Project Report

BAN 675

Text Mining and Social Media Analytics

Review-based Popularity Analysis of the Online Fashion Industry

Team Members:

Antony Paulson Chazoor

Komal Jaiswal

Muhtasin Rahman

Shweta Patil

Table of Contents

SUMMARY

Popularity among customers drives sales and profits for businesses and reviews greatly influence popularity of products. especially online. The project aimed to analyze customer reviews in order to predict popularity of clothing from a retail website. The products with highest number of reviews formed the basis of the analysis as they were denoted the most popular. A corpus of all their reviews was assembled by removing the stop words and punctuations. Part of speech tagging was implemented to identify the adjectives. For each clothing, occurrence of the most frequent adjectives was considered and the resulting data was partitioned into training and validation. Logistic regression and decision trees were chosen as models and trained with 50 adjectives as predictors and “Popular” as a binary outcome. Ultimately, the models were successful in predicting, based on the presence of certain key adjectives, whether the clothing was popular with 96-98% accuracy.

1 INTRODUCTION

In most business scenarios, majority of profit is driven by the products which consistently generate high sales. These few highest selling products are the company's "bread and butter", that is their most revenue generating items. The "fast fashion" industry is no stranger to this phenomenon. It is an extremely profitable arena; where products made quickly using cheap labor and materials are treated similarly by the customers as clothes are bought as fast as they are discarded in favor of new ones. Perceived current "fashion value" of such items are dependent on word of mouth. Popularity is what builds perception.

The difference between quality and popularity, however, must be clarified. In many cases, popularity is not necessarily correlated with quality. There are products and businesses e.g. Walmart, fast food, 3rd party manufactured electronics which are not deemed as high quality but are still popular among consumers and sell better than their alternative, good quality products e.g. Whole Foods Market, healthy food, original company manufactured electronics. For the general public, price is a major factor in their buying decisions. As much as quality is important to them, the value or "bang for buck" proposition matters more. Accordingly, retailers and businesses who are profit-driven are more focused on what's popular among buyers and sell well rather than what is perceived as high quality.

In the online retail space, where customer satisfaction and feedback are of paramount importance in dictating popularity, reviews are extremely influential. Since there are no physical trials, even more so. Almost all online retail websites have a section for customer reviews which is populated by input from the customers. These words are a reflection of how the product is perceived by the majority of the customers: What works, what doesn't, how they used it, tips for others etc. These in turn influence would be buyers in their purchasing habits which dictate sales.

The goal of this project is to analyze review text from a large online retail clothing store in order to predict whether a certain apparel was popular or not based on the presence of certain key words. Implementing such analyses into everyday business proceedings would allow the retailer to know trending consumer preferences and help them decide future product design, stock, promotions etc. in accordance. Capitalizing on the demand and popularity of these items is supremely important for continued business success.

2 RESEARCH QUESTIONS

Since the principal aim is to analyze clothing popularity based on reviews, there are some interesting questions that could be answered using the data set content.

- What are the most popular items in data set?
- Which words do reviewers tend to use the most?
- Do popular products have characteristic key words?
- How to accurately predict popularity based on review content?

3 IMPLICATIONS OF RESEARCH

With a successful model, the retailer can implement a system to analyze the initial reviews from customers for new products and predict whether they will prove popular based on historical standards. Unpopular products could be changed during the production process or advertised differently in line with consumer preferences. Consumer feedback would enable the retailer to adjust their marketing and research initiatives.

Other possible uses for this research include:

- Recognize trends and consumer buying preferences.
- Understand which perceived qualities majorly influence apparel sales.
- Improve or change low selling items.
- Implement customer service policies to curb consumer criticism.
- Engender customer engagement through website design and promotions.
- Decrease return rate and avoid restocking fees which greatly affect profit margins for

4 DATA DESCRIPTION

The data set was taken from the data repository website, Kaggle. It includes 23486 rows and 10 feature variables. It has numerical ratings as well as text reviews for clothing items for women. Each row corresponds to an individual customer review and includes the following variables:

- **Clothing ID:** Integer categorical variable that refers to the specific piece being reviewed.
- **Age:** Positive integer variable of the reviewers age.
- **Title:** String variable for the title of the review.
- **Review Text:** String variable for the review body.
- **Rating:** Positive ordinal integer variable for the product score granted by the customer from 1 (Worst) to 5 (Best).
- **Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count:** Positive integer documenting the number of other customers who found the review positive.
- **Division Name:** Categorical name of the product hierarchical division.
- **Department Name:** Categorical name of the product department.
- **Class Name:** Categorical name of the product class.

Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

Figure 4.1 Sample Data Screenshot.

5 EXPLORATORY DATA ANALYSIS

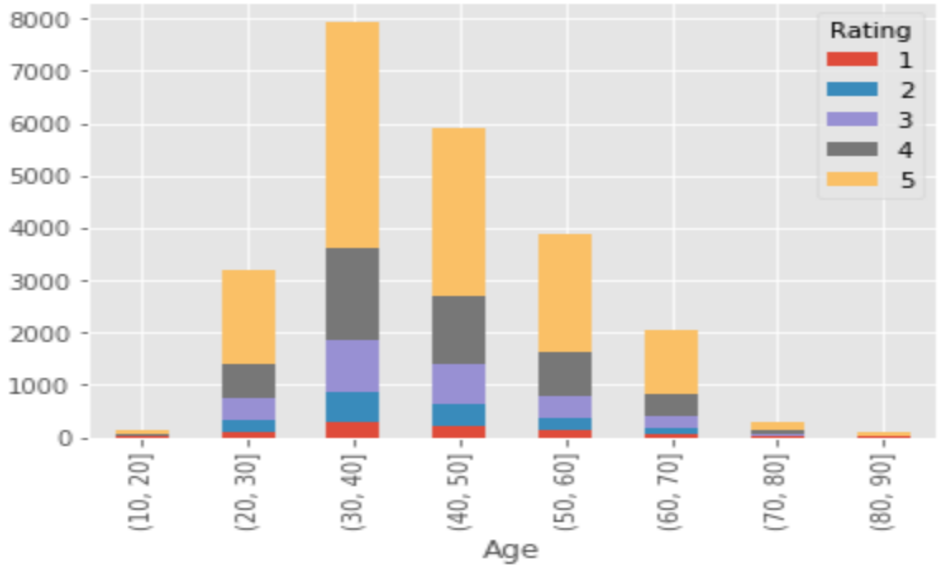

Figure 5.1 Reviewer Age Distribution with Product Ratings.

From the above bar plot, it can be observed that the customers aged 10-20 do not leave reviews. In this age group teenagers generally don't care about reviewing apparel online. The age group 30-40 gave more high ratings of 5 compared to all the other age groups. In fact, this is the age group that left the most reviews followed by the 40-50 age group who also generally rate products highly. In terms of writing reviews, then come to 50-60 and 60-70 age groups who are not as prolific as the previous ones. Similar to teenagers, consumers who are 70 years and older barely write any reviews as they have other engagements in their life.

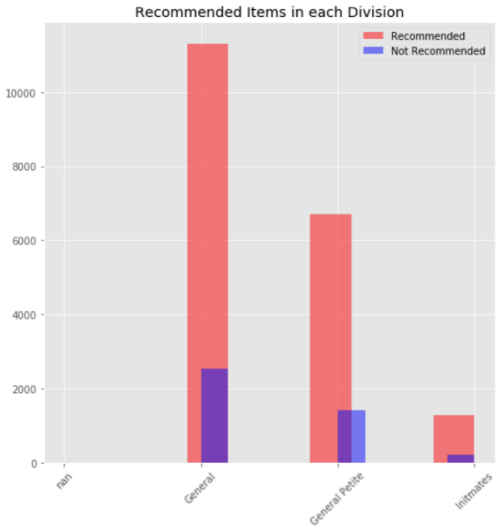

Figure 5.2 Recommendations by Division.

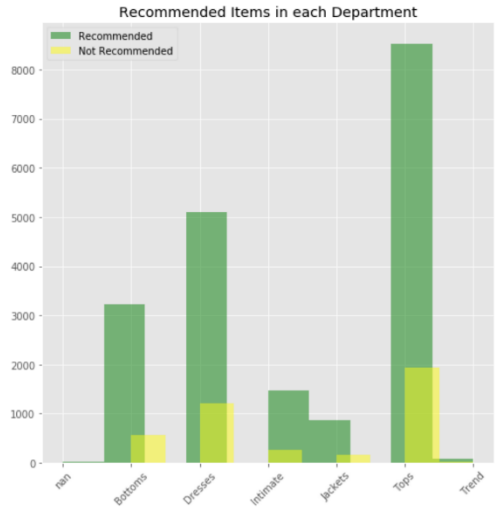

Figure 5.3 Recommendations by Department.

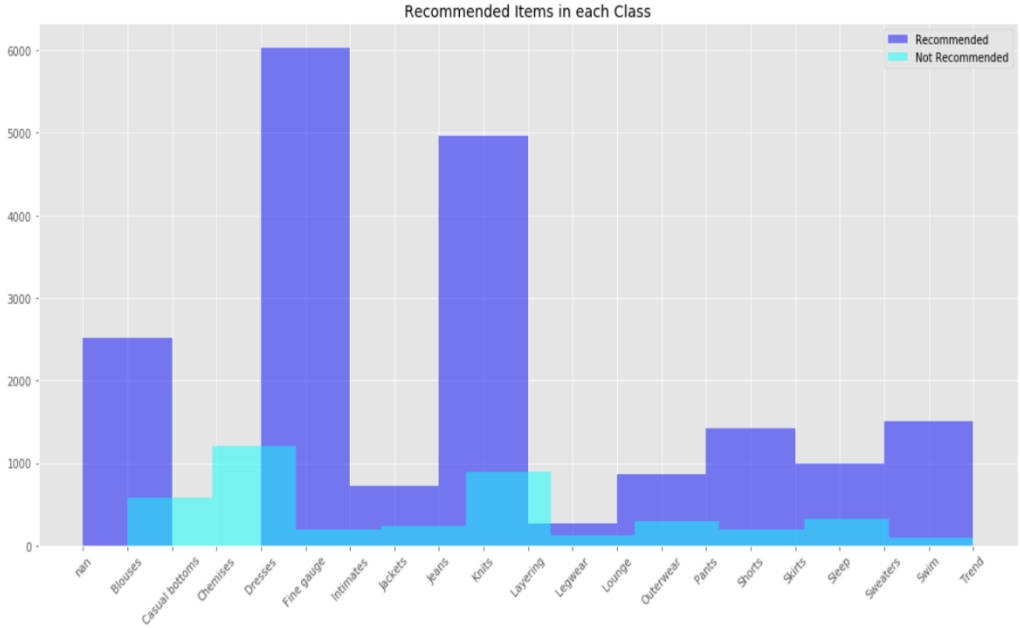

Figure 5.4 Recommendations by Class.

Observing the preceding histograms, we can conclude the following:

- The most recommended apparels belong to the General division, followed by the Everyday Petite and Intimates division.
- Tops are recommended the most, followed by Dresses and Bottoms.
- For classes, Dresses, Fine Gauge and Intimates are most recommended, followed by Jeans and Knits and then Blouses.
- Customers are much more likely to recommend something rather than not recommend it.

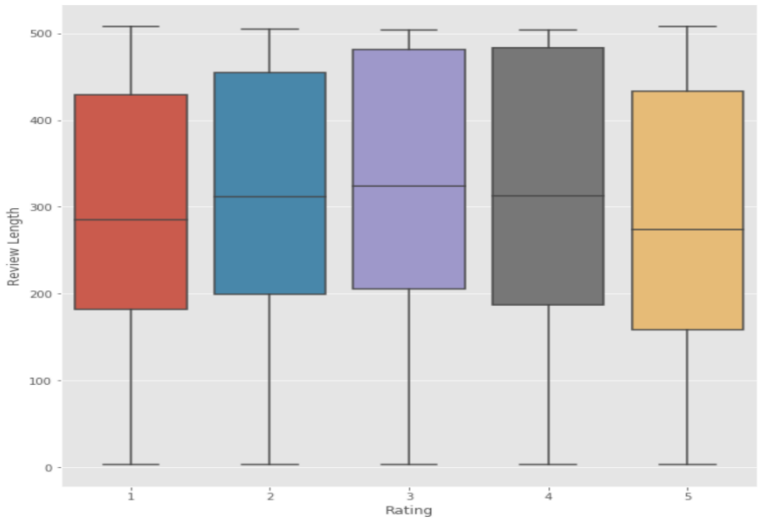

Figure 5.5 Review Length vs. Ratings.

From the above boxplot, its clear that reviews are almost always more than 200 words in length. Customers giving ratings of 3 and 4 tend to write more than those who give ratings of 2. Word length for reviews for products with the highest (5) and lowest (1) are on average, less than 300.

6 METHODOLOGY

The initial dataset contains 1205 unique clothing items and their associated data points. The following steps were taken in order to explore the data even more and make a final customized dataset of popular products in which the intended analyses could be done.

Step 1 Identification of Popular Clothing Items

We created a data frame with only the clothing IDs and their reviews. The clothing ID's where no reviews were present were removed from our analysis. The top 100 clothing ID's with maximum reviews were considered for further analysis. These formed the basis of the most popular clothing items.

	Clothing ID	Count
1078	1078	1024
862	862	806
1094	1094	756
1081	1081	582
872	872	545
829	829	527
1110	1110	480
868	868	430
895	895	404
936	936	358

Figure 6.1 10 Most Popular Clothing IDs with Review Count

Step 2 Word Cloud Visualization for Popular Words

All the reviews of each popular items were combined into a massive text corpus. After removing all irrelevant data such as stop words and punctuation, the final review corpus consisted of 482,899 words. In the generated word cloud below, its observed that dress, fabric, top, look, color, love, one, shirt are some of the words that occur the most.

Figure 6.2 Word Cloud of the Review Corpus of the 100 Most Popular Items

Step 3 Identification of the Popular Adjectives

The part of speech tagging method was used to tag each word in the review corpus with the part of speech for that word. This was further utilized to identify all the adjectives in the corpus since descriptive terms are crucial in identifying review sentiment. We included the 'JJ', 'JJR' and 'JJS' POS tags for the adjectives. There were roughly around 124,607 unique ones. The top 50 adjectives which have at least 500 recurrences were selected for the next step. The graph below shows the chosen adjectives and their number of repetitions in the corpus. The adjective 'top' has the highest frequency, followed by 'great', 'small', 'little', 'soft' and 'fit', 'wear' and many others.

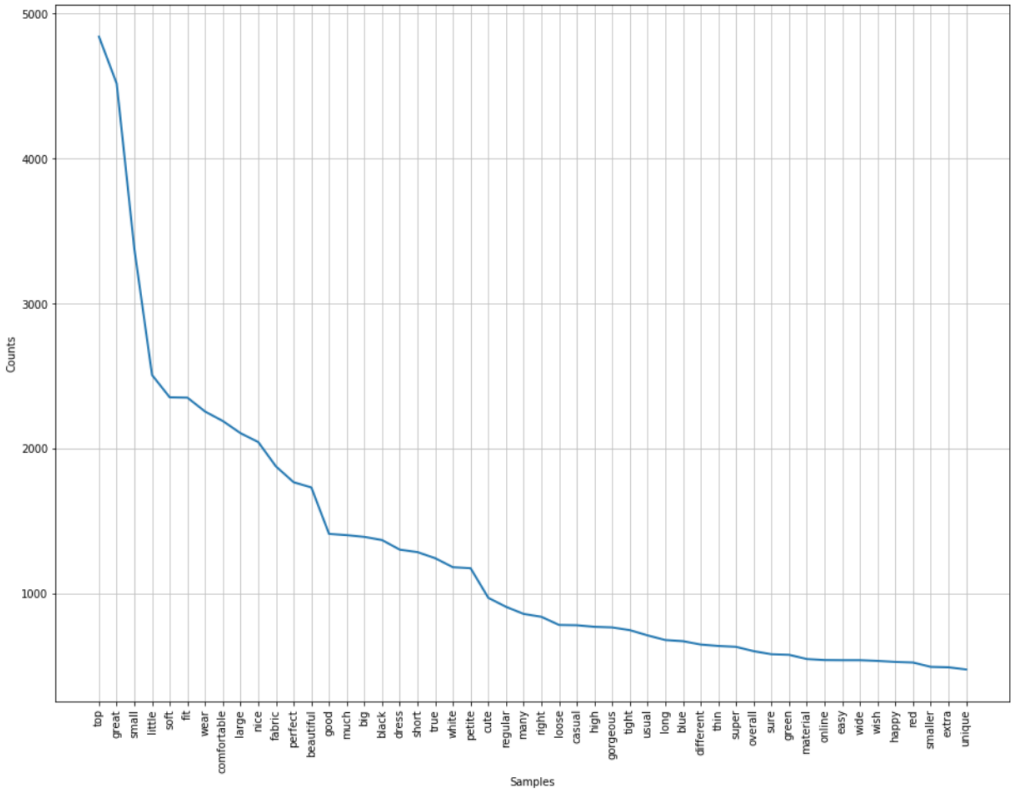

Figure 6.3 Top 50 Adjectives with Most Repetitions

Step 4 Building a Classification Dataset

We transposed each of the chosen 50 adjectives into their own columns thus creating 50 unique columns. The new dataset was updated with clothing ID, reviews and 50 individual adjective columns, each showing their frequency of appearance in the reviews. The binary variable 'Popular' which was previously created based off the number of reviews for each clothing ID was added to the dataset. The final dataset consisted of columns such as Clothing ID, review corpus, 50 columns for each top adjective and Popular column denoting whether the item is popular or unpopular.

Step 5 Data Partition

The newly created dataset was split into training and validation sets in the ratio of 68:32 respectively for the purposes of running analyses.

Step 5 Classification Model Selection

We chose two different classification techniques: Logistic Regression and Decision Tree. These models will be applied on the dataset to classify the 'Popular' binary variable based off the 50 adjective frequency columns as predictor variables. After the two models are run, their classification of popularity results will be compared and the model with the higher predictive accuracy rate will be denoted as the best.

7 ANALYSIS AND RESULTS

First, we are considering logistic regression to predict whether the clothing ID is popular or not based on the presence of the adjectives within each review. Thus, our model has 50 unique predictors, the top 50 adjectives. The model was built on the training set consisting of 824 records and validated on the validation set consisting of 384 records.

```
: #Creating the confusion Matrix
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
print(confusion_matrix(y_Val, pred))
print (accuracy_score(y_Val, pred))

conf = pd.DataFrame(confusion_matrix(y_Val, pred), index = ['Unpopular', 'popular'],
                    columns = ['Predicted Unpopular', 'Predicted Popular'])
conf
```

```
[[319   6]
 [  7  23]]
0.9633802816901409
```

Figure 7.1 Logistic Regression Code Snippet and Results

	Predicted Unpopular	Predicted Popular
Unpopular	316	4
popular	9	26

Figure 7.2 Logistic Regression Confusion Matrix Table

The model's prediction accuracy was calculated based on the confusion matrix. The logistic regression was able to predict results with 96.3% accuracy. We misclassified only 13 items. The false negatives where the actual clothing item is popular but was classified as unpopular are only 9. Similarly, 4 unpopular items were misclassified as being popular by the model. Still, the high accuracy shows that more often than not, the logistic regression model was able to correctly classify a clothing items' popularity based on the presence of the adjectives in its review content.

Then we move on to decision trees. The best feature of this classification tool is that the results are easy to interpret. Therefore, we are considering the decision tree to predict whether the clothing ID is popular or not based on the presence of the adjectives in each review.

```
#Decision Tree model and its confusion matrix

from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier()
dt.fit(X_train,y_train)
pred2 = dt.predict(X_Val)

from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
print(confusion_matrix(y_Val, pred2))
print (accuracy_score(y_Val, pred2))
conf = pd.DataFrame(confusion_matrix(y_Val, pred2), index = ['Unpopular', 'popular'], columns =
                    ['Predicted Unpopular', 'Predicted Popular'])
conf
```

```
[[321   4]
 [  3  27]]
0.9802816901408451
```

Figure 7.3 Decision Tree Code Snippet and Results

	Predicted Unpopular	Predicted Popular
Unpopular	321	4
popular	3	27

Figure 7.2 Decision Tree Confusion Matrix Table

The Decision Tree model was able to predict results with 98% accuracy. We misclassified only 7 items. The false negatives where the actual clothing item is popular but the model predicts as unpopular are only 3. Only 4 actual unpopular items are misclassified as being popular. Just like the logistic regression model, using decision trees allowed us to correctly classify the popularity of an item based off its review content quite consistently. But comparing the two, the accuracy of the test data is 98% in this case which is a little higher when compared to Logistic Regression model. Thus, the decision tree model was identified as the best model for the purposes of our project.

8 CONCLUSION AND FURTHER IMPROVEMENTS

With this project we sought to correctly classify an item as 'Popular' or 'Unpopular' based on the presence of certain adjectives within the content of the item's review. For each clothing item on sale, the frequency of reviews was used to detect whether the item was popular or not. Using part of speech tagging on the review corpus, we were able to identify unique adjectives for training our model. Utilizing these adjectives as predictors, we utilized two models in predicting whether a clothing item fell under the category "Popular" or "Unpopular". Both the Logistic Regression and Decision Tree classification models were very successful in predicting the popularity of the clothing items but the model with highest precision values is the Decision Tree model with accuracy of 98%.

Model	Accuracy
Logistic Regression	96.3%
Decision Tree	98%

Figure 8.1 Models with Accuracies

Interestingly, both models learned to classify unpopular items better (300+ in the confusion matrices) than popular items (around 26). So, in order to further improve this model to predict popular items better, perhaps more records of popular products could be collected and used for training. Additionally, incorporating other features from the original dataset such as Recommendation, Ratings Score etc. would result in a more holistic prediction model which would provide even better results.

9 REFERENCES

1. Aggarwal, C., & Zhai, C. (2012). *Mining text data*. New York: Springer.

2. Text Preprocessing in Python: Steps, Tools, and Examples. (2019). Retrieved from https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908

3. Ultimate guide to deal with Text Data (using Python) - for Data Scientists and Engineers. Retrieved from https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

4. Women's E-Commerce Clothing Reviews. (2019). Retrieved from https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews