*University of Essex*

## Department of Mathematical Sciences

MA981: DISSERTATION

Sentiment Analysis of Women’s E-Commerce Clothing Reviews with Detailed EDA

### Sayali Kambli

Supervisor: Dr Lisa Voigt

Dec 02, 2021

Colchester

## Contents

1. [Introduction](#_bookmark0) 5
2. [Literature](#_bookmark3) Review 7
3. [Methods](#_bookmark6) 8
   1. [Data](#_bookmark7) Analysis 8
   2. [Data](#_bookmark8) Retrieval 8
4. [Conclusions](#_bookmark9) 9

**List of Figures**

## List of Tables

**Abstract**

Every E-commerce company ask their customers to participate in survey or ask them for ratings and reviews. Customers also like to give brief review about how they received their product. These reviews are purely human language filled with various emotions such as happy, sad, satisfied, anger, etc. As these reviews hold very important information about product it has to be processed well with various techniques. Sentiment analysis is way to analyze this written human language reviews and extract the main emotions review wants to give out. So that company can identify whether their product has reached to customer satisfaction or not and they can improve it as needed.

In this project I aim to fulfil 3 goals. First is to perform complete Exploratory Data Analysis of the data set. It will help me show the relationship between different attributes. Second goal is to perform Sentiment analysis on the dataset’s Reviews attribute to find out how are they overall reviews are displaying the emotions of customers and how it may help in product development in future. Third goal is to use machine learning algorithms to predict the recommendation values of the clothing attribute.

## Introduction

Sentiment analysis has become a very trendy topic in today’s life. Reason behind that is sentiment analysis can be used in all fields where customer reviews are retrieved. All companies take survey, feedback or reviews from their customers. Sometimes it is only about rating but with rating people like to give written comment about product where they can express their feelings about product. As they put effort in writing such feedback it is important to extract correct interpretation from the result. Sentiment Analysis is used to extract such emotions in correct manner so that result can be analyzed further and used it for companies beneficial.

Sentiment Analysis is study of emotions in human language. Sometimes it is needed to use the software to analyze what human wants to depicts through the writing. We human can easily tell the emotions by reading the reviews of a person but in case of industry where company wants to launch its product it becomes very important to read the mind and brain of their customers in order to give best product. Sentiment Analysis helps at this point. It can differentiate between negative and positive emotions which is very helpful in the company perspective. If company can use sentiment analysis on reviews of customers, they can identify whether customers have liked the product or not and till what extent they have liked it.

Sentiment analysis is being used any many sectors in business. They are used to identify what emotions are used in books by using it on words in any book. Which helps strongly because there is no need to read entire book to know what are the main plot is,like if it is tragedy or comedy or happy or sad. Then this study can help give feedback to book and decide its best readers list. Another application is social media study. People now a days use Twitter a lot to express their feelings toward any trending topic. These posts can be happy client or angry common man or any teenager expressing their issues. Use of sentiment study in twitter can help exploring in many new ideas. Like weather prediction, how people in particular area have reacted to particular situation. There is one more main application of sentiment analysis and that is studying reviews of any product given by valid customers.

We see now everywhere when we buy anything from store, online platform like clothes, software even games we are asked to give review. Initially it was just about ratings but now every platform has included a section called review. Review is basically a set of words with a certain limit of count of words which will tell how customer has described the his or her feelings towards the particular product. This review can be positive like saying good things about the product and how it helped the user and how he or she will recommend it to others. And sometimes it can be negative also which will say how it is not safe to use this product for others and they should avoid buying it or buying it from that particular place.

There are also rating surveys can be done like rating from 1-10 scale can be used for analyzing customers reviews which makes computer easy to analyze and give better results. As input is in numerical form machine can work easily with logic and algorithms and give output of detailed analysis. But whenever company offers their customers to write a written review, customers express their feelings through them. And these reviews later on help others to use that product. If customer has written a good review and has recommended the product then more customers will read that review and buy that product more often. Also, once they like the product they will also be encouraged to write good review and it will go on. On the other hand, if customers give bad reviews about any product and if company does not acknowledge it on time, it will help spreading bad reputation about product and company also. Competitors brands will use this opportunity to bring this company down.

That’s why it is very important to recognize the use of analyzing customers reviews. And for any product it is obvious that review will be on large scale count so it is not possible to read and analyze every review of customer about that product so company needs help of this data science concept called Sentiment Analysis. In sentiment analysis there are many algorithms, tools and packages are available to help identifying the emotions of words used in any sentence. So basically, there are words saved in these algorithms which has associated emotions attached to it. Example is I am very hurt. Here hurt word is obviously a negative emotion and also used to show negative feeling towards any situation. Similarly When we say, ‘I liked this product’ we want to show our gratitude and as we know like is positive word, sentiment algorithms will identify this as positive feedback.

Following are some algorithms which can be used to detect emotions in sentiments:

1.Recurrent Neural Network (RNN):

Recurrent Neural Network is a part of artificial neural network. It is mostly used to identify hand writing and speech recognition.

2.Syuzhet Package:

This package was developed by NLP group in Stratford. It has 4 dictionaries with words and their associated emotions which can be used to detect a emotion attached with any word in the sentence.

**2.Literature Review**

### Data Retrieval

The one dataset I have used for this project is found on Kaggle by Nicapotato. This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

1. Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
2. Age: Positive Integer variable of the reviewers age.
3. Title: String variable for the title of the review.
4. Review Text: String variable for the review body.
5. Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
6. Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
7. Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
8. Division Name: Categorical name of the product high level division.
9. Department Name: Categorical name of the product department name.
10. Class Name: Categorical name of the product class name.

### Data Analysis

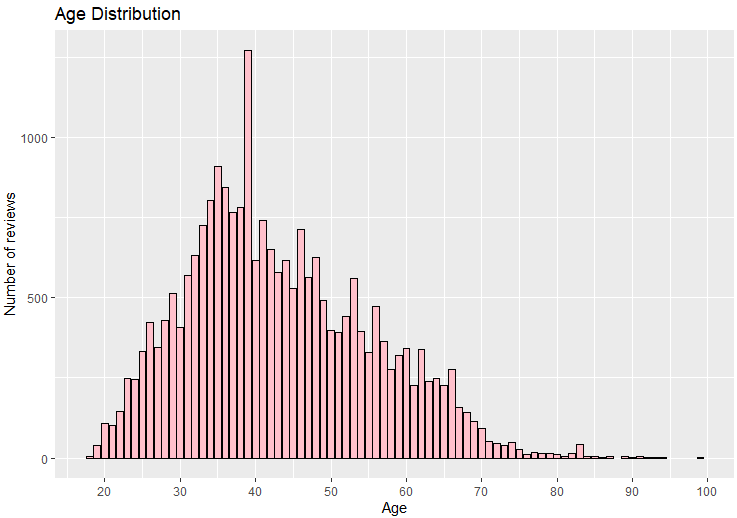
|  |  |  |  |
| --- | --- | --- | --- |
|  | Min | Max | Average |
| Age(Yrs) | 18 | 99 | 43.2 |
| Positive Feedback Count | 0 | 122 | 2.536 |
| Rate | 1 | 5 | 4.196 |

The given table is showing the Minimum, Maximum and the Average values of main 3 numerical attributes needed for reviews analysis in the dataset.

Age is varying from 18-99 years and most of the customers who submitted reviews are near age 43 years. Positive feedback count has minimum value of 0 and maximum of 122 with average of 2.536 count. Rating system has range of 0-5 with mean 4.196 rating.

|  |  |
| --- | --- |
| Department | Bottoms, Dresses, Intimate, Jackets, Tops, Trend |
| Division | General, General Petite, Initmates |
| Class | Blouses, Casual bottoms, Chemises, Dresses, Fine gauge, Intimates, Jackets, Jeans, Knits, Layering, Legwear, Lounge, Outerwear, Pants, Shorts, Skirts, Sleep, Sweaters, Swim, Trend |

Above table gives list of available values of general attributes in data set.



Graph shows Age distribution of the customers in the dataset. As seen in the graph, number of reviews are increasing with the age increment till 1000 reviews. There is sudden raise in reviews near age 40 years and then there is more decrease in reviews after age around 45 years.

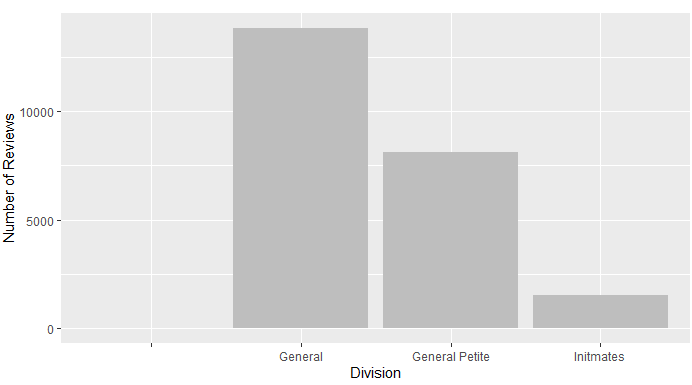


Figure shows number of reviews in different Divisions of clothes. General division has highest number of reviews and General petite has around half number of reviews compared to General division. Initmates division is having lowest number of reviews. There is blank division type because of incompleteness in dataset. This needs to be removed.

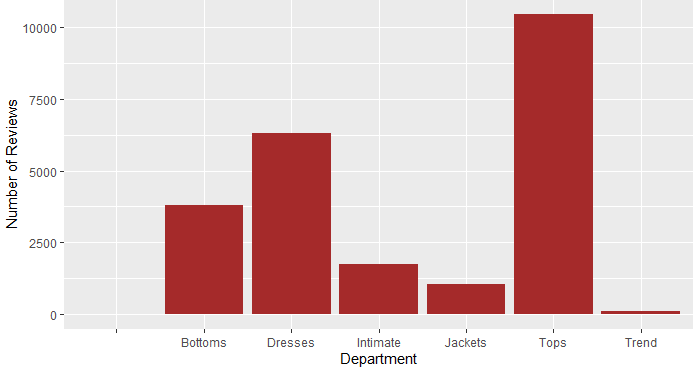


Figure shows relationship between count of reviews and Department of clothes.

There is blank department present in graph due to the incompleteness of dataset.

Trend department is having very low count for number of reviews. Jackets and Intimate department is having more number of reviews than trend department but around same.

Bottoms department is having around 3750 reviews which is quite good. After Bottoms department there is Dresses department with higher count of reviews. Tops department is having highest number of reviews i.e. more than 10000 reviews.

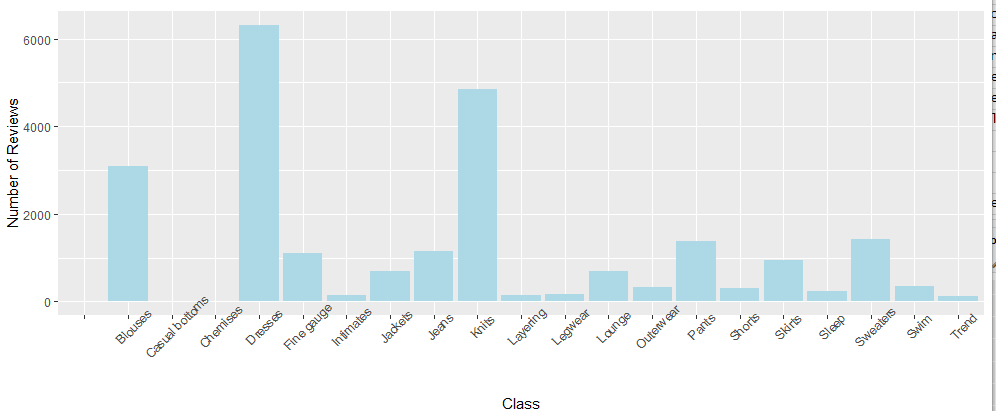
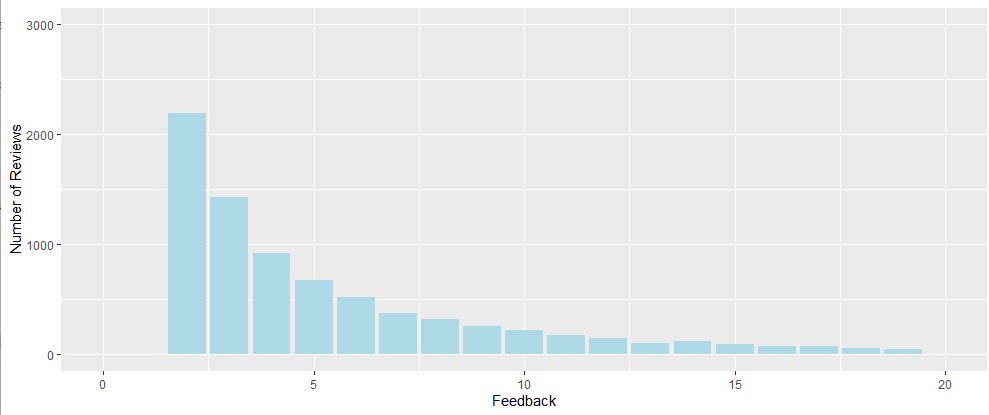


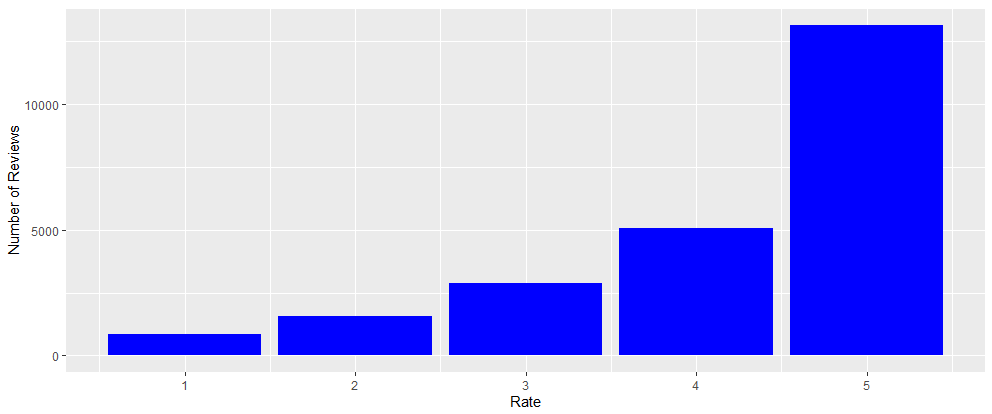
Figure shows relation between Class and Review count in dataset. Dataset has so many different types of classes. There are some classes in dataset which has very lowest number of reviews less than 1000 reviews i.e. Casual bottoms, Intimates, Layering, Legwear, Lounge, Outwear, Shorts, Jackets, Skirts, Sleep, Swim, Trend.

Fine gauge, Jeans, Parts, Sweaters have slightly higher reviews i.e. more than 1000 reviews.

Blouses and Knits classes have very good count of reviews and the Dresses class has highest number of reviews i.e. more than 6000 reviews.



Above figure shows Relation between Feedback and the number of reviews given in the dataset. For Feedback from 0-5 there are higher reviews like between 1000-3000 reviews and after feedback 5 it is getting decreased rapidly.



This figure shows rating and its count of reviews. There are 5 ratings ranging from 1 to 5. From graph it is clearly seen that rating 5 has highest reviews in the data set that is about more than 10000 reviews. And then there is large difference between rate 5 and 4 and their count of reviews. Rate 4 has almost 5000 reviews. Products with rate 3 has almost 2500 reviews and then rating with 1 and 2 are very low in count.

