Predicting Next Review Sentiment

Gloria P. Moore

DSC680_Applied Data Science

Bellevue University

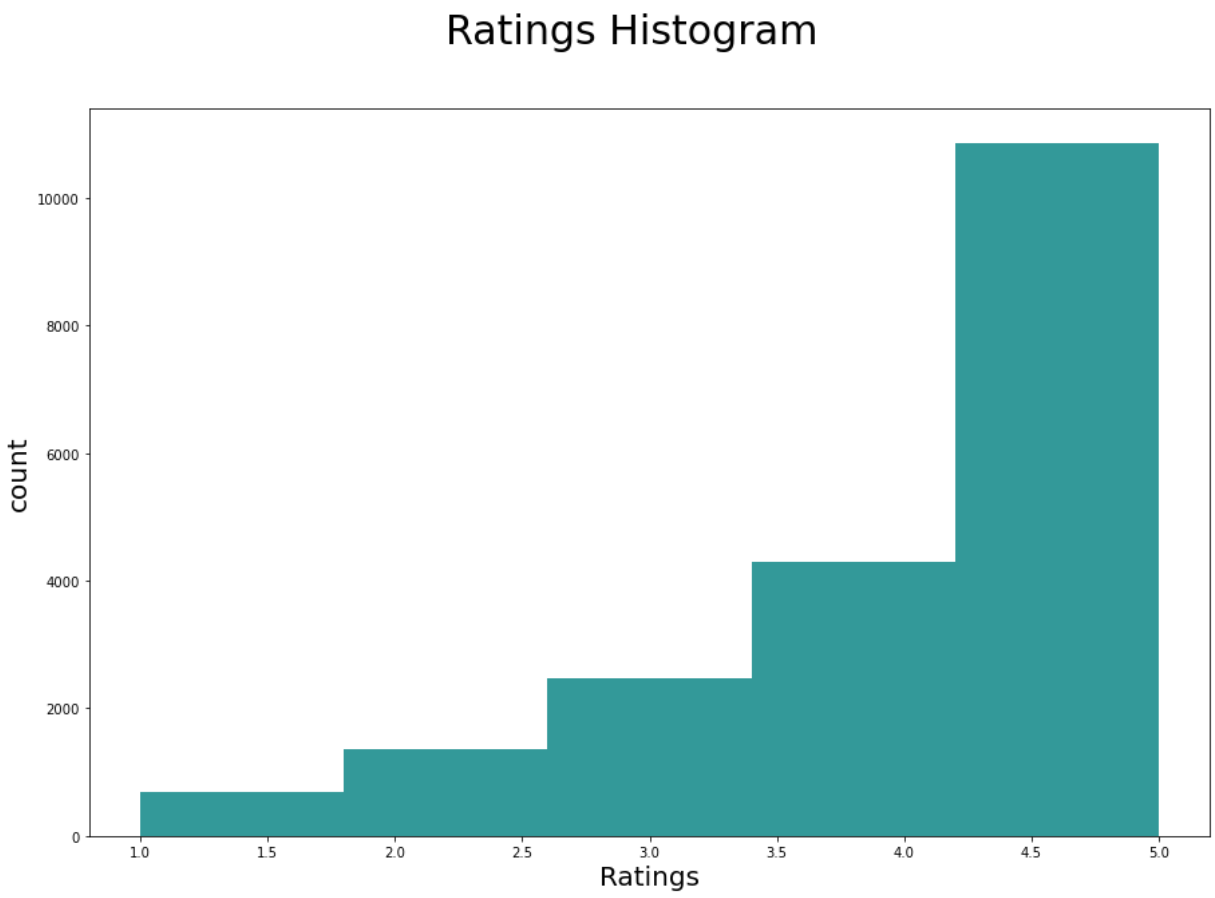
Bellevue University Data Science Master's Program

Abstract

The following project is based on the sentiment analysis of reviews of more than 19000 clients of a clothes store, these reviews have corresponding ratings according to the client experience, and the goal of the following work is based on these reviews, to build a model to predict sentiment of the next review, and check if this model can be used with reviews of other line of business. The clothes store review is also classified by product and department, this time graph analysis was done over some of the variables giving us as a general result that the store has more positive reviews than negative, Vader from the NLTK library was used to determine sentiment of the reviews, these results were translated to positive, negative and neutral results, and used as labels to build a naïve bayes model to classify and predict probability of the sentiment of the next review. Other models like KNN and clustering might be used in later for this project to compare results

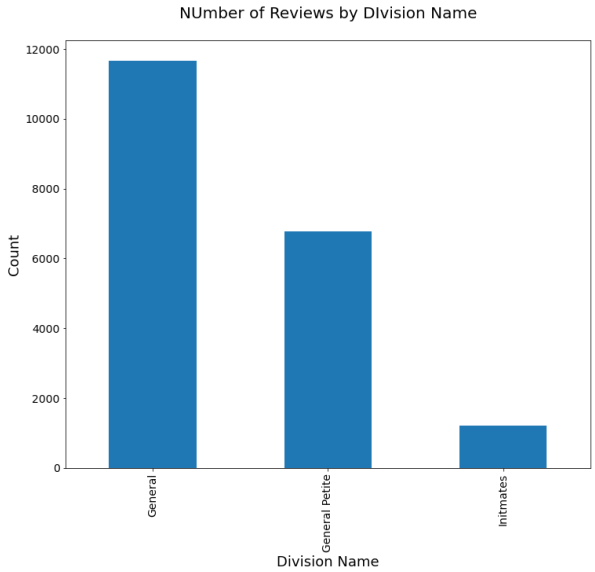
Method

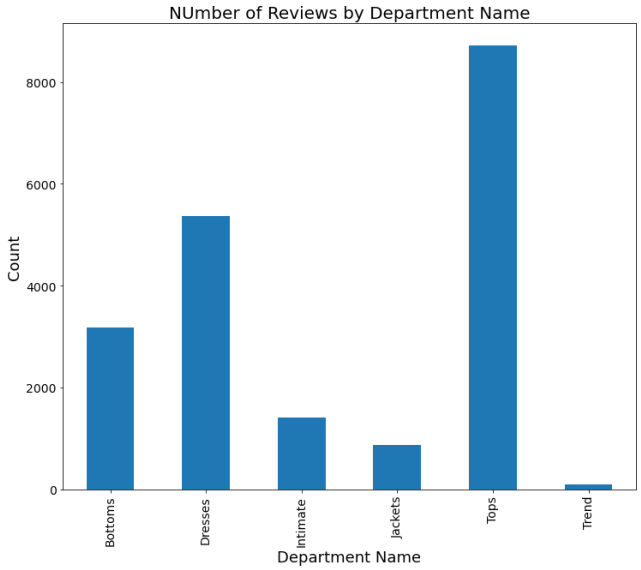
The method used so far for this project has been Vader sentiment analysis, the data set was read and transform to a data frame were the reviews were under the “ reviews” column, following this, missing values were check, finding around 2000 missing reviews, these observations were dropped, ending with a data frame of 19662 rows. Graph analysis was done for some of the variables:

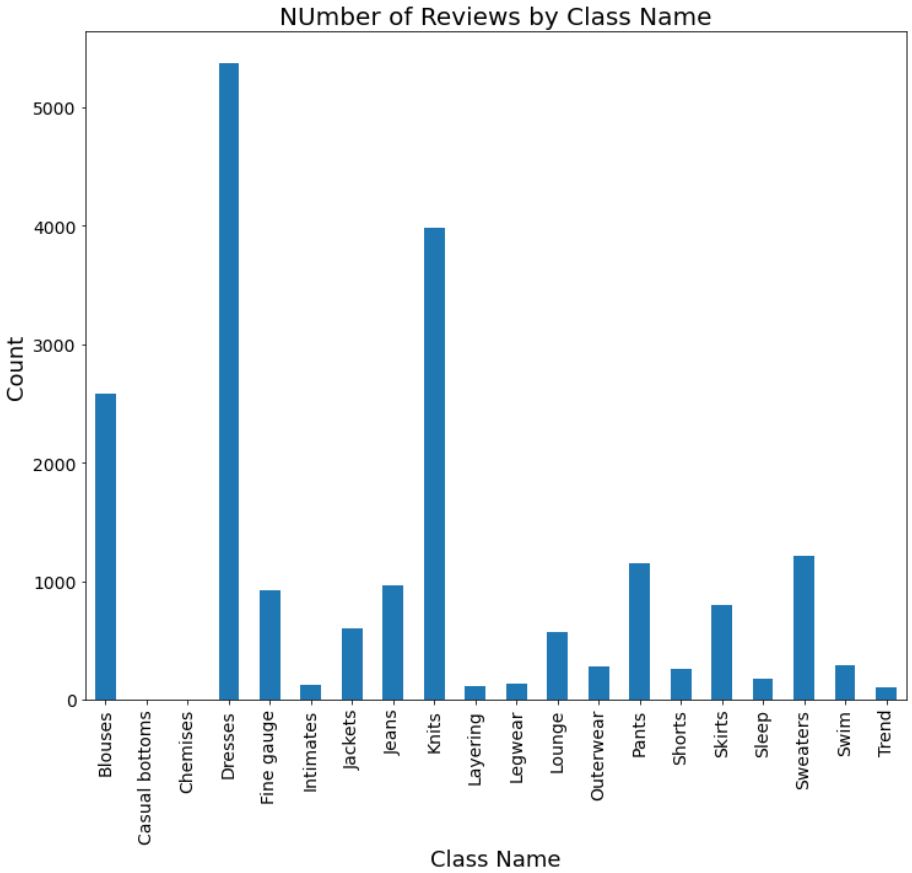


From this histogram of our target variables, we can see a skewed distribution showing that this store have more positive reviews than negative

Following the number of reviews by department, division and class name







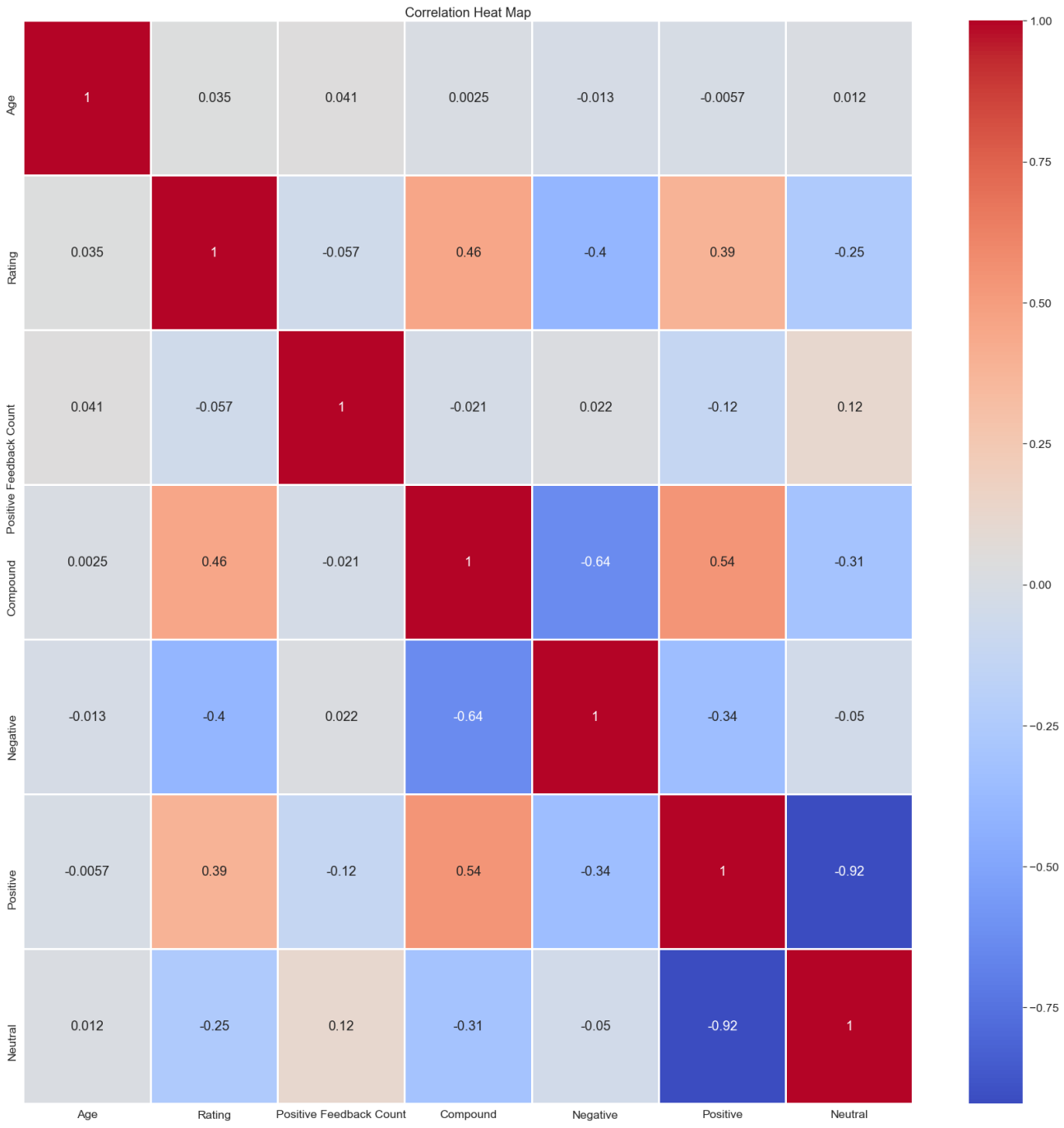
The graphs before shown how most of the reviews are made by General Division, the reviews are also focused on Tops department and dresses, blouses and trends are also the most criticized. For details about the reviews by department please refer to Appendix 1

For this analysis only a set of variables were used:

- 'Age'
- 'Title'
- 'Review Text'
- 'Rating'
- 'Positive Feedback Count'
- 'Division Name'
- 'Department Name'
- 'Class Name'
- 'Compound'
- 'Negative'
- 'Positive'
- 'Neutral'
- 'Analysis'

For sentiment analysis, the article published in Start it Up by Keppler. E [1], Vader sentiment analysis was used to determine sentiment of the reviews available on the data set, Neutral, Positive, Negative and Compound were calculated and added to the data frame (Appendix 2). After these results, another column was added to the data frame, the new column " target" contains the final result of the review based on the Compound final score, everything over 1 will be positive, all observations equals 0 would be Neutral, and everything else is negative, this column would work as label for our predictions as well as target variables

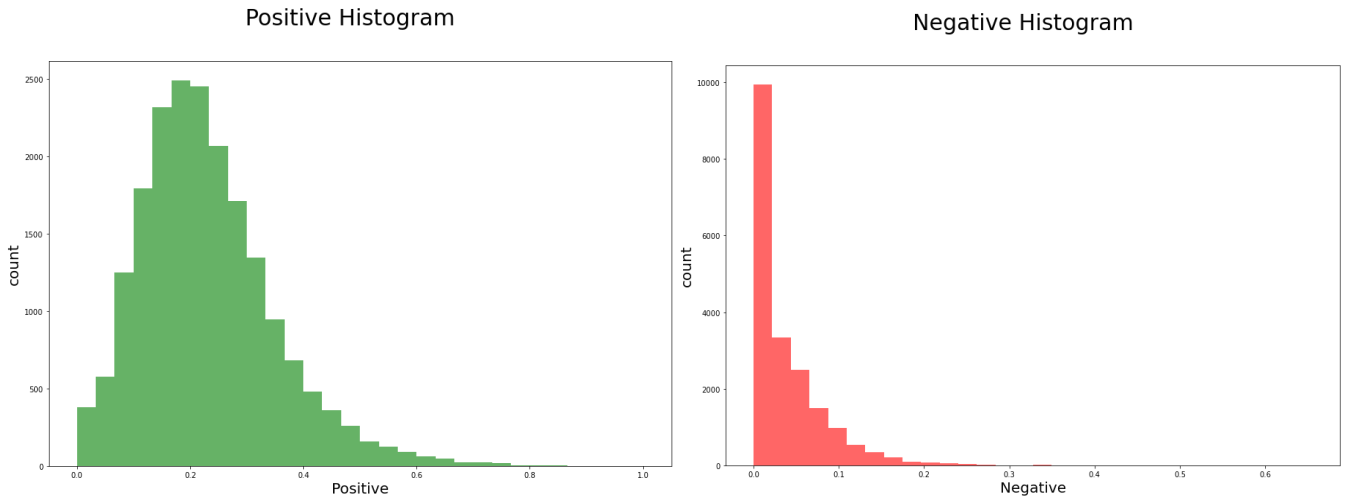
For more analysis, correlation values were calculated having the following results:

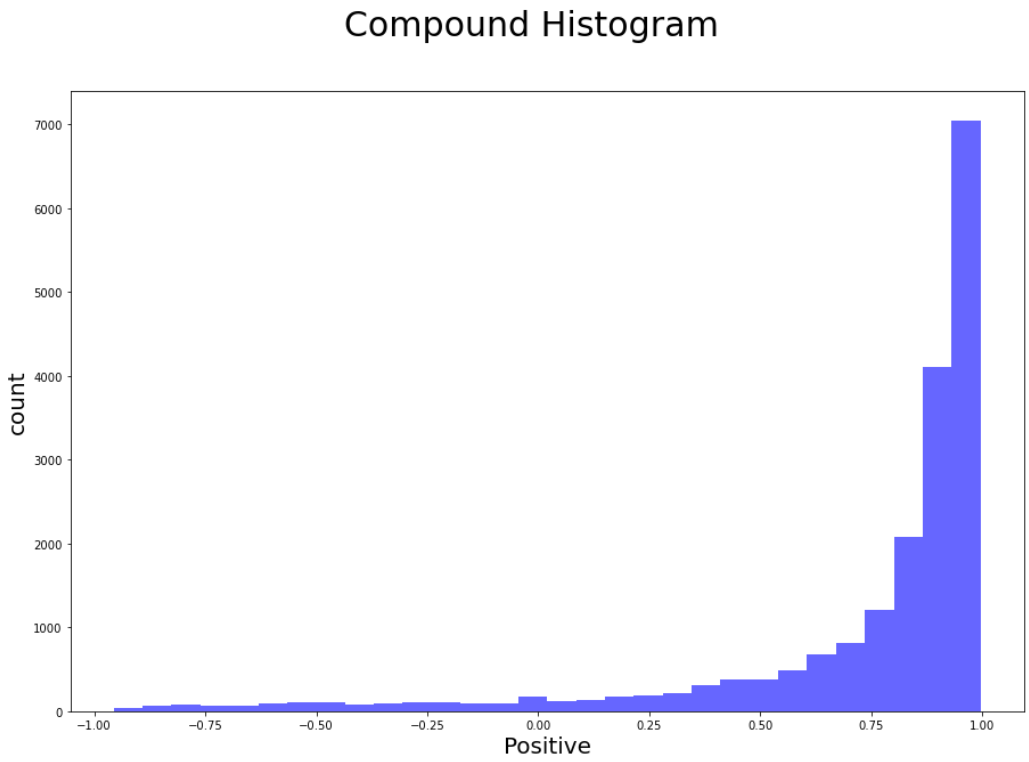


From this graph we can see how the positive results of the sentiment analysis are positive correlated with the ratings, this would be an obvious correlation, as well as the compound result,

the age does not show a high value of correlation for the reviews as well as for either of the other variables.

Histograms of the Sentiment results:





These analyses correspond with the first highlight about that this store have a positive perception of the clients.

To predict sentiment of next review I used Naïve Bayes classifier, resulting accuracy of this model:

```
Accuracy of Naive Bayes classifier = 91.78 %
```

In the text few days, I am going to apply TextBlob for sentiment analysis, random forest, and KNN to compare with these first results.

Conclusion

Based on the analysis shown before, sentiment analysis can be hard to apply because of complexity of the semantic of the paragraph, for these there are other options to use, there also different types of sentiment analysis depending on how detailed we would like our results [2]. For this project my goal is to build a model that can predict sentiment of next comment or review independent of the topic of the review, however, to build such a general model can take too long and for the time frame for this project, I will build a model to apply to the clothes review and to the Alexa reviews and compare both. TextBlob will be used as another method to perform sentiment analysis and other classifiers will be used as well. So far Naïve Bayes model accuracy was 91%, and it represent a good model so far, however there is an important factor to take in consideration like the skewed distributions, and this will be detailed in the next days

References

[1] Kepple. E (2020). Simple Sentiment Analysis for NLP Beginners and Everyone Else using VADER and TextBlob. Retrieved August 07, 2021. from https://www.watermelonwebworks.com/google-analytics-users-vs-sessions-vs-pageviews/

[2] MOnkeyLearn. Sentiment Analysis: A Definitive Guide. Retrieve on August 03, 2021 from: Sentiment Analysis: The Go-To Guide (monkeylearn.com)

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6

https://medium.com/swlh/simple-sentiment-analysis-for-nlp-beginners-and-everyone-else-using-vader-and-textblob-728da3dbe33d

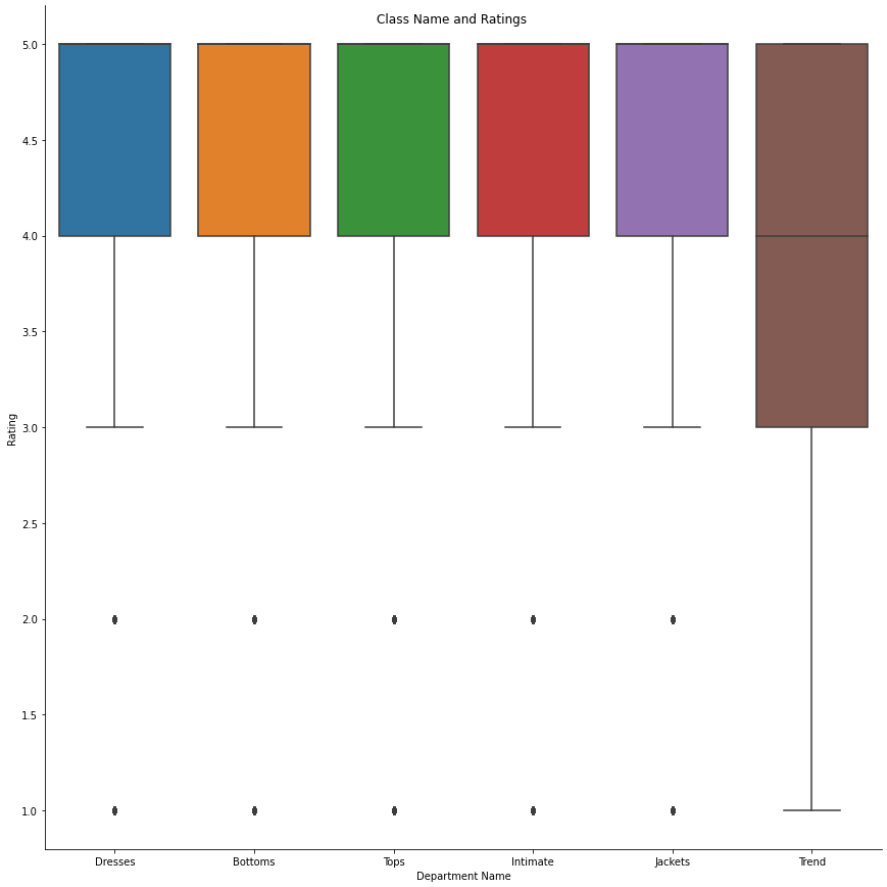
https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews

https://www.kaggle.com/therohk/million-headlines

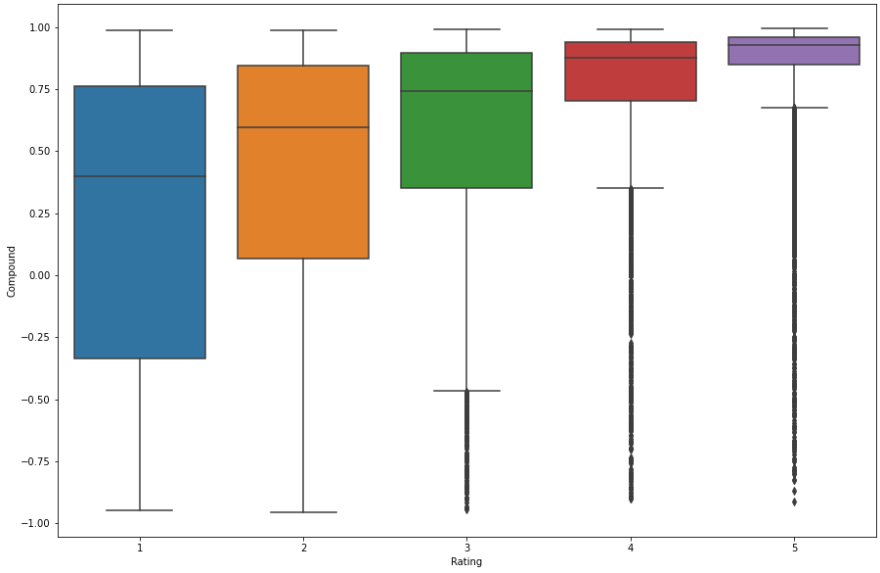
https://www.kaggle.com/sid321axn/amazon-alexa-reviews

Appendix 1

Box Plots Reviews by Department and Division



Rating and Compound



Appendix 2

ned: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	Compound	Negative	Positive	Neutral	Analysis
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses	0.9427	0.027	0.181	0.792	Positive
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants	0.5727	0.226	0.434	0.340	Positive
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses	0.9291	0.000	0.300	0.700	Positive
5	1080	49	Not for the very petite	I love tracy reese dresses, but this one is no...	2	0	4	General	Dresses	Dresses	0.9419	0.000	0.147	0.853	Positive
6	858	39	Cagrcoal shimmer fun	I aded this in my basket at hte	5	1	1	General Petite	Tops	Knits	0.8004	0.023	0.096	0.881	Positive

QUESTIONS

1. **Is Naïve Bayes the Best Classifier for these types of problems?**
2. **Why do not predict next word on the reviews?**
3. **How can include a detailed word by word analysis to improve the sentiment analysis?**
4. **Are there other approaches to reach the goal? Which one?**
5. **Is sentiment analysis useful nowadays? Who does use it?**
6. **Is there any method to improve accuracy?**
7. **If the review is in another language, how can you deal with this?**
8. **What other predictions can be done using this data set?**
9. **Is it possible to use an ensemble for this case?**
10. **What is the most popular field where sentiment analysis is applied?**