# **Project Deliverable 2**

# **Statement of Problem**

In this project, we will be analyzing a Women's Clothing E-Commerce dataset from Kaggle. The dataset is an anonymized real commercial data which contains data about reviews of various categories of products written by customers at different ages. The dataset also includes the rating scores for each item. By exploring the nine supportive features of the dataset, we can obtain a general idea about the relationship between customers and how they rate the different classes of clothes. By analyzing the review texts, we can discover the more profound insights which help us to find the trends in the customer reviews and extract actionable plans to improve online e-commerce of the retailer. Following are the questions we would like to answer with our analytics.

1. What are the patterns of purchases, customer reviews, and ratings? (i.e., What type of product in which class has the highest rank? What is the relationship between reviewers' age and the number of reviews? What factors affect the ratings most? What is the biggest problem of the products?)
2. How did the customers like their purchases? (i.e., What are the most common words in reviews? Whether the ratings lead to a recommendation of a cloth? What are the emotions expressed in the reviews, positive or negative?)
3. **Additional Data Processing**

We will use the bag of words approach where we clean the data and extract text. Specifically, we will create a corpus, clean the text, and generate a matrix derived from term frequencies.

1. Create a corpus from the variable ‘Review.Text’
2. Use tm\_map to clean the text
   1. transform text to lowercase
   2. remove punctuation
   3. remove English stop words using the following dictionary tm::stopwords(’english)
   4. remove whitespace
3. Create a dictionary
4. Use tm\_map to stem words
5. Create a DocumentTermMatrix

After creating the document term matrix, we will remove infrequently occurring terms that appear in fewer than 10% of the reviews and convert it to a data frame. Next, we will use the dictionary created before to complete the stems.

We evaluate the frequency of the tokens and find out “dress”, “fit”, “love” occurs more frequently in the corpus.

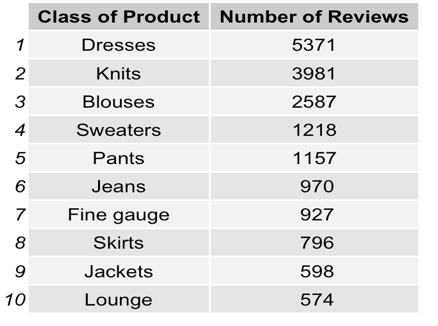


1. **Select Analytical Technique**
2. Descriptive Analysis

We are using descriptive analysis to analyze the relationship about reviews by class, ratings by class and reviews age group.

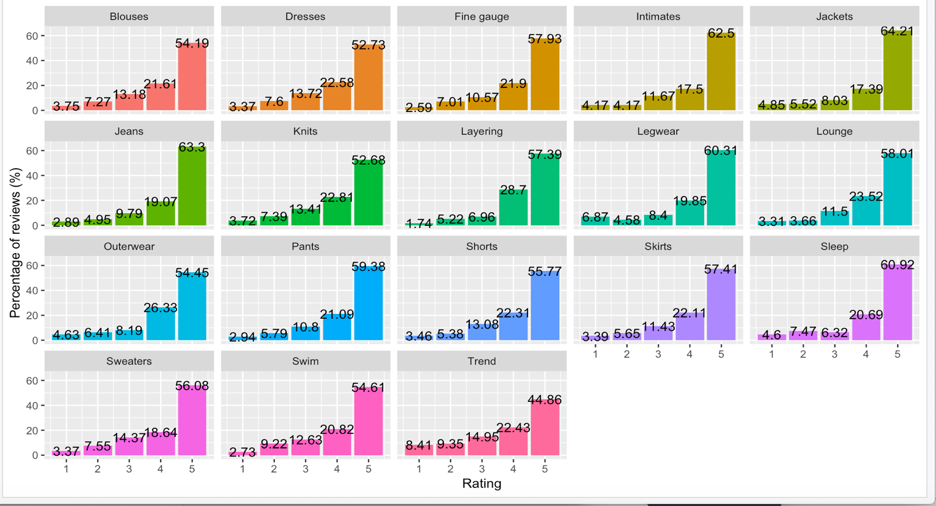
**Reviews by class**

We see that Dresses top the list followed by Knits and Blouses. It might because women need a dress for many situations. Women prefer to buy tops than pants.

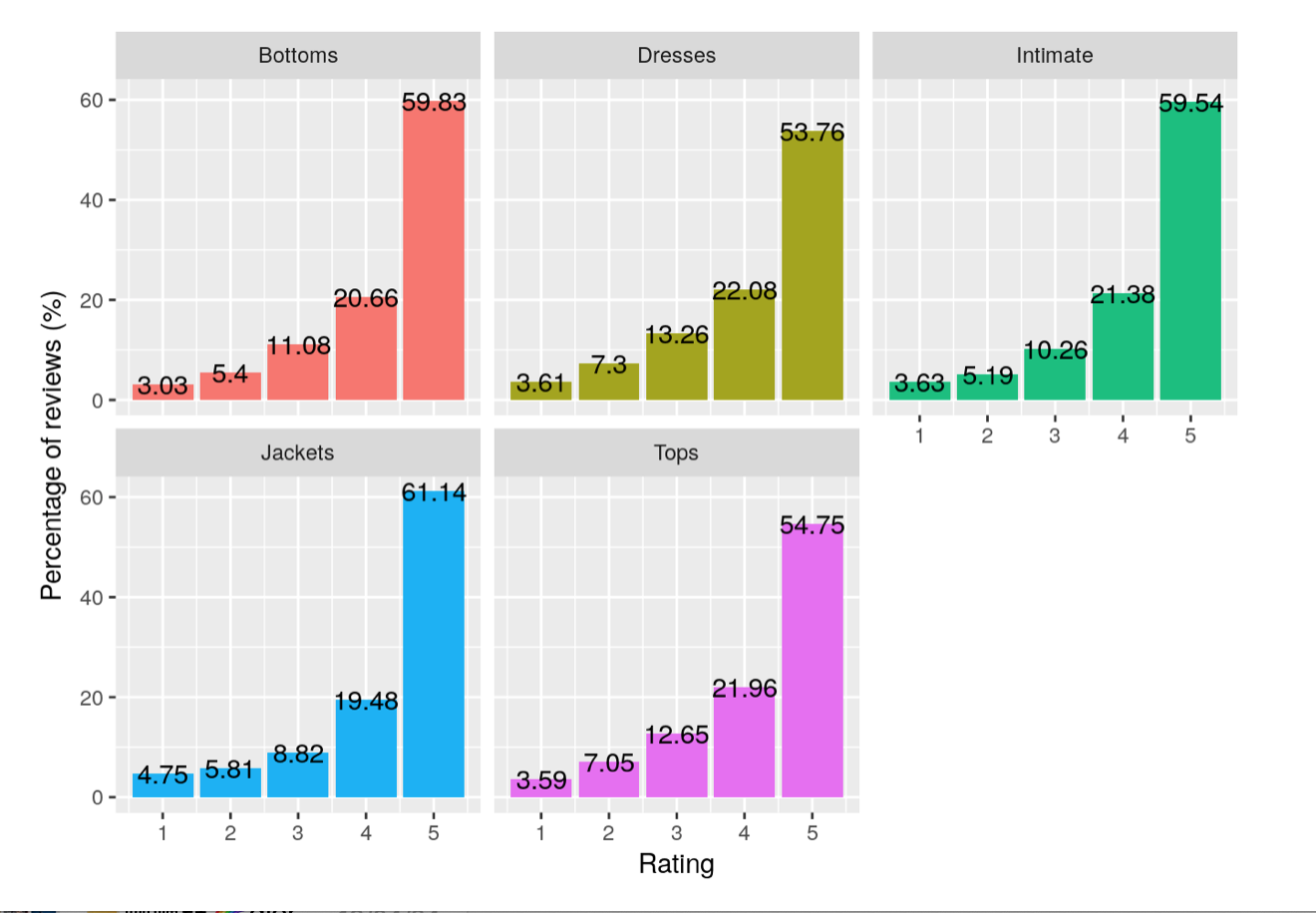


**Ratings by Class and Departments**

For ratings by Class, we will be excluding "casual bottoms and chemises" in Class as it contains only one review in each category. It does not have a big effect on data analysis. we will focus on the rest of the 16 class.

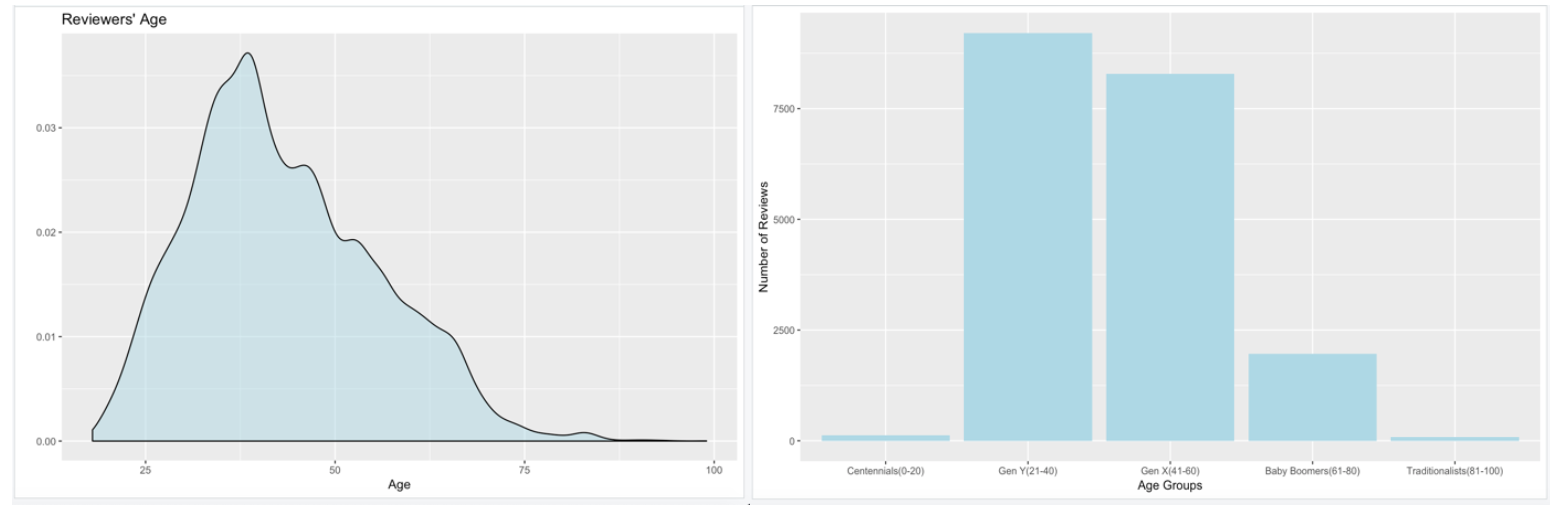


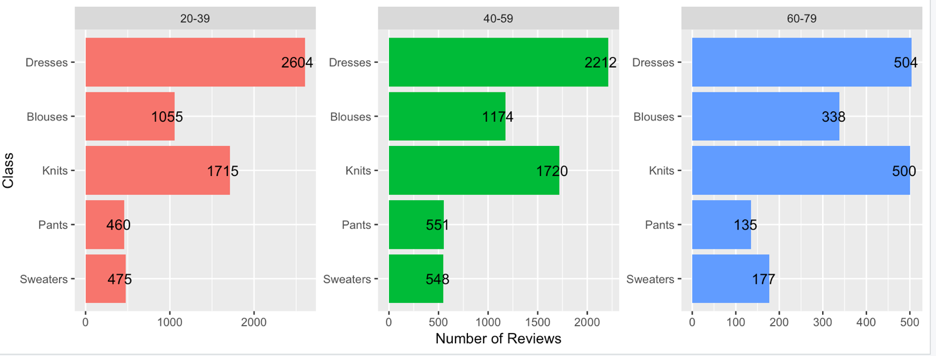
For ratings by Departments, we will be excluding ‘Trend' as it contains a mix of clothes that can fit in the other categories of Dept. They also represent 0.5% of the dataset, so we don't expect a significant effect on the data analysis. we will focus on five departments: Bottoms, Dresses, Intimate, Jackets, and Tops. Now, most of the reviews/ratings are for Tops and the least, for Jackets.



In each Class and departments, the dominant rating given is 5-stars. **Jacket has the highest number of 5-star ratings within its class**. As far as we can tell, jackets may be a good investment as consumers seem to give more 5-star reviews. Our speculation: The fit of the apparel may have something to do with this. Dresses and tops tend to be tricky, especially to purchase online, as body shape varies from person to person. What looks great on one person may feel too tight on the other. The jackets may use flexible material, or the customer may only require a loose fit. There may be fewer ways for jackets to go wrong than for tops and dresses.

**Reviews Age Group**



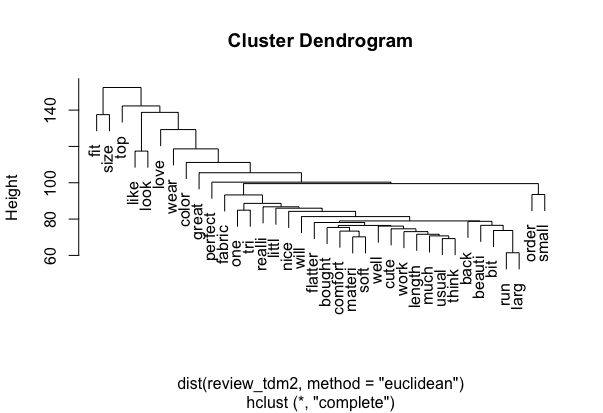
The youngest reviewer is aged 18, and the oldest is 99. We can see a pattern here, in today's world, people from the age group of **21-40** seem to shop more on the internet, followed by **41-60**. The bigger surprise is the 100 odd reviews in the **81-100** age group, and you would have thought that the **Traditionalists** may not be shopping online, let alone leaving reviews after making the purchases. It is surprising that there are few reviewers in age 0-20. It seems younger people aged 0-20 do not shop on the internet very often. It may not be true because the dataset is not big enough and the youngest reviewer is aged 18. 

We exclude age 0-20 and 80-100 because there are few reviews in that age range. we only include the top five classes (which have the most reviews) because the top five classes contain 14314/19660= 72.8% of data. Every age group has similar patterns but slightly different. Age group 20-40 and 40-60 are more likely to buy dress than age group 60-80. Age group 60-80 have a higher percentage of buying knits and sweaters than other groups.

**B. Clustering**

We use clustering method to identify word groups used together based on complete distance and visualize the word clusters with dendrograms. The technique gives us a quick idea about customer evaluations in the beginning process of exploratory data analysis.

Based on the dendrogram, we can see that the most common word groups used together are like "fit, size", "run, large", "order, small". This may reflect the problem that customers usually complain about the size of the clothing they purchased in the reviews. Thus, the retailer should provide more detailed sizing information on the shopping page of their website or improve their free return and exchange customer service to dispel customer's concern about sizing when they shop online.



**C. Sentiment Analysis**

Sentiment analysis is the process of identifying the emotions behind words. It is handy to learn insights and find the meanings of each attitude and opinion from a large number of reviews, comments, feedbacks emerging online. For this case of women's clothing reviews, we decide to apply sentiment analysis to gain a better understanding of the reviews.

**Words in reviews**

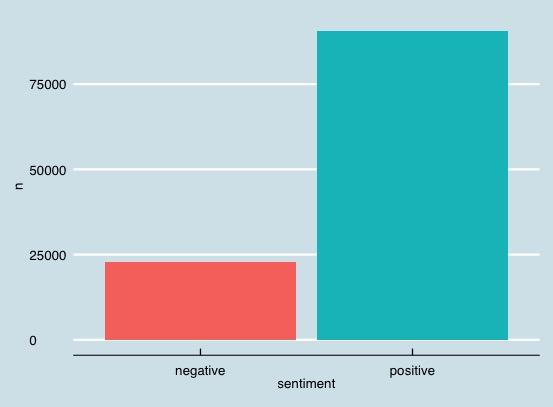
We begin with the counting of the words in each review and the total words. After using the dplyr and tidytext packages, we find there are 1,226,595 words totally in the reviews.

**Bing Lexicon**

Bing lexicon, as one of many word lexicons, is used to classify words as being positive and negative. Here, we decide to firstly use bing dictionary to determine the tones of each review.

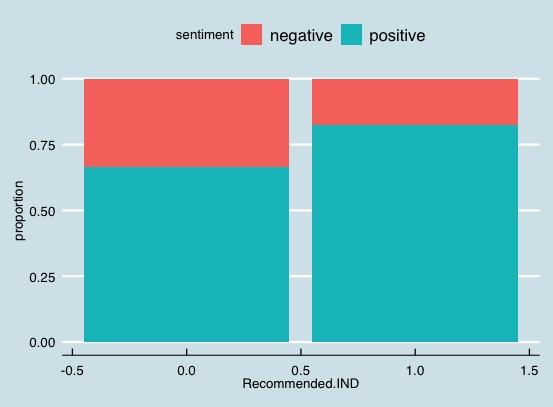
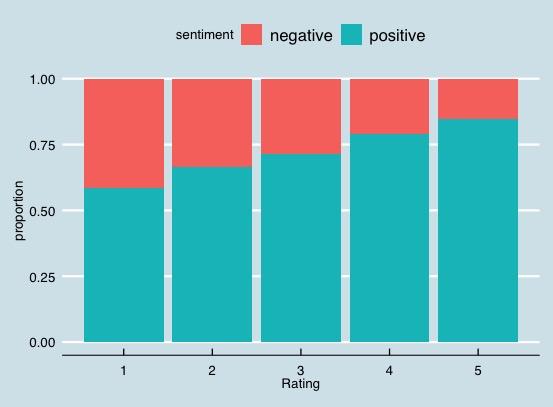
**Positive and negative words in reviews**

With bing lexicon, we find there are 22,938 negative and 90,474 positive words in reviews.



**The proportion of positive reviews, rating score and recommendation.**

We want to check whether a higher proportion of positive reviews come with a higher rating? We can see that a higher rating score comes along with a higher percentage of positive sentiment. Also, it is interestingly founded that only comparatively higher proportion of positive reviews mean recommendation.



**Correlation between positive words and review helpfulness**

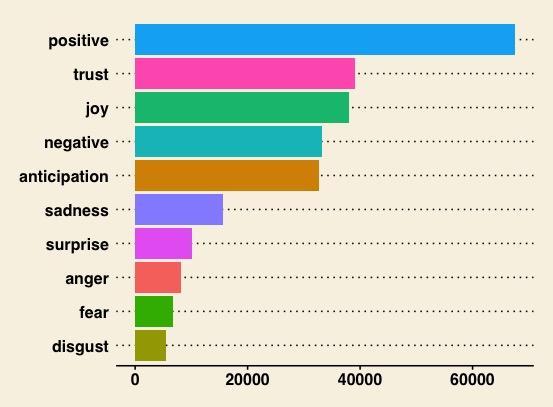
With bing lexicon, we scan the positive and negative words in each review and calculate the positivity of each review. Further, we check the correlation between positivity and rating. The correlation is **0.363**, indicating a weak positive linear relationship between positivity and rating.

**Nrc Lexicon**

The nrc lexicon will categorize words by emotion

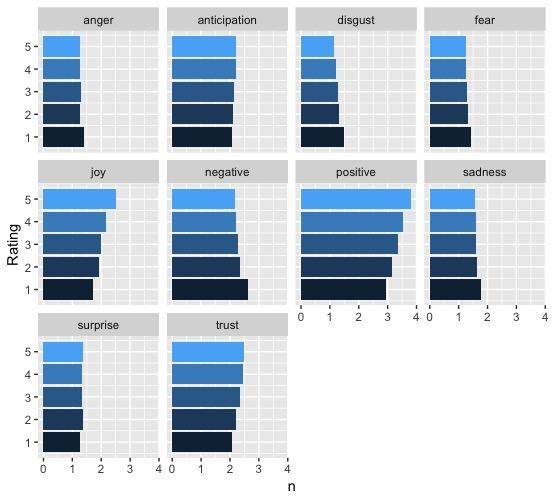
**Emotions in women’s clothing reviews**

With nrc lexicon, we will check the emotions in the reviews. Our findings are as shown in the plot below.

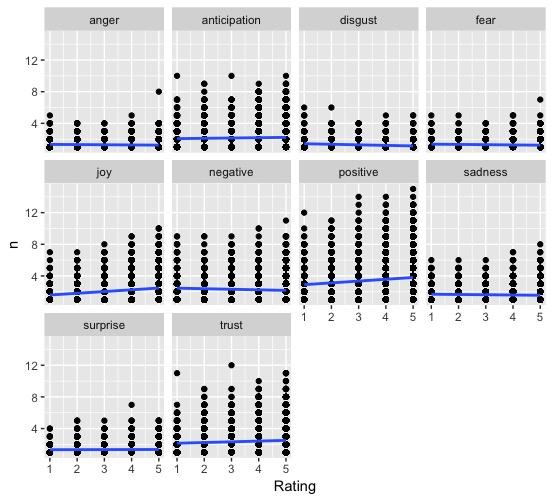
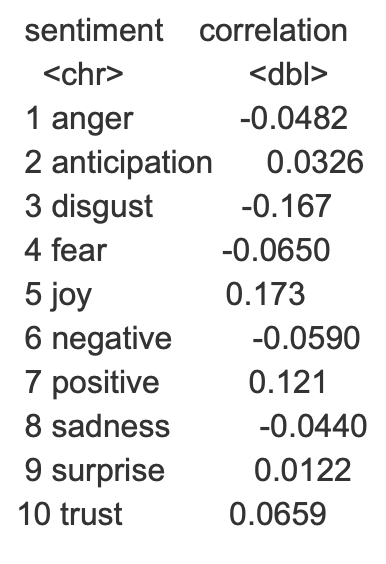


**Ratings of all Reviews based on Emotion Expressed**

There are 10 emotions under nrc lexicon. For each emotion, we want to check the relationship between the number of words and the review rating score.



Next, we want to see the correlation between emotion expressed and review.

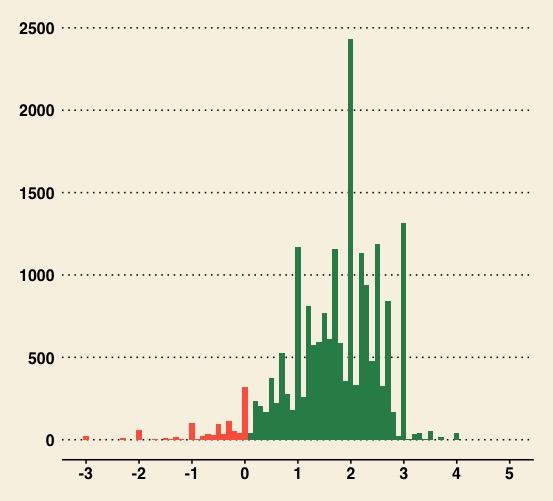


Although the correlation is very weak, we can still find that good emotion such as anticipation, joy, positive, surprise, and trust have a positive relationship with review ratings. Whereas, bad emotions such as anger, disgust, fear, negative, and sadness have a negative relationship with review ratings.

**Afinn Lexicon**

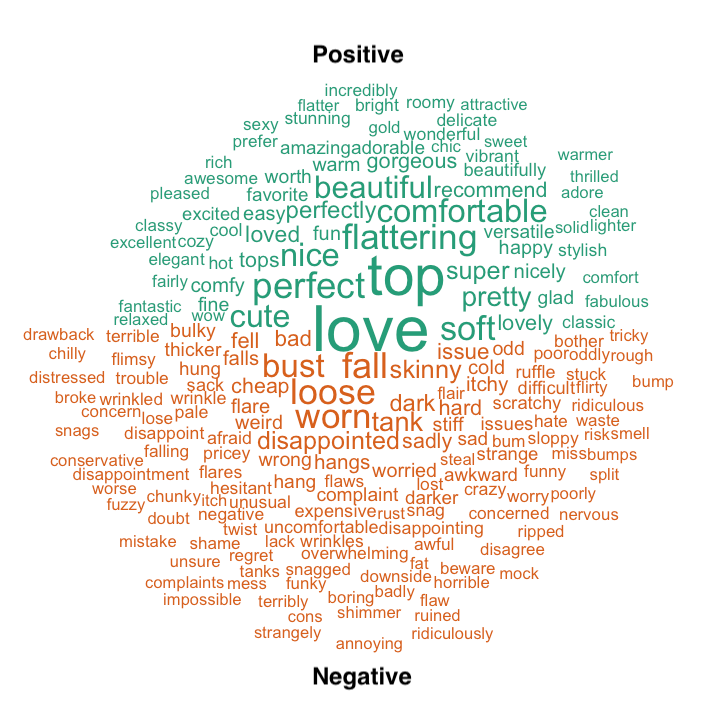
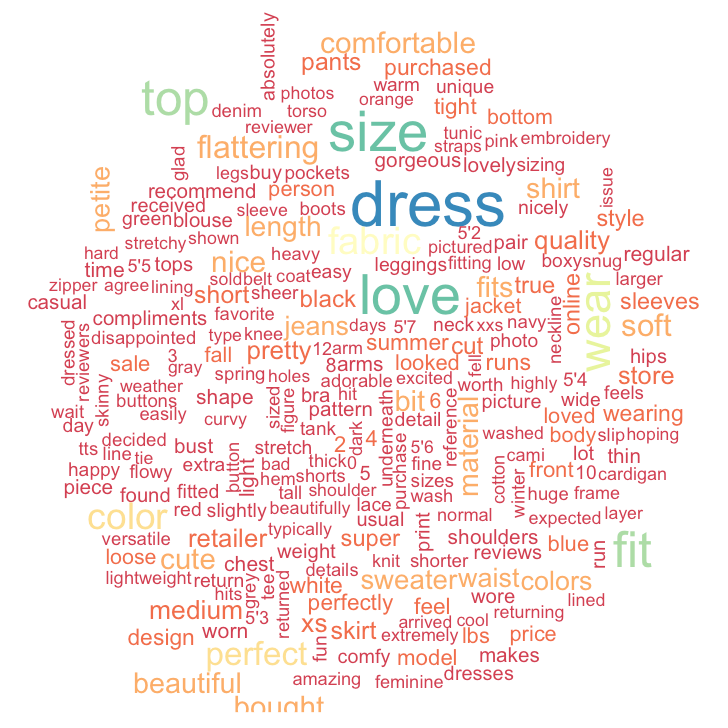
**Examine the sentiment of all reviews**

With afinn lexicon, we want to examine the sentiment of all the reviews. From the plot, we notice that there are much more positive words than negative words in all the reviews.



**Word Cloud**

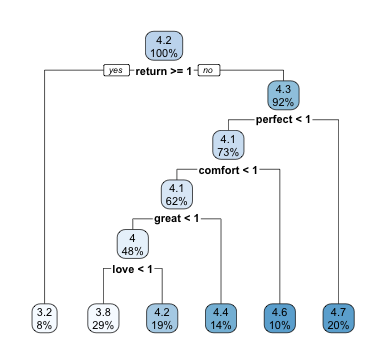
A word cloud is used to get some insights from all the words. Then, we decide to use a comparison to see the positive and negative words in reviews. We find that positive words such as "love", "perfect", and "beautiful" are widespread. In contrast, "loose", "worn", and "bust" are negative words.



**D. Predictive Analysis**

In this part, we are using the linear regression model and decision tree model to examine the predicting power of some frequently occurring terms. The goals of this analysis are two-fold. First, we want to build a machine learning model that predicts the rating using frequently occurring tokens in reviews. Secondly, we aim to develop a CART model which accurately predicts the sentiment of reviews. Those models can then be utilized to predict the ratings of other online reviews.

**Part I. Predicting review rating using tokens**

We start with data preparation using the codes in the previous section. After cleaning and tokenizing, we generate a matrix derived from term frequencies. We only keep terms that appear in at least 5% of documents. Then we split the data containing rating and term frequencies into a train sample with 70% of the data and test sample with the remainings. 

We first use a CART model to predict review rating using the terms. The plot below shows that the reviews containing the term “return” are rated lower than those without this term. On the other hand, reviews that contain “perfect”, “comfort”, “great” and “love” are rated higher than those without these terms.

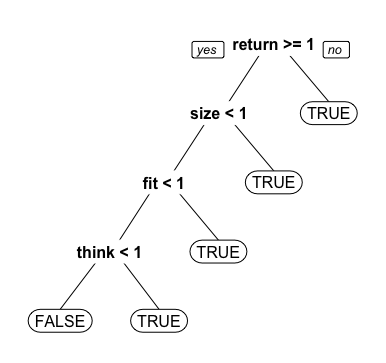
We then tried a linear regression model predicting the rating based on frequently occurring terms. Its summary shows that the most frequently occurring term "dress" is not predictive of review rating, but it also indicates many words such as "great", "love", "perfect" and "comfort" are predictive and are positively related to review rating, and the word "return" can lead to lower rating.

The **RMSE** is ***1.007508*** for the **CART model** and ***0.9183897*** for the **linear regression model**. Therefore, the linear regression model produces more accurate prediction than the CART model (lower RMSE).

**PART II. Predicting the sentiment of reviews**

First of all, we define the positive and negative sentiments by rating scores. We use ratings rather than sentiment lexicon as a predictor because sentiment lexicon may cause inaccuracy. For example, sometimes sarcasm may lead to positive words.

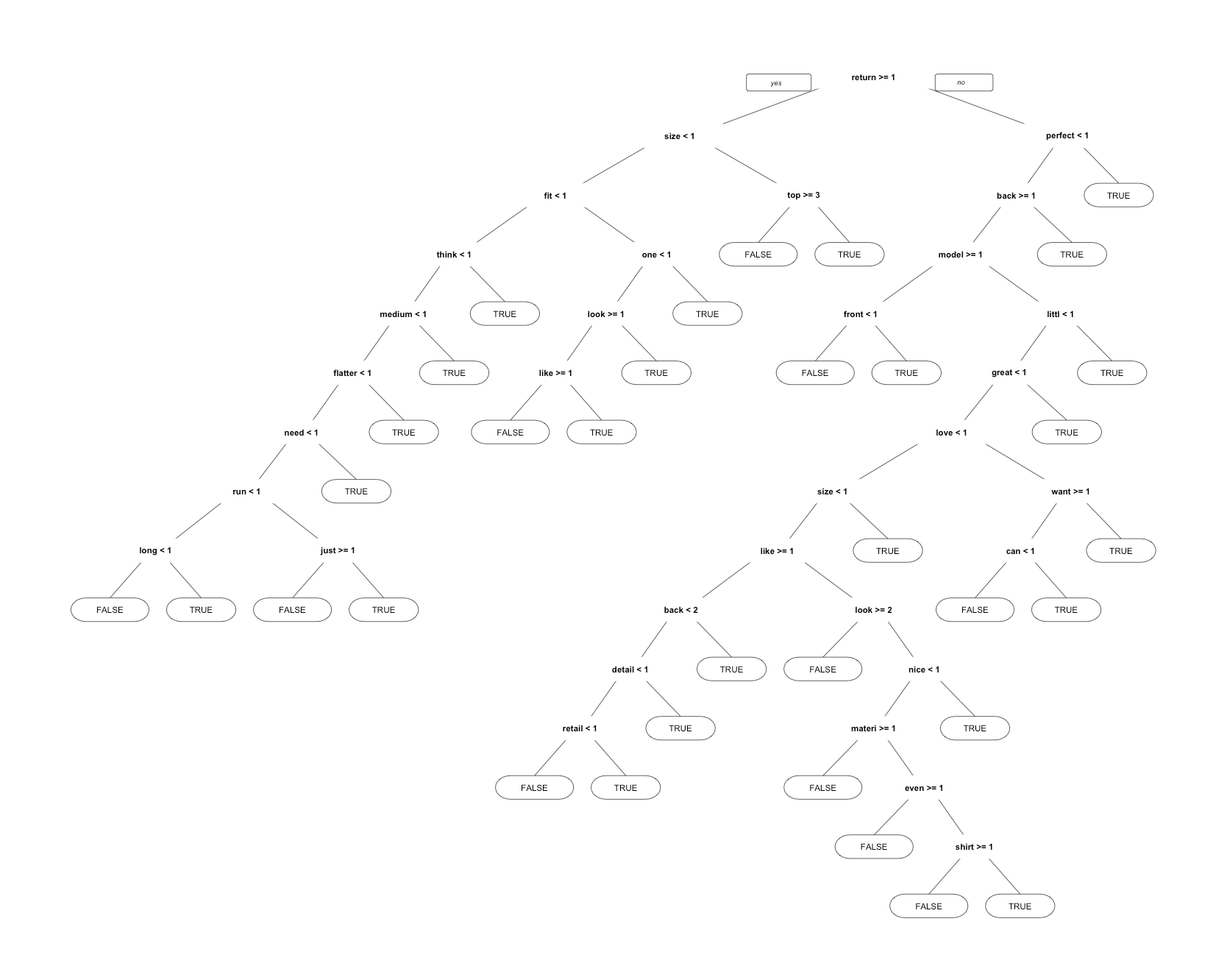
The rating score ranges from 1 to 5, so we associate 4 and 5 with a positive sentiment and 1 and 2 with a negative sentiment. The scoring 3 is likely to be neutral so we discard it so that it wouldn't affect our model accuracy. Then we create a binary variable for positive sentiment called "positive". The following steps of data preparation are similar to the previous part.

Before building a model, we examine the baseline accuracy that our model needs to surpass. The baseline accuracy is around **88% (0.880721)**, which means the dataset is highly biased towards positive reviews. Next, We build a CART model. The plot on the left shows that if a review contains the word "return" at least once, it will be labeled as negative. While words such as "size", "fit" and "think" lead to positive reviews.

The accuracy of this model is **0.8867846**, which has 0.6% over the baseline model.

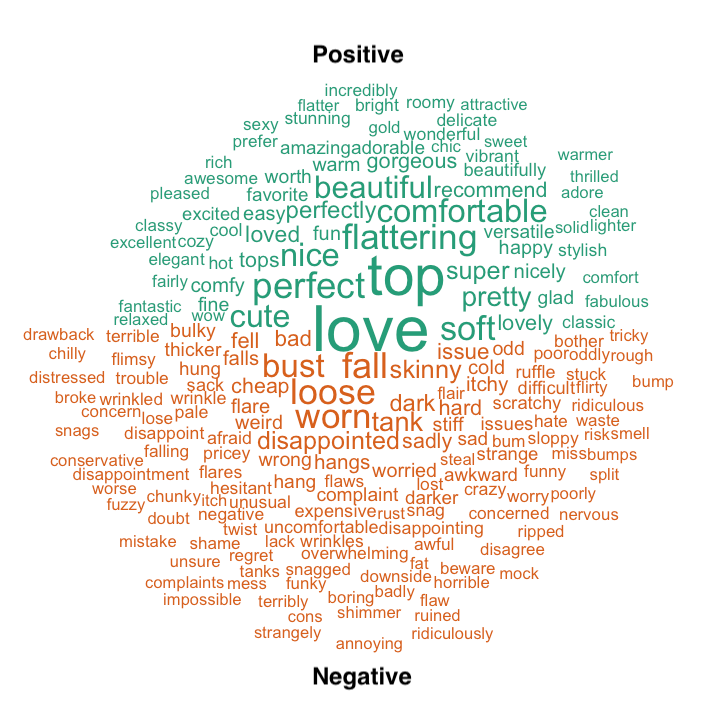
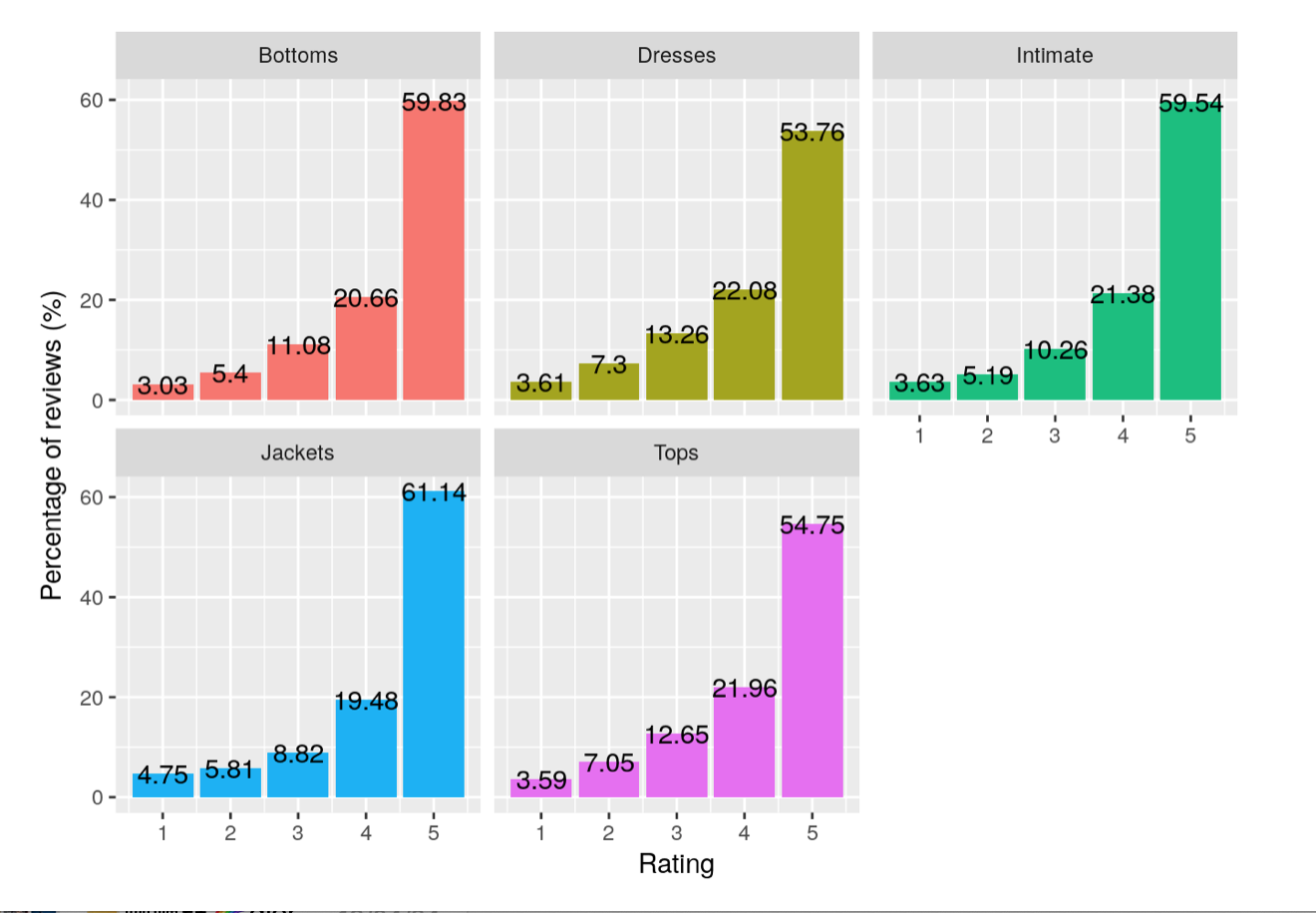
We then enhanced the model by introducing cross validation which helps us pick the optimal number of splits. The optimal parameter it gave was cp=0.002 which we will use to improve the CART model.

The improved tree has more splits that provide an overview of which words contribute most to the positive and negative sentiment.

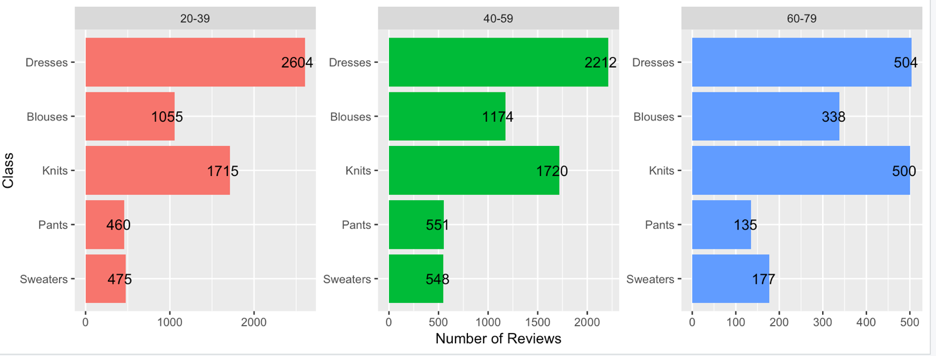
The improved model has an **89.7% (0.8970845)** accuracy which improved almost 2% over the baseline.

**5. Conclusion**

With all the analysis we used, we've found some interesting patterns of purchases. Women are more likely to buy dresses, knits, sweaters, and blouses online. However, customers might not be very satisfied with the tops and dresses purchased online. Comparatively, the rating scores for jackets bought are better, indicating a higher satisfaction level. After scanning the associated reviews, we find that negative words such as "worn", "loose", "bust", and "fall" show frequently. This may be due to the more flexibility in shapes and materials of jackets than tops or dresses. Usually, the fit of dresses and tops doesn't always tolerate the change of body shapes of different people.

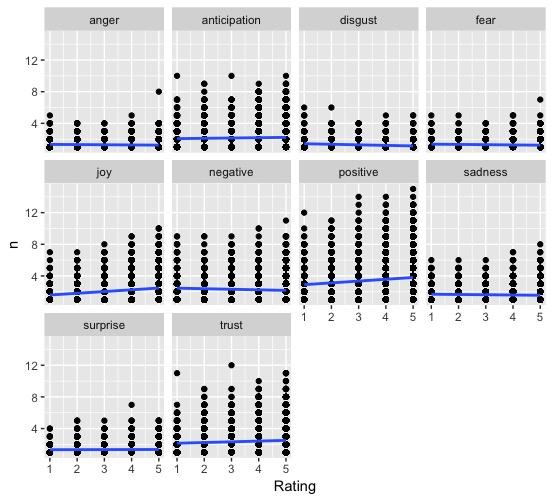
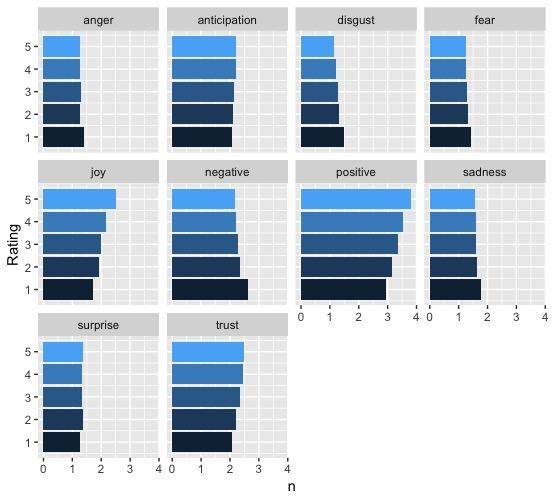


After exploring the relationship between age group and purchases, we've found that women aging from 20 to 60 tend to be more likely to buy dresses. Women aging from 60 to 80 prefer sweaters.

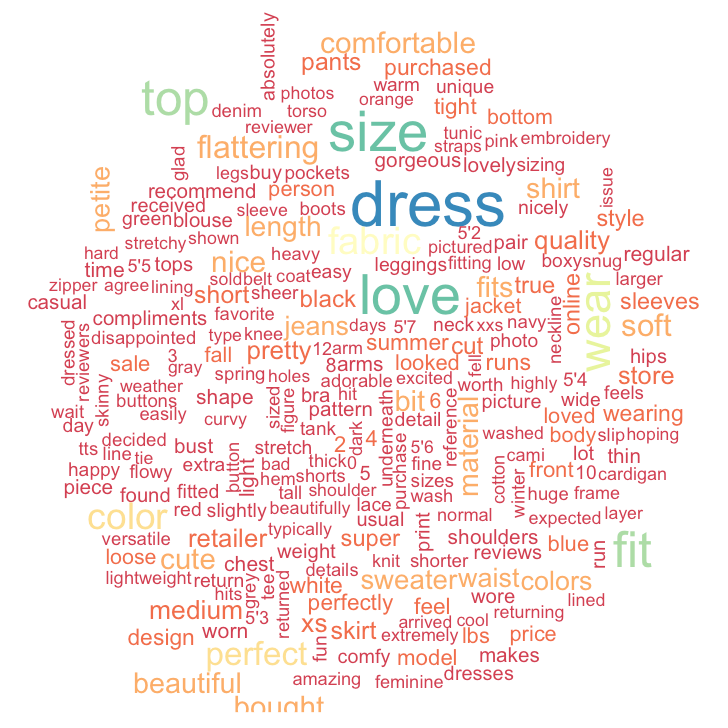
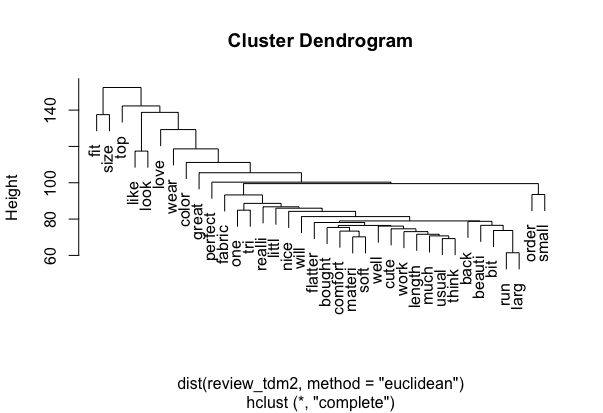


With sentiment analysis, we are also able to detect some emotions behind the words in review texts. With the simple bing lexicon, we've found that a higher proportion of positive words means better rating score. Also, a higher percentage of positive words is more likely to secure a recommendation on the purchased item.

With nrc lexicon, we can categorize emotions into more specific groups. We detect a very weak but existing correlation between emotions and rating scores. Good emotions such as anticipation, joy, positive, surprise, and trust have a positive relationship with review ratings. In contrast, bad emotions such as anger, disgust, fear, negative, and sadness have a negative relationship with review ratings.



With clustering and word cloud, we find that the word "size" appears very frequently, indicating that the size problems are closely related to satisfaction and review of the items. Well-Fitted clothes will always have better ratings.

We conducted predictive analysis which includes building models for predicting review rating and sentiment of reviews. The results of our predictive analysis indicate the predictive power of certain terms on rating and sentiment, and therefore the retailers can use this model to predict the rating of products.

To sum up, we would recommend the online women clothing retailers to improve customer satisfaction according to the results we got from this analysis. First, they need to provide more detailed sizing information of their products on the website or improve returning and exchange services. Second, they can improve the effectiveness of marketing campaigns by targeting customers of different age groups. Third, retailers can offer more selections for the most frequently purchased products like dresses and improve the low-rated products. It's essential for the online retailers to make sense of the unstructured text and draw actionable insights from it to enhance the customer experience.