National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name:	Tej Rup Sai Munagal		
Student ID:	19196628		
Programme:	Master of Science in Data Analytics	**Year:**	2020
Module:	Domain Application of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	23/08/2020		
Project Title:	Sentiment Analysis and Predicting Recommendations Based On Customer Reviews		
Word Count:	2706		

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Tej Rup Sai Munagala
Date:	23/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1.Please attach a completed copy of this sheet to each project (including multiple copies).
2.**You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

SENTIMENT ANALYSIS AND PREDICTING RECOMMENDATIONS BASED ON CUSTOMER REVIEWS

Tej Rup Sai Munagala
Master of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x19196628@student.ncirl.ie

***Abstract*— In this paper, we see the impact of predictive algorithms in the application of predictive analytics. For this purpose, a dataset "Women's E-commerce clothing reviews" is chosen, this data set contains reviews of customers and these reviews are used to prescribe a product that is relevant to customers. The proposed model helps in segregating the reviews and also recommends the customers with appropriate products. Techniques of data analysis like resampling the data, removing unnecessary data and transforming the dataset into a utilizable form are used along with some machine learning applications like recommendation the system through naive Bayes algorithm and the analysis of the review data is done by sentiment analysis. The qualitative analysis is validated by checking at the performance of the model with the confusion matrix.**

Keywords— Predictive algorithms, Analytics, E-commerce, classification, recommendation, Preprocess techniques, EDA, sentiment analysis, naive Bayes, confusion matrix

I. INTRODUCTION

Innovative technology has transformed the world's traditional businesses into virtual businesses. This transformation gave birth to new form of technology 'E-Commerce', this new technology providing platform to numerous businesses are being developed by cutting edge technology [1]. The word ‘E-commerce’ stands for electronic commerce and E-commerce is the activity of electronically buying or selling of products on online services or over the internet, this is made by trading the data between the users who purchase the products and the organization which sells the products [2].

The relation between a business organization and its customer has been affected by E-Commerce, numerous benefits are being served to customers by using E-Commerce [3]. Some market investigators anticipate 4 trillion dollars yearly online use in 2020 [4]. Online sales in Europe increased from 15% in 2014 to 19% in 2017 by using AI in digital marketing [5].

From 2019 AI is observed to have the most influence on the E-Commerce generating revenue by understanding customer reviews and number of products sales. Retailer’s stock size is being affected by AI. Retalon stated that AI can lower the cost up to 25-40 per cent on Inventory Management and increases the sales turnover by 350 per cent [6].

The numbers in online sales of a product are dependent on the customer reviews and recommendations to the customer, customers get to know and decide to buy a product based on product recommendation and product review [7]. Firstly, the right recommendations may lead the customer to explore more products. Secondly, good reviews of a product increase sales [8].

Algorithms for sentiment classification based on logistic regression, Neural Network, Support Vector Machine and the recommendation algorithm like naive Bayesian (NB), Pearson r correlation are used for text analysis and recommendations

A data related to women’s E-commerce clothing reviews was acquired from Kaggle repository is used for this research this data set consists of 11 columns and 23468 attributes.

II. ETHICAL CONCERNS

The dataset comes with the names of the customers. In valuing the customer’s information, random digits are used instead of the customers’ names. Along with the names of customers, some other details of the company are replaced by a type of department name and with a division of which product it comes under.

As we are working on Kaggle repository which is an open-source of data there won’t be any legal obligations faced on this research

Ethical concern should never be neglected especially when a dataset that has customer details is shared in the supply chain market. As the information flows through different stages of supply chains, data must be restricted when passed between stages to maintain ethical decency.

With proper data that is gathered from the site, the behaviour of customers can be predicted using machine learning algorithms. These are useful in recommending a specific type of product to customers. However, the purchase history of an individual can be misused [9].

Customer behaviour can be analyzed and categorized based on predictive analytics and sentiment analysis. Sometimes categorizing an individual based on their behaviour could be inappropriate, for example, some users find it offensive if they see alcohol products in their recommendations, this recommendation could be based on their reviews or search history but still, it is not appreciated by the user [9].

Dataset carries a lot of sensitive information related to customers like card details, personal details. In a supply chain, the data has to pass through several stages so only verified members should have access to sensitive data [9].

The organization that designs predictive algorithms in E-Commerce has to be aware of ethical risks to its customers. A right amount of caution is needed when designing the algorithms so that efficiency and ethical concern are delivered properly [10].

III. BUSINESS VALUE

To generate sustainable profit from the challenging environment, e-commerce, online e-business needs to give more importance to customers (B2C) in many areas of industry. There are three major factors like Human-computer interaction (HCI), behavioural, and consumerist orientations.

Human-computer interaction (HCI): E-business will concentrate primarily on designing a user interface that is user-friendly, simple to use, and productive. The information content like website visual appeal, web design, ease of access are the most significant factors for a shopping experience that can satisfy.

Behavioural approach: The trust elements of online shopping must be the priority of e-business. There should also be a good sense of trust between the online retailer and the customer since nearly all transactions occur. So, it is necessary to have sufficient confidence in sellers.

Consumer characteristics: The conduct of online shopping also depends on user characteristics such as attitude, demographics, and profiles. Website friendliness and the level of consumer satisfaction also have an influential impact on the market.

These are the three most important factors influencing the growth of an online business, which can bring a radical change in the e-commerce environment by spending enough time on them.

IV. OBJECTIVE

This project aims to analyze customers' behaviour. To analyze that we have to look at the following objectives:

- Analyze and get sentiment scores from many customers of various products on the website
- Recommend products to customers based on the sentiment analysis of reviews

V. PLANNED STRATEGY

A. *Dataset*

The dataset selected to conduct this study is the "Women's E-commerce clothing review[1] ". It has a total of nine features with text data and has multiple dimensions like clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, Class Name.

Most of the features are divided in the following ways:

- Clothing ID and Title has anonymized user information.
- Division Name, Department Name and Class names are about the details of the products.
- Review Text, Rating and Positive Feedback Count has customers' viewpoints

B. *Exploratory Data Analysis (EDA)*

To start with exploratory data analysis (EDA), we looked at individual variable distribution by using univariate and multivariate distribution and then models like sentiment analysis and navies bayes algorithms are applied then they are checked for quantitative and qualitative analysis.

All the individual variables are analysed by using both univariate and multivariate analysis to filter out the variables.

B.1 Univariate analysis:

The Fig. 1 States the age group 10-20 gave lower ratings. Teenagers usually do not care about online shopping and comments. However, a high number of the 5-star rating was granted for the age group of 30-40 compared with all other age groups. It is just the category that gives more ratings and feedback as opposed to other age groups. At least, when we look at the age group of 70 and beyond, online shopping did not matter.

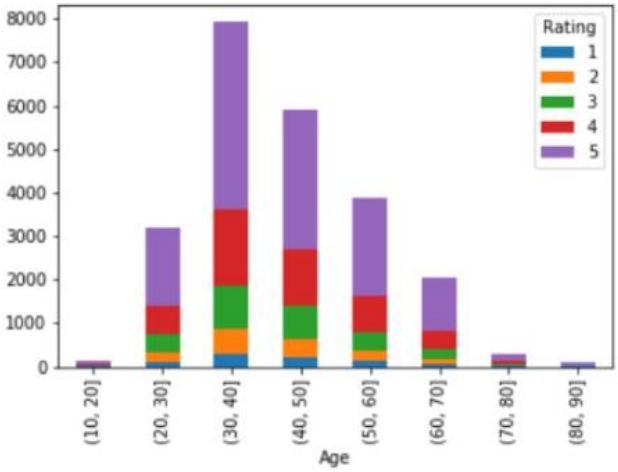

Fig 1-Plot between age group and rating

Pareto distribution states that 20% population holds 80% wealth, this rule of thumb is applied to the review dataset. Fig 2 tells that 47% of positive feedback belongs to top 20% which means that Pareto distribution is effective on our dataset and most of the reviews are following this rule.

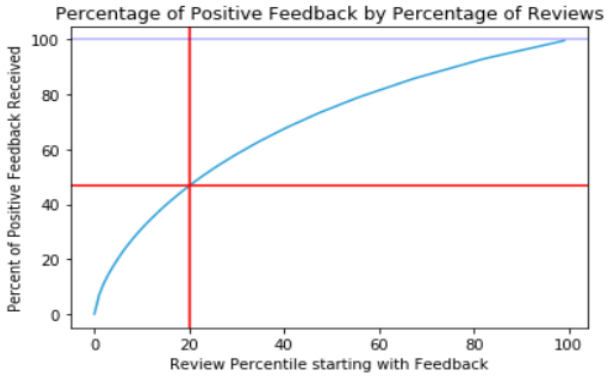

Fig 2- Percentage of Positive Feedback by Percentage of Reviews

B.2 Multivariate analysis:

Here we are analyzing multiple variables on bases of four categories, able to understand the correlation and influence of individual variable on other variables.

B.2.1 Categorical variable by categorical variable:

[1] Women's E-commerce clothing review-https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Here we analyzed two continuous different categorical variables, The Fig 3. Is a 3-Dimensional graph that shows the numeric value of the type of class and average rating with the respective number of sales. While swimwear has the lowest average rating, Layering has the highest average rating. Blouses, dresses, jackets are observed to have a high number of sales when compared to swim and skirts.

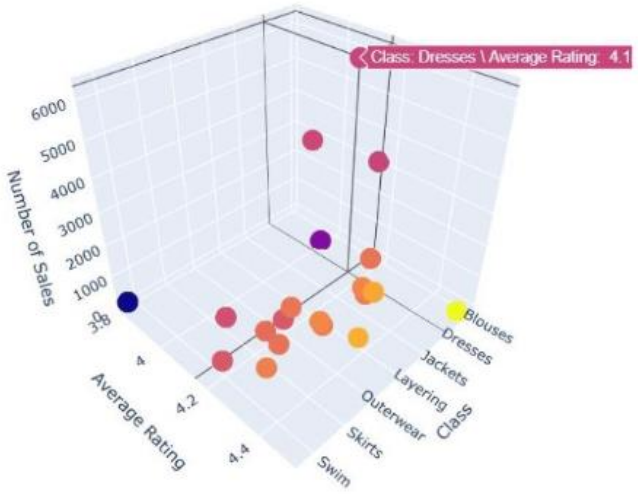

Fig 3-Plot between average rating, classes and number of reviews

B.2.2 Continuous variable by categorical variable:

Here we analyzed two different continuous and categorical variables, Fig 4 is about the ranking ratings; the plot discusses the contrasting interest between the customers ' personal experience with the goods. The company does not recommend only items with lower ratings. However, due to its high number of views, the products with higher rank are not shown in the recommendation list. The 3-star rating products are mainly the products recommended by the business. A business wants to promote its stock and shed light on these product categories as well.

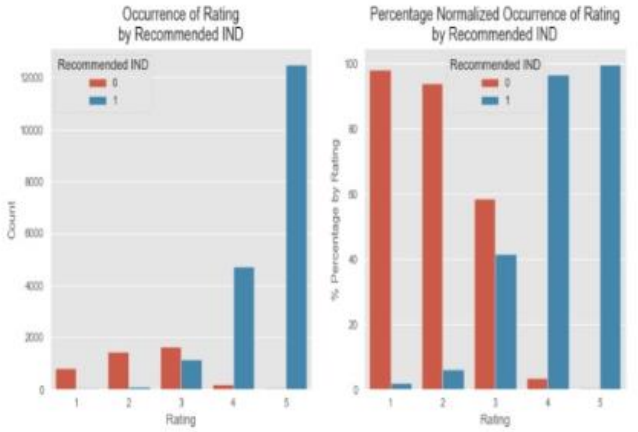

Fig 4-Rating by recommendations

B.2.3 Continuous variable on continuous variable:

Analysis of two continuous variables, Fig 5 is a combination of both the business recommendations and reviews for all the product categories issued. Most products that have tested with high ratings are not recommended, the plot shows the average items that have tested like the trend, bottoms are most recommended for consumers to sell their goods.

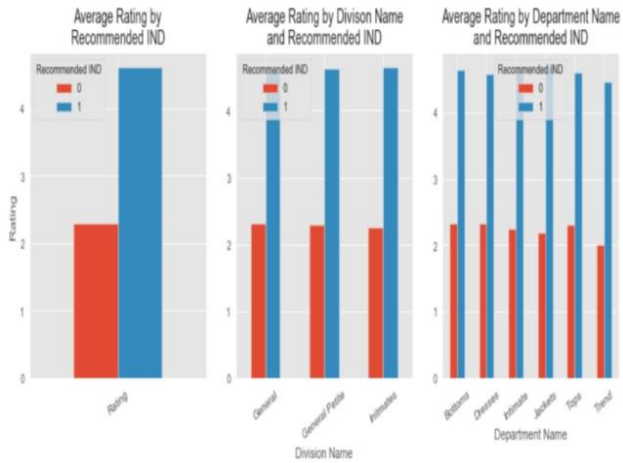

Fig 5-Average ratings and recommendations

B.2.4 Percentage standardize distribution plot:

Standardization by percentage is applied on categorical variables as these variables are mostly unbalanced, Fig 6 provides information on what kinds of ratings are issued to all clothes in the form of a department and the names of the section. We can see that most of the items have 5-star ratings, and the highest-rated clothes under department sessions are of all jackets and intimates. Intimates are highly rated in the group session, and all three category groups are equally ranked close by the close.

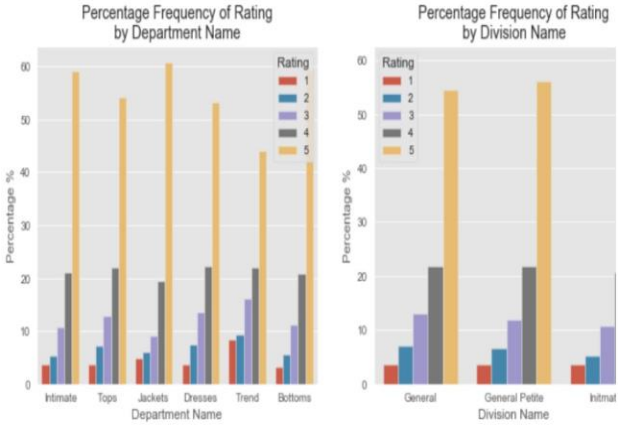

Figure 6-Rating by department and division name

B.2.5 Multivariate analysis and descriptive statistics:

Here we are looking at the averages of different variables and other descriptive statistics act each other, Fig 7 is the frequency distribution plot for rating, recommended ID, and labelling. Here most reviews with a score of five out of five are incredibly positive. It demonstrates most retailers are doing well in the marketplace. Review binary classification, i.e. good = 1 and bad = 0. Great reviews are more numerous than bad reviews. From the part of the label, we can suggest that many of the recommended products have three and more reviews.

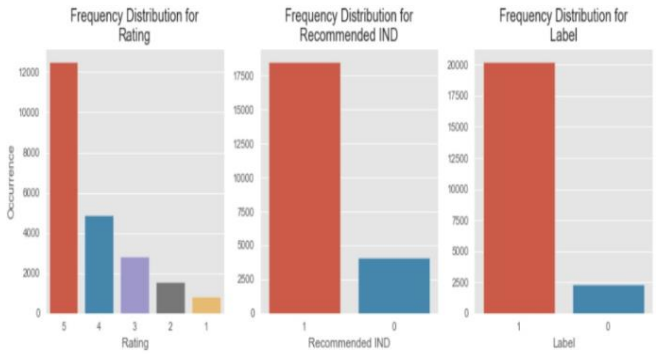

Fig. 7-Distribution of rating, recommended IND, labels

VI. TECHNIQUES EMPLOYED

Our aim is to analysis the reviews and recommends the products according to customer preference. Firstly, customer reviews are analyzed by using the sentiment analysis.

A. Sentiment analysis:

The preprocessing part of text analysis is done by using NLTK package, here we converting all the words into lowercase and tokenized all words to a single word. All the punctuations stop words and irrelevant symbols are removed from the review dataset. The processed words are analyzed and segmented into three categories based on the polarity score of the individual score. Neutral, negative and positive are the three categories that are employed to segment the words.

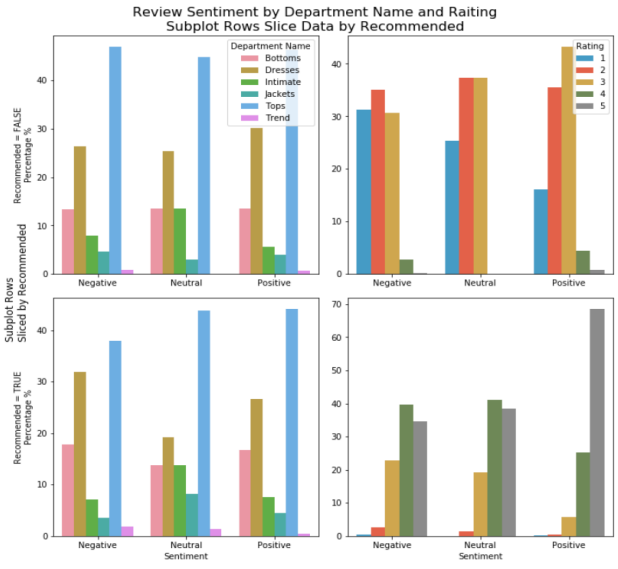

Fig. 8-Sentiment analysis by department name and rating

Fig 8 summarizes that the positive sentiments are the most reviewed, neutral and negative reviews are the least reviewed by the customers. Positive reviews are increasing because the high rating is increasing and department variable doesn't have any effect on the recommendations. However, the rating variable is completely effecting the recommendations.

A.1 Visualization of reviews:

The word frequency is counted the processed word dictionary and the setting function parameter is defined to create graphical plot by using matplotlib function. Following plots are created to see which words are highly repeated.

Fig. 9-High rated comments

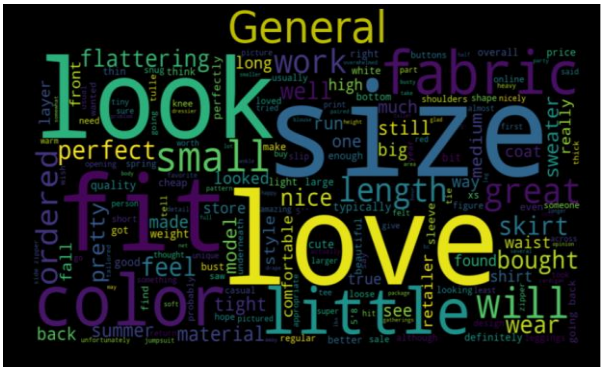

Fig. 10-Word count in general department

From figure 9 and 10 we can infer that the most common words that are used in class name type names are design, cute and major. In the high rated comment section, the most repeated words are dress, love and size. While three types of departments are intimate, general, and general petite. In all these division the most high-frequency words are size, love, look, fit, colour and wear.

B. Recommendation features by N-Grams:

Here we are categorizing the recommendations in two ways i.e. non-recommended items and recommended items. The categorization of words is done by using N-grams, these n-grams are applied for both positive and negative reviews. The n-gram value of 5 is selected so that all the words are separated accordingly. These n gram separations will be helpful for the prediction of recommendations. Fig 11 is a sample of recommended items that are under positive reviews.

	1-Gram	Occurrence	2-Gram	Occurrence	3-Gram	Occurrence	4-Gram	Occurrence	5-Gram
0	dress	8591	true size	1243	fits true size	264	compliments every time wear	46	34b 26 waist 36 hips
1	love	8017	love dress	657	fit true size	192	26 waist 36 hips	32	get compliments every time wear
2	size	7561	5 4	622	received many compliments	163	34b 26 waist 36	28	5 2 currently 33 25
3	fit	5995	usually wear	588	runs true size	143	looks great skinny jeans	25	115 lbs 30 dd 26
4	top	5846	looks great	574	love love love	138	get compliments every time	23	2 currently 33 25 37

Fig. 11-Sample of n-gram separation for recommended items

B.1 Intelligible supervised learning:

We used a supervised learning methodology for the recommendation of products. Fig.12 tell about Naive bayes is the supervised learning algorithm that is for recommendations, naive bayes is a probabilistic model which depends on the bayes theorem. This theorem helps in

categorizing the words by computing the probabilities of words and their occurrence at different classes. This model considers both the good and bad reviews which means the model isn't biased to one category.

```
labtext= list(zip(df.tokenized, (df["Recommended IND"])))

def find_features(document):
    words = set(document)
    features = {}
    for w in word_features:
        features[w] = (w in words)

    return features

featuresets = [(find_features(text), LABEL) for (text, LABEL) in labtext
len(featuresets)


training_set = featuresets[:15000]
testing_set = featuresets[15000:]
```

```
classifier = nltk.NaiveBayesClassifier.train(training_set)

print("Classifier accuracy percent:",(nltk.classify.accuracy(classifier, testing_set))*100)
print(classifier.show_most_informative_features(40))
```

Fig. 12-Navies bayes model

The independent variables are word choice in review and dependent variables are weather or not review was recommended. Tuples are created on bases of customer comments and customer rating labels, then each word is treated as a variable. Now, model notes all the words that are present in the comment section as, for computational complexity, we consider only the top 500 common words that are used by customers in the dataset. We consider only the one-gram words for recommendations and these one-gram tokens are used to categorize based polarization values. Fig 12 shows the results of one-gram words with appropriate recommendations.

```
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
                    bc = True               0 : 1      =      2.5 : 1.0
                  pair = True               1 : 0      =      2.5 : 1.0
```

Fig. 13-Results from navies bayes model

VII. MODEL EVALUATION:

The navies bayes algorithm has shown an overall accuracy of 82% in recommending the products to customers. Fig 14 tells about the model has shown different results on the train and validation sets, the training set has 88% accuracy and the validation sets has shown an overall of 85% accuracy which means the navies bayes is good at recommending to customers.

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

Validation Set Accuracy: 0.8915156871409633
Validation Set ROC: 0.8423407868220376
```

Fig. 14-Model accuracy results

Fig.15 is about the precision, recall and f1-score values of the navies bayes classification algorithm. Here, precision tells about the ratio between the correctly predicted from the overall predicted classes, our model has 56% of accurate recommendations and 96% of non-recommendations. The recall is about the mistakenly predicted values out of the correct predictions. 78% accuracy value of recall means we less number of wrong recommendations.

	precision	recall	f1-score	support
0	0.56	0.78	0.65	597
1	0.96	0.91	0.94	3929
micro avg	0.89	0.89	0.89	4526
macro avg	0.76	0.84	0.79	4526
weighted avg	0.91	0.89	0.90	4526

Fig. 15-Precision recall and f1-score results

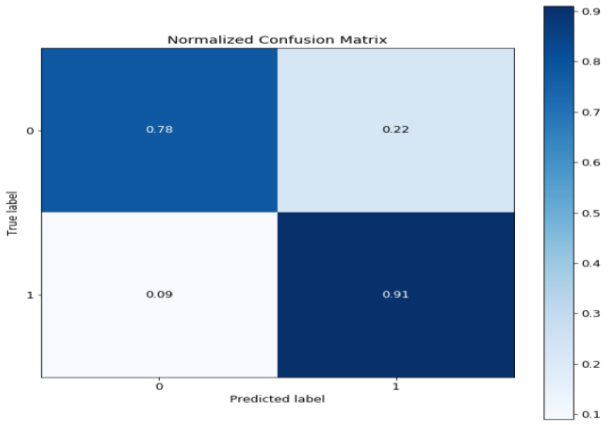

Figure 16 confusion matrix

Fig 16.tells about the confusion matrix tells about the classification accuracy development in the given navies bayes model. The model has shown a 76% efficiency in predicting the true positives and 91% accuracy in filtering the false negatives. Overall, this model performed well in offering recommendations to the customers.

VIII. Conclusion

Predicting the recommendations and analysing the reviews of the data is done by using the sentiment analysis. In our model the supervised algorithm approach i.e. navies bayes algorithm helped in recommending the products with an accuracy of 82%. Most of the products that are recommended by this algorithm have good positive reviews from the customers.

The sentiment analysis on the review data has helped for the recommendation algorithm to recommend good quality products for the customer. It also helped to filter out the products that have bad reviews.

This model can be applied in any online E-commerce retailing business can figure out the potential products that are being sold in the market and recommend the appropriate product according to customer preference.

Further studies will require by including a few new variables like the feedback of customers, negative reviews can help the algorithm to predict more accurately. Implementing the unsupervised algorithms like clustering and word vectorization can result in more precise prediction for the recommendation model.

IX. References

[1] R. Singh and J. Kaur, "Service Quality of Online Shopping Portals: A Review of Literature," Journal of General Management Research, vol. 5, no. 1, pp. 45–63, Jan. 2018.

[2] V. Kaushik, A. Khare, R. Boardman, and M. B. Cano, "Why do online retailers succeed? The identification and prioritization of success factors for Indian fashion retailers," Electronic Commerce Research and Applications, vol. 39, p. 100906, Jan. 2020, doi: 10.1016/j.elerap.2019.100906.

[3] R. Singh and J. Kaur, "Service Quality of Online Shopping Portals: A Review of Literature," Journal of General Management Research, vol. 5, no. 1, pp. 45–63, Jan. 2018.

[4] "Online Shopping Statistics You Need to Know in 2020," OptinMonster, Nov. 06, 2019. https://optinmonster.com/online-shopping-statistics/ (accessed Jun. 26, 2020).

[5] Z. Ruan and K. Siau, "Digital Marketing in the Artificial Intelligence and Machine Learning Age," AMCIS 2019 Proceedings, Aug. 2019, [Online]. Available: https://aisel.aisnet.org/amcis2019/treo/treos/96.

[6] L. Parker, "The AI Revolution is Now: Artificial Intelligence (AI) is capturing and scanning reams of sales data for predictive analytics, tracking consumer behaviour and changing retail as we know it," Earnshaw's Review, vol. 104, no. 2, pp. 8–30, Feb. 2020.

[7] "Featuring Mistakes: The Persuasive Impact of Purchase Mistakes in Online Re...: Discovery Service for NCI Library..." http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=1a96b241-beed-4fbf-aa46-25ddb6a030f3%40sessionmgr4008 (accessed Jun. 26, 2020).

[8] S. H. Choi, S. Kang, and Y. J. Jeon, "Personalized recommendation system based on product specification values," Expert Systems with Applications, vol. 31, no. 3, pp. 607–616, Oct. 2006, doi: 10.1016/j.eswa.2005.09.074.

[9] K. E. Martin, "Ethical Issues in the Big Data Industry," in Strategic Information Management, 5th ed., R. D. Galliers, D. E. Leidner, and B. Simeonova, Eds. Routledge, 2020, pp. 450–471.

[10] D. J. Yates, G. J. J. Gulati, and J. W. Weiss, "Understanding the Impact of Policy, Regulation and Governance on Mobile Broadband Diffusion," in 2013 46th Hawaii InternationalConference on System Sciences, Wailea, HI, USA, Jan. 2013, pp. 2852–2861, doi: 10.1109/HICSS.2013.583.

[11] "Examining Online Purchase Intentions in B2C E-Commerce: Testing an Integrated Model - ProQuest." https://search.proquest.com/docview/215884472/fulltext/805AEFE7A234D03PQ/1?accountid=103381 (accessed Jun.27, 2020)