Predicting Next Review Sentiment

Gloria P. Moore

DSC680_Applied Data Science

Bellevue University

Bellevue University Data Science Master's Program

Abstract

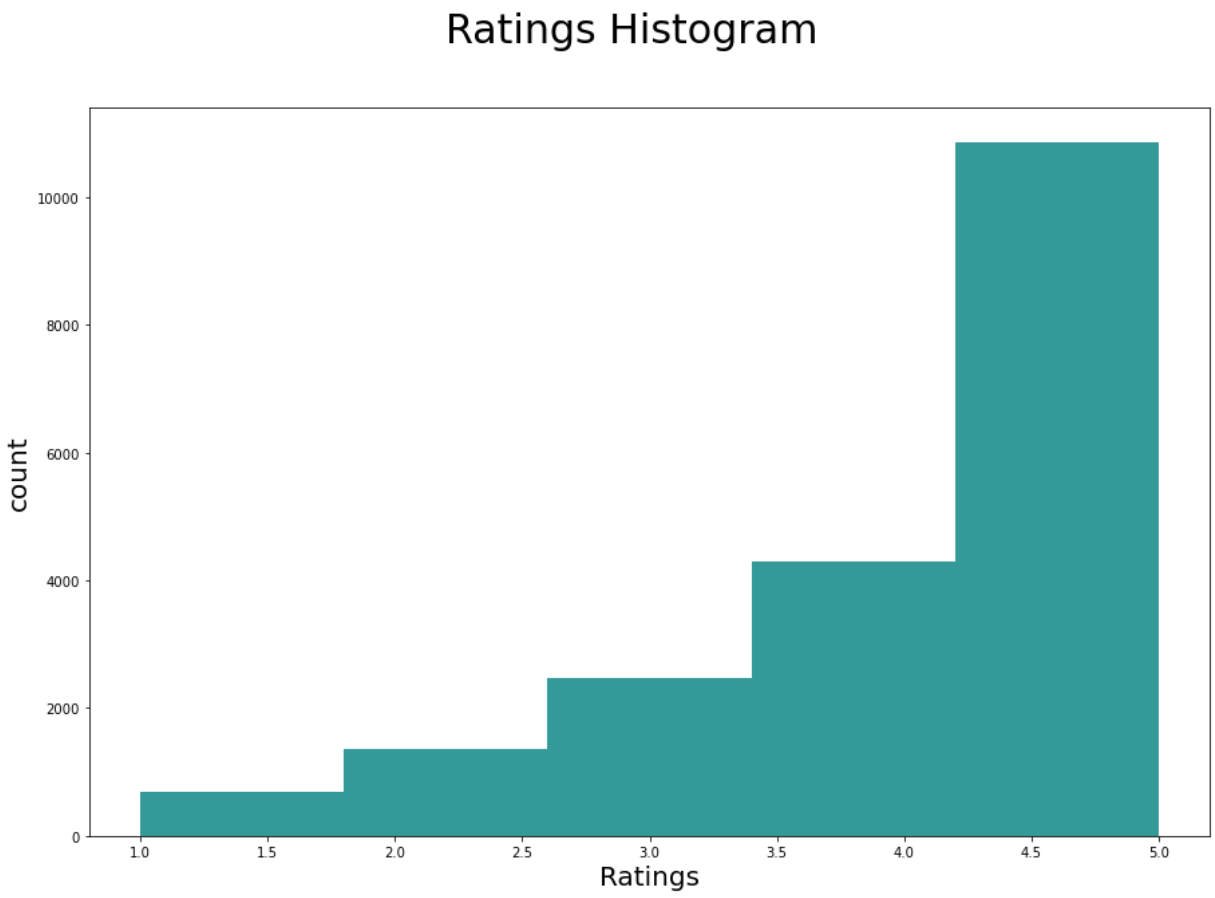
The following project is based on the sentiment analysis of reviews of more than 19000 clients of a clothes store, these reviews have corresponding ratings according to the client experience, and the goal of the following work is based on these reviews, to build a model to predict sentiment of the next 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 and TextBlob 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, KNN and Random Forest. The model selected for the clothes review sentiment prediction was Random Forest Classifier, with an accuracy of 0.97.

Introduction

Sentiment analysis is used to analyze human behavior, to support marketing strategies, is defined by Gupta (2018) as the process that use contextual mining of text and identifies and extracts subjective information in source material. Positive Negative or Neutral nature of a comment or review. This type of analysis is widely used for a lot of companies in different fields like, stock markets, sports, medicine, etc. These analysis gives insights to companies about their product or services, allowing them to take decisions based on the perception that the public is posting on social media or other communication form [1]. Having information about how the client feels about different area of a companie, or about different products or service is great information for decision making process, it allows to executives to address data driven efforts to strategic areas of the business. Predicting these sentiments, however, will give a wide step forward in a competitive way. In the following project we are going to explore sentiment analysis and predict future reviews in terms of sentiment.

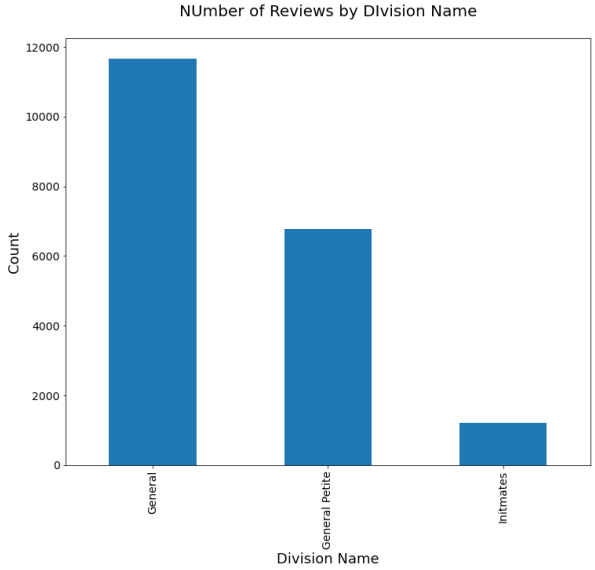
Method

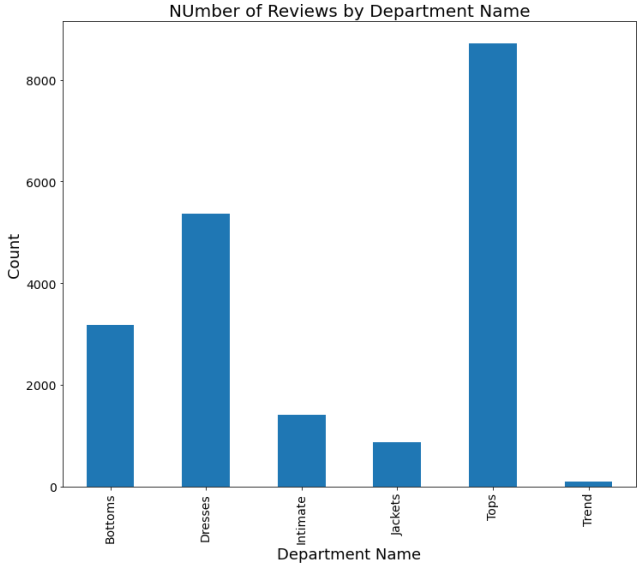
The method used for this project has been Vader sentiment analysis, and TextBlob libraries to apply sentiment analysis, the data set was read and transform to a data frame were the reviews were under " Review Text" column, following this, missing values were check, finding around 2000 missing reviews,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

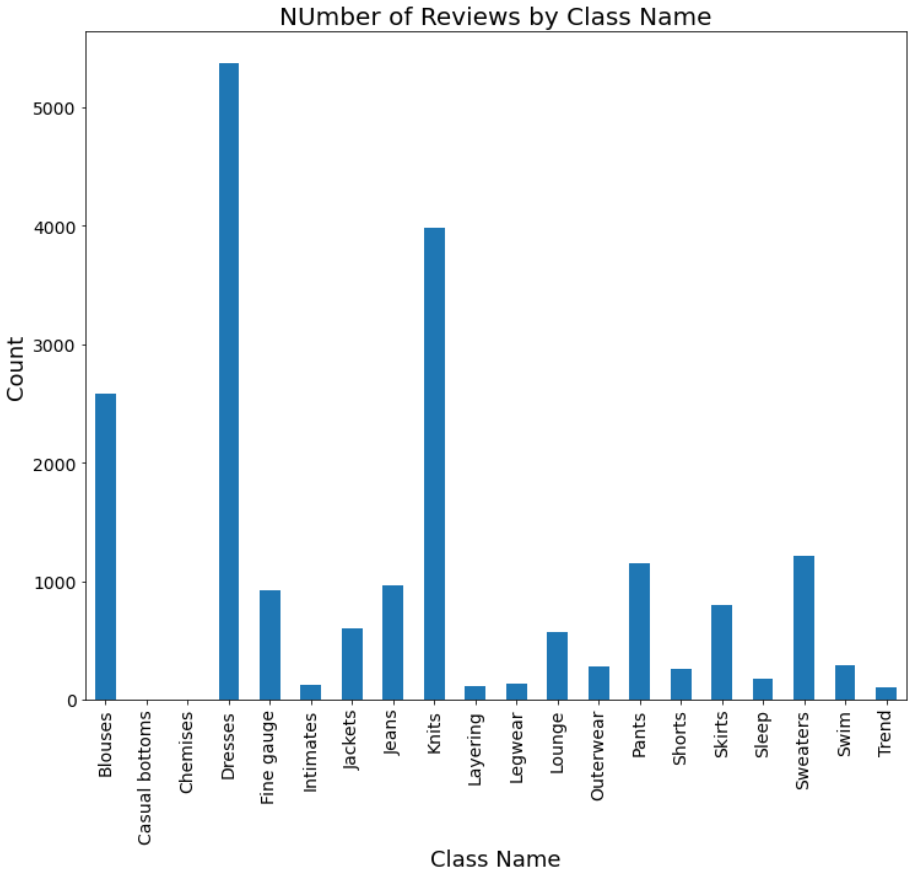


From this histogram of our target variables, we can see a skewed distribution showing that this store has more positive reviews than negative

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

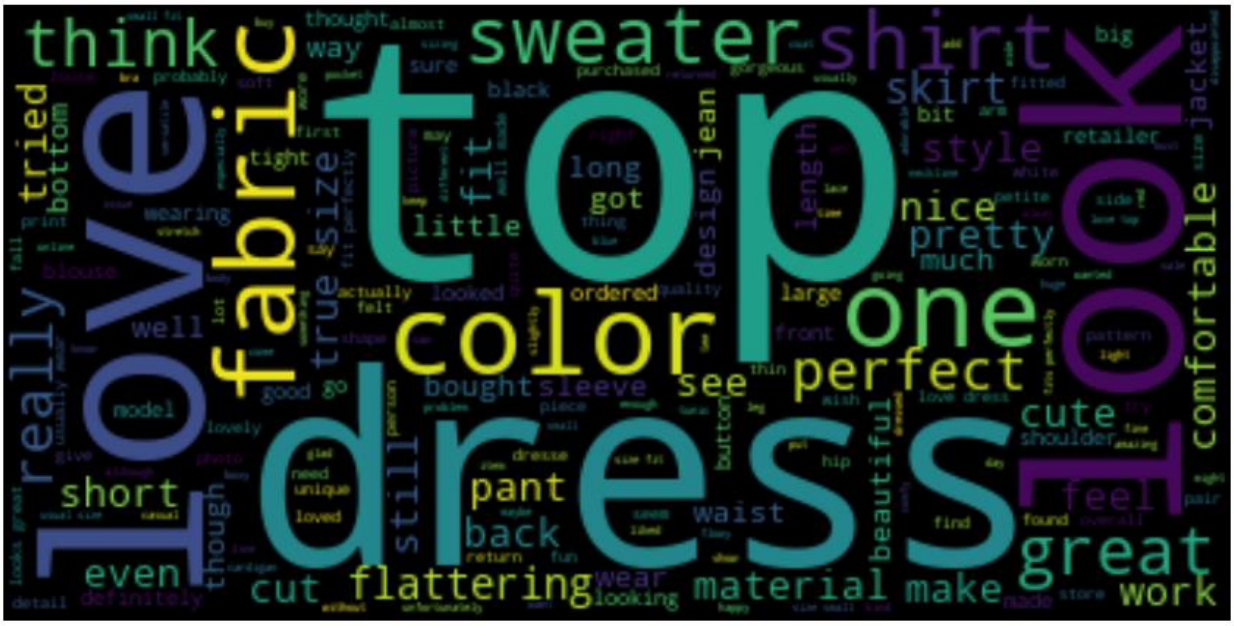






The graphs before shown how most of the reviews are made by General Division, the reviews are also focused on Tops department and dresses, blouses and trends are also the most criticized. For details about the reviews by department please refer to Appendix 1.

After graph analysis of the features, format and processing of the column “Review Text” is performed: removing special punctuation signs, transforming all words to lower case and removing stop words using mainly the NLYK library. A new column was created with no stop words to apply sentiment analysis using Vader and TextBlob. Word graph analysis was made to verify common words in the reviews



We can see how positive words like “love”, “perfect”, “great”, and others are noticed in the previous graph. This graph supports the previous histogram of Ratings, were most of the ratings are 5.

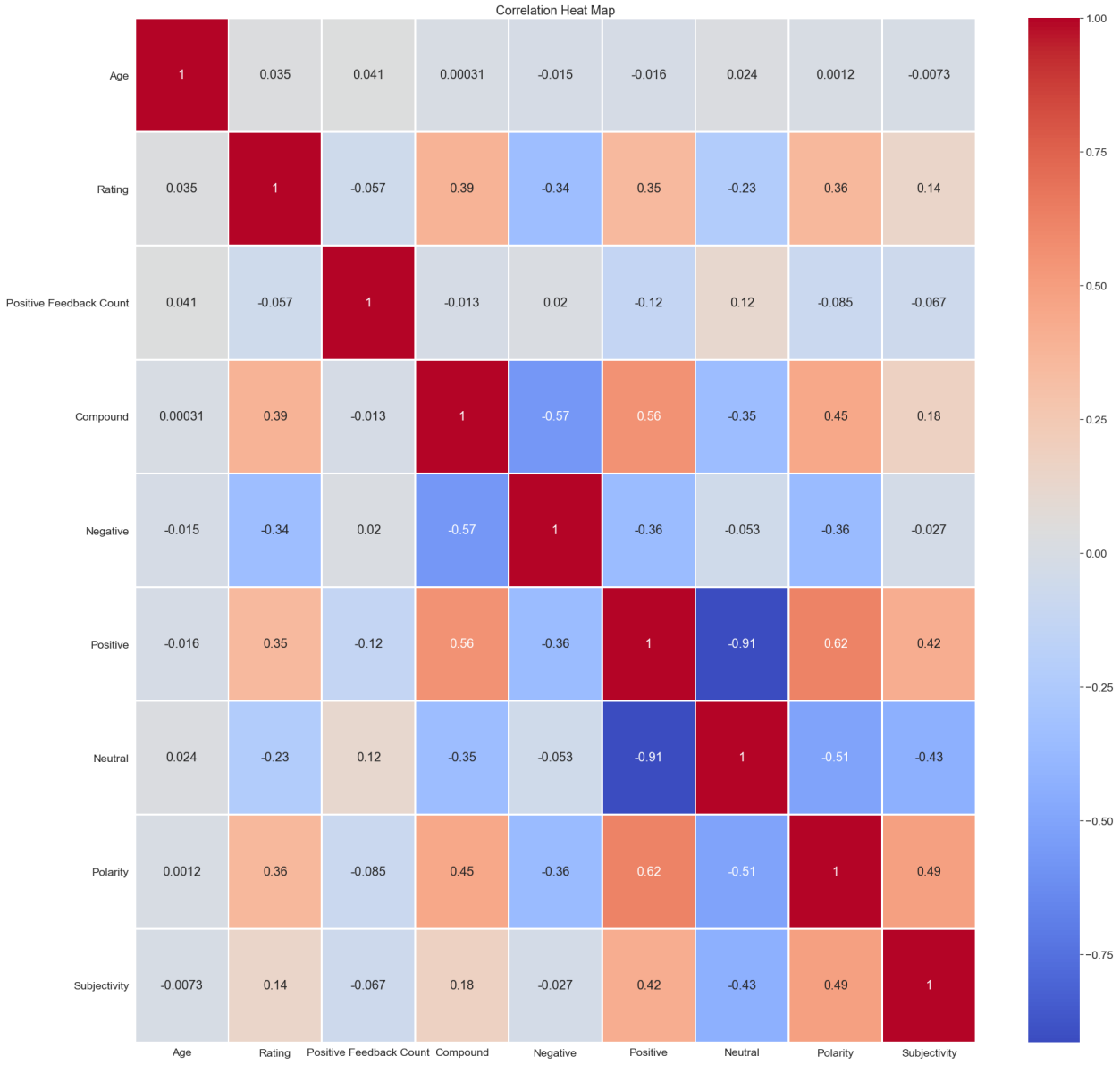
For sentiment analysis, the article published in Start it Up by Keppler. E [2], Vader sentiment analysis was used to determine sentiment of the reviews available on the data set, Neutral, Positive, Negative and Compound were calculated and added to the data frame. TextBlob was also applied to the review column, TexBlob analysis gives two values, one representing Polarity (Sentiment) and the second representing Subjectivity (Positive or Negative) [3]. After applied both methods to the review column, another column was added to the data frame, the new column " Analysis" 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 (See Appendix 2)

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'
- 'Polarity'
- 'Subjectivity'

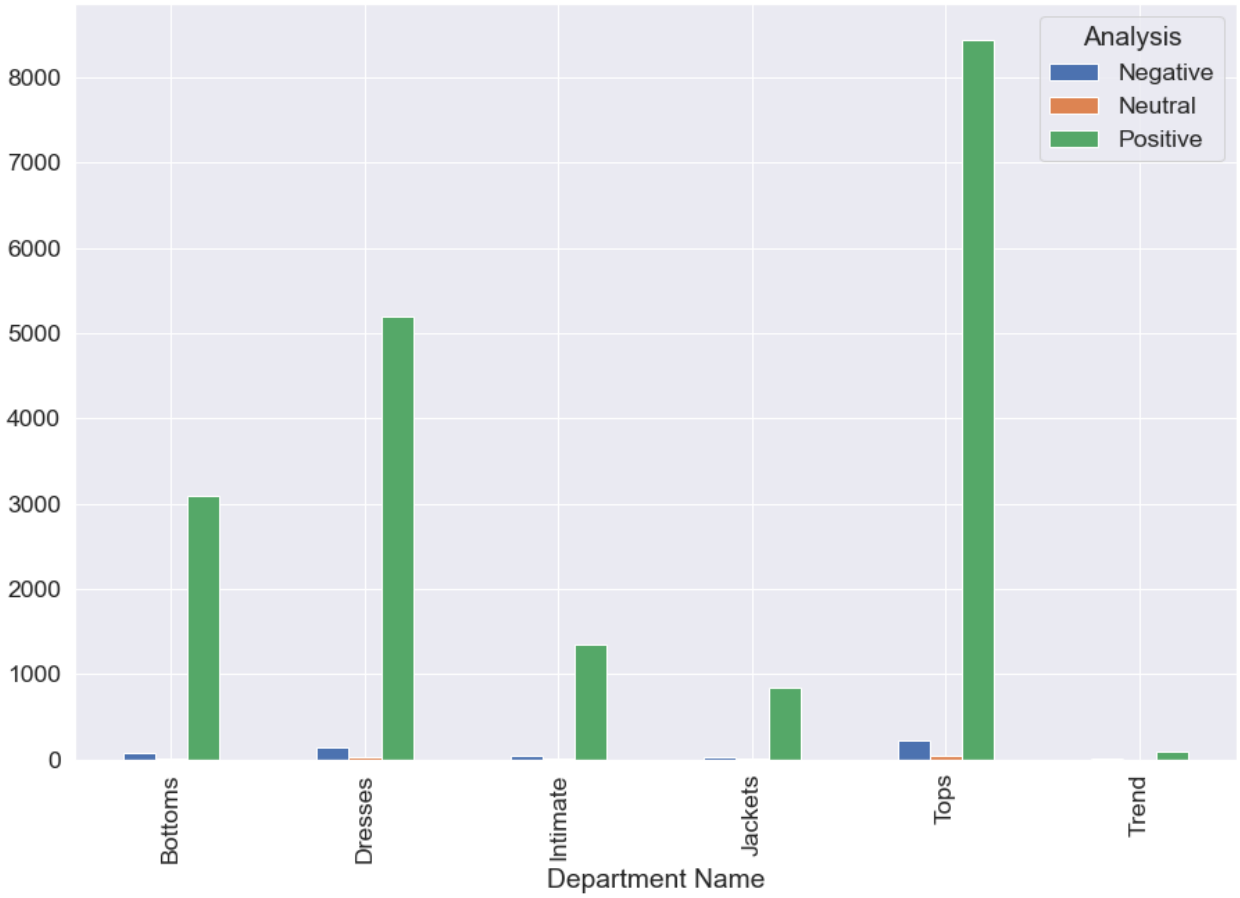
Target variable was defined as "Analysis", and the labels to build the classification models were the values in this column: "Positive", "Neutral", "Negative"

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

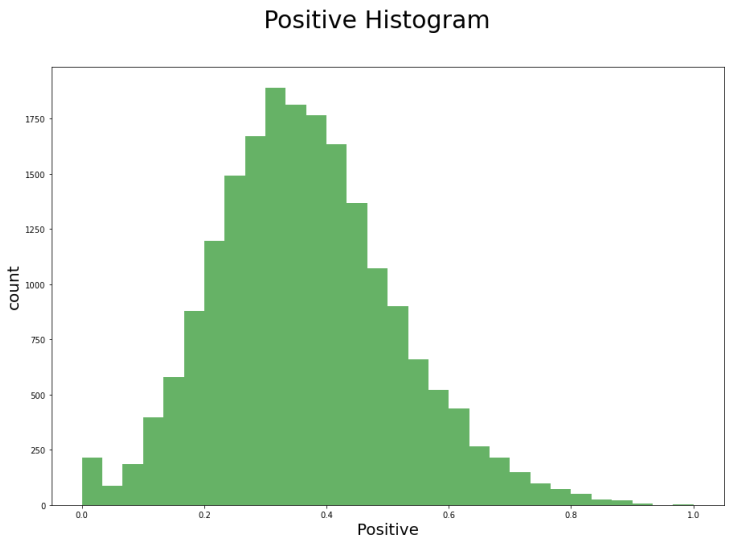


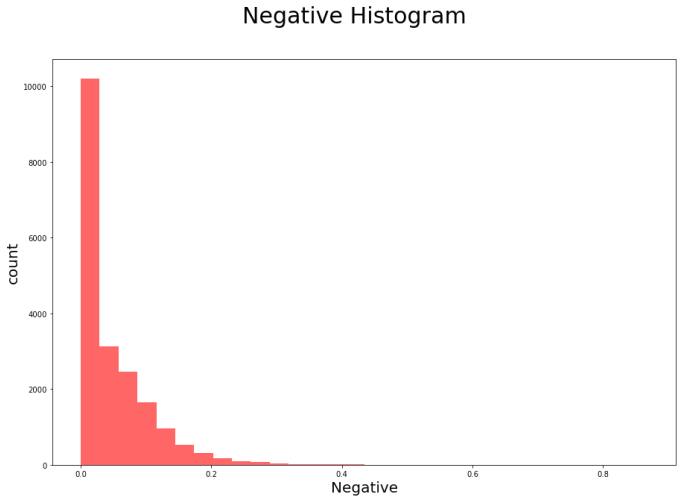
From this graph we can see how the positive results of the sentiment analysis are positive correlated with the ratings, this would be an obvious correlation, as well as the compound result, the age does not show a high value of correlation for the reviews as well as for either of the other variables.

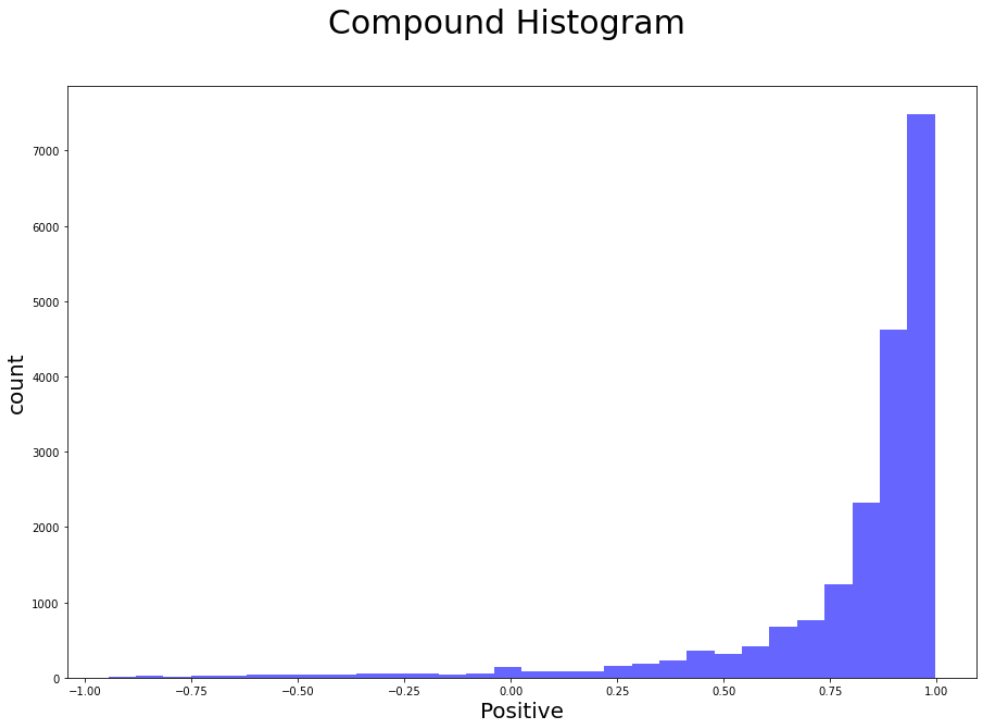
Bar Chart of the Vader Sentiment results by Department:



Histograms for Vader Sentiment Analysis:







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

Modeling

Naïve Bayes

For this model, the data set was split 50% for training and 50% for validation, the results are the following:

```
              precision    recall  f1-score   support

    Negative       0.03      0.09      0.04       281
     Neutral       0.02      0.02      0.02        63
    Positive       0.96      0.90      0.93      9487

    accuracy                           0.87      9831
   macro avg       0.34      0.34      0.33      9831
weighted avg       0.93      0.87      0.90      9831

0.8736649374427831
```

It is important to highlight that for this model, only the Values of the Vader analysis were used

Random Forest

Random forest model was trained using the whole data set including the TextBlob values, reviews were vectorized and data was split 50% for training and 50% for validation, the following are the results:

```
              precision    recall  f1-score   support

    Negative       0.36      0.01      0.03       281
     Neutral       0.75      0.05      0.09        63
    Positive       0.97      1.00      0.98      9487

    accuracy                           0.97      9831
   macro avg       0.69      0.35      0.37      9831
weighted avg       0.95      0.97      0.95      9831

0.9650086461194182
```

KNN

Knn was build following the same methodology as Random Forest, results:

```
              precision    recall  f1-score   support
```

```
    Negative       0.31      0.02      0.03       281
     Neutral       0.00      0.00      0.00        63
    Positive       0.97      1.00      0.98      9487

    accuracy                           0.96      9831
   macro avg       0.43      0.34      0.34      9831
weighted avg       0.94      0.96      0.95      9831

0.9641948937035907
```

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 [4]. For this project my goal was 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. The EDA made on this data set shows how this store have more positive reviews than negative, it has a very good acceptance rate and rating. Predictive models applied show a better metrics when positive reviews are predicted than negative, this is because the high number of positive data in the sample. The model selected was Random Forest based on accuracy 0.965. It is recommended for future analysis to balance classes before train the model

References

[1] Insight Thematics. The Complete Guide to Sentiment Analysis. Retrieved on ugust 11, 2021 from https://getthematic.com/insights/sentiment-analysis/

[2]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/

[3] Pernling Frödin, Y. (2018). Sentiment Analysis with TextBlob. Retrieved August 12, 2021. From Sentiment Analysis with TextBlob. This week I will take on a somewhat… | by Youn Hee Pernling Frödin | Medium

[4] 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

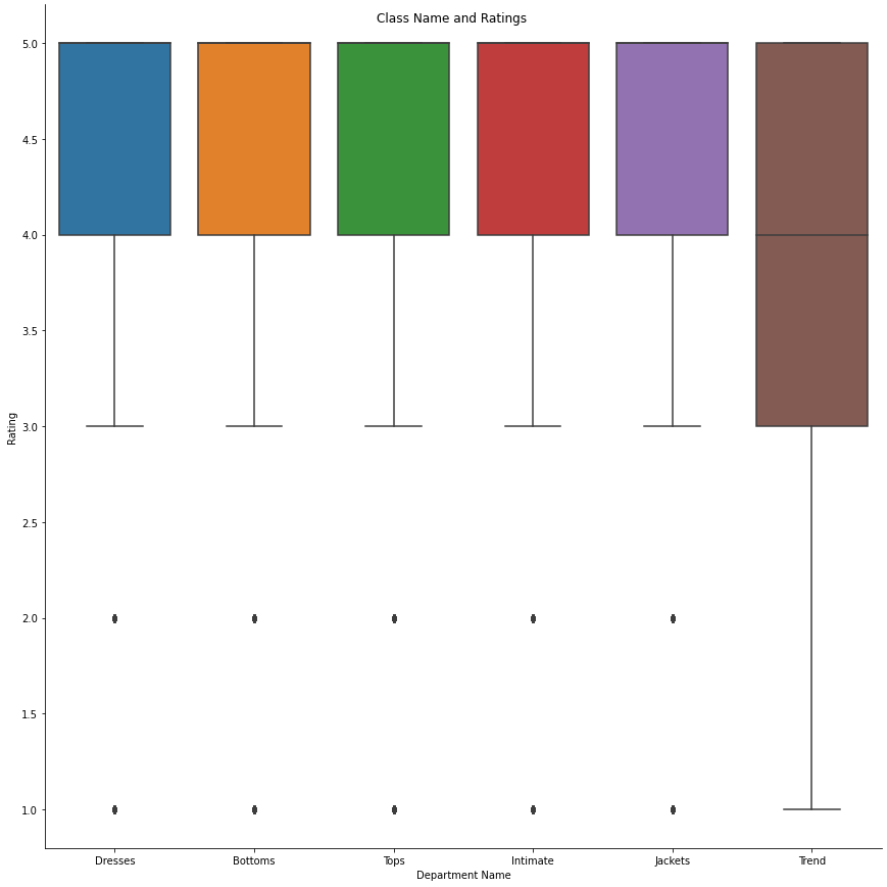
https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews

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

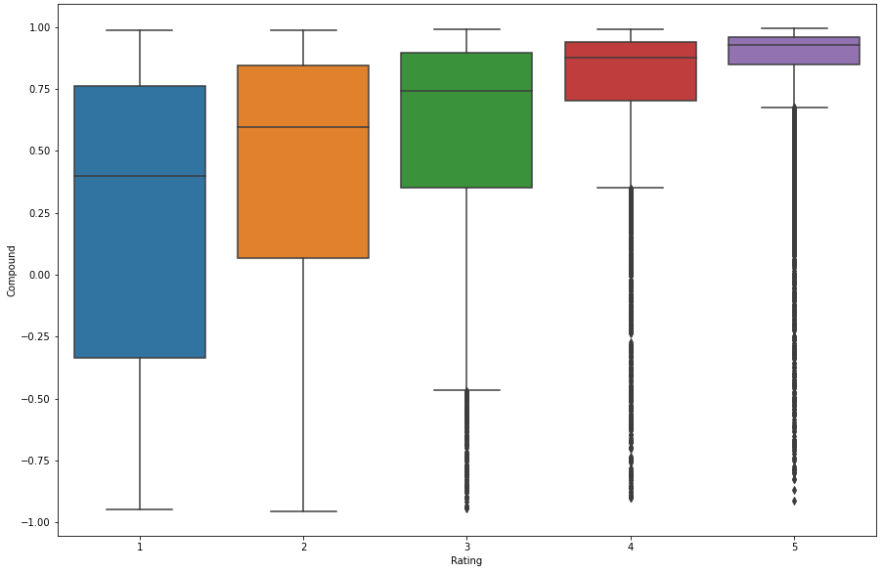
https://www.kaggle.com/sid321axn/amazon-alexa-reviews

Appendix 1

Box Plots Reviews by Department and Division



Rating and Compound



Appendix 2

ıg	Positive Feedback Count	Division Name	Department Name	Class Name	rev_no_stopw	Compound	Negative	Positive	Neutral	Analysis	sentiment_textblob	Polarity	Subjectivity
3	0	General	Dresses	Dresses	high hopes dress really wanted work me. initia...	0.9117	0.036	0.259	0.705	Positive	(0.07945887445887445, 0.3458658008658009)	0.079459	0.345866
5	0	General Petite	Bottoms	Pants	love, love, love jumpsuit. fun, flirty, fabulo...	0.9511	0.163	0.631	0.206	Positive	(0.5499999999999999, 0.625)	0.550000	0.625000
5	6	General	Tops	Blouses	shirt flattering due adjustable front tie. per...	0.9213	0.000	0.523	0.477	Positive	(0.6171875, 0.6583333333333333)	0.617188	0.658333
2	4	General	Dresses	Dresses	love tracy reese dresses, one petite. 5 feet t...	0.9153	0.000	0.257	0.743	Positive	(0.15, 0.5428571428571428)	0.150000	0.542857
5	1	General Petite	Tops	Knits	aded basket hte last mintue see would look lik...	0.8439	0.000	0.159	0.841	Positive	(0.1605, 0.5293333333333333)	0.160500	0.529333

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. **Is it possible to apply weights to the Textblob results compared with the Vader results?**