. It was seen that none of them same reviews (Table 4). This means that when we delete duplicate reviews, short reviews that are coincidentally

same will not be deleted because we do not have any exactly same short reviews in the dataset. Before dropping duplicate reviews, there were 567950 reviews. After dropping duplicates, remaining number is 392718 (Table 5). This shows that there are considerable number of duplicate reviews.

Table 4

Reviews that are shorter than 4 words

Reviews	Reviews
excellent product	idea growing black
excellent	peanuts delicious order
sorry say like	much else say
ok good pumpkins	price couldf lower
seller accomadating helpful	worth
favorite excellent item	item easy use
try youll like	great make time
good little dry	bottle
Number of Reviews:	16

Table 5

Comparison of number of reviews

	Number of Reviews
Before dropping duplicates	567950
After dropping duplicates	392718

After this point, I checked reviews for extreme values for different features. There were not a suspicious or anormal situation. Reviews looked quite normal and authentic (Table 6). This shows that by dropping duplicate reviews completely (without keeping any of them), we could get rid of some problematic data points.

Table 6

Reviews that has more than 900 words

Reviews
excellent christmas gift birthday time gifti w...
green tea ingredient slows breast cancerantiox...
see pics ruler show size came double boxed pac...
weight loss benefits green teabr drink green t...
serious gopher problems years effective point ...
eden dried montmorency cherries perfection del...

Number of reviews: 6

After this point, we will examine the data more closely from different perspectives.

Examining Number of Numeric Values

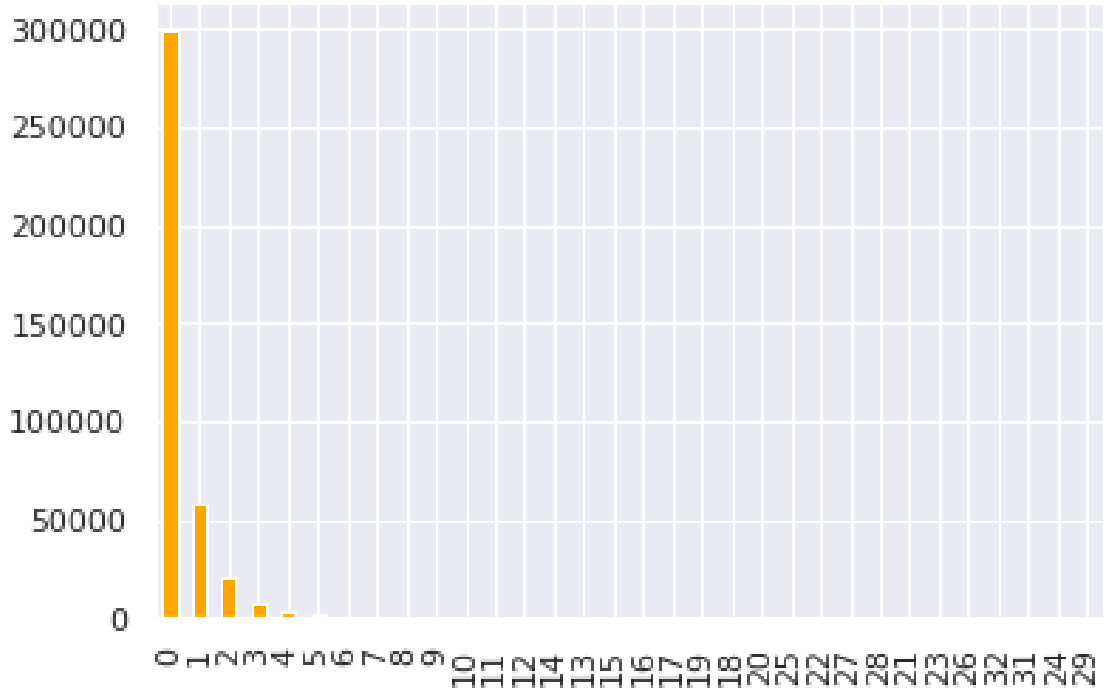

Figure 4: Bar plot of distribution of numeric value number for reviews

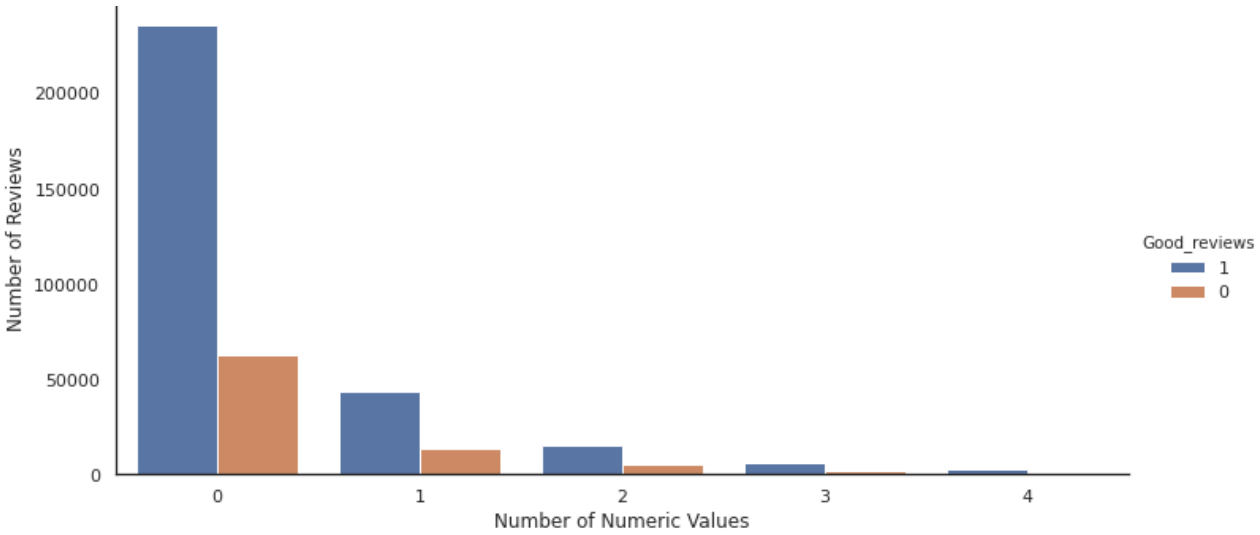

Figure 5: Number of numeric values depending on Good Reviews category.

In Figure 2, it can be observed that majority of reviews do not have numeric values. Then, it is hard to see numeric values more than 4 even though there are tweets that have 29 numeric values. In order to look at this feature more closely, we can examine it until 4 numeric values (Fig. 3). At the Figure 3, we see that number of numeric value's distribution is similar to distribution of Good Reviews (Fig. 1). Keep in mind that number of numeric values was obtained before text cleaning. In the cleaned data which we are using now, we cannot see any numbers in reviews.

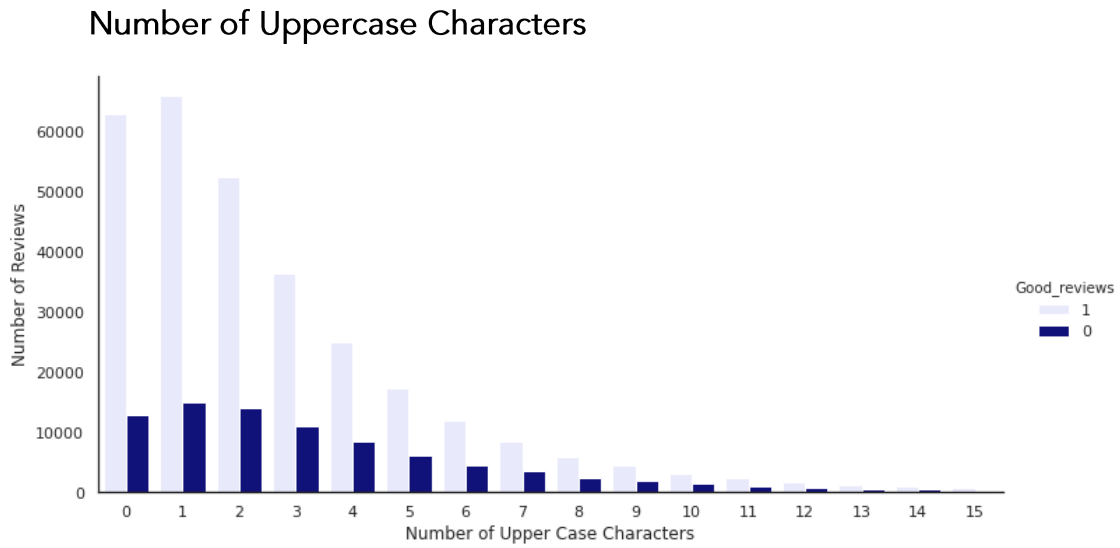

Figure 6: Number of upper-case characters depending on Good Reviews category.

Uppercase characters are capital letters. Using capital letters to indicate strong feeling may be the most famous example of typographical tone of voice. It can express a wide range of emotions. However, as we can see from the figure 5, most of the reviews have less than 5 upper case characters. There can be different explanations for this. Jess Joho (2019) states that for younger all generation all lower case is cool, joyous, faster to read and write. It is casual and friendly. For older generation, this can also be true for writing a casual Amazon review. It would not be surprising if people are not paying much attention to grammar rules on Amazon website.

Examining the Time

In the dataset, we have 'Time' feature that tells day, month and the year of reviews. First, I wanted to see how number of reviews changed in time. We see that number of reviews are increasing considerable in time. I want to note that the dates in the dataset starts from 1999. However, between 1999 and 2005, number of reviews is so small compared to rest (Fig 8). Therefore, Figure 7 starts with the year 2005.

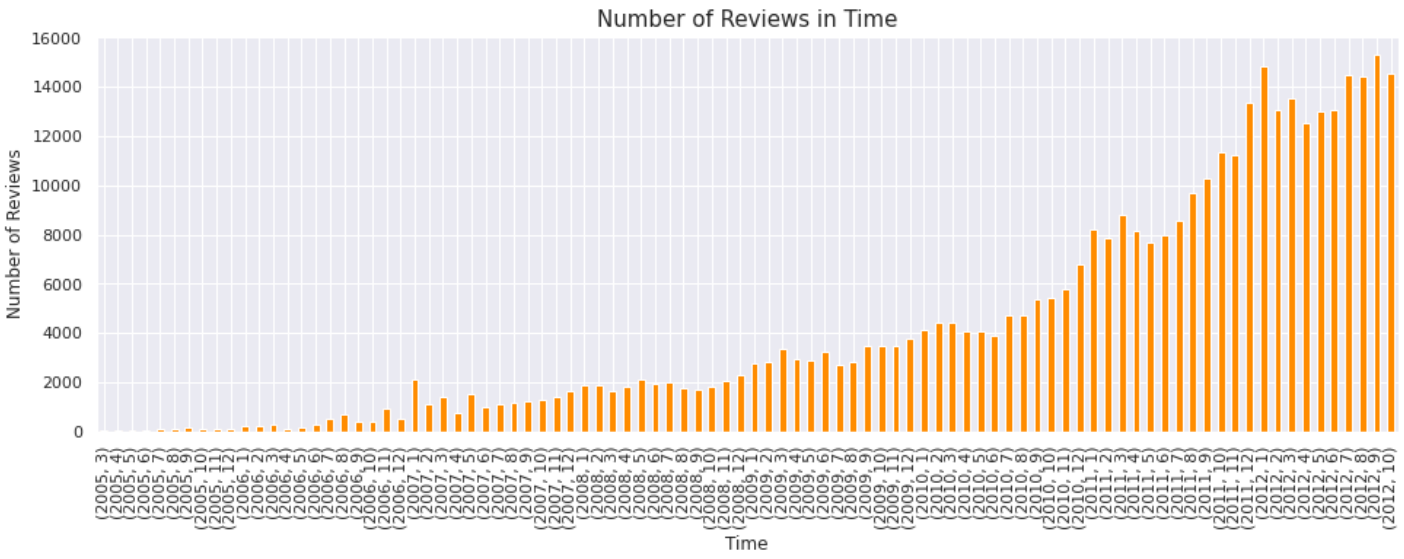

Figure 7: Change in number of reviews in time from 2005 to 2012

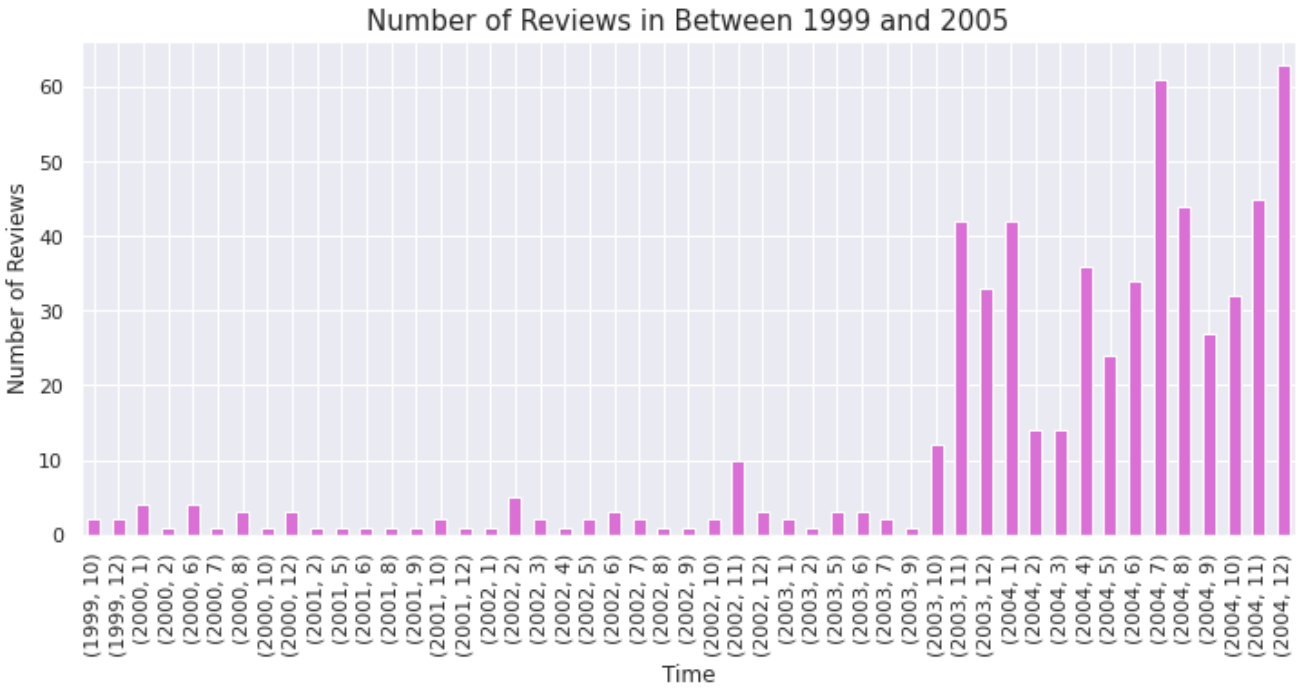

Figure 8: Number of reviews before 2005

Secondly, I wanted to check whether if the 'Time' has a visible effect on the 'Good Reviews' rating. This affect can be investigated in a couple ways. We can investigate whether is there

a specific month that people write more reviews. According to Figure 9, there is no a favorable month in this aspect. Mean of 'Good Reviews' for each month so close to each other. It not possible to claim that there is a significant increase in the positive reviews for a specific month. In order to investigate this further, we can check Figure 10. Figure 10 tells that average good review score is around 0.8. This holds true for years. Again, we didn't include the dates between 1999 and 2005 because there are so limited number of reviews in this time frame (Fig. 8). Small number of reviews may give produce extreme values for the mean of *Good Reviews*.

Keep in mind that '*Good Reviews'* feature has two category 0 and 1. '0' is for 1, 2 and 3-star ratings, and '1' is for 4 and 5-star ratings.

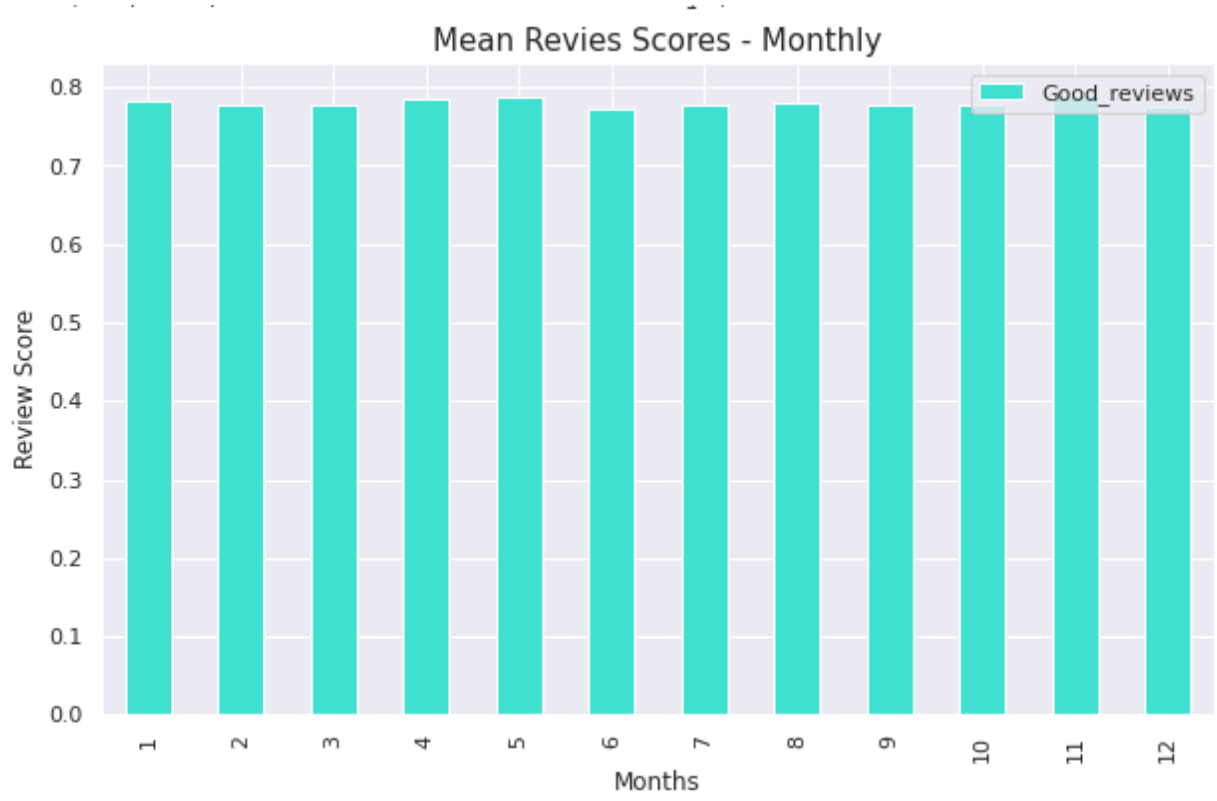

Figure 9: Mean Good Review score depending on Month between 1999-2012

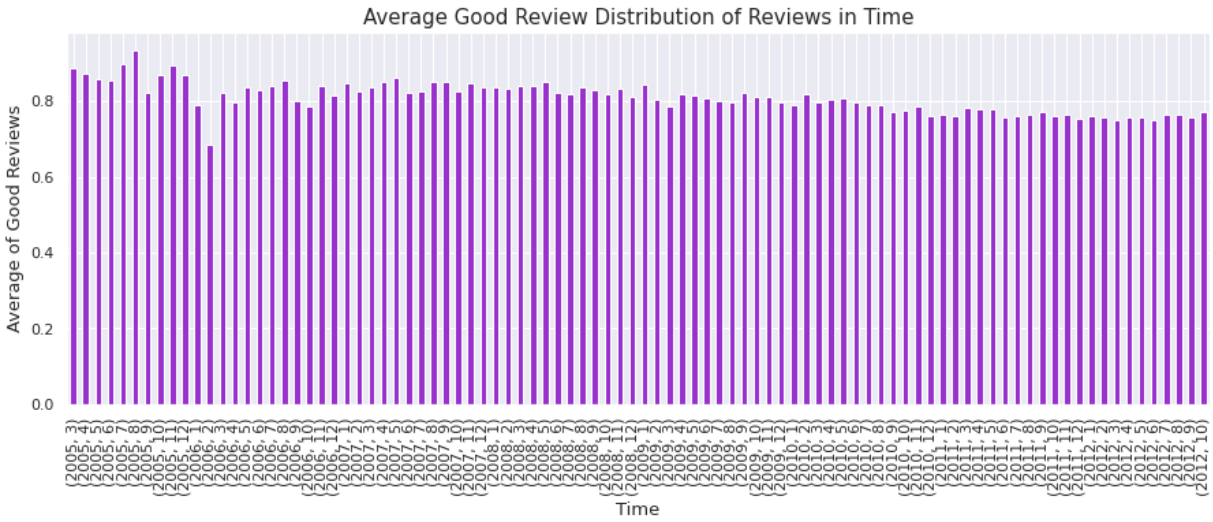

Figure 10: Mean Good Review score in time

Most Frequent Words

For checking most frequent words, word clouds offer useful visuals. It is a great visual for conveying information. Figure 11 represents most frequent 100 words in the reviews.

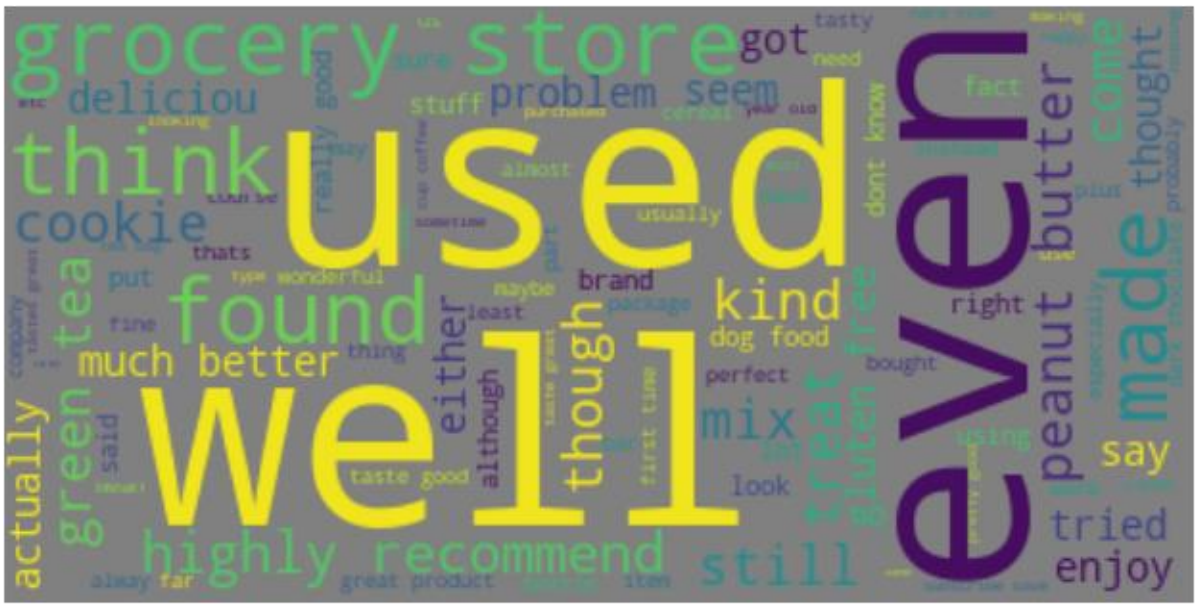

Figure 11: Most frequent 100 words in reviews

Sentiment Analysis

Before making predictions based on machine learning models, we need to understand the data better. The reviews are unstructured, in other words it's unorganized. Sentiment analysis, however, helps us make sense of all this unstructured text by automatically tagging it. In addition, sentiment analysis helps us process huge amounts of data in an efficient and cost-effective way. That's why, sentiment analysis will be performed in this study.

In this study two main sentiment classifiers were used:

1. Polarity
2. Subjectivity

The TextBlob package for Python is a convenient way to perform many Natural Language Processing (NLP) tasks. For this study, Textblob package was used for sentiment analysis. When calculating sentiment for a single word, TextBlob takes average for the entire text. For heteronym words, Textblob does not negotiate with different meanings. In the other words, only the most common meaning of a word in entire text is taken into consideration. For making all these calculations, Textblob uses WordNet Database of the Princeton University. WordNet is a large lexical database of English. In this database, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations (Fellbaum, 1998).

Polarity

Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Figure 12 shows the distribution of polarity score in reviews. Most of the reviews are on positive side of the plot (Fig. 12).

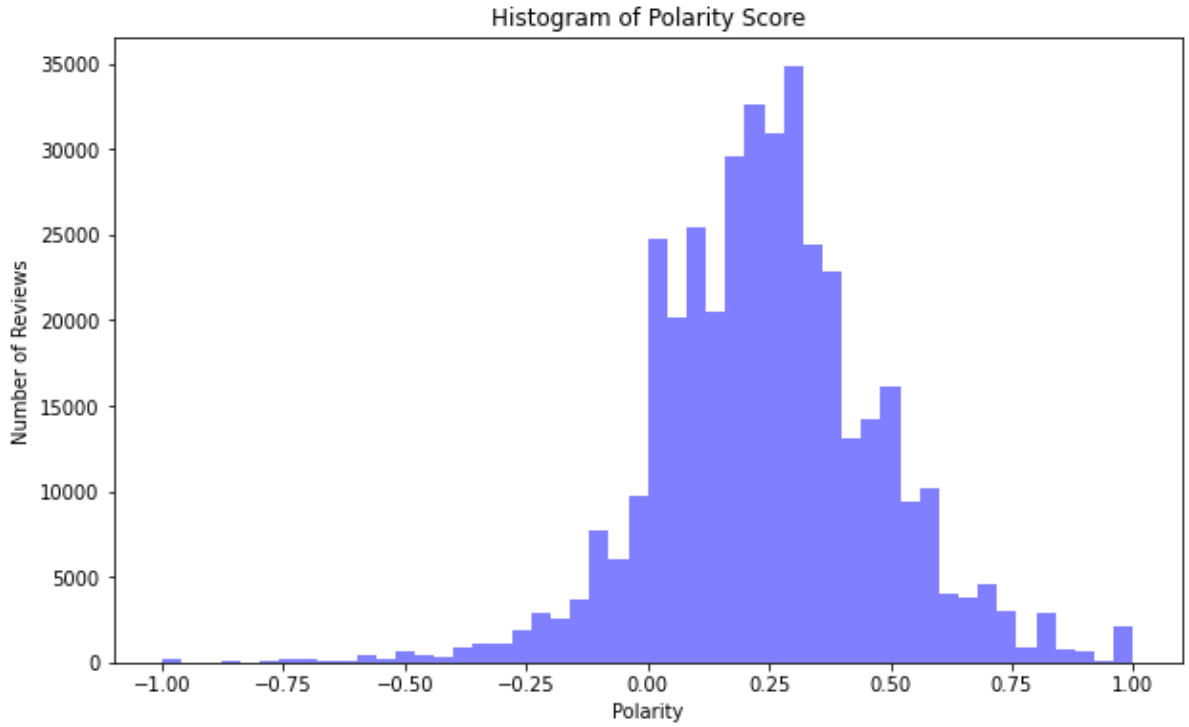

Figure 12: Distribution of polarity score for reviews

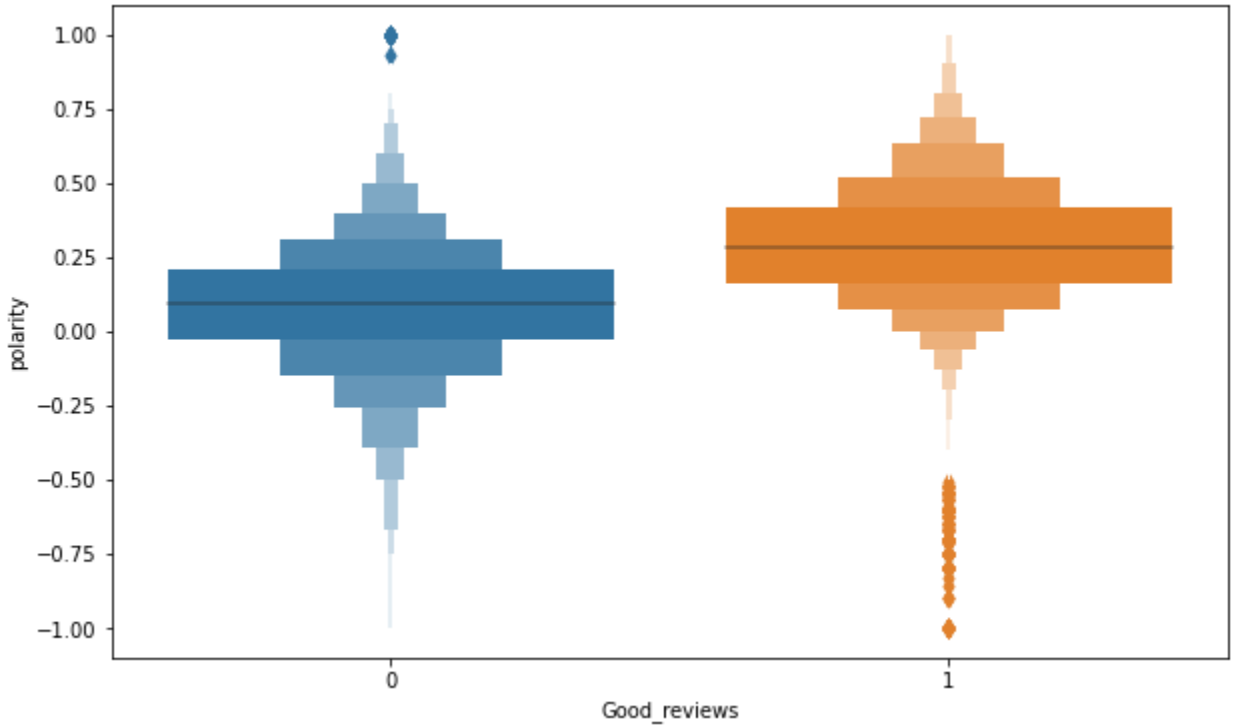

Figure 13: Boxplot of distribution of polarity regarding good or bad review.

In Figure 13, it can be observed that good reviews (Good reviews =1) have higher polarity compared to bad reviews. On the other hand, good reviews also have higher number of negative polarity reviews. This is an unbalanced data and number of good reviews are higher than bad reviews. Therefore, it is not much surprising to see more number of extreme values in this category.

As we can see from this box plot, we have some good reviews that has very low polarity (very negative), and some bad reviews that has high polarity (positive statement). Let's check some of them:

Table 6

Reviews that have polarity is 1 (most positive), Good review is 0 (bad reviews)

Reviews
‘product received advertised strawberry bags pack',
‘expecting terms companys reputation excellent home delivery products',
‘bought allot different flavors happens one favorites getting soon',
'deal healthiest salt use box last family year problem iodized sea salt raise blood pressure regular salt',
'trouble finding locally delivery fast hunting flour aisle local grocery stores',
'took one two get used pickle taste aim hooked keep bottle hand',
'found crisps local walmart figured would give try yummy may never go back regular chips big chip fan anyway problem eat entire bag one sitting give crisps big thumbs',
'use product daily provides steady stream energy get jittery crash helps practice portion control acts slight appetite suppressant',
'put husbands stocking christmas hit'

When Table 6 is examined, it can be seen that some of the reviews are actually positive but somehow got bad review score. Keep in mind that these are extreme case reviews and it is not surprising to see that it does not make much sense.

Punctuation vs Polarity: From Figure 14, we can see that polarity comes close to zero when score of extreme cases for polarity is shirking. A possible explanation for this is people who are paying more attention to punctuation tend to be more balanced in

their product evaluation. Despite outliers, average polarity score is almost a line, and it is around 0.25. This information is consistent with Figure 12.

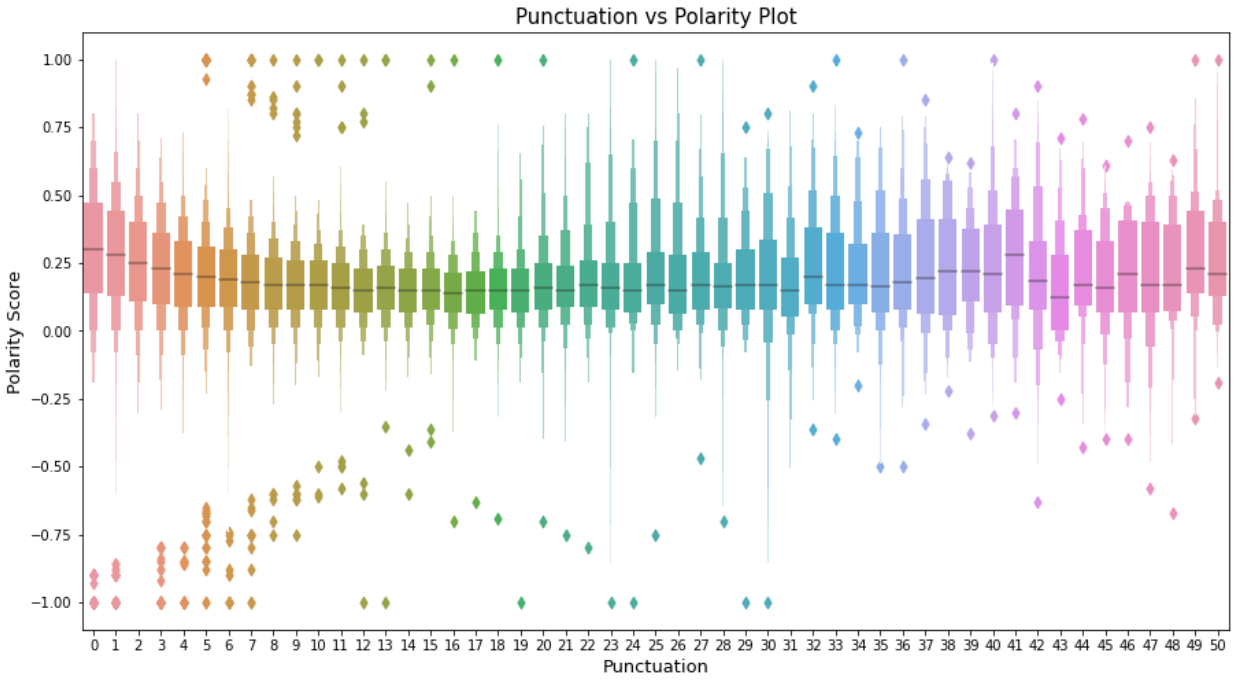

Figure 14: Number of punctuation and polarity score

Helpfulness vs Polarity: Figure 15 presents relation between *helpfulness* and *polarity* in *Good Reviews* category. There are interesting outliers. For example, some reviews have lowest polarity (most negative) but good rating (good review is 1) and helpfulness is more than 3. This is an interesting combination. In the Table 7, we can see that those reviews are not using negative words for the purchase. Those negative expressions are for comparison with other purchases. For now, NLP cannot handle with this kind of contextual usage of words.

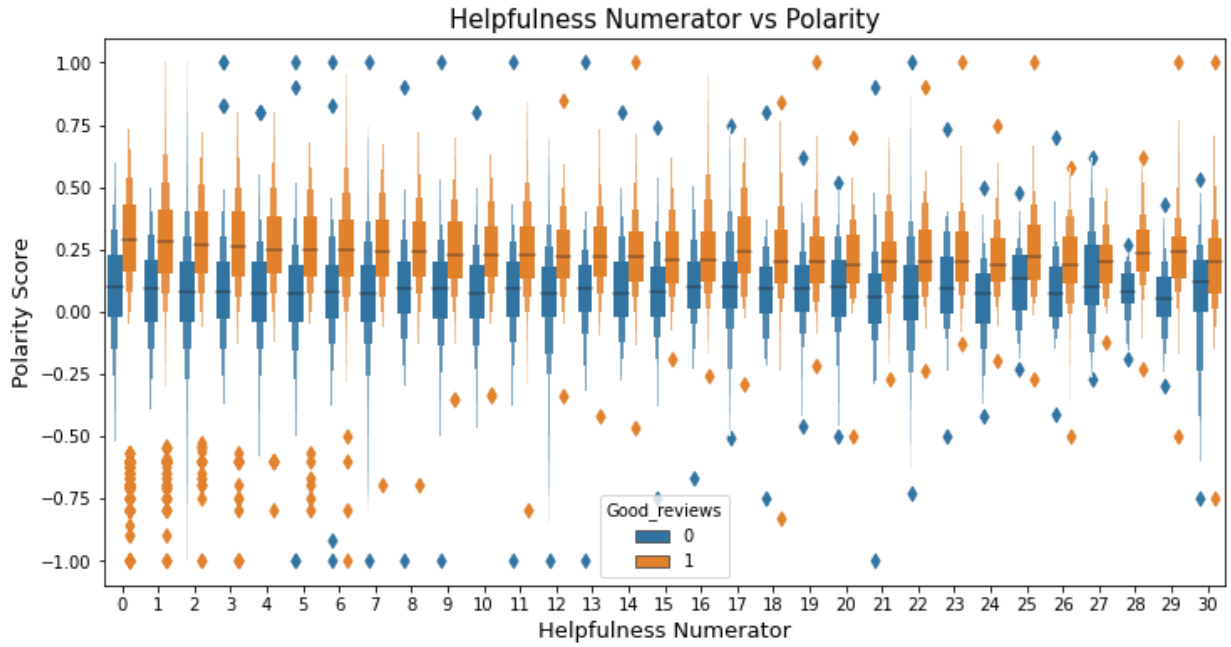

Figure 15: Helpfulness and Polarity in category of Good Reviews

Table 7

Reviews that have polarity is -1 (most negative), helpfulness score is more than 3, and Good review is 1.

Reviews
'brotherinlaw got hooked bariani olive oil terrificbr use almost every day like storebought brands almost gooey awful tasting recommend everyone well everyone wouldnt',
'forget highpriced energy tsps anything give energy youve ever imagined shocking',
'coffee greatthe price awful get thing bed bath beyond use one coupons',
'sf syrups taste awful one taste like expect bravo',

Polarity vs Number of Words: Number of words represent how many words exist in the review. In this aspect, Figure 16 offers a very interesting understanding. In Figure 16, I categorized number of words into 3 categories. If number of words is more than 50 and lower than 150, it tends to have low polarity score and bad review rating (Good review is 0). There can a couple reasons behind this. First, it looks like people do not bother to spend effort for writing too long for a bad review (more than 150 words). In

addition, writing review requires dedication of a reviewer because long reviews are time and effort consuming. This may be the reason for polarity scores near zero. On the top of that, if people write a bad review, they spend some effort on it give details to a certain degree. Thus, number of words can be between 50-150 for a negative review.

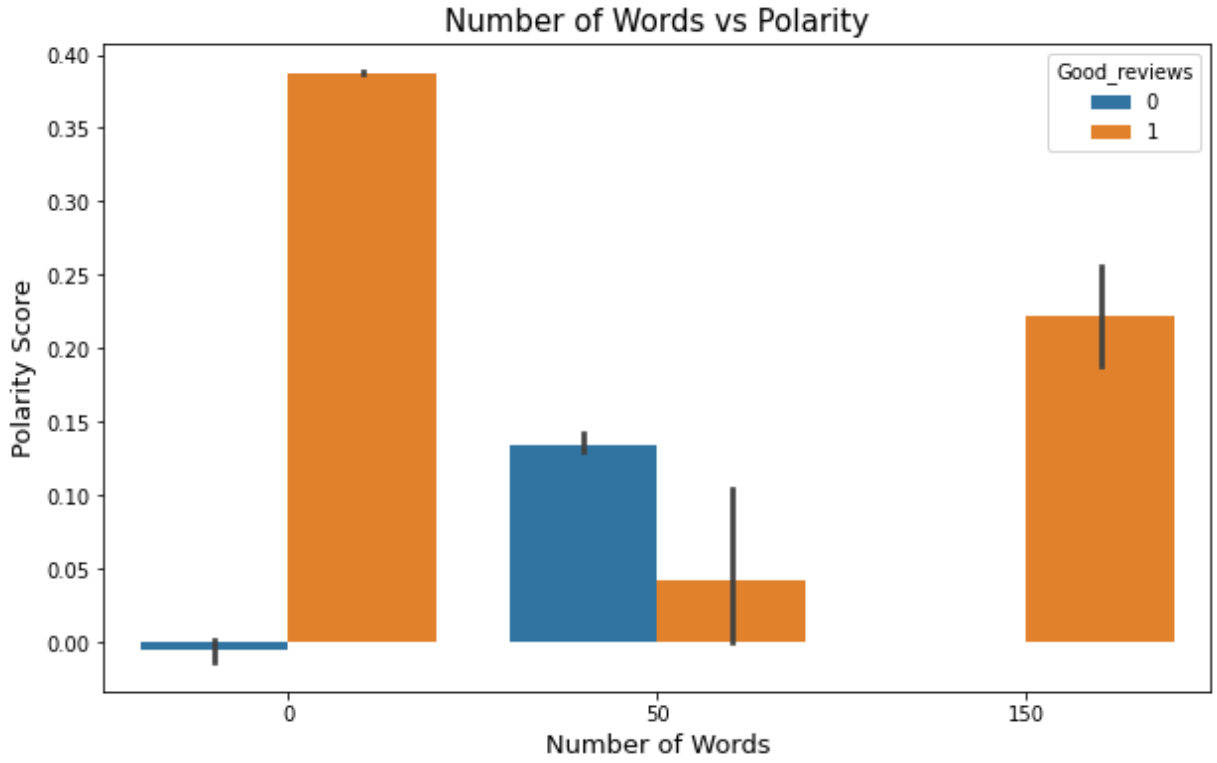

Figure 16: Polarity score and number of words hued to good reviews category

Subjectivity

Subjectivity used for individual sentences to determine whether a sentence expresses an opinion or not. In terms of subjectivity, textual information in the world can be broadly categorized into two main types: facts and opinions. In other words, Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties (Liu, 2010).

In sentiment analysis, *subjectivity* is also a float which lies in the range of [0,1]. When it is close to 0, it is more about facts. When it increases, it comes close to be an opinion. Distribution of subjectivity scores for reviews are similar to normal distribution (Fig. 17). When we look more deeply and add *polarity* and *Good Reviews* features to the plot, we get Figure 18. Figure 18 tells that subjectivity and polarity shows a funneling patterns to a certain degree. It can also be observed that low subjectivity scored reviews are also neutral reviews in terms of polarity.

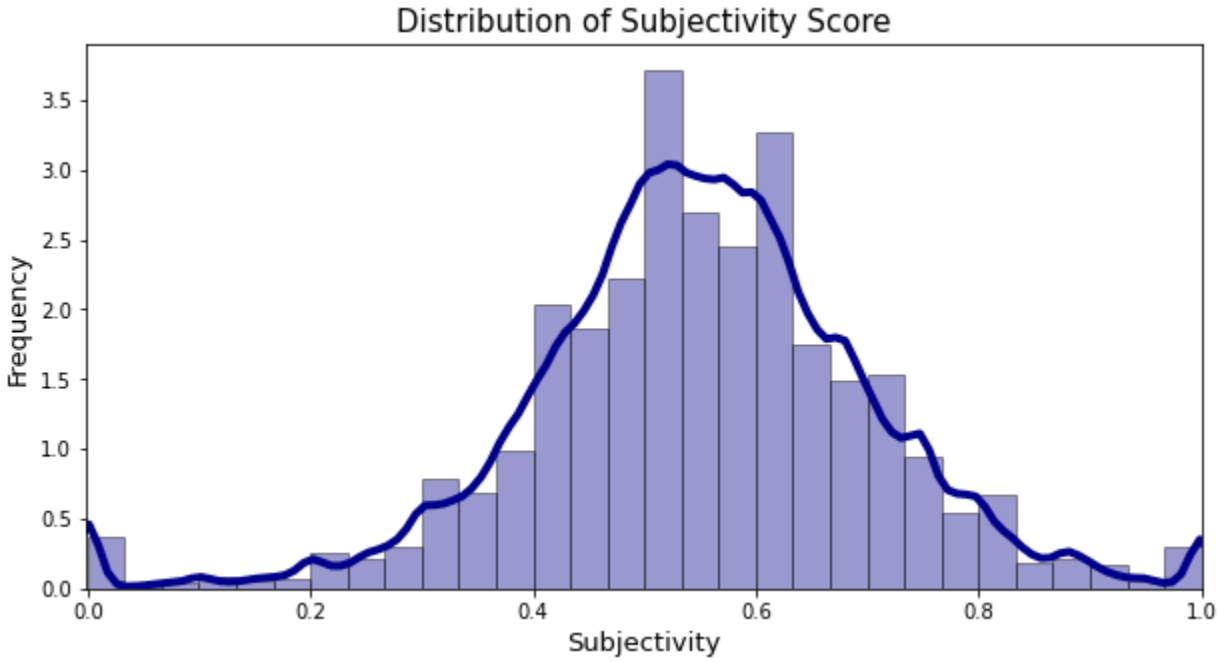

Figure 17: Distribution of subjectivity scores in Amazon reviews

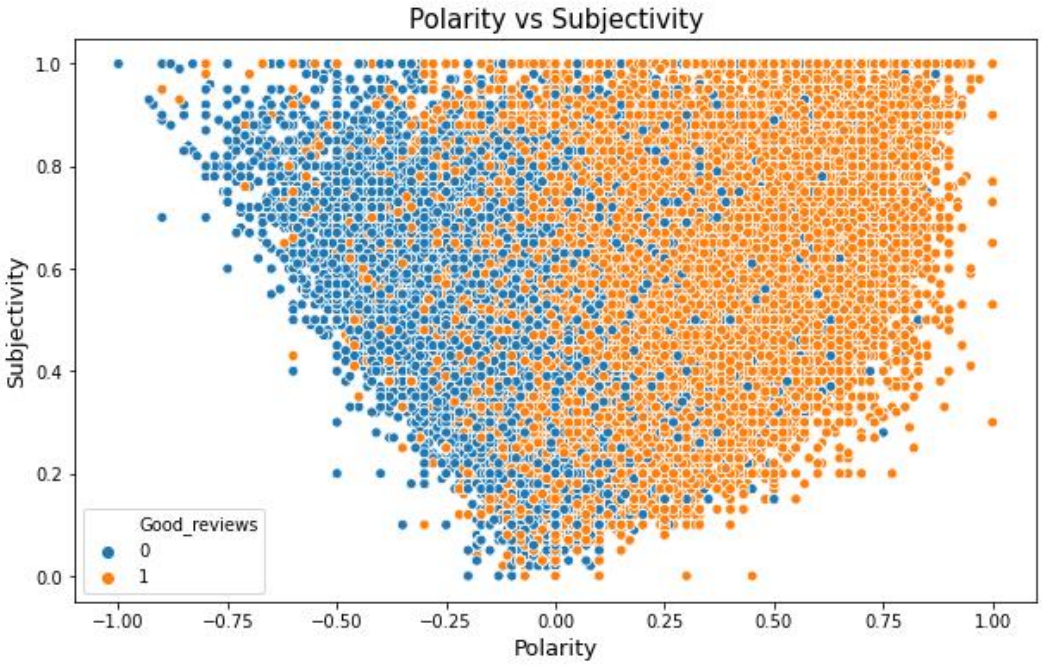

Figure 18: Subjectivity and polarity scores hued to good reviews category

When we check number of words and subjectivity (Fig. 19), it is hard to observe a relation between these two criteria. On the other hand, it can be observed that mean subjectivity score is slightly higher in positive reviews (Good reviews is 1).

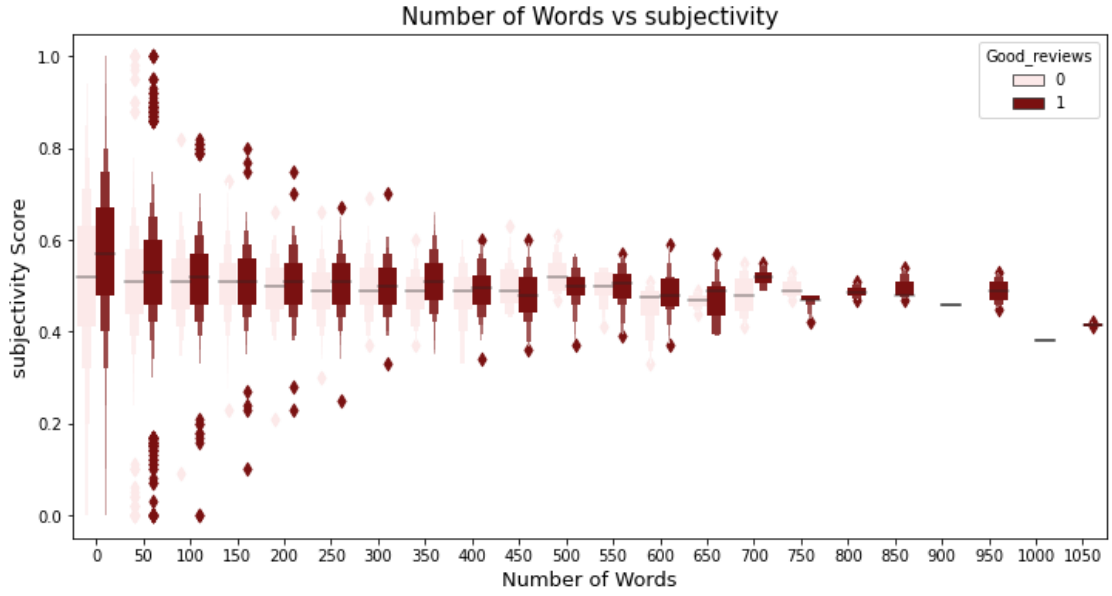

Figure 19: Number of Words and Subjectivity depending on Good Review Category

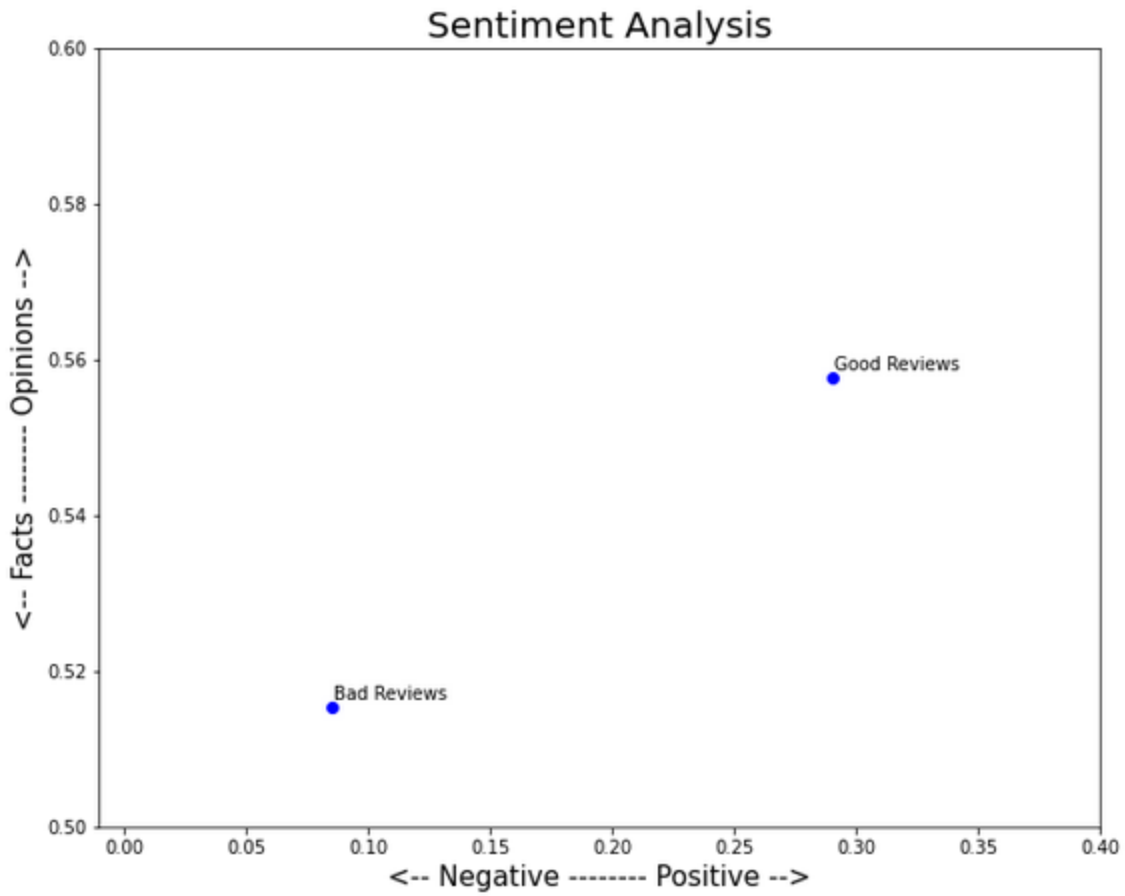

Figure 20: Mean value of polarity and subjectivity scores for review ratings

Figure 20 is a presentation of how polarity and subjectivity is affected by rating of the reviews (Good review feature). While reading this plot, we need to keep in mind that y-axis is in a very small range. Mean subjectivity score difference between two groups is negligible.

Examining extreme reviews: There are some reviews that can be considered as extreme case. For example, Table 8 presents 10 reviews that have highest polarity (most positive sentiment) but 'good review' value is 0, and most subjective (opinion). These tweets are hard to score for sentiment analysis algorithms. It is not surprising that they have most positive score (polarity =1).

Table 8

Reviews that have polarity is 1 (most positive sentiment), subjectivity is 1 (opinion) and bad rating (good review is 0)

Reviews
expecting terms companys reputation excellent home delivery products',
'used icicle two hours taking break plug laptop wont recognize everything could think fix even hours searching google nothing process return wouldve perfect worked',
'description tea lists organic ingredients tea received one ingredients organic say amazon awesome responding refunding think tea listed organic',
'damaged cans well eight cans dented box perfectly packaged undamamaged',
'havent quite determined yogi detox cause since drinking three days lots dizziness going drink week see relationship symptom lots ingredients tastes wonderful may interacting medication',
'flavors tried excellent one isnt wouldnt suggest',
'bought oatmeal target trying save money starbucks oatmeal price perfect diet trying cut sugar oatmeal would perfect separated fruit nuts sugar cinammon add depending people preference',
'like itbut impressive waywill stick grain bread use company',
'sardines excellent cautionthese packed lemon sauce lemon flavored olive oil variety without oil packed tomato sauce',
'product expired months didnt see anywhere product information impressed'

In order to understand how the data is shaped and how the sentiment analysis works, let's examine more tweets with different criteria (Table 9 and 10).

Table 9

5 sample reviews that have the highest polarity (most positive sentiment) and 'good review' value is 1, and most subjective (opinion):

Reviews
would thought would order candy mail product arrived heat september damage ordered candy favors wedding perfect',
'grateful find espresso capsules amazon used lost supplier wont go anywhere else excellent stuff',
'absolutely wonderful tasty filling two points ww plan flavors three staple house',
'wife tea drinker says tea flavorful blend vanilla spice tastes smells wonderful',
'product excellent baking size convenient well use oatmeal cookies trail mix etc etc'

Table 9

5 sample reviews that have the lowest polarity (most negative sentiment) but 'good review' value is 1, and most subjective (opinion):

Reviews
'son loves like cheese pufftype thing leave crumbs everywhere brother banned daughter resulting poop quite terrible',
'forget highpriced energy tsps anything give energy youve ever imaginedbr shocking',
'helps recovery time running keeps nasty lactic acid away',
'terrible awful dont buy means availability coffee need say please please dont buy',
'say chick biskit crackers kids raised em course nasty squirt cheeseyum'

Topic Modeling

Topic modeling is another popular text analysis technique. The ultimate goal of topic modeling is to find various topics that are present in your corpus. Each document in the corpus will be made up of at least one topic, if not multiple topics.

In this part, I will be covering the steps on how to do Latent Dirichlet Allocation (LDA), which is one of many topic modeling techniques. LDA was specifically designed for text data.

To use a topic modeling technique, you need to provide (1) a document-term matrix and (2) the number of topics you would like the algorithm to pick up. Our goal with topic modeling to find a theme across reviews, and discover hidden topics. In topic modeling, I will create different models and compare them. In the end, I will choose the topic model that makes most sense.

During LDA Topic modeling, researcher is the one who decides number of groups in the output. However, we do not know what is the best number of groups. Therefore, different number of groups will be obtained. Then, the one that makes most sense will be decided.

Once the topic modeling technique is applied, the researcher's job as a human is to interpret the results and see if the mix of words in each topic make sense. If they don't make sense, we can try changing up the number of topics, the terms in the document-term matrix, model parameters, or even try a different model.

Topic Modeling with Good reviews (Good reviews =1)

In this part, topic models in good ratings will be examined.

Topic modeling - Attempt 1 (with all the review data): No text filtering was applied in this process. By looking at Table 10, it can be said that Topic group 1 from Modeling with 2 Topics make most sense. The first group is about beverages, and the second group is about reactions. For now, this will stay here. Table 10 will be reconsidered after completing further steps of topic modeling.

Table 10

Topic Modeling with All Text Data

Topic Modeling With 2 Topics
Topic 1:
'0.020*"coffee" + 0.011*"like" + 0.010*"tea" + 0.010*"flavor" + 0.009*"good" + 0.009*"taste" + 0.007*"one" + 0.007*"great" + 0.006*"use" + 0.006*"cup"'
Topic 2:
'0.010*"great" + 0.009*"like" + 0.009*"good" + 0.007*"love" + 0.007*"food" + 0.007*"one" + 0.007*"product" + 0.005*"taste" + 0.005*"get" + 0.005*"amazon"'
Topic Modeling With 3 Topics
Topic 1:
'0.016*"coffee" + 0.011*"like" + 0.010*"good" + 0.010*"great" + 0.008*"flavor" + 0.008*"taste" + 0.006*"use" + 0.006*"one" + 0.005*"love" + 0.005*"chips"'
Topic 2:
'0.013*"like" + 0.013*"tea" + 0.010*"taste" + 0.010*"good" + 0.009*"coffee" + 0.009*"flavor" + 0.008*"one" + 0.008*"great" + 0.008*"chocolate" + 0.006*"love"'
Topic 3:
'0.012*"food" + 0.010*"product" + 0.009*"great" + 0.007*"one" + 0.007*"love" + 0.007*"amazon" + 0.007*"good" + 0.006*"dog" + 0.006*"price" + 0.006*"get"'
Topic Modeling With 4 Topics
Topic 1:
'0.010*"like" + 0.008*"great" + 0.008*"good" + 0.007*"use" + 0.007*"sauce" + 0.006*"flavor" + 0.006*"taste" + 0.006*"oil" + 0.005*"make" + 0.005*"seeds"'
Topic 2:
'0.017*"food" + 0.008*"dog" + 0.008*"product" + 0.007*"one" + 0.007*"like" + 0.005*"use" + 0.005*"great" + 0.005*"good" + 0.004*"cat" + 0.004*"milk"'
Topic 3:
'0.012*"great" + 0.010*"good" + 0.009*"love" + 0.009*"amazon" + 0.008*"like" + 0.008*"find" + 0.008*"price" + 0.007*"get" + 0.007*"one" + 0.007*"product"'
Topic 4:
'0.029*"coffee" + 0.014*"like" + 0.014*"tea" + 0.012*"taste" + 0.011*"good" + 0.011*"flavor" + 0.008*"one" + 0.008*"cup" + 0.008*"great" + 0.007*"drink"'

Topic Modeling - Attempt 2 (Nouns only): In this step, only nouns were were used for creating topics by using the LDA method.

Table 11

Topic Modeling with Nouns

Topic Modeling With 2 Topics
Topic 1:
0.016*"product" + 0.013*"taste" + 0.011*"flavor" + 0.010*"chocolate" + 0.008*"price" + 0.007*"cookies" + 0.007*"chips" + 0.006*"eat" + 0.006*"use" + 0.006*"store"
Topic 2:
0.040*"coffee" + 0.017*"tea" + 0.015*"food" + 0.014*"taste" + 0.013*"flavor" + 0.010*"product" + 0.009*"cup" + 0.009*"water" + 0.009*"price" + 0.008*"dog"
Topic Modeling With 3 Topics
Topic 1:
0.034*"food" + 0.020*"product" + 0.013*"dog" + 0.010*"treats" + 0.010*"dogs" + 0.008*"price" + 0.008*"mix" + 0.007*"seeds" + 0.007*"use" + 0.006*"cat"
Topic 2:
0.015*"taste" + 0.012*"flavor" + 0.012*"tea" + 0.011*"product" + 0.010*"chips" + 0.010*"cookies" + 0.010*"bag" + 0.009*"price" + 0.008*"order" + 0.007*"snack"
Topic 3:
0.056*"coffee" + 0.022*"flavor" + 0.021*"taste" + 0.015*"chocolate" + 0.013*"cup" + 0.012*"water" + 0.011*"tea" + 0.010*"sugar" + 0.010*"product" + 0.009*"use"
Topic Modeling With 4 Topics
Topic 1:
0.029*"product" + 0.017*"price" + 0.014*"chocolate" + 0.012*"cookies" + 0.011*"use" + 0.011*"store" + 0.011*"taste" + 0.009*"flavor" + 0.009*"order" + 0.008*"amazon"
Topic 2:
0.016*"milk" + 0.010*"nuts" + 0.010*"peanuts" + 0.009*"candy" + 0.007*"chocolate" + 0.006*"cake" + 0.005*"almonds" + 0.004*"work" + 0.004*"mint" + 0.004*"party"
Topic 3:
0.060*"coffee" + 0.026*"tea" + 0.025*"food" + 0.014*"cup" + 0.013*"taste" + 0.012*"flavor" + 0.012*"dog" + 0.010*"treats" + 0.010*"dogs" + 0.008*"use"
Topic 4:
0.022*"taste" + 0.020*"flavor" + 0.015*"chips" + 0.014*"sugar" + 0.011*"water" + 0.011*"snack" + 0.010*"flavors" + 0.008*"bag" + 0.008*"juice" + 0.008*"salt"

Topic Modeling - Attempt 3 (Nouns and Adjectives): In this step, only nouns and adjectives were used for creating topics by using the LDA method.

Table 12

Topic Modeling with Nouns and Adjectives

Topic Modeling With 2 Topics
Topic 1:
0.020*"coffee" + 0.015*"great" + 0.015*"good" + 0.011*"food" + 0.010*"product" + 0.008*"price" + 0.007*"taste" + 0.006*"time" + 0.006*"flavor" + 0.006*"amazon"
Topic 2:
0.014*"tea" + 0.011*"good" + 0.011*"taste" + 0.010*"flavor" + 0.009*"great" + 0.008*"hot" + 0.008*"chocolate" + 0.007*"sugar" + 0.007*"water" + 0.006*"product"
Topic Modeling With 3 Topics
Topic 1:
0.017*"great" + 0.014*"good" + 0.014*"product" + 0.009*"price" + 0.008*"chocolate" + 0.008*"cookies" + 0.008*"taste" + 0.008*"amazon" + 0.006*"order" + 0.006*"time"
Topic 2:
0.035*"coffee" + 0.017*"tea" + 0.014*"good" + 0.014*"flavor" + 0.013*"taste" + 0.010*"great" + 0.010*"cup" + 0.008*"water" + 0.006*"use" + 0.006*"hot"
Topic 3:
0.018*"food" + 0.012*"great" + 0.011*"good" + 0.008*"dog" + 0.007*"chips" + 0.006*"product" + 0.006*"eat" + 0.006*"bag" + 0.006*"little" + 0.006*"healthy"
Topic Modeling With 4 Topics
Topic 1:
0.055*"coffee" + 0.013*"good" + 0.013*"cup" + 0.012*"great" + 0.011*"flavor" + 0.010*"tea" + 0.009*"taste" + 0.007*"product" + 0.006*"use" + 0.006*"price"
Topic 2:
0.015*"good" + 0.014*"tea" + 0.013*"taste" + 0.013*"great" + 0.011*"chocolate" + 0.010*"sugar" + 0.010*"flavor" + 0.009*"milk" + 0.009*"water" + 0.008*"product"
Topic 3:
0.017*"great" + 0.015*"good" + 0.012*"flavor" + 0.011*"taste" + 0.010*"chips" + 0.008*"snack" + 0.007*"bag" + 0.007*"salt" + 0.006*"delicious" + 0.006*"oil
Topic 4:
0.023*"food" + 0.014*"product" + 0.010*"great" + 0.010*"dog" + 0.010*"good" + 0.008*"price" + 0.007*"time" + 0.007*"treats" + 0.007*"dogs" + 0.006*"old"

Now, in the final stage, Table 11, Table 12 and Table 13 needs to be evaluated. We need to ask our self 'which group makes more sense?'. Out of the 9 topic models we looked at, nouns only, 3 topic one made the most sense to me (Table 11). I see three distinct groups here: (1) pet items, (2) cookies and snacks, and (3) beverage.

So, let's pull that down here and run it through some more iterations to get more fine-tuned topics.

Table 13

Final Topic Modeling with Fine Tuned Parameters with Nouns Only

Topic Modeling With 3 Topics
Topic 1:
'0.053*"coffee" + 0.025*"tea" + 0.020*"taste" + 0.019*"flavor" + 0.013*"cup" + 0.011*"chocolate" + 0.011*"water" + 0.011*"product" + 0.009*"price" + 0.008*"use"
Topic 2:
0.034*"food" + 0.016*"product" + 0.013*"dog" + 0.012*"cookies" + 0.010*"dogs" + 0.009*"treats" + 0.008*"price" + 0.008*"mix" + 0.008*"milk" + 0.007*"seeds"
Topic 3:
0.016*"flavor" + 0.014*"taste" + 0.014*"chips" + 0.013*"product" + 0.009*"price" + 0.008*"bag" + 0.008*"snack" + 0.007*"order" + 0.007*"store" + 0.007*"salt"

In Table 13, I see similar groups in different orders. By considering all the steps of topic analysis with LDA method, it can be concluded that good reviews (Good Reviews is 1) can be categorized into three main topics: (1) beverages, (2) pet items, and (3) cookies and snacks.

For this study, same steps of topic modeling also run for bad reviews (Good reviews is 0). In the end, out of 9 topics, nouns only, 2 topics model made most sense. The prevailing topics are (1) pet items, and (2) beverages. As a result of the topic modeling, it can be seen that reviewers are complaining and praising for almost same products because prevailing topics are same for both good reviews and bad reviews.

Machine Learning Models

This section used machine learning models for the data analysis. The data set is a supervised data which refers to fitting a model of dependent variables to the independent variables, with the goal of accurately predicting the dependent variable for future observations or understanding the relationship between the variables (Gareth, Daniela, Trevor, & Tibshirani, 2013). In relation to the data set, literature suggests below listed methods can be appropriate.

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bernoulli Naive Bayes
- Complement Naive Bayes
- Logistic Regression

I will apply four different methods in Naive Bayes Model. Evaluating these different Naive Bayes models is also suggested if time permits (Pedregosa et al., 2011). Therefore, I will evaluate different Naive Bayes models. As performance evaluation metrics, accuracy, precision and recall rate will be calculated. Receiving Operating Characteristic (ROC) Curve will be drawn and models' performance will be compared.

Interpretation of Metrics

Before starting to create models, here is a brief definition for model evaluation metrics accoding to Scikit-Learn website (Pedregosa et al., 2011):

Accuracy: Accuracy classification score. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

Precision: Intuitively the ability of the classifier not to label as positive a sample that is negative. Precision is the estimated probability that a randomly selected retrieved document is relevant (Cakir et al, 2019).

Recall: Intuitively the ability of the classifier to find all the positive samples. Recall is the estimated probability that a randomly selected relevant document is retrieved in a search (Cakir et al, 2019).

F1 Score: The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

The precision-recall curve: This shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

Gaussian Naive Bayes

This model assumed that the likelihood of the features is to be normal distribution. For our dataset, this assumption does not hold. On the other hand, checking Gaussian Naive Bayes model does not hurt our study. Table 14 presents model performance of the model. Accuracy is very low for our unbalanced data.

Figure 21: Confusion matrix for Gaussian NB model

Table 14

The Results of the Gaussian Naive Bayes

Evaluation	Score (%)
Accuracy	77
Precision	94
Recall	95
f1	77

Multinomial Naive Bayes

Multinomial Naive Bayes implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice (Pedregosa et al., 2011).

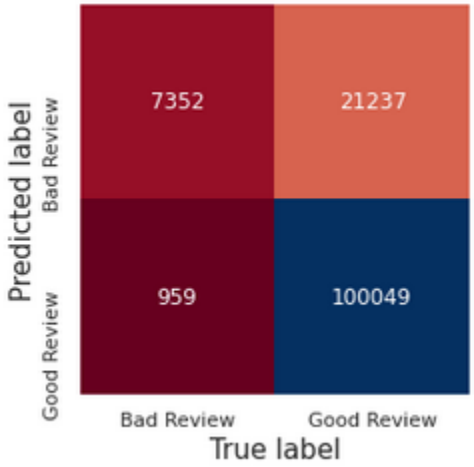

Figure 22: Confusion matrix for Multinomial NB model

Table 15

The Results of the Multinomial Naive Bayes

Evaluation	Score (%)
Accuracy	83
Precision	82
Recall	99
f1	83

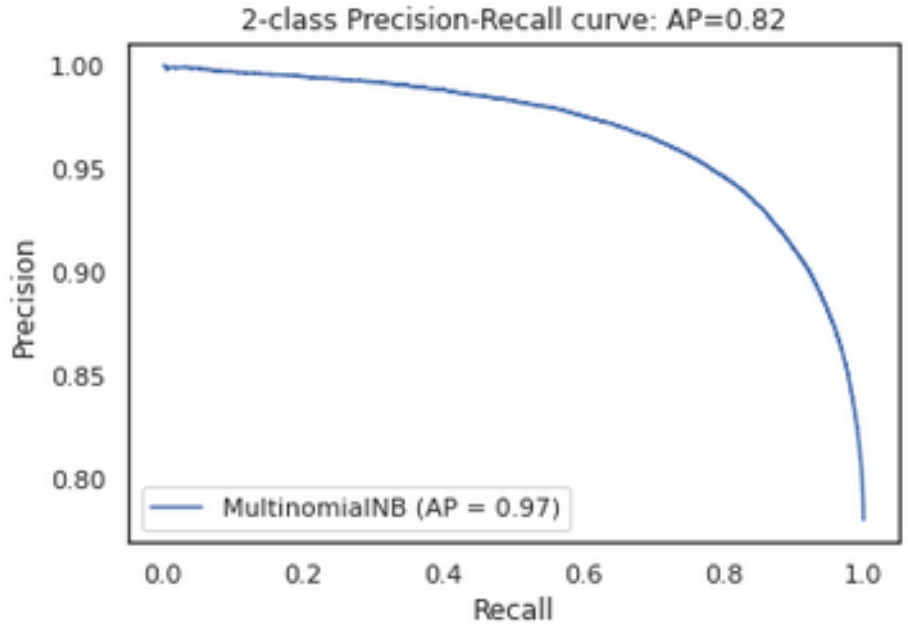

Figure 23: Precision-recall curve for Multinomial NB

Bernoulli Naive Bayes

Bernoulli Naive Bayes implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors (Pedregosa et al., 2011). In this aspect, we can expect that Bernoulli Naive Bayes model can show a good performance. Table 16 supports our expectation to a certain degree.

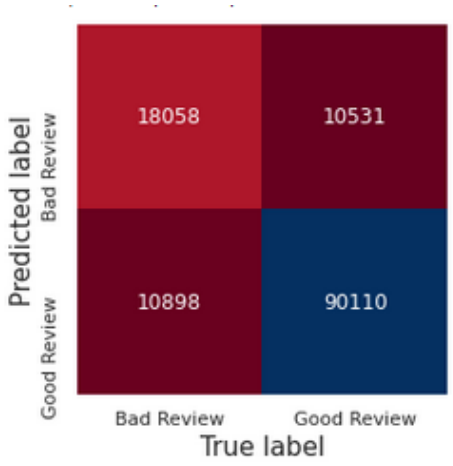

Figure 24: Confusion matrix for Bernoulli NB model

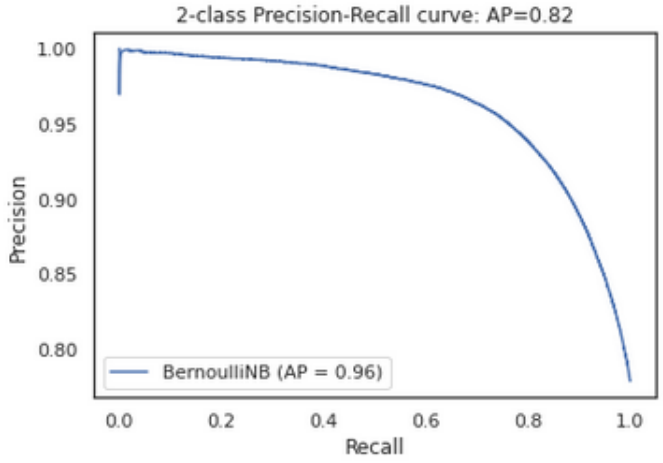

Figure 25: Precision-recall curve for Bernoulli NB

Table 16

The Results of the Bernoulli Naive Bayes

Evaluation	Score (%)
Accuracy	83
Precision	90
Recall	89
f1	83

Complement Naive Bayes

The Complement Naive Bayes classifier was designed to correct the "severe assumptions" made by the standard Multinomial Naive Bayes classifier. It is particularly suited for imbalanced data sets (Pedregosa et al., 2011).

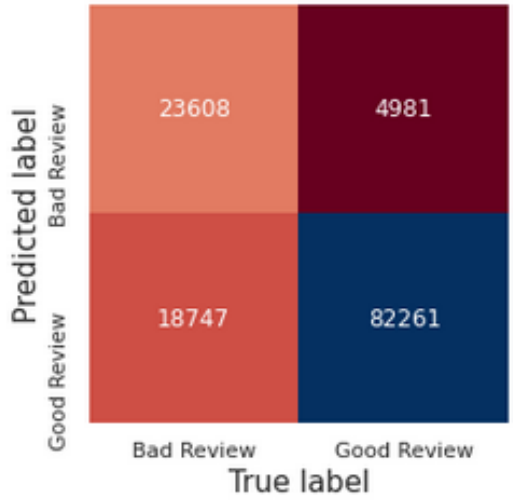

Figure 26: Confusion matrix for Complement NB model

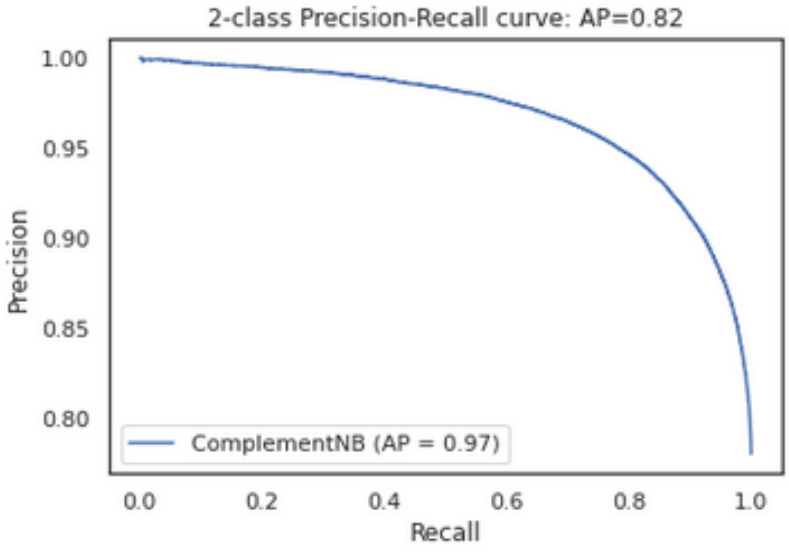

Figure 25: Precision-recall curve for Complement NB

Table 17

The Results of the Complement Naive Bayes

Evaluation	Score (%)
Accuracy	82
Precision	94
Recall	81
f1	82

Logistic Regression

When we have binary response category such as Yes and No, logistic regression models the probability that Y belongs to a particular category (James et al., 2013).

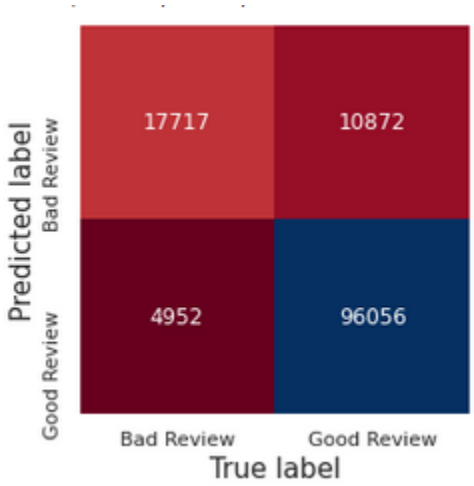

Figure 28: Confusion matrix for Logistic Regression model

Table 18

The Results of the Gaussian Naive Bayes

Evaluation	Score (%)
Accuracy	88
Precision	90
Recall	95
f1	88

Comparison of Models

When we examine Table 19, it can be seen that Logistic Regression Model shows a good classification performance overall. Performance of Naive Bayes models also close to logistic regression. Even in some metrics, Naive Bayes models outperforms the logistic regression. After this point, Receiver Operating Characteristic (ROC) curve will be investigated. Final model selection will be done based on ROC curve.

Table 19

Comparison of Model Performances

Measure (%) / Model	Gaussian NB	Multinomial NB	Bernoulli NB	Complement NB	Logistic Reg
Accuracy	77	83	83	82	88
Precision	94	82	90	94	90
Recall	95	99	89	81	95
f1	77	83	83	82	88

Receiver Operating Characteristic (ROC) Curve

This metric evaluates classifier output quality. ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better (Pedregosa et al., 2011).

Table 20

Area Under ROC Curve (AUC)

Evaluation	Score (%)
Gaussian NB	84
Multinomial NB	90
Bernoulli NB	88
Complement NB	90
Logistic Reg	92

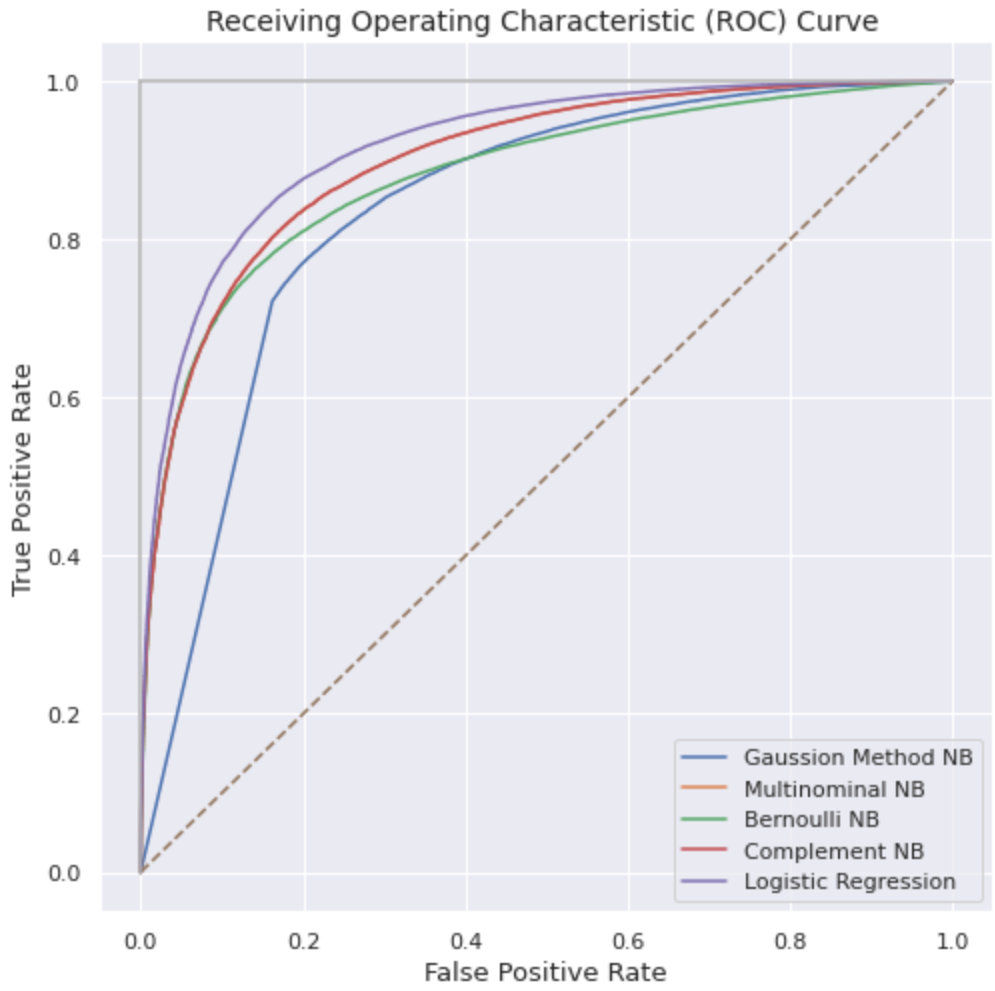

Figure 26: ROC curve for all models used in classification

Now, interpretation of ROC curve helps us to determine which model is the best. In Figure 26, the "steepness" of ROC curves is important, since it is ideal to maximize the true positive rate while minimizing the false positive rate. In this figure, it can be observed that highest true positive rate and lowest false positive rate is performed by logistic regression model. This can also be seen by comparing AUC scores. Logistic regression model has highest AUC score: 92 (Table 20).

Multinomial and Complement Naive Bayes models' performance are same. Their Area Under Curve (AUC) scores are 90 (Table 20). That's why they overlap on the plot (Fig. 26). We can also see the color of the Complement NB.

Sanity Check

Running sanity checks is an important part of the data analysis process. The final analysis is only as accurate as your data, which means it's worth it to spend some time to validate the data's accuracy and completeness. For sanity check, I will take random sample from data and check their predicted and actual *Good Reviews* value. While making sanity check, the best performing model, logistic regression model, were used for prediction. In other words, predicted values were gathered from logistic regression.

In Table 21, we can observe that model prediction could solve hard to understand reviews successfully. On the other hand, we see that there is a wrong prediction in ten random review sample. In Table 22, there are random samples of wrong predictions that their actual value is 0, but and Predicted Value is 1. In table 24, there are random samples of wrong predictions that their actual value is 1, but and Predicted Value is 0. Examining these wrong predictions may help us to understand what went wrong with model, what affected the model's performance negatively.

Table21

5 Random Reviews with Their Actual and Predicted Good Review Values

Actual Value	Predicted Value	Review
1	1	product tried true although delicious cereals astounding property calming effect literally anyone eats find enjoying bowl gorilla munch simply laughing jabs may thrown edge another daybr know say box table keeps jimmies stablebr recommended anyone',
1	1	'great product good price fast shipping product retails stores much higher price thats lucky enough find great deal would highly recommend seller',
1	1	'excellent pine apple juice price higher amazon whole foods sells lakewood pine apple juice sale compared amazon price reason per bottle price difference',
1	1	'green mountain french roast closest whole bean ground get k cup tried nearly k cup french roast favorite distinct flavor smoky taste true french roast consider breakfast coffee hot milk love latte bolder dinner satisfy taste coffee mountain coffee french roast kcups keurig brewersa',
1	0	'product nice flavor however somewhat granular regular splenda dont think would baking purposesbr overallnot bad though'

Table22

3 Random Reviews with Their Actual Value is 0 but and Predicted Value is 1

Actual Value	Predicted Value	Review
0	1	'lavazza coffee crema e gusto goodbut expires monthbr bought four knowing thisyou post expiration dates foodnow three blocks expired coffee great folks',
0	1	'based reiviews purchased creamer loved idea refridgeration tastes great exactly like refrigerated liquid version untill use purchase againi picky morning coffee',
0	1	'live middle nowhere want try new food products read often come amazoncom find products arrive felt duds recently read raving review hungry girl product decided give shot wow good may made corn tasted like rice cake good one plus cracker wrinkly brain appearanceits kind creepy wanting judge gave husband try loves snack foods eat anything tasted couple said dont need bother threw rest box away arent worth calories',

While obtaining the data in Table 22, I run this query many times with random samples and examined the reviews. Here are some examples from the query, and my explanation for what possibly went wrong:

Table23

Examples of Reviews that are Mistakenly Predicted as Positive

Review	Possible Reason for Wrong Prediction
'sweet chocolate taste tried various family members appeal would definately recommend'	The buyer's rating mistake
'bought looking good glutenfree cereal theyre decent great little sweet going take get many boxes'	The buyer's rating mistake
'decent coffee ridiculously overpriced starbucks ashamed charging buck half cup coffee made home'	A hard case for the model
'soup thick pleasant way bland one least favorite cups dr mcdougalls'	A hard case for the model
expecting spicier regular wasabi peas actually less spicy good flavor would order'	A hard case for the model
make mistake chocolate bar sweet maybe creamy overwhelmed sweetness couldnt taste anything else according label theres sugar within per chocolate decide'	A hard case for the model

Table24

3 Random Reviews with Their Actual Value is 1 but and Predicted Value is 0

Actual Value	Predicted Value	Review
1	0	'celsius green tea raspberry acai vitamin enriched beverage contains good stuff none bad stuff sugar aspartame preservatives high fructose corn syrup artificial colors flavors however tradeoff simple flavor drink weak tasting tart aftertaste doesnt taste bad doesnt taste good either drink claim burn calories per give energy think primarily vitamins cant confirm actually burns calories however say drinking product several daysi drank one morningi lose couple pounds bad effects like get energy drinks',
1	0	'received three english breakfast teas instead assortment ordered great tea wrong description kept normally one return things',
1	0	'read reviews bought tasted sample variety pack good taste youre reading youre interested prob taste buy',

Table 24 shows random samples of wrong predictions that their actual value is 1, but and Predicted Value is 0. Similar to Table 22, while obtaining the data in Table 24, I run this query many times with random samples and examined the reviews. And, I tried to understand the reason behind the wrong prediction. Table 25 offer my explanation for the wrong predictions.

Table25

Examples of Reviews that are Mistakenly Predicted as Positive

Review	Possible Reason For Wrong Prediction
'needs tried brands taste disgusting one actually tastes like candy'	Urban language
product taste good prefer product glass container instead plastic product made listed label disappointed'	A hard case for the model
'happy productit advertised larger size actually customer service processed refund issued apology'	Returned item, forgiving buyer
'love amazonbut product showed less months dating hot cocoa tastes great shocked short date even upsetting see nonreturnable item make sure read return policy make sure ready drink lot hot cocoa short period time',	A hard case for the model
'product joint rescue enjoyed beloved pekinese however expiration date soon disappointing'	Forgiving buyer
received three english breakfast teas instead assortment ordered great tea wrong description kept normally one return things'	Forgiving buyer

In Table 23, we see that buyers sometimes give bad ratings even though they are satisfied with purchase. This can cause error in prediction in text analysis. In Table 25, we also that even though buyers write a negative review, they be sometimes forgiving and give good ratings. In some cases, they are changing their rating if they get refund. One common reason for misclassification is that the reviews can contain both positive and negative sentiments. Text analysis algorithms can have seriously hard to deal with this kind of complicated reviews.

In English, some words can be used for both good and bad things such as 'awful', 'crazy' etc. Classification also could not understand contextual meaning of this kind of words. Lastly, sarcastic language was another reason for the misclassification. In sarcasm, buyers meant the opposite meaning of the words. However, NLP tools that was used in this study is vulnerable to sarcastic usage of words.

Limitations of the Study

During data cleaning, all the emojis and emoticons were cleaned. However, emijos and emoticons may provide a substantial understanding on text analysis. Linguistic and socio-linguistic interpretation of emojis may be useful in sentiment analysis (Guibon et al., 2016). In this aspect, there is a room for improvement at this point of the analysis.

Another suggestion for further studies is that stemming and lemmatization can be used. For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. But, of course stemming and lemmatization have a big drawback: they change the text itself. This hurt some studies seriously. For this study, I didn't prefer to apply these processes. However, for the further step, they can be applied and results can be compared.

For dealing with misspelled words, I used TextBlob.correct() method during data cleaning. However, there were still misspelled words in the data. I believe there is a room for improvement for spelling correction.

Lastly, I believe that it is possible to get good classification by using features other than *Text* in the dataset such as *number of words, helpfulness numerator, subjectivity, polarity.* I did my predictions only using text data and focused on NLP based methods. However, random forest, lasso, ridge, xgboost may also give good classification by using other features available in the dataset.

References

Cakir, F., He, K., Xia, X., Kulis, B., & Sclaroff, S. (2019). Deep metric learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1861-1870).

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345-354.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009, April). How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web* (pp. 141-150).

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, *49*(10), 1407-1424.

Dhanasobhon, S., Chen, P. Y., Smith, M., & Chen, P. Y. (2007). An analysis of the differential impact of reviews and reviewers at Amazon. com. *ICIS 2007 Proceedings*, 94.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
Guibon, G., Ochs, M., & Bellot, P. (2016, June). From emojis to sentiment analysis.

Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, *42*, 161-168.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer.

Joho, J (2019, 08). *the surprising reasons we turn off autocaps and embrace the lowercase.* mashable. https://mashable.com/article/disable-auto-caps-lowercase-texting-online-communication/

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627-666.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-commerce*, *11*(2), 23-25.