WOMEN CLOTHING – PREDICTING RATINGS BASED ON REVIEWS

APRIL 16, 2020

SUBMITTED TO: PROFESSOR ROBERT SIMIONE

Table of Contents

Description of the Data

This dataset is about the buyer's reviews of woman's clothing in e-commerce. The dataset contains the details of customer reviews based on 10 different variables. Along with the review written by the customer, there is demographic information as well. For example, the age of the customer. The specificity of the reviews i.e. the dimensions the customer reviews the clothing item (recommend or not, division name, department name), will make it easier to analyze the data on different dimensions.

This dataset includes 23,486 rows and 10 feature variables.

Data Source: Kaggle CC0: Public Domain, Owner - https://www.kaggle.com/nicapotato

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

Description of the Variables

1. **Clothing ID:**
   Clothing ID is an integer categorical variable that refers to the specific piece being reviewed.
2. **Age:**
   A positive integer variable of the reviewers age. The data contains ages of reviewers which range from 18 to 99-year-old women.
3. **Title:**
   Title is a string variable for the title of the review. Titles of review are like "Perfect Fit!", "Runs Big", "Love it". The title characters vary from 0 to 12 words. From the title, it is easier to figure out what the review text will be about.
4. **Review Text:**
   The review text is string variable for the review body. The review text is a detailed description of the ***Title.*** In the review text the reviewer is explaining the title e.g.
   Review Title: Zipper is weird
   Review Text: "I ordered the dress on-line. this dress looked pretty, the material was soft and comfortable, the length was perfect, but the zipper was totally out of place. it bulged out in a weird way which made it look like the dress was torn. I had to return the dress. I normally don't see a problem at this level in retailer clothes."
5. **Rating:**
   Rating in the dataset is positive ordinal integer variable for the product score granted by the customer from 1 Worst, to 5 Best. Based on the review, the customer allocates a rating. The table below contains the number of customers that gave a particular rating integer out of the total 23,486 customers:

Rating	Percentage of Customers
1	3.59
2	6.66
3	12.22
4	21.62
5	55.91

6. **Recommended IND:**
Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended. In the dataset, out of the total customers, 4,172 customers did not recommend the product and 19,314 recommended the product.
7. **Positive Feedback Count:**
Positive Integer documenting the number of other customers who found this review positive. The positive feedback count ranges from 0 to 122. Hence, the maximum number of positive feedback count on a review is 122, the minimum being 0.
8. **Division Name:**
Categorical name of the product high level division. There are 3 divisions; general, general petite and intimates. The table below contains the details of division and the number of reviews in a particular division:

Division Name	Number of Customers
General	13850
General Petite	8120
Itimates	1502

9. **Department Name:**
Department name is the categorical name of the product department name. There are 6 departments namely bottoms, dresses, intimate, jackets, tops and trend.

Department	Number of Reviews
Bottoms	3799
Dresses	6319
Intimate	1735
Jackets	1032
Tops	10468
Trend	119

10. **Class Name:**
Class name is the categorical name of the product class name. There are a total of 20 classes for the company.

Questions/Problem being Addressed:

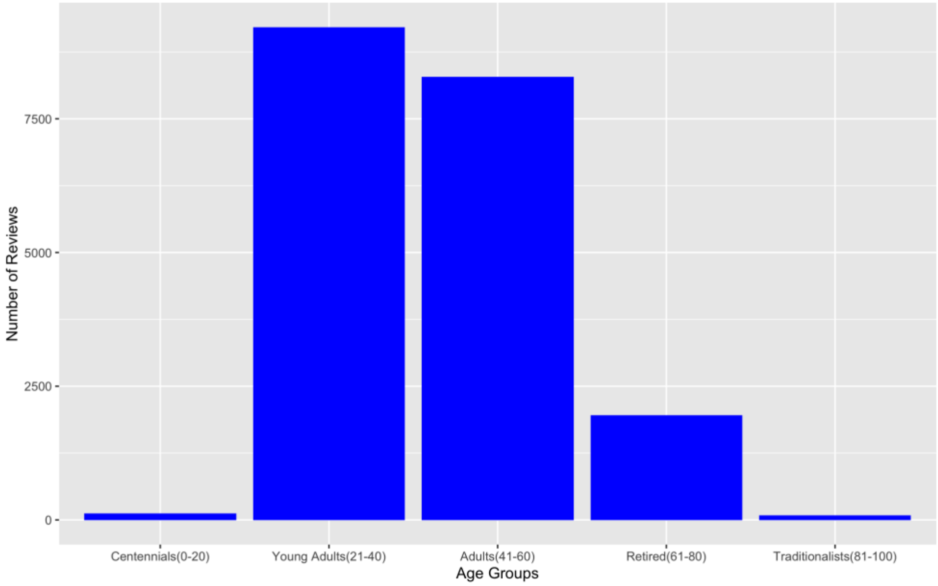
By processing and analyzing the data, we developed graphical representations of the characteristics of the text reviews. Further, we built a predictive model, by splitting our data set into train and test and predicted the rating of the clothing items, on a scale of 1-5. The prediction of the rating of the clothing items will help the e-commerce website make marketing and sales strategy to improve the performance and increase the revenues. We also performed sentimental analysis and applied text mining techniques.

Analytical Techniques:

Section 1: Exploratory Analysis of Different Variables

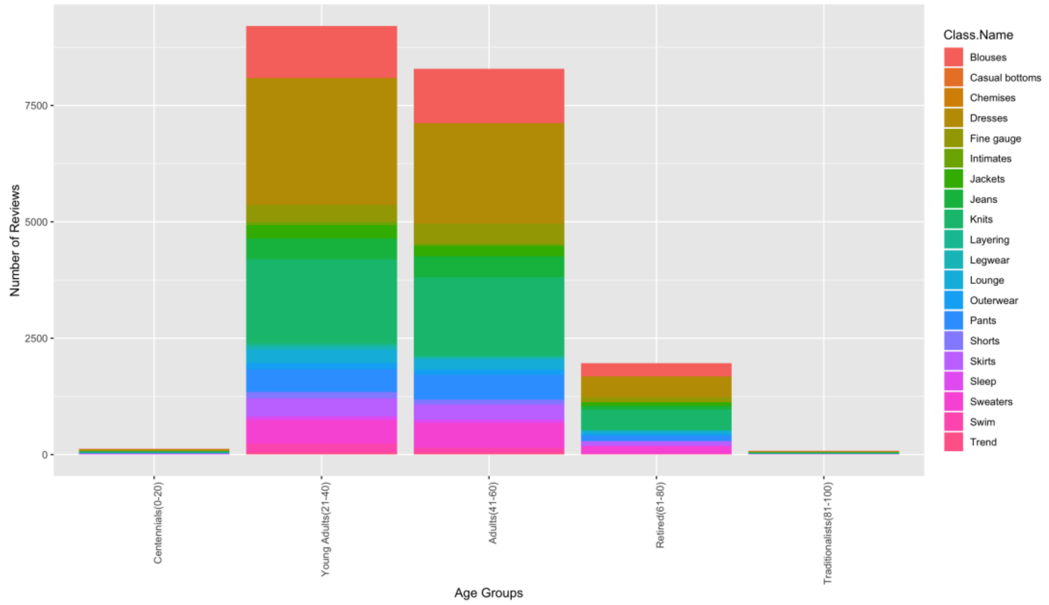
Part 1: Number of Reviewers by Age Group

By visualizing the number of reviewers by age group, it is evident that the age group between 21-40 have given the greatest number of reviews i.e. 9,209 out of the total reviews which are 19,662. The age group below 20 and from 81-100 have given the lowest number of reviews which are 123 and 82, respectively. Below is the visualization of the reviews by age group:



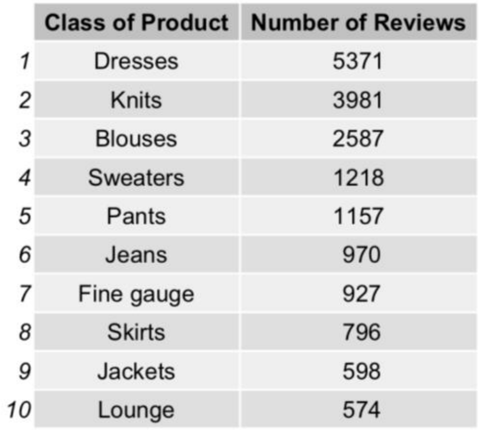
Part 2: Distribution of Age Group with Shopping Department

In both young adults and adults, aged 21-40 and 41-60, dresses are the most commonly purchased items following by knits and blouses among the 20 clothing categories.



Part 3: Most Review Products by Class of Product

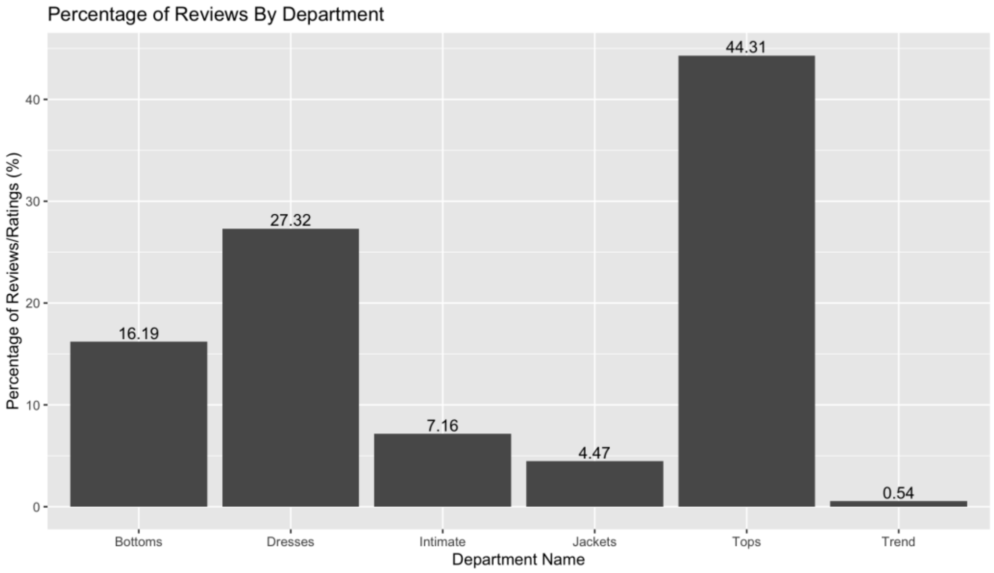
According to the number of reviews by class of product, dresses and knits along with being the most purchased items, have the highest number of reviews as well.



	Class of Product	Number of Reviews
1	Dresses	5371
2	Knits	3981
3	Blouses	2587
4	Sweaters	1218
5	Pants	1157
6	Jeans	970
7	Fine gauge	927
8	Skirts	796
9	Jackets	598
10	Lounge	574

Part 4: Distribution of Reviews by Department

Out of the five departments, tops have the highest percentage of reviews and ratings i.e. 44.31% followed by dresses 27.32%. The jackets and trend department received the lowest number of ratings and reviews.



Section 2: Exploratory Analysis of Text Column

To understand the statistical features and distribution, the numeric column 'Rating' and text column 'Reviews.Text' was analyzed.

Part 1: Ratings- Mean and Median

The average rating for all the ratings given by customers is 4.18 and the median rating is 5. This signifies that most of the ratings are positive and customers have positively related the products. Hence, good ratings.

Part2: Review.Text- Characters, Words and Sentences Count for Reviews, Length and Ratings Correlation

The mean and median of character count, word count and sentences count are similar which implies that the counts distribution is highly symmetric and the skewness is low across the individual counts.

The correlation between ratings and characters, words, and sentences are -0.05478506, -0.05622374 and 0.01813276, respectively. The figures show that there is a very low correlation and hence brings us to the conclusion that the length of the review does not impact the rating given.

Correlation: Ratings and Upper-Case Letters

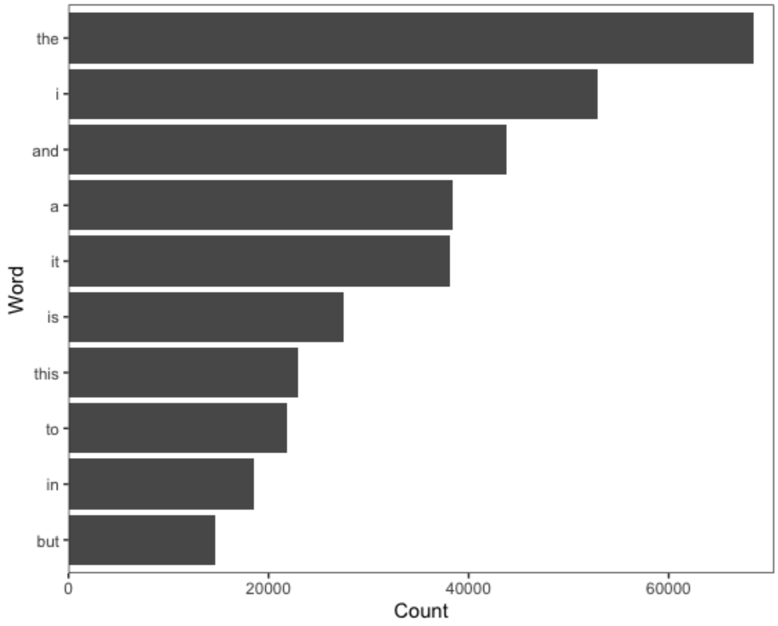
The correlation of 0.05779606 between ratings and upper-case letters means that upper-case letters do not affect the ratings give.

Correlation: Ratings and Exclamation Marks

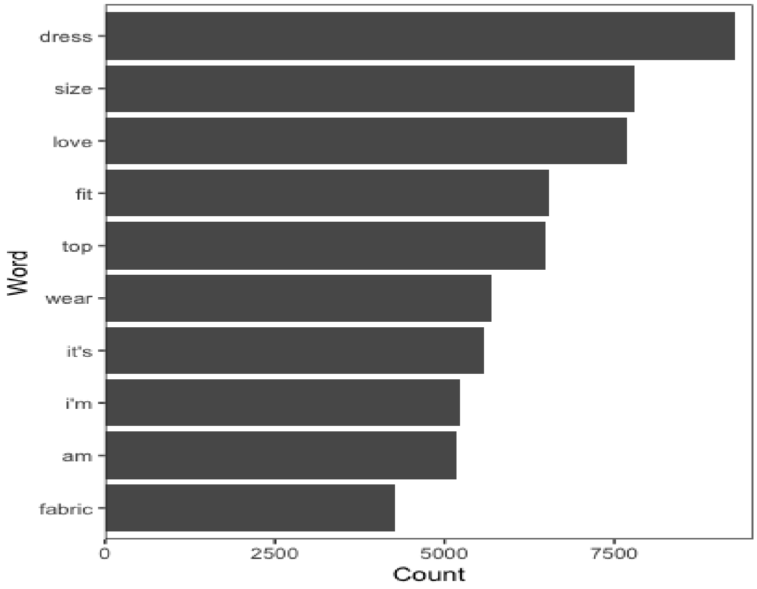
Between ratings and exclamation marks the correlation is 0.1776584 which is also very low. But impacts ratings more than upper-case letters do.

Part 5: 'Review.Text'- Most Common Words in Reviews

Without moving the stop words the 10 most common words that are found in the reviews are:



After moving stop words, the 10 most common words are:

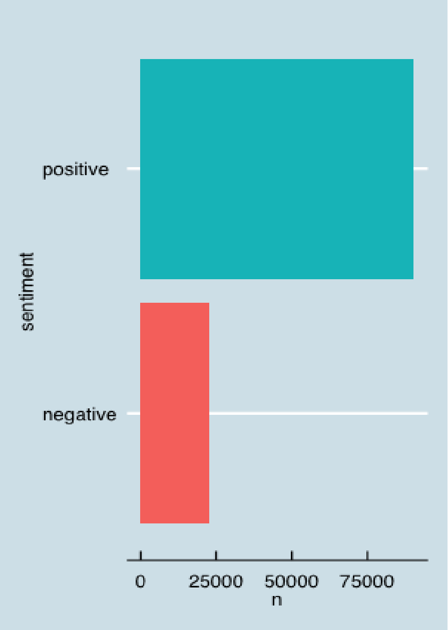


Dress is the most commonly used word which was used 9,263 times in reviews followed by size which was used 7,796 times. A point to be noted here is that dresses were a product type that was purchased the most by customers, a product that was also reviewed most by customers, and the most commonly used word in reviews.

Section 3: Sentiment Analysis on Review.Text and Formation of Wordclouds

Part 1: Binary Sentiment (Positive/Negative) – Bing Lexicon

The number of positive words used in the reviews are 90,474 whereas the number of negative words is 22,938. This shows that approximately 80 percent of the words used in the reviews are positive and hence also justifies the higher ratings given by customers.

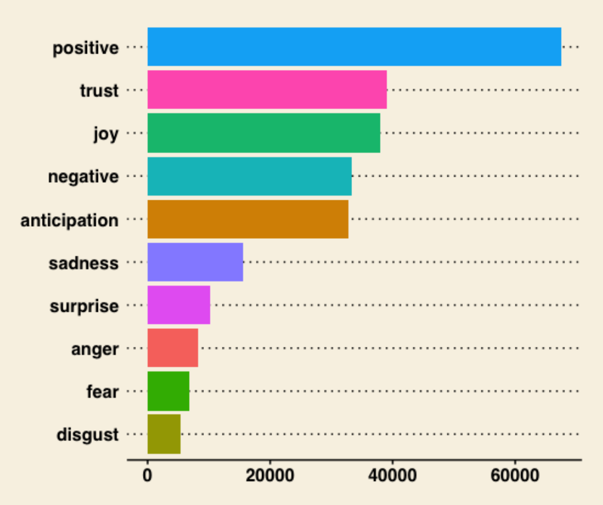


Moreover, the correlation between positive words and ratings was 36 percent which indicates that positive words used in a review does not necessarily mean a good rating was given.

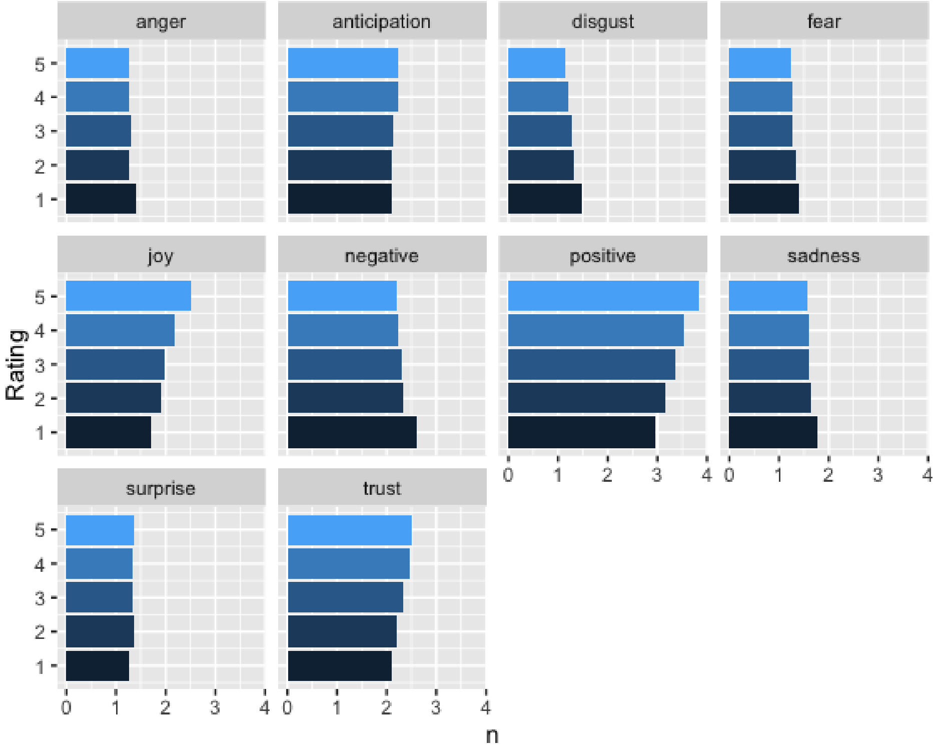
Part 2: NRC Sentiment Polarity Table – Lexicon

According to the NRC sentiment polarity table the count of '-1' words = 31221 and '1' words = 63759.

Part 3: Emotion Lexicon - NRC Emotion Lexicon



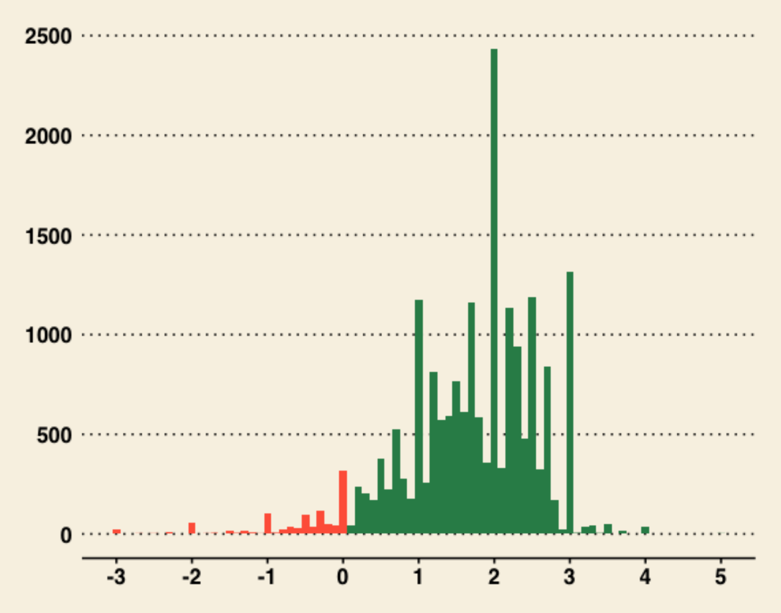
According to the emotion lexicon, the positive emotion has the highest count followed by trust. Hence, positive emotions are mostly used in reviews.



As displayed in the above graph, there is a rise in the number of positive words and emotions as rating goes up and drop in the number of negative words e.g. disgust words as the rating goes up.

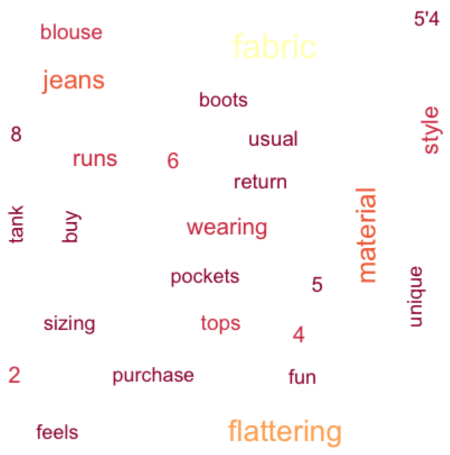
Part 4: Sentiment score Lexicons - Afinn Lexicon

Sentiment score lexicon shows that the lowest sentiment score for any 'Review.Text' is -3 and the maximum is 5. The mean sentiment score is 1.71 and the median is 1.85.

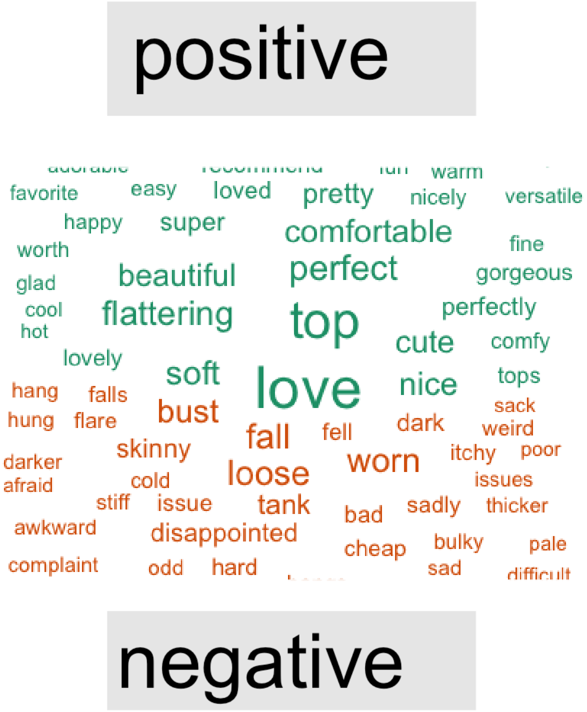


Part 5: Word Cloud of 100 words (Except Stop Words)

Below is a word cloud of 100 Words.



The following visualization word cloud shows positive negative words. Green are positive words and red are negative.



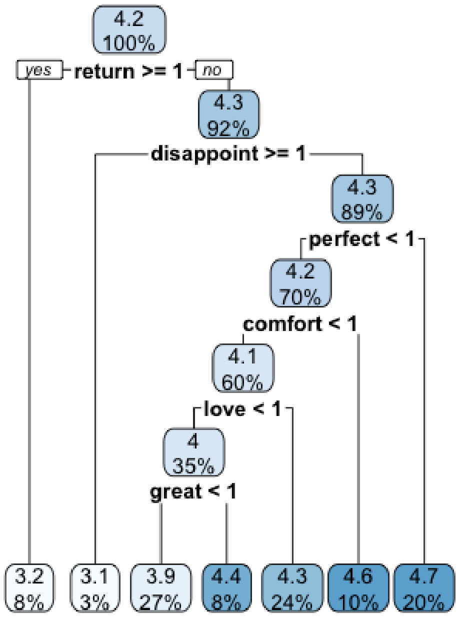
Section 4: Data Preparation for Predictive Modelling (TF, TF-IDF of text columns 'Review.Text' and 'Title'), and Exploratory Analysis from Corpus for 'Review.Text'

1. Checked for astronomical variables
2. Checked for outliers
3. Determined correlation among variables
4. There are only 16% NA in title, 4% in Review.Text, and 0.06% in Division.Name, Department.Name and Class.Name respectively. Given the low percentage of NA entries, we removed the NA values and did no imputation.
5. In order to prepare the data for sentiment analysis and rating prediction, we did the following
   - Created a corpus from the variable ‘Review.text’
   - Used tm_map to transform text to lower case
   - Removed English stop words using the following dictionary tm::stopwords(‘english’)
   - Removed punctuation
   - Removed whitespace
6. Created a dictionary
7. Used tm_map to stem words
8. Created a document term matrix
9. Removed sparse items (words that appeared in less than 5% of the reviews)

Section 5: Predictive Modelling

According to the CART method, the RMSE was 1.009915 on the Reviews.Text and for Title is 1.075686. According to the Regression method, the RMSE for Titles is 1.06697 and for Reviews.Text is 0.9013822. In both methods, Reviews.Text is a better predictor of the ratings given as it has a lower RMSE in both methods.

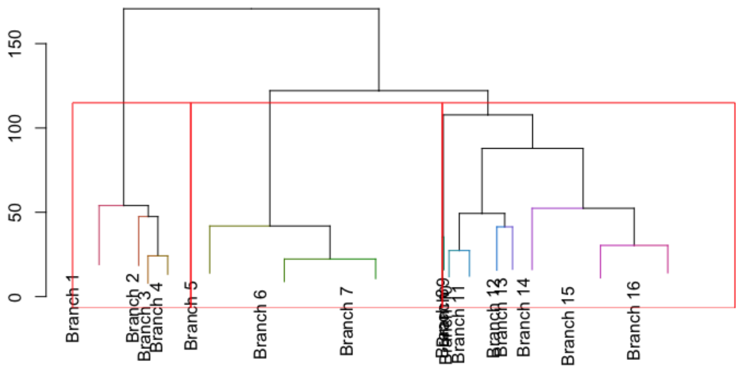
The regression tree below shows the words and how the words used in reviews impact whether the item was returned or not. For example, if the review has the word ‘disappoint’, there is a 92% chance the item was returned.



Both methods, i.e., TF and TF-IDF give the exact same RMSE for both Review.Text and Title. Review.Text always gives lower RMSE than any method used for 'Title'. So, we should use Review.Text as a predictor going forward. For best RMSE, we need to use the regression method of predictive modelling but might need to compare results from TF and TF-IDF methods.

Section 6: Clustering and Predictive Modelling

First, the data was prepared for cluster analysis. Simple regression was run and since cluster analysis is sensitive to scale, the data was normalized. Then a hierarchical and k-means cluster analysis was run. Based on the plot achieved, a 3-cluster solution looks good.



After doing the k-means clustering, total SSE plots, ratio plot and silhouette plot, the results were applied to the test set to compare the results. For this purpose, based on cluster membership, the data was split into train and test. The predication was done for each cluster and then the results were combined. The results indicated the following: SSE on Entire data = 2262.32, SSE on Clusters = 1643.99. Therefore, prediction using clusters is more accurate, as the standard error is less.

Moreover, prediction using Trees was also done and the results were as following: SSE on Entire data = 2262.07, SSE on Clusters = 1643.25. In Trees also, the prediction using clusters is more accurate, as the standard error is less.

All in all, the lowest Error is achieved when cluster analysis is done with Trees to predict.

Conclusion:

Our goal for this project was to predict the ratings for the reviews that will help the company develop graphical representations of the characteristics of the text reviews, build a predictive model, by splitting our data set into train and test and predict the rating of the clothing items, on a scale of 1-5. All of this was to predict the ratings of the clothing items that will help the e-commerce website make marketing and sales strategy to improve the performance and increase the revenues.

The results of the analysis include all the goals listed in deliverable 1. We successfully developed graphical representations of text reviews, determined correlations between words and ratings/text reviews, determined positive and negative words and most common words/characters. Moreover, we built a predictive model using TF, TF-IDF, Regression Cart and Trees method and the individual ratings of for the reviews were predicted using all the methods listed. The lowest RMSE was achieved using the Trees method and hence, we have come to the conclusion to predict using clusters with the Trees method which also has the lowest standard error.