1. Introduction

Understanding customer sentiments is an importance in marketing strategies today. Companies are starting to understand their customers in order to improve their product or services performance. The volume of information circulating in a typical enterprise continues to increase. Knowledge hidden in the information however, is not fully utilized, as most of the information is described in textual form (as sentences). As a part of this movement, text analysis has become an active field of research in computation of sentiments and natural language processing. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. One of the most popular problems in the text analysis field is text classification, a task which attempts to categorize documents to one or more classes. As growth of this interest sentiments of statements found in social media, reviews, and discussion groups. This task is known as sentiment analysis, a computational process that uses statistics and natural language processing techniques to identify and categorize opinions expressed in a text, particularly, to determine the polarity of attitude (positive, negative, or neutral) of the writer towards a topic or a product. This task is now widely used by companies for understanding their clients through their customer support in social media, or through customer or client reviews.

In this paper, researcher had attempt to analyse the women customer reviews on clothing by employing statistical analysis and sentiment classification. This project was concerned with using the Python programming language and Natural Language Processing technology to explore trends in the customer reviews from an women's clothing store platform, and extract actionable plans to improve company's performance. Researcher first analysed the non-text review features (e.g. age, class of dress purchased, etc.) found in the dataset, as an attempt to find any connection between them and customer recommendation on the product and reviews. Then, we implement a predictive modelling with Naïve Bays classifier and logistic regression for predicting whether a review text recommends the purchased product or not, and used sentiment analysis for classifying the user review sentiment towards the product. In a world where we generate 2.5 quintillion bytes of data every day, sentiment analysis has become a key tool for making sense of that data. Researcher has used sentiment analysis in this paper to know most dominating sentiment.

2.Theoretical review of the Concepts used for study:

2.1 Text Mining Process

Text Mining is also known as Text Data Mining. The purpose is too unstructured information, extract meaningful numeric indices from the text. Information can extract to derive summaries contained in the documents. These activities are used in this study:

a. Text Pre-processing

It involves a series of steps as shown in below:

- **Text Clean-up**

Text Clean-up means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.

- **Tokenization**

Tokenizing is simply achieved by splitting the text into white spaces.

- **Part of Speech Tagging**

Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text.

- **Stemming and Lemmatization**

Also Called Word Normalization to convert the word into its root form.

b. Text Transformation (Attribute Generation)

A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:

i. Bag of words

ii. Vector Space

c. Feature Selection (Attribute Selection)

Feature selection also is known as variable selection. It is the process of selecting a subset of important features for use in model creation.

d. Data Mining

At this point, the Text mining process merges with the traditional process. Classic Data Mining techniques are used in the structured database. Also, it resulted from the previous stages.

e. Evaluate

Evaluate the result, after evaluation, the result discard.

2.2 Natural Language Processing:

NLP is one of the oldest and most challenging problems. It is the study of human language. So those computers can understand natural languages as humans do. NLP research pursues the vague question of how we understand the meaning of a sentence or a document. What are the indications we use to understand who did what to whom? The role of NLP in text mining is to deliver the system in the information extraction phase as an input.

3. Methodology:

3.1 Machine Intelligence Library:

As for the data pre-processing and handling, the numpy and pandas Python libraries were used. For text analysis Natural Language Processing (NLTK) Package used. The Modelling section includes nltk's sentiment analysis module, which can determine the mood of text, NLTK's N-grams, and gensim.models's word2vec. Lastly, for the data visualization, the matplotlib and seaborn Python libraries were used.

3.2 The Dataset:

The Women's Clothing Reviews was used as the dataset for this study. This dataset consists of reviews written by real customers; hence it has been anonymized, i.e. customer names were not included, and references to the company were replaced with "retailer"

The Spice Clothing dataset includes 22641 rows and 10 feature variables. Each row contains customer review, and includes the variables as:

Attributes	Description
Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed.
Age	Positive Integer variable of the reviewers age.
Review Text	String variable for the title of the review
Title of 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.

Recommendation 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 this review positive.
Division Name	Categorical name of the product high level division.
Department Name	Categorical name of the product department name.
Class Name	Categorical name of the product class name.

Table.1 Details About Variables

Researchers goal was to understand what it is the customers appreciate and dislike about their purchases. The Raw Dataset has been collected from Clothing Store which was present in MS Excel file.

Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
767	33		Absolutely	4	1	0	Initmates	Intimate	Intimates
1080	34		Love this c	5	1	4	General	Dresses	Dresses
1077	60	Some maj	I had such	3	0	0	General	Dresses	Dresses
1049	50	My favorit	I love, love	5	1	0	General Pe	Bottoms	Pants
847	47	Flattering	This shirt i	5	1	6	General	Tops	Blouses
1080	49	Not for th	I love trac	2	0	4	General	Dresses	Dresses
858	39	Cagrcoal s	I aded this	5	1	1	General Pe	Tops	Knits
858	39	Shimmer,	I ordered	4	1	4	General Pe	Tops	Knits
1077	24	Flattering	I love this	5	1	0	General	Dresses	Dresses
1077	34	Such a fun	I'm 5"5' ar	5	1	0	General	Dresses	Dresses
1077	53	Dress look	Dress runs	3	0	14	General	Dresses	Dresses
1095	39		This dress	5	1	2	General Pe	Dresses	Dresses
1095	53	Perfect!!!	More and	5	1	2	General Pe	Dresses	Dresses
767	44	Runs big	Bought	5	1	0	Initmates	Intimate	Intimates
1077	50	Pretty par	This is a ni	3	1	1	General	Dresses	Dresses
1065	47	Nice, but r	I took thes	4	1	3	General	Bottoms	Pants
1065	34	You need	Material a	3	1	2	General	Bottoms	Pants
853	41	Looks grea	Took a cha	5	1	0	General	Tops	Blouses
1120	32	Super cute	A flatterin	5	1	0	General	Jackets	Outerwear

Figure. 1 Snapshot of Dataset

Column	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
Unique	1172	77	13984	22621	5	2	82	3	6	20
Missing	0	0	2966	0	0	0	0	0	0	0

Figure. 2 Frequency count

Table 2 shows the frequency distribution of dataset features and label (Recommended IND) and insights about missing values. There are approximately 3000 missing values, which

represents 1% of the dataset, but the dataset will not get trimmed further since the review text body is the only variable that must be complete. Amongst the categorical variables, the high unique count of Clothing ID and Class Names will require non-visual exploratory methods.

	mean	std	min	25%	50%	75%	max
Clothing ID	919.695908	201.683804	1.0	861.0	936.0	1078.0	1205.0
Age	43.282880	12.328176	18.0	34.0	41.0	52.0	99.0
Rating	4.183092	1.115911	1.0	4.0	5.0	5.0	5.0
Recommended IND	0.818764	0.385222	0.0	1.0	1.0	1.0	1.0
Positive Feedback Count	2.631784	5.787520	0.0	0.0	1.0	3.0	122.0
Word Count	60.211950	28.533053	2.0	36.0	59.0	88.0	115.0
Character Count	308.761534	143.934126	9.0	186.0	302.0	459.0	508.0
Label	0.895263	0.306222	0.0	1.0	1.0	1.0	1.0

Figure. 3 Statistical Summary of Attributes

The table describe statistical summary of each variable as mean, standard deviation, minimum, maximum etc. This distribution shows that the data is normally distributed so we can apply various techniques to get the results.

3.3 Data Analysis:

3.3.1 Analysis of Univariate Distributions:

(1) Frequency Distribution of Age and Positive Feedback variable:

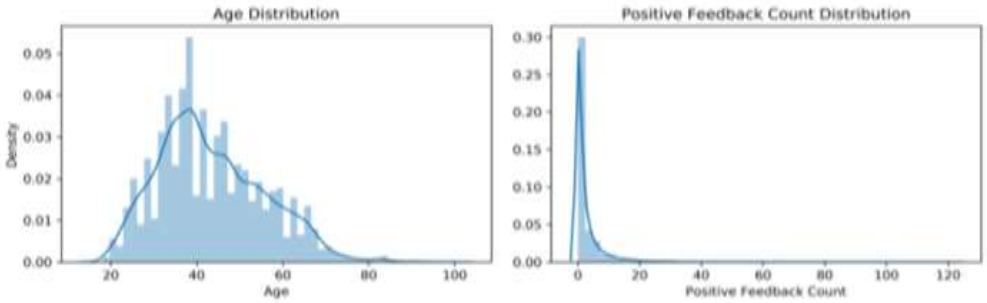


Figure.4 Frequency Count for Age and Positive feedback count

Age and Positive Feedback Count: Figure 4 reveals that the most engaged customers in reviewing purchased products were in the age range of 35 to 44. In addition, the figure suggests that they have the most positive reviews on their purchased products. From this, we have two points to consider: (1) the said age group is the most satisfied group in the range of customers,

thus, the e-commerce at review must focus on maintaining this segment, and (2) the e-commerce entity can explore why other age groups are comparatively less satisfied than the age group 35 to 44.

(2) Frequency Count of categories in Department Name:

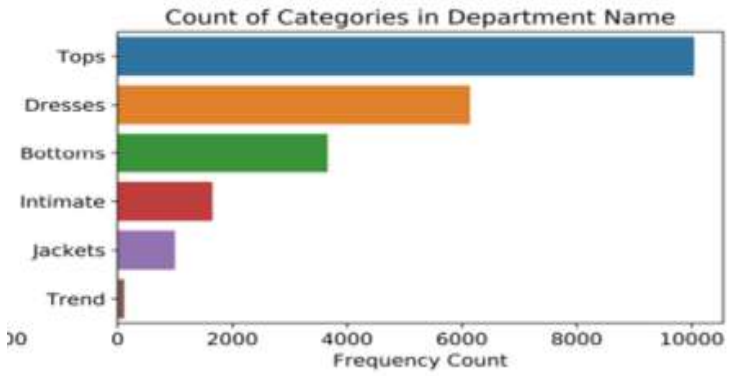


Figure.5 Frequency Count for Department name

This high-level feature describes had three categories: General, Petite, and Intimates. This offers some insight into the clothing sizes of the customers leaving reviews. It is notable to observe that *Tops and Dresses* are the most commonly reviewed products. It would be interesting to investigate the motivation of leaving a review in the first place.

(3) Frequency Count of Clothing ID with Top 30 to 60 ID's and Class Name:

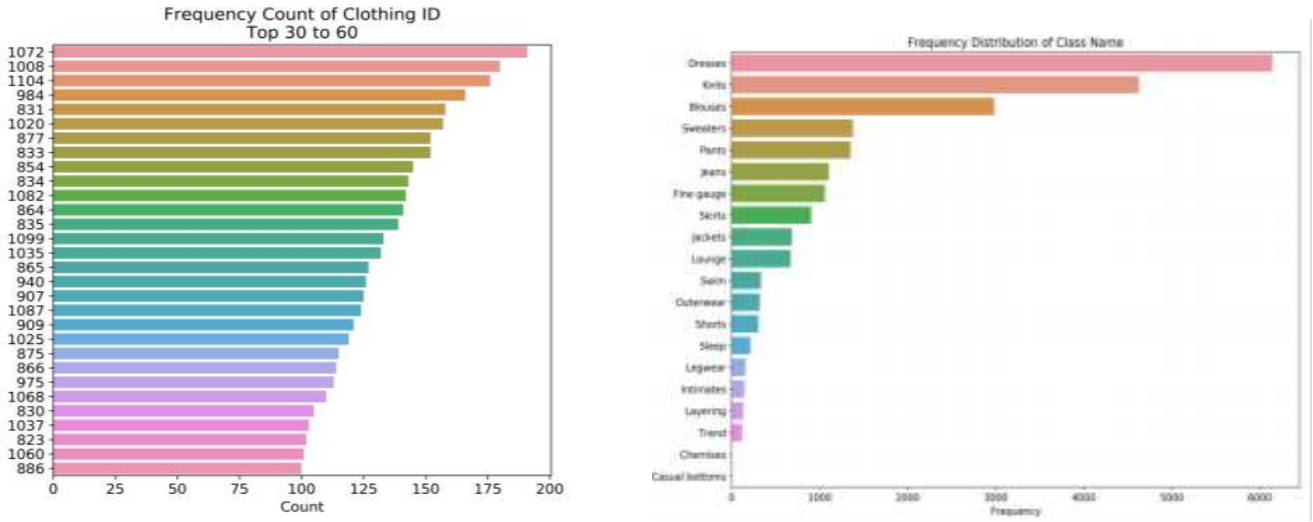


Figure. 6 Frequency Distribution for Clothing Id and Class

From figure we clearly see that Clothing id 's 1072, 1008, 1104 general division of cloth is most reviewed. Exploring the class variable suggests that the most popular clothing types are:

Petite and Anthro, Dresses, Blouses, and Cut and Sew Knits. The distribution of reviews is fairly constant, suggesting that there are not negative nor positive outliers.

(4) Frequency Distribution of Rating and Recommendation variable:

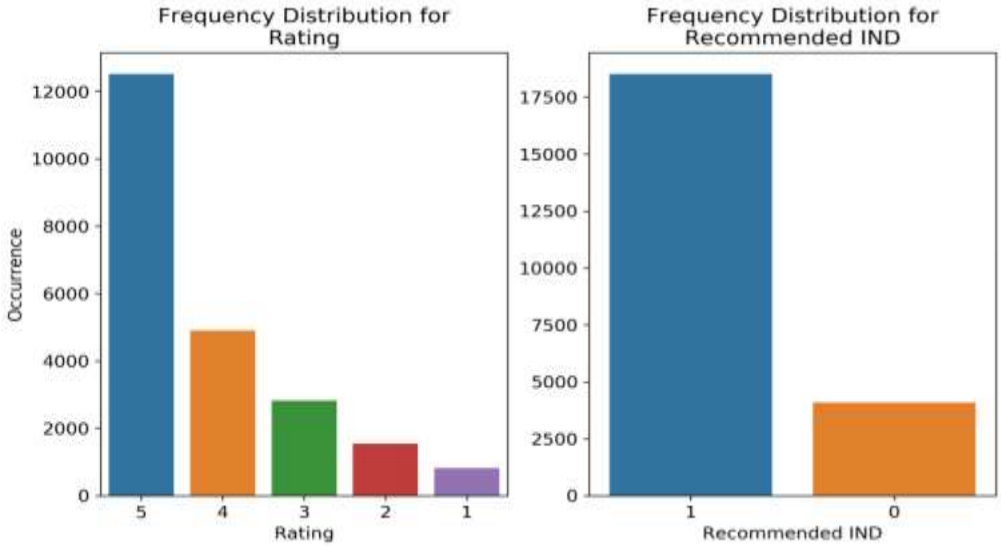


Figure. 7 Rating and Recommendation Frequency

The vast majority of reviews were highly positive, with a score of five out of five. This suggests that this retail store is performing fairly well. Competitor reviews may be scraped and analysed. It is important to note that these reviews are subjective, and some negative reviews may a outcome of a bad day, instead of constructive feedback.

3.3.2 Multivariate Distribution:

(1) Word Count Distribution of Rating, Department and Recommendation variable:

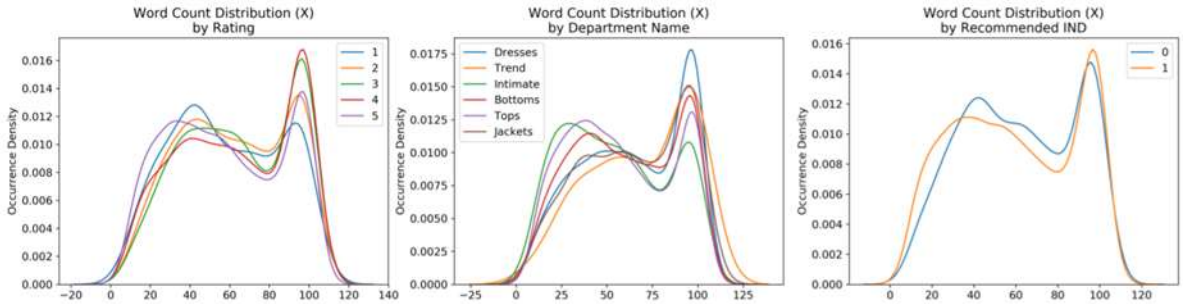


Figure.8 Multivariate Distribution

Unlike Positive Feedback Count, Age has not been transformed into a logarithm. For these reasons, slight noise between the age distribution by these features are nothing to worry about. Age doesn't seem to receive influence on these dimensions.

(2) Occurrence of Rating by Recommendations:

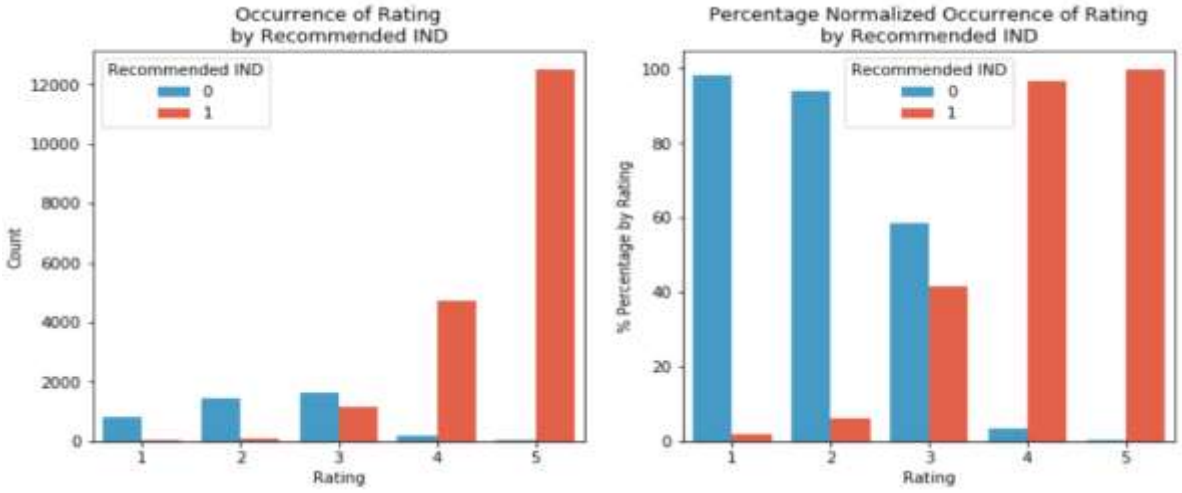


Figure. 9 Rating Occurrence with Recommendations

There is a conflicting interest between the customers personal interaction with the product, such as the personal size fit, experience, and other personal synergies, and what the customer would invasion for other customers. Looking at the data, it appears like five-star ratings are void of non-recommendations, but low rated products are recommended a small amount of the time.

(3) Frequency of Recommendations by Department Name and by Division Name:

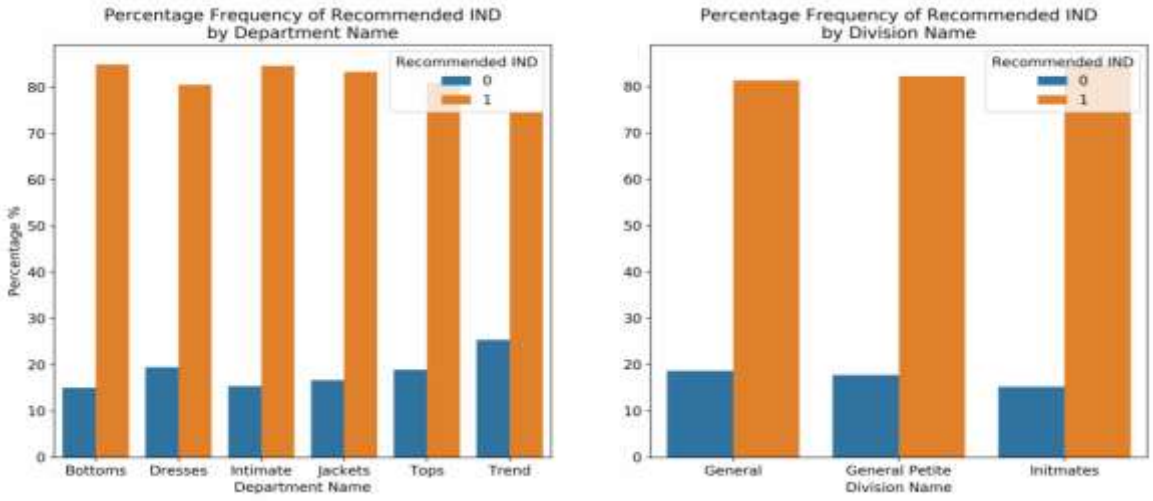


Figure. 10 Recommendations by Department

Here we get a closer glimpse at the breakdown of specific clothing types. This plot wraps up the interplay between Blouses, Dresses, and Knits by showing that most reviews revolve around the normal sized version of the products. It is interesting to note that Dresses attract higher proportion of "Petite" sized customers.

3.3.3Multivariate Analysis of Division categories and Department Categories using Heatmap:

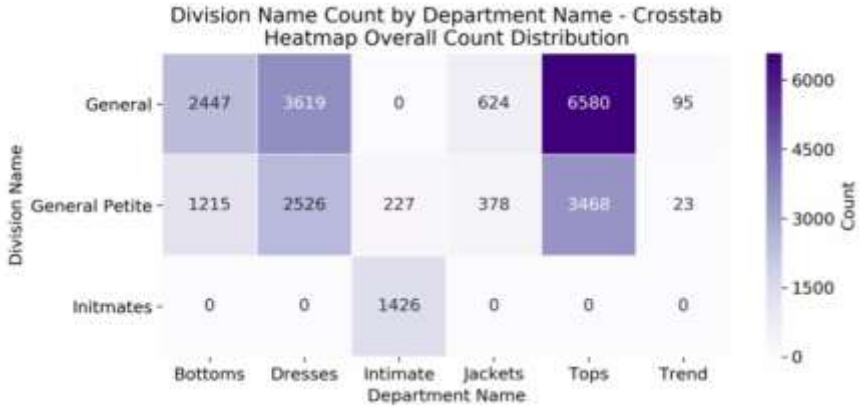


Figure. 11 Heatmap for overall Distribution

The dominance of the General size is consistent across the various categories within Department Name. There a notable overall between General Petite and Department Name. The above Heat map shows Dominance of **General Tops** and **General Dresses** are also getting more reviews. The heatmap represent all department name on x axis and all Division name on y axis as this we can use heatmap for analysing the multiple variables

3.3.4 Descriptive Statistic with correlation matrix of all variables in the dataset:

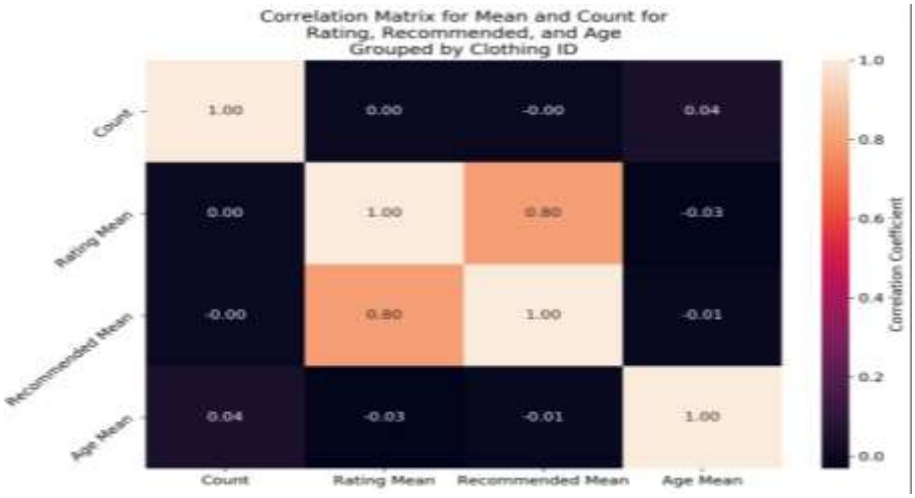


Figure. 12 Correlation Matrix

The more stress on the *Grouped by Clothing ID* aspect of this analysis. This is a different lens of analysis than merely running a correlation on *all customers reviews*. This correlation heatmap suggest that there is in fact no correlation between count and average value, which means that the popularity of the item does not lead to differential treatment when it comes to average scoring. The age variable behaves in this same as well. However, there is a strong positive correlation of .80 between rating and recommended IND mean.

4.Text Analysis:

4.1 Text Pre-Processing in python:

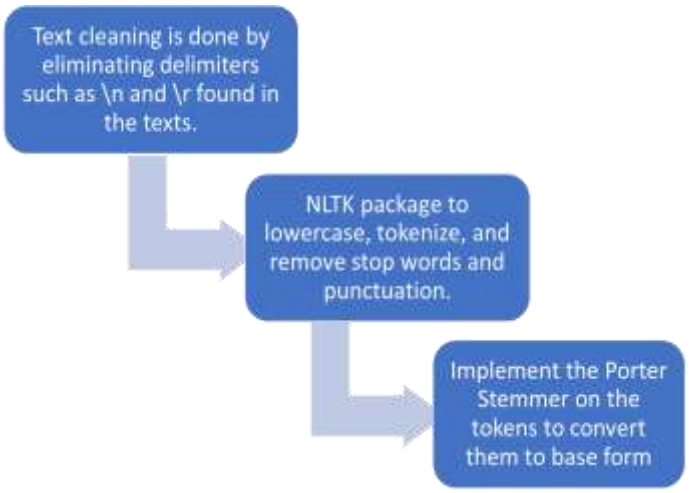


Figure. 13 Text Pre-processing with Python

In order to process the data set's centrepiece, the review body, utilized the NLTK package to lowercase, tokenize, and remove stop words and punctuation.

4.2 Sentiment Analysis Process and steps to generate Sentiments:

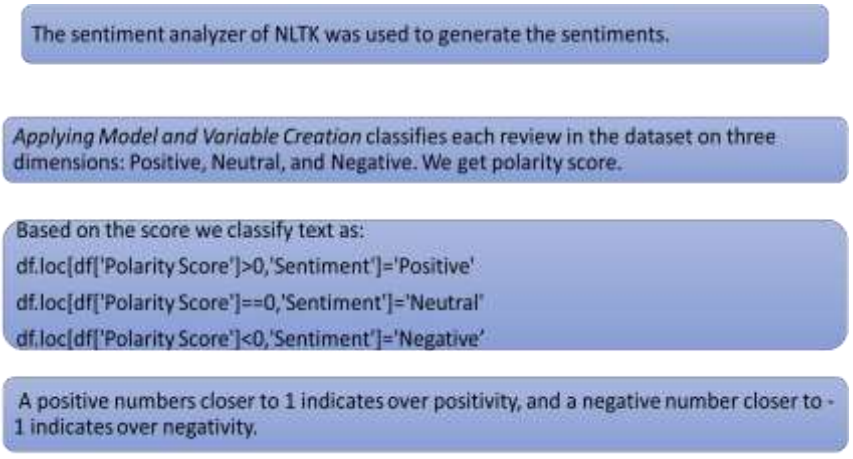


Figure. 14 Sentiment Analysis

For understanding the customer reviews is to see how the textual sentiment relates to the rating scores. Paper will explore the interaction between sentiment score and following:

- Rating
- Recommended
- Positive Feedback Count

Neutral/Negative/Positive Score: Indicates the potency of these classes between 0 and 1.

Polarity Score: Measures the difference between the Positive/Neutral/Negative values, where a positive numbers closer to 1 indicates overwhelming positivity, and a negative number closer to -1 indicates overwhelming negativity.

Generated Sentiments in the Dataset:

othing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	Word Count	Character Count	Label	Polarity Score	Neutral Score	Negative Score	Positive Score	Sentiment	tokenized
767	33	NaN	Absolutely wonderful - silky and sexy and comfortable	4	1	0	Initmates	Intimate	Intimates	8	53	1	0.8932	0.272	0.000	0.728	Positive	[absolut, wonder, silki, sexi, comfort]
1080	34	NaN	Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's petite. i bought a petite and am 5'8". i love the length on me- hits just a little below the knee. would definitely be a true midi on someone who is truly petite.	5	1	4	General	Dresses	Dresses	62	303	1	0.9729	0.664	0.000	0.336	Positive	[love, dress, sooo, pretti, happen, find, store, glad, bc, never, would, order, onlin, bc, petit, bought, petit, 5, 8, love, length, hit, littl, knee, would, definit, true, midi, someon, truli, petit]
1077	60	Some major design flaws	I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual size) but i found this to be outrageously small. so small in fact that i could not zip it up! i reordered it in petite medium, which was just ok. overall, the top half was comfortable and fit nicely, but the bottom half had a very tight under layer and several somewhat cheap (net) over layers. imo, a major design flaw was the net over layer sewn directly into the zipper - ...	3	0	0	General	Dresses	Dresses	98	500	1	0.9427	0.792	0.027	0.181	Positive	[high, hope, dress, realli, want, work, initi, order, petit, small, usual, size, found, outrag, small, small, fact, could, zip, reorder, petit, medium, ok, overal, top, half, comfort, fit, nice, bottom, half, tight, layer, sever, somewhat, cheap, net, layer, imo, major, design, flaw, net, layer, sewn, directli, zipper, c]
1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!	5	1	0	General Petite	Bottoms	Pants	22	124	1	0.5727	0.340	0.226	0.434	Positive	[love, love, love, jumpsuit, fun, flirti, fabul, everi, time, wear, get, noth, great, compliment]
847	47	Flattering shirt	This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to wear with leggings and it is sleeveless so it pairs well with any cardigan. love this shirt!!!	5	1	6	General	Tops	Blouses	36	192	1	0.9291	0.700	0.000	0.300	Positive	[shirt, flatter, due, adjust, front, tie, perfect, length, wear, leg, sleeveless, pair, well, cardigan, love, shirt]

Figure. 15 Snapshot of Sentiment Score

4.2.1 Exploring the sentiments Generated by Sentiment analyser:

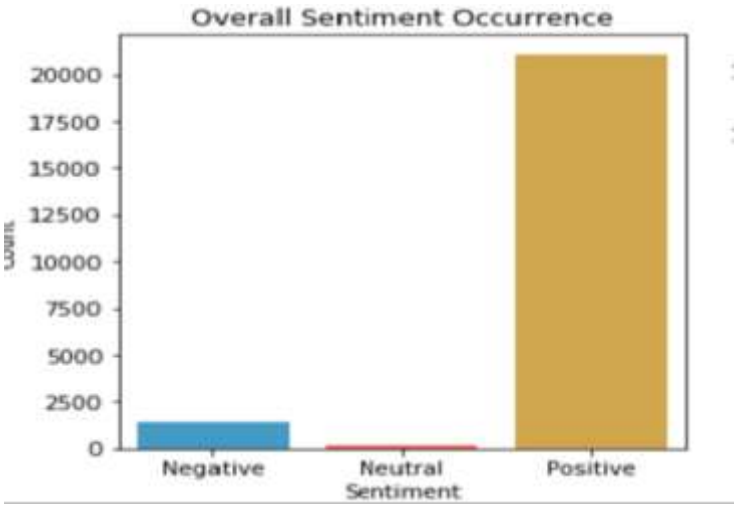


Figure. 16 Sentiments Exploration

In the above figure the 3 types of sentiments are shown with the help of graph to know the frequency of each sentiment. From total 22641 records 20000 are positive reviews, 1000 are neutral reviews and remain 1641 are Negative. Most reviews have a positive sentiment. Unlike the distribution of rating, there is a lower occurrence of neutral rating is lower in proportion to the occurrence of medium ranged ratings.

4.2.2 Occurrence of sentiments with respect to Recommendations and Rating:

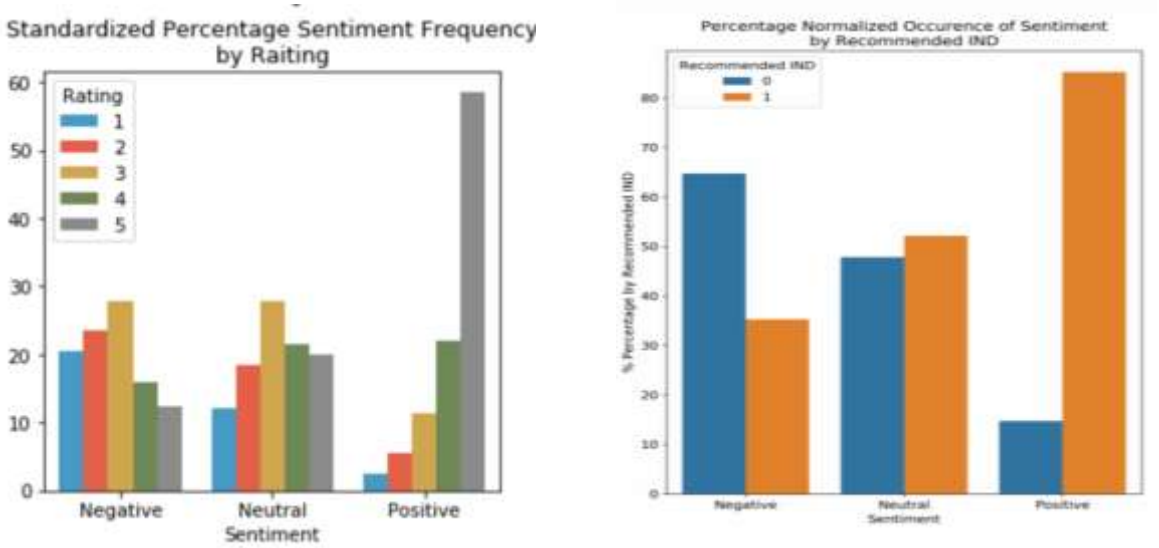


Figure. 17 Sentiments with Rating and Recommendations

Like the distribution of rating, most reviews have a positive sentiment. Unlike the distribution of rating, there is a lower occurrence of neutral rating is lower in proportion to the occurrence of medium ranged ratings.

The plot on the bottom right tells and interesting story. The rating of positive sentiment reviews have an increasing occurrence as the rating gets higher.

4.3 Word Distribution and Word Cloud

Word clouds are generated by using word cloud package which gives frequency of occurrence of words and shows most dominant word. The bigger size of word and darker colour shows the most frequent occurred word. The paper contains the word clouds for:

- Title
- Highly Rated comments

Word Count Table:

Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	Word Count	Character Count	Label
The quality and fabric are fabulous but it's j...	2	0	0	General	Dresses	Dresses	24	137	0
I'm not normally one to spend $89.99 on a dres...	5	1	1	General	Dresses	Dresses	94	502	1
This will be a great summer staple. good fit ...	5	1	0	General	Tops	Knits	19	102	1

Figure. 18 Snapshot of Word Count

Wordcloud for Titles:

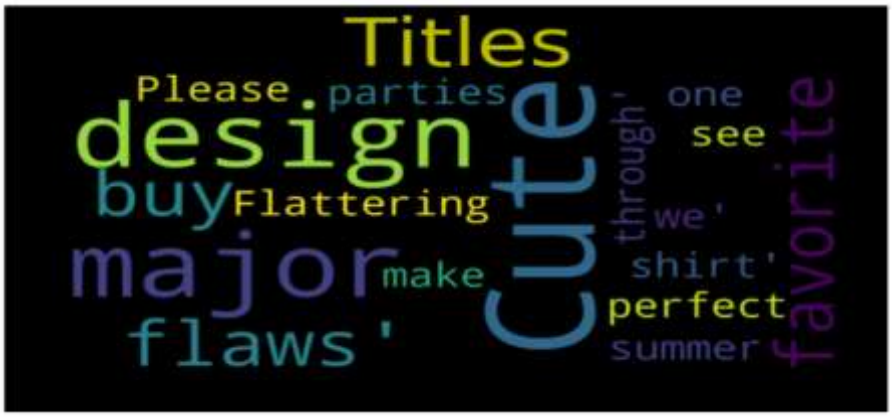


Figure. 19 Wordcount for Title

Wordcloud for titles in above figure shows that design is the word most repeated in the title of the review which indicates that design is very much matter in the clothing product according to customers. Similarly, cute , perfect, major these are the words mostly used in the title of the reviews.

Wordcloud for Highly Rated Comments:

Figure. 20 Wordcount for Highly Rated Comments

The above figure shows the wordcloud for high rated comments in that most highlighted words are like fit, small from that inference can be drawn, that small fitted dress are highly recommended by the women's. So company can focus on the fitting for clothing products. Similarly, in this wordcloud we see the words like store, medium and also top, where tops are more trending in women's reviews. Here figure also shows that never word is becoming highlighted and also wanted is the word highlighted, from such words sometimes the actual meaning is not getting so that instead of single words two words or three words are necessary to get some meaning, for that here researcher has used N-gram technique to generate required M-grams.

4.4 N Grams by Recommended and Non-Recommended Items:

Sometime single words in wordcloud can't give exact meaning for that we require N-grams. In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech.

The central themes in the product reviews brought to light by the n-grams are:

N- Grams for Recommended items:

	1-Gram	Occurrence	2-Gram	Occurrence	3-Gram	Occurrence	4-Gram	Occurrence
0	dress	8591	true size	1243	fits true size	264	compliments every time wear	46
1	love	8017	love dress	657	fit true size	192	26 waist 36 hips	32
2	size	7561	5 4	622	received many compliments	163	34b 26 waist 36	28
3	fit	5995	usually wear	588	runs true size	143	looks great skinny jeans	25
4	top	5846	looks great	574	love love love	138	get compliments every time	23
5	wear	5678	fit perfectly	553	usually wear size	107	love love love dress	22
6	great	5584	well made	531	every time wear	81	115 lbs 30 dd	22
7	like	5368	love top	524	ordered usual size	79	usually wear size 4	21

Figure. 21 N-Grams for Recommended Items

Here, positive reviews are void of criticism, and are preoccupied with confirming fit and sharing social experience with the clothing. “True Size”, “Fit Perfectly”, “Fit like a glove”, on top of the multiple 2-grams with customer’s height suggest that a large part of positive reviews is employed to confirm product fit according to certain size. The high occurrence of this review suggest that height and size is usually a big issue, which this retail managed to consistently satisfy.

N- Grams for Non-Recommended items:

In the negative reviews, customers express their disappointment in the product, stating that they “really wanted to love” the item. This signifies that the product did not live up to the customers’ expectations. This occurred for multiple reasons. “Order wear size” and “Usual wear size” suggest that the fit did not suit their typical universal body size. Perhaps if better product dimension information could be provided, then the likelihood of this negative response could decrease.

	1-Gram	Occurrence	2-Gram	Occurrence	3-Gram	Occurrence	4-Gram	Occurrence
0	dress	1976	wanted love	243	really wanted love	70	really wanted love dress	15
1	like	1780	going back	215	wanted love dress	65	looked like maternity top	10
2	top	1572	looked like	187	really wanted like	40	really wanted like dress	9
3	would	1348	looks like	153	made look like	29	really wanted like top	9
4	fit	1327	really wanted	151	wanted love top	28	5 4 120 lbs	8

Figure. 22 N-Grams for Non-Recommended Items

“Too much fabric” and “Looks nothing like” suggest inconsistency with retailer presentation and actual product. These reviews are especially destructive, since they damage the reputation of the store product quality, which is a biggest asset.

4.5 Predictive Modelling Supervised Learning:

Supervised learning requires features (independent variable) and a label (dependent variable). Currently the independent variable is the entire comment. However, in order to the Naïve Bayes Algorithm to work, each word must be treated as a variable. Here to conduct predictive modelling first researcher has used Naïve Bays modelling technique in which Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. P(c|x) is the posterior probability of class (target) given predictor (attribute).

4.5.1 Naive Bayes

```
Classifier accuracy percent: 82.48557944415312
Most Informative Features
                  cheap = True              0 : 1      =     12.3 : 1.0
                   glad = True              1 : 0      =      5.4 : 1.0
                 bummer = True              0 : 1      =      5.0 : 1.0
                    net = True              0 : 1      =      4.6 : 1.0
                   idea = True              0 : 1      =      4.4 : 1.0
                 pencil = True              1 : 0      =      4.3 : 1.0
                perfect = True              1 : 0      =      3.8 : 1.0
               charcoal = True              1 : 0      =      3.7 : 1.0
               shimmer = True               1 : 0      =      3.7 : 1.0
                    fun = True              1 : 0      =      3.4 : 1.0
                  later = True              1 : 0      =      3.0 : 1.0
                   sooo = True              0 : 1      =      2.6 : 1.0
                    ton = True              1 : 0      =      2.5 : 1.0
```

Figure. 23 Naïve Bays Performance

For predicting recommendations based on review.

Independent Variable: Word choices in Reviews

Dependent Variable: Whether or not review was Recommended

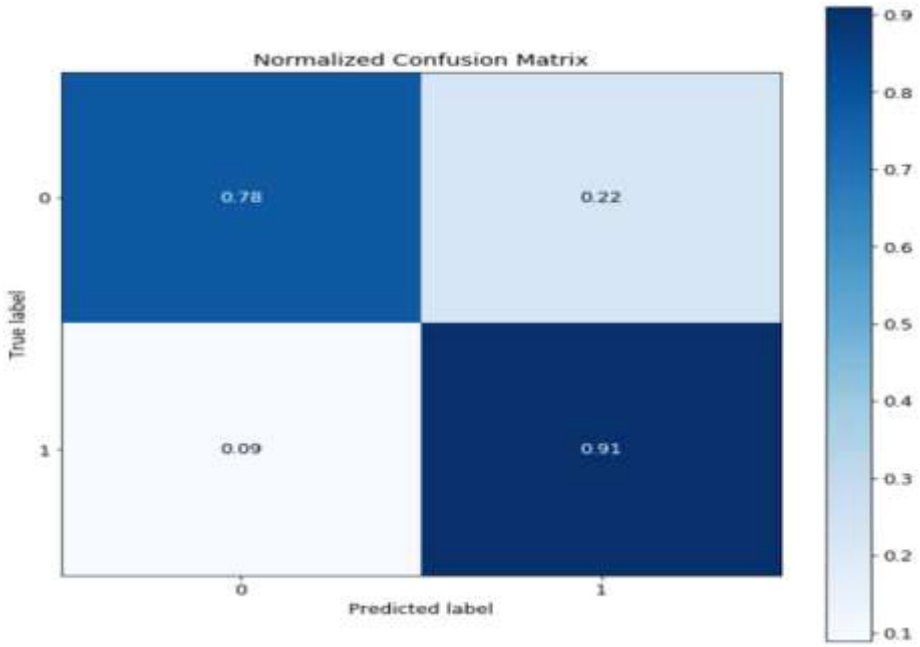
- Naïve Bays Classifier is used to train the model.
- But the accuracy score for this is 82% which is low
- 1st Column -word, 2nd displays – Recommendation (1:0) Not Recommended (0:1)
- Word Cheap – who's presence is 12.3 times more likely to be negative than positive

As the accuracy score got by the Naïve Bays model is low so, we can go for another modelling technique that is Logistic Regression.

4.5.2 Logistic Regression Modelling:

Logistic Regression model is used as Naïve Bays gives less accuracy. In Logistic regression from multiple independent variables it will predict the dependent variable.

Ax1+bx2+dx3+c=Y




```
Train Set Accuracy: 0.9129377969285162
Train Set ROC: 0.8889105042166693
```

Figure. 24 Confusion Matrix for Logistic Regression

The above figure is Confusion Matrix for Logistic Regression and its accuracy score.

- Logistic Regression Gives Score of 91% which is good fit model for prediction.
- The Confusion matrix shows that for non-recommendations **0.78 predictions are true and 0.22 are wrong**
- For recommendations **0.91 predictions are correct and 0.09 are false.**

So, the logistic Regression model is best fit on this data for prediction of recommendations based on customer reviews.

5. Conclusion: