A TEXT MINING APPROACH TO ANALYZE THE CUSTOMER's SATISFACTION FROM THE ONLINE PRODUCT REVIEWS (OPRs)

A SUMMER PROJECT REPORT

Submitted by

KARTHICK RAJ S

1913050

in partial fulfillment of Summer Project for the award of the degree

of

POST GRADUATE DIPLOMA IN MANAGEMENT

THIAGARAJAR SCHOOL OF MANAGEMENT
PAMBAN SWAMY NAGAR, THIRUPARAKUNDRAM
MADURAI – 625005

OCTOBER - 2020

THIAGARAJAR
SCHOOL OF MANAGEMENT
(An Autonomous College affiliated to Madurai Kamaraj University)
Accredited by NAAC with 'A' Grade

Established in 1962
Pamban Swamy Nagar
Thirupparankundram
Madurai – 625 005
Tel: +91 452 248 4099
Tel: +91 452 248 6900
URL: www.tsm.ac.in

CERTIFICATE

Certified that the Summer Project report ***"A text mining approach to analyze the customer's satisfaction from the Online Product Reviews (OPRs)"*** is the bonafide work of **KARTHICK RAJ.S, 1913050** (PGDM) in Thiagarajar School of Management, Madurai carried out under my supervision during April 2020 to October 2020.

Place: Madurai

Date:

Dr. V. Senthil

Faculty Supervisor

Associate Professor

DECLARATION

I certify that

a. The work contained in this Summer Project report is original and has done by myself under the supervision of my faculty supervisor.
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in writing the report.
d. I have confirmed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Wherever I have used materials (data, theoretical analysis and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references.
f. Wherever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Karthick Raj S

ACKNOWLEDGEMENT

I use this opportunity to express my deepest gratitude to everyone who supported me throughout the course of this PGDM project. The Project opportunity I had was a great chance for learning and professional development.

I would like to express my special thanks of gratitude to our Director **Dr. Murali Sambasivan**, our beloved Principal **Dr. Selvalakshmi M**, and our Dean **Dr. Goutam Sutar**, for providing golden opportunity and valuable guidance for the project.

I would like to extend my sincere thanks to **Dr. Senthil V**, for being my Project Guide. He put all his valuable experience and expertise in directing, suggesting and supporting me throughout the project to bring out the best.

I perceive this opportunity as a big milestone in my career development and wish to thank my family and friends for their consistent support. I will strive to use gained skills and knowledge in the best way, and I will continue to work on it, to attain desired career objectives.

Karthick Raj S

EXECUTIVE SUMMARY

Text mining is the process of analyzing large volumes of documents/text to discover new information. Text mining helps to identify the facts, relationships and assertions that are buried in the mass volume of textual data. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP). Combining NLP with Machine Learning helps to develop a learning framework to clean, process and analyze the textual data. Sentiment analysis is to get real voice or opinion of people towards specific product, persons, services, organizations, news, events, issues and their attributes. It is used to identify the polarity (positive, negative and neutral) of opinions, emotions and evaluations.

The objective of the project is to understand the customer satisfaction in the E-commerce sites. In this project from a particular product category the product's online review data are extracted from the online store's website (www.amazon.in). Data is then preprocessed as the data will contain some noise such as emoticons, punctuations, etc. Once the data is preprocessed it will be analyzed in word level, sentence level and aspect level sentiment analysis. The sentiment analysis can be done through the Machine Learning classifiers through the help of scikit - learn in Python. By doing the opinion mining on the reviews given by the customers, we can come to know about the opinion of the people towards the products, brands. A descriptive text analysis can done for text statistics. Through the analysis of the reviews, the products can be recommended based on the customer satisfaction.

From the analysis we have found that there is a relationship between the customer satisfaction and the customer ratings. Support vector machines is the best suitable algorithm to classify the sentiments of the customers. The opinion of the customer is based on the sentiment polarity of the product. The most recommended product can be identified based on the opinion and the score for the product. The future scope of the project is that it can be used for other text classification applications in different sectors.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	-	Natural Language Processing
OPRs	-	Online Product Reviews
ML	-	Machine Learning
NB	-	Naïve Bayes
DT	-	Decision Trees
K-NN	-	K-Nearest Neighbors
SVM	-	Support Vector Machines
ME	-	Maximum Entropy
OLS	-	Ordinary Least Square
SA	-	Sentiment Analysis
JST	-	Joint Sentiment Topic
eWOM	-	Electronic Word Of Mouth
TF	-	Term Frequency
IDF	-	Inverse Document Frequency
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative

CHAPTER I: INTRODUCTION

1.1 Introduction

The customer satisfaction is the mental state between the perceived quality of the product/service and post purchase performance. The customer satisfaction and customer trust impacts the customer behavior to repurchase and revisit the store. The customer satisfaction can be measured through the customer feedback and ratings. Online reviews has both positive and negative effects to the customer behavior. E-Commerce is buying and selling of goods and services or transmitting of funds or data, over an electronic networking, primarily the internet. The E-commerce sites in India has experienced a remarkable growth along with the rapid growth of internet. With the growth of internet users and internet penetration in India, the online retailing market is still evolving and certainly has room for growth.

According to statista, the digital population of India is 687.6 million and the e-commerce penetration is around 74%. The e-commerce sites offers convenience to the customers. The customers can order the products from their home and the payments are simplified through the online payments via credit cards, debit cards. The e-commerce sites are much user friendly in terms of product selection, payment methods, accurate product description and delivery. Due to the rise of e-commerce, many physical stores also started online business to compete with the online stores. The main challenge to both these online stores is to provide customer satisfaction.

There is a vast amount of data generated daily in the world. There are tools to analyze these data and obtain the knowledge. The text data can be analyzed with the help of data mining techniques and NLP to extract the knowledge about the customer satisfaction. Sentiment analysis is to get real voice or opinion of people towards specific product, persons, services, organizations, news, events, issues and their attributes. It is used to identify the polarity (positive, negative and neutral) of opinions, emotions and evaluations. The Machine learning algorithms are useful in implementing the sentiment analysis classification of reviews with highest accuracy. Some of the machine learning algorithms are Naïve-Bayes, Decision Tree, K – Nearest Neighbors and Support Vector Machines.

The methodology that we have formulated consists of 10 steps. It includes hypothesis, Sentiment polarity identification, Descriptive statistics. The hypothesis of our methodology focus on the relationship between the customer sentiment and customer satisfaction. The descriptive statistics and the visualization of the dataset helps to understand the dataset in a precise manner.

The sentiment polarity is an important factor as it is the factor that deals with the customer satisfaction. It is predicted through the ML algorithms.

The techniques used in the methodology are NLP techniques, ML algorithms, Statistical test and Ordinary Least Square Regression. These techniques will be described in detail in the chapter 3. The tools used for the analysis are Python for Data preprocessing, Visualization, Statistics, and implementation of ML algorithms through scikit-Learn package. The IBM SPSS is used for the statistical analysis.

1.2 Objectives

The objectives of this study are

To analyze the customer satisfaction from the online reviews

To study and select the best method for finding the customer sentiment

To identify the sentiments of the customers based on the reviews

1.3 Scope

The textual data in the social media and websites contains valuable information. Due to the increase in the internet usage and the growth of internet penetration, many e-commerce sites have been started and their product reviews are not taken into account for any insights. The online product reviews contains information useful to the product development and customer needs. It helps business to formulate business strategies. The strength and the weakness of a product can be identified.

1.4 Research Questions

The research questions for this study are

RQ 1: What are the impacts of customer sentiment on the customer satisfaction?

RQ 2: Which classification algorithm will be best suited for the prediction of polarity of the reviews?

RQ 3: What is the opinion of the customers towards the products?

RQ 4: Which products can be recommended to the customers?

The report is classified into 5 parts. The Chapter 2 contains the review of the literature. The Chapter 3 gives an theoretical background and framework for our methodology. The chapter 4 analyze the data, discuss the findings and answers the research questions. The chapter 5 concludes the study with suggestion, future research and limitations.

CHAPTER: II REVIEW OF LITERATURE

2.1 Literature Review

The customer satisfaction has a positive impact on the customer behavior such as revisit of the store/website, word of mouth and repurchase. The study identified that the customer satisfaction has high impacts on the revisit and the customer trust has high impacts on the WOM (Word Of Mouth)[17]. There is a difference in value perceived by the managers and the actual value of the experiences. Through the regression analysis the variation of customer ratings can be explained by the customer sentiment[4]. The pre-purchase stages and post-Purchase stages of also contributes to the overall customer satisfaction. The product selection, and customer service are two of the important components in customer satisfaction. On the other hand on-time delivery has significant impacts on the customer satisfaction[2].

The customer trust and commitment plays a significant role in the spread of eWOM. The customer trust and commitment can only arise from the customer satisfaction[19]. The sentiment polarity and the subjectivity of the reviews has influence on the customer satisfaction because the emotions of the customers are expressed on the reviews. The reviews combined with the ratings can be very useful in finding the customers overall experience and perception[22]. The online review mining is useful to understand the attitude of the reviewers from the polarity. The sentiment analysis and topic mining can be applied to the opinion mining of the reviews. The WSTM (Word-pair Sentiment Topic Model) captures the sentiment and the topic information[21].

Traditionally the customer needs are analyzed with the help of survey, interviews and focus groups. Sentiment analysis is helpful in finding the customer needs and to help the product design[7]. Sentiment analysis uses NLP, text analysis and Machine learning to automate the extraction of the polarity. The sentiments can be structured, semi-structured or unstructured[6].

Social media, Microblogging and E-commerce sites have huge amount of data generated by the users in the form of text, videos, photos and audio. These can be taken advantage of analyzed using the sentiment analysis to get valuable information. The methods used in sentiment analysis are Lexicon-based approach and Machine learning method. The ML approach uses different algorithms and needs labelled data for training the classifier[1]. The stages of sentiment analysis are data collection, data pre-processing, classification and evaluation. It uses three methods of

classification Naïve Bayes, Decision Tree and random forest. The evaluation is based on the accuracy, precision and F1 – measures[3].

The two main challenges of sentiment analysis are the review structure and accuracy. The topic nature and review structure determines the challenges in SA reviews. The second comparison is based on the accuracy rate of sentiment analysis challenges. To obtain the highest accuracy and eliminate the challenges in SA, These four has to be met (1)Research area (2)challenges in accuracy (3)use of sentiment analysis technique and (4)relationship among domain dependence, lexicon type and accuracy result[6]. Learning from the positive and negative labelled examples will not model the accurate classifier. The neutral examples are also important as much as positive and negative examples to classify[8].

The sentiment polarity of the machine model is significantly accurate and not time consuming compared to the human model[7]. The online reviews are useful for both the companies and also for the customers to make a decision. This study deals with the sentiment analysis for amazon product reviews. It compares the accuracy of three machine learning classifiers (NB, SVM, ME) with unigrams and weighted unigrams. The selection of the ML algorithms can be done through the evaluation measures. The SVM is the accurate one among the three classifiers in this study[15].

The sentiment analysis for BHB supplement products customer reviews in amazon in terms of brands, flavor and packaging. The sentiment analysis is based on the lexicon based approach. The two factors flavor and packaging are taken into account for product distribution. The complexity analysis is also done for text statistics. The feedback of the customer is compared with the different brands, flavor, and product type of the BHB products[9]. The products can be ranked and recommended to the users based on the polarity and product score[16].

The business impacts of online reviews are consumer purchase decisions and product sales through the analysis of numerical and textual reviews of the product. The product's numerical and textual reviews influences the sales of the product. A Joint Sentiment Topic (JST) model is used for mining the text reviews topics and their sentiments. The regression analysis and mediation analysis are used for finding the relationships between the sales and both numerical and textual reviews[10]. The business implications of the research are the opinion from large volume of reviews, customer relationship management, product design, marketing and consumers shopping

decision based on the comparison of product on different features. The traditional sentiment analysis methods fail to identify context-sensitive sentiment polarity. A semantic Knowledge based sentiment analysis model is required for sentiment analysis at an fine graded level[18].

S.No	Paper	Theoretical/ Technical	Domain	Methodology
1	Comparative study of machine learning approaches for Amazon reviews	Technical	E-Commerce sites	Machine Learning (NB, SVM, ME)
2	Text mining datasets of β-hydroxybutyrate (BHB) supplement products' consumer online reviews	Technical	E-commerce sites	Lexicon Based Approach
3	Application of data analytics for product design: Sentiment analysis of online product reviews.	Technical	E-Commerce sites	Machine Learning (NB)
4	Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm.	Technical	Microblogging sites	Machine Learning (NB, DT, Random forest)
5	Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining.	Technical	Microblogging sites	Lexicon Based Approach

6	short text sentiment-topic model for product reviews	Technical	E-Commerce sites	WSTM (Lexicon dictionary & Topic model)
7	The effect of online reviews on product sales: A joint sentiment-topic analysis.	Technical	E-Commerce sites	JST
8	Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level.	Technical	E-Commerce sites	Rule based method (Lexicon Dictionary)
9	A survey on sentiment analysis challenges.	Theoretical	Multi-Domain	Systematic Review
10	Sentiment Analysis in Social Media and Its Application: Systematic Literature Review.	Theoretical	Multi-Domain	Systematic Review
11	Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level.	Technical	E-Commerce sites	Fuzzy product ontology mining

Table 1. Gives the overview of the literature in terms of types, domain and methodology used in the papers

CHAPTER III: THEORETICAL BACKGROUNDS AND RESEARCH METHODOLOGY

3.1 Theoretical Backgrounds

3.1.1 Customer satisfaction

The customer satisfaction is the mental state between the perceived quality of the product/service and post purchase performance. The customer satisfaction and customer trust impacts the customer behavior to repurchase and revisit the store. Higher the customer satisfaction higher the customer commitment, this creates a long term relationship with the store. The customer satisfaction can be measured through the customer feedback and ratings. The post purchase performance leads to either customer satisfaction or dissatisfaction. The customers can write feedback about the product/service after the post purchase performance to show their satisfaction. Potential customers can be increased through the sharing of online reviews and ratings.

3.1.2 Online Reviews and Ratings

Online reviews/eWOM (Electronic Word of Mouth) has both positive and negative effects to the customer behavior. Positive reviews and ratings are based on customer satisfaction. When the customer receive higher satisfaction, they tend to provide positive review towards the product. The online product review mining helps to understand the attitude of the customer towards the product. The physical stores has the advantage of showing the product to the customers. Thus the quality of the product can be seen and also touch the product. Whereas in online stores, reviews are the only way to know about the quality of the product. The reviews helps the customers to make a decision about the product. There are three types of review formats in which the customer opinions are expressed. The formats are pros/cons type, detailed type and free type. The pros/cons type of review contains the product likes and dislikes. The detailed format contains the detailed opinions of the product. The free format is a mix of both the review types. The reviews can either be subjective or objective towards the product. The reviews can also be implicit or explicit about the features.

The rating gives a numerical to the customer satisfaction towards the product. The ratings of the product scales from one star to five star. The five being the highest and one being the lowest.

The review data that is to be analyzed can be classified into three types-Structured, Semi-Structured and Unstructured data. Mostly text data are presented in unstructured format.

3.1.3 Data mining

There is a vast amount of data generated daily in the world. There are tools to analyze these data and obtain the knowledge. Data mining or Knowledge Discovery from Data (KDD) is the process of obtaining knowledge by mining the data. It involves the following process data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation.

Data mining utilizes techniques from other domains also. Some of the techniques that are used in this methodology are statistics, Machine Learning, Visualization, Algorithms, Applications and pattern recognition.

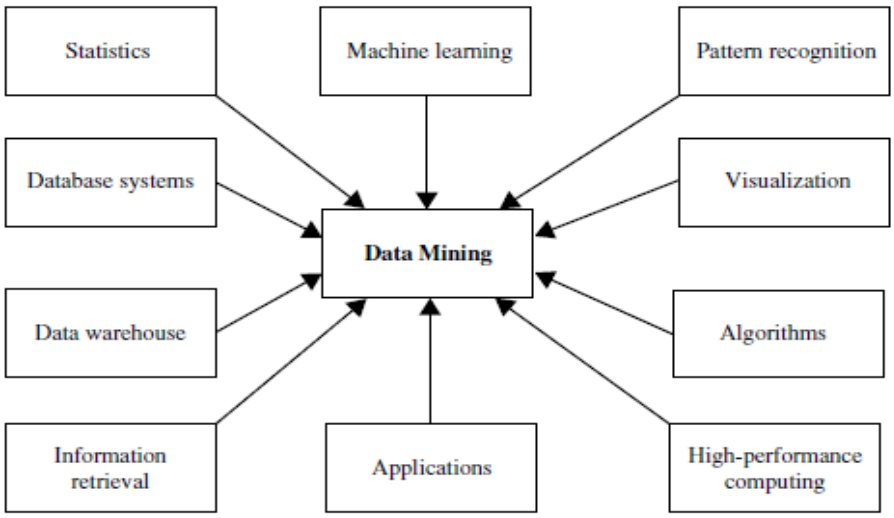


Fig 1.Techniques used in Data mining [5]

3.1.4 Text mining

There are many kinds of data present in the world. The text data is the only kind of data that we are interested in this project. Text mining/ Text analysis is the process of transforming the unstructured text data into meaningful information and insights. Some of the text mining techniques such as Natural Language Processing (NLP), Sentiment analysis will be utilized in our methodology.

3.1.5 NLP

Natural Language Processing is the process of allowing computers to analyze and understand the human language. The NLP open source library used in the methodology is Natural Language Toolkit (NLTK). NLTK is a python library that provides modules for processing text, classifying, tokenizing, stemming, tagging, parsing, and more.

NLP techniques

Some of the Natural Language Processing Techniques that are used are explained below

TF – IDF

The words that occur too frequent in the reviews but are not important can be eliminated by giving each words weightage based on the frequency. For this purpose the Term Frequency Inverse Document Frequency is used.

TF: Term Frequency defines how frequent a word occurs in a document.

$$\text{TF} = \frac{No.\ of\ words\ appear\ in\ the\ document}{Total\ no.\ of\ words\ in\ the\ document}$$

IDF: Inverse Document Frequency defines how significant that term is in the whole corpus.

$$\text{IDF} = \log_{10} \frac{Number\ of\ document}{Number\ of\ document\ in\ which\ word\ appear}$$

The TF ranks the words based on the frequency and the IDF ranks the words based on the relevancy of the word. The product of both gives the TF-IDF.

Stemming

The goal of stemming is to reduce the inflectional forms into common base forms. The example of stemming is that the words consultant, consult, consulting will be reduced to consult.

The porter stemmer is one of the popular stemming algorithms used frequently. It is also very effective in stemming the words.

Stopwords

Stopwords are the words that are used frequently. They are the most common words that are not important. Examples of stopwords are A, An, The and etc…

3.1.6 Sentiment analysis

Sentiment analysis is a branch of text mining. The term analysis of sentiments was first presented by Nasukawa and Yi in the paper **"Sentiment Analysis: Capturing Favorability Using Natural Language Processing"** and the first term opinion mining emerged by Dave et al., in his paper **"Mining the peanut gallery: opinion extraction and semantic classification of product reviews"**.

It is a technique used for analyzing the polarity (Positive, Negative and Neutral) of the text. The sentiment analysis finds the sentiment of the text/review by classifying the review's sentiment through Natural Language Processing (NLP), text analysis and other computational techniques. Sentiment analysis is also called as opinion mining as it involves the mining of opinions of the reviews. The challenges in sentiment analysis are based on the topic nature and the review structure.

The sentiment analysis can be done in three levels. They are

Document-level analysis:

Determines the overall sentiment orientation of a document. Such an approach is useful if the document contains information about a single feature.

Sentence-level analysis:

Each sentence is first checked for being subjective or objective. Then, in case of having a subjective sentence, its positive or negative orientation is determined.

Aspect or feature-level:

Techniques perform finer-grained analysis to first find the feature-opinion pairs in a given sentence and then ascertain their sentiment classifications.

There are 2 main methods of sentiment analysis have been identified which is a machine learning approach and lexicon-based approach. The lexicon based approach classifies the sentiment based on the frequency of positive and negative words while the Machine learning approach utilizes algorithm to classify the sentiments.

The lexicon based approach doesn't cover all the challenges and it is not that much effective as the quality is based on the quality of the lexicon used. The machine learning approach uses supervised machine learning algorithms. Some of the commonly used algorithms are SVM and Naïve Bayes.

Different algorithm performs differently based on the accuracy, working methods and process time. The evaluation of different classifiers will be based on the accuracy, precision and other such measures.

3.1.7 Machine Learning

Machine learning is a technique that is evolved from the study of pattern recognition and computational learning theory in Artificial Intelligence. Machine learning allows the computer to learn from the algorithms and make predictions on data based on the algorithms.

Machine learning can be broadly classified into Supervised Learning, Unsupervised Learning and Semi-Supervised Learning. Supervised Learning has labelled examples of inputs and outputs to achieve its goal of predicting the data. The Unsupervised Learning doesn't have any labelled outputs whereas it uses the data to find its own structure and predict the data. The Semi-Supervised learning deals with both the labelled and unlabeled datasets.

3.1.8 Classification

The data mining is broadly classified into two functionalities: Descriptive and Predictive. Descriptive data mining examines the properties and characteristics of the data while the Predictive data mining learns from the current data and make predictions based on it.

The prediction of data can be done in two ways Regression and Classification. In regression the target values are continuous values. In classification the target values are discrete classes. The classification will be used by the sentiment analysis to find the polarity of reviews. The classification can be linear or multi class.

3.1.9 Text classifiers

Data classification contains two steps Learning step and classification step. In the learning step, a model or classifier is constructed to predict the categories.

Training data → Classification Algorithm → Classification rule

Classification rule → Evaluation data → Evaluation

→Test data → predicting categories

Some of the classification algorithms that has been implemented in the methodology are Naïve Bayes, Support Vector Machines, Decision tree and K-Nearest Neighbors.

K-Nearest Neighbors

The K-Nearest Neighbor is based on learning from the neighbors in the training dataset and labelling the test dataset. "Closeness" is the distance between the test data from its neighbors in the training data. K-NN is a conventional nonparametric classifier. The most commonly used measure is the Euclidean distance.

The Pythagorean Theorem can be used to calculate the distance between two points, as shown in the figure below. If the points (x1,y1) and (x2,y2) are in 2-dimensional space, then the Euclidean distance between them is $\sqrt{(x2-x1)^2+(y2-y1)^2}$.

$$Dist(X1, X2) = \sqrt{\sum_{i=1}^{n}[x1(i) - x2(i)]^2}$$

The first circle represents the k = 3 whereas the second dotted circle represents k=5. The value of k will be determined based on the error rate of the classifier.

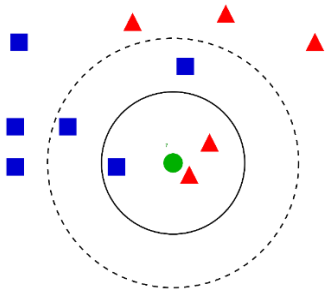


Fig 2. Example for K-Nearest Neighbor algorithm [5]

Decision Tree

Decision tree is a flowchart-like tree structure, where each internal node (non –leaf node) denotes a test on an attribute, each branch represents an outcome of the test and connect to the next node, and each leaf node (or terminal node) holds a category. The topmost node in a decision tree is the root node.

An example of a decision tree is given below. The decision tree explains whether a person can be accepted for loan or not. The outcome of each leaf node will be either Yes or No.

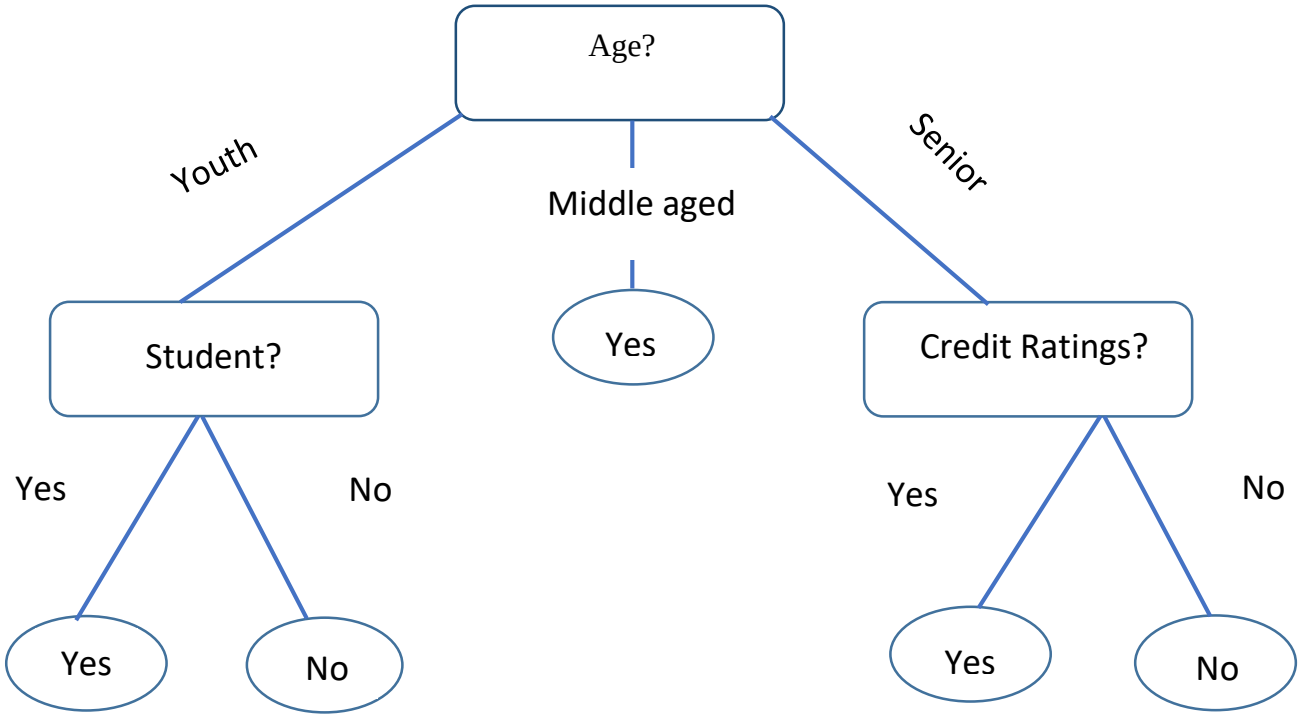


Fig 3. Example for decision Tree algorithm

It is easily prone to overfitting of data.

Naïve Bayes Classifier

Naïve Bayes is a simple and effective classification method. It has an error rate of 20%.It's called "naive" because its core assumption of conditional independence (i.e. all input features are independent from one another).The NB classifier is based on the Bayes theorem

The Bayes theorem states that,

$$P(B|A) = [P(A) * P(A|B)]/P(A)$$

This method will produce the highest probability of the category

In our case we have three category/Class i.e. Polarity of sentiments (Positive, Negative and Neutral)

The probability of a review r being in class c is computed as

$$P(C_k|X_{1,2,\ldots..n}) = P(C_k) *[P(X_{1,2,\ldots..n}| C_k)/ P(C_k)]$$

where $P(X_i|c)$ is the conditional probability of feature vector (x_1, …, xn) occurring in a review of class c

P(c) is the prior probability of a review occurring in class c.

According to the naïve bayes independent assumption of features

$$P(X_{1,2,\ldots..n}|\ C_k) = P(X_i|C_k)$$

$$P(C_k|X_{1,2,\ldots..n}) = P(C_k) * \prod_{i=1}^{n} P(X_i|C_k)/P(C_k)$$

$$P(C_k|X_{1,2,\ldots..n}) \propto P(C_k) * \prod_{i=1}^{n} P(X_i|C_k)$$

$$=> \tilde{Y} = argmax_k\ P(C_k) * \prod_{i=1}^{n} P(X_i|C_k)$$

To avoid underflow, log probabilities can be used

$$\tilde{Y} = argmax_k\ [\ln\{P(C_k)\} * \ln\{\sum_{i=1}^{n} P(X_i|C_k)\}]$$

Prior probability $P(C_k) = N_c/N$

N_c is the number of reviews in class c in training

N is the total number of reviews in training.

Support Vector Machines

SVM can be used for both linear and non-linear data. It uses an non-linear mapping to classify the data into new dimension and it finds a linear optimal separating hyperplane within the new dimension. The hyperplane will be formed based on the support vectors and margins. Eventhough they are accurate and classify complex non-linear decision boundaries they are slow. The overfitting of data is lower when compared to other models.

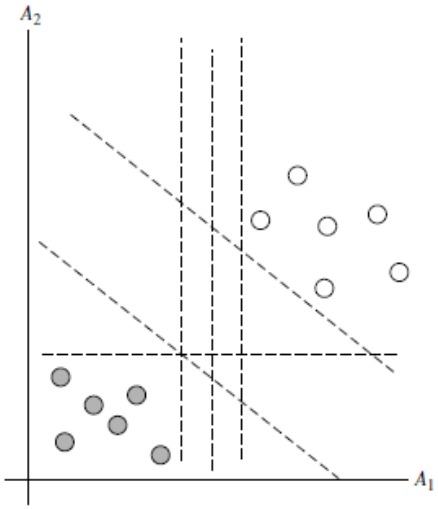


Fig 4. Example for Support Vector Algorithm [5]

The above diagram is an example of linearly separable SVM. There is an infinite number of decision boundary for the above data. The minimum classification error decision boundary will be taken as the hyperplane.

The nonlinear kind of data also follows the same procedure to find the hyperplane but it cannot be separable by a linear line. The SVM will add another dimension and finds the hyperplane. This can be done by using the kernel. For an linear seperable data, linear kernel can be used. There are also other types of kernel such as RBF and etc…

3.1.10 Evaluation

There are many classifier that can be used for the classification of polarity of the reviews. The best classifier model will be selected based on the measures such as accuracy, precision and etc.. These measures will be generated based on the confusion matrix. The confusion matrix consist of the TP, TN, FP and FN

True Positive: These refer to the Positive labels that were correctly labeled by the classifier.

True Negative: These refer to the Negative labels that were correctly labeled by the classifier.

False Positive: These are the Negative labels that were incorrectly labeled as Positive.

False Negative: These are the Positive labels that were incorrectly labeled as Negative.

The TP, TN, FP and FN will be used for the construction of confusion matrix.

CM	**Predicted Label**		
True Label		**YES**	**NO**
	YES	TP	FN
	NO	FP	TN

Table 2. Confusion matrix

The above table gives the confusion matrix of two class classifier.

Let P be the Total positive labels and N be the Total Negative labels.

Accuracy: It is the ratio of correctly labelled predictions to all predictions.

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

Precision: It is the ratio of correctly labelled positive predictions to all positive predictions. It gives the percentage of all positively predicted labels that are actually positive. It is the measure of exactness.

$$\text{Precision} = \frac{TP}{TP+FP}$$

As the precision is high, it has low false positive rate.

Recall: It is the ratio of correctly labelled predictions to all labelling to the class. It is the percentage of positive labels that are predicted. It is the measure of completeness. It is otherwise called as sensitivity or True Positive Rate.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{TP}{P}$$

F1-Score: It is the weighted average of precision and recall. The F1-score is the measure of combining both precision and recall.

$$\text{F1-Score} = \frac{2*Precision*Recall}{Precision+Recall}$$

The Precision, Recall and F1-Score can also be calculated classwise for each category.

3.1.11 Statistical tests

The data mining makes use of techniques and applications from other domains. One of the important domain is Statistics. Some of the few common statistical tests are Regression, Chi-square, Comparison of means and Non parametric tests. The tests can be used for any data but the importance of using these tests is knowing when and where to use these tests.

3.2 The Methodology

The methodology that we have framed for this study contains the following ten steps

STEP 1: Hypothesis formulation.

The Alternate and Null hypothesis are formulated in this step.

STEP 2: Extraction of Product Reviews.

The Online Product Reviews (OPRs) are extracted for the further analysis of the study.

STEP 3: Data cleaning.

Before analyzing the data, the data has to be preprocessed. As the reviews are text data, NLP techniques are used for the cleaning of data.

STEP 4: Evaluation of Classifiers.

This is the most important step in the methodology. The classifier are trained on the learning data and evaluated based on accuracy.

STEP 5: Polarity of the reviews.

The polarity of the test data are predicted using the trained classifier. This gives the sentiment polarity.

STEP 6: Product score.

The product score is calculated based on the age, length, polarity, subjectivity and rating of the reviews.

STEP 7: Descriptive Statistics of Reviews.

This gives the textual analysis and the summary of the reviews

STEP 8: Hypothesis testing.

The formulated hypothesis is tested based on the statistical test.

STEP 9: Relationship between Customer Ratings and Customer Sentiments.

The linear regression is used for finding the relationship between the variables.

STEP 10: Recommendation of the product.

The product are ranked and recommended based on the product score and % of polarity.

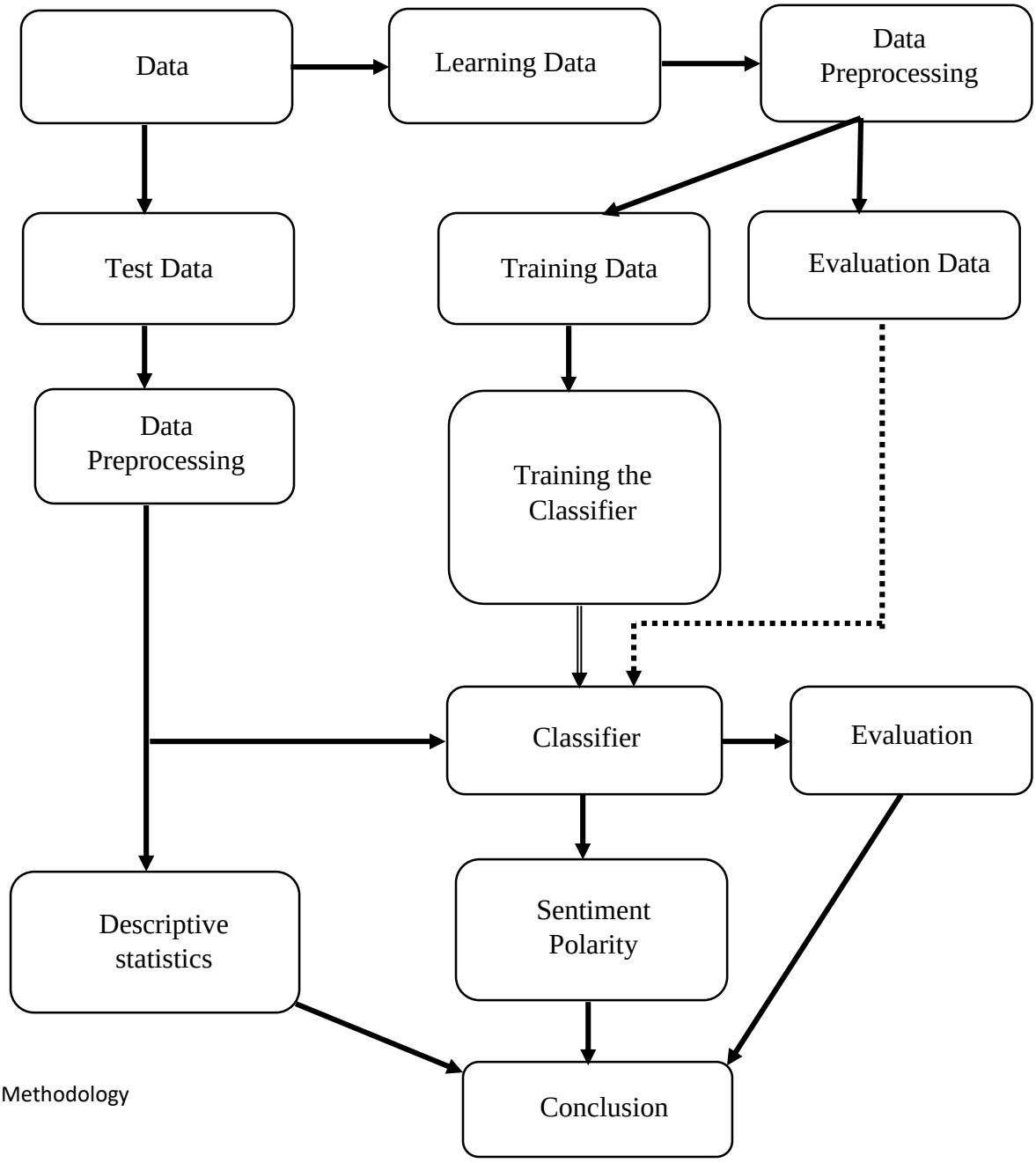


Fig 5. Methodology

CHAPTER IV: RESULTS AND DISCUSSION

4.1 Discussion

4.1.1 Data Description

The total dataset used for the study are 29912

The dataset contains the reviews of six different categories of amazon product. The categories are Books, Mobiles, Camera, Household, Entertainment and watches.

Ratings	Learning data	Test data
1-Star	2406	1525
2-Star	779	436
3-Star	1448	740
4-Star	3524	2035
5-Star	10266	6753
Total	18423	11489
Total dataset	29912	

Table 3. Description of total dataset

4.1.2 Hypothesis formulation

Customer satisfaction may be measured by the gap between perceived quality of the product or service, and pre-purchase quality expectations.

The Customer satisfaction can be expressed in the form of Ratings.

H1: The Customer Sentiment Polarity has a positive impact on the Customer Ratings.

H0: There is no relation between Customer Sentiment Polarity and Customer Ratings.

4.1.3 Extraction of Product Reviews

This step includes three sub steps

- Choosing the website
- Choosing the product
- Extraction of the reviews

The amazon is one of the popular online marketplace for variety of products. The amazon (www.amazon.in) has been chosen as the e-commerce website for the extraction of product reviews.

The amazon has a huge variety and different brand of products. The products are classified into different categories. Six most commonly used and ordered product category are selected. The categories are: Books, Camera, Mobiles, household, Entertainment and Watches.

From these six categories, seven products has been chosen for the analysis. The products present in the test dataset are

Books – The Intelligent Investor

Camera – Nikon D3500

Mobiles – Apple iphone 11 Max Pro and Samsung Galaxy M31

Household – Vim dishwash gel

Entertainment - Echo Input Portable Smart Speaker

Watches - Fastrack Reflex 2.0 Activity Tracker

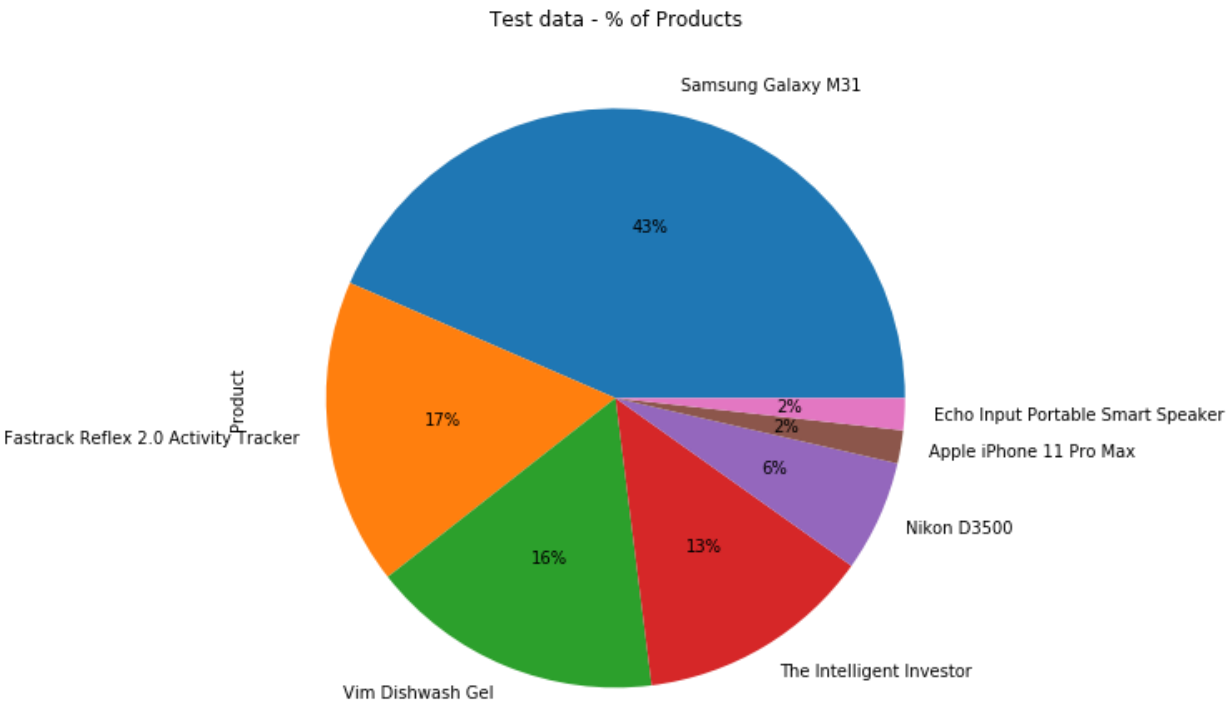


Fig 6. % of test data product wise

The product reviews are scraped from the amazon website with the help of Webscraper extension of google chrome. The extracted data contains the Product name, Reviewer name, Review title, Review content, Review rating and Date of the review.

4.1.4 Data cleaning

The data cleaning process contains steps. They are

- Removing Unwanted data
- Removing Unwanted text data/Noise
- Removing the stopwords
- Word normalization

The unwanted contents are removed from the data. Only the data useful for the analysis are taken such as Review title, Review content, Review rating and Date of the review. The name of the reviewer are removed.

Other than the alphabets and numbers, The reviews also contains unwanted text data or noise such as Punctuations, Emoticons and etc.. These are removed in this process.

The stopwords (a, an, the) are removed from the reviews using the stopwords in NLTK package of python.

The word normalization is done through Stemming of words. Stemming is the process of converting words to their root form. The porterstemmer in the NLTK package is used for the stemming of words in reviews.

4.1.5 Evaluation of Classifiers

This step involves the following steps

- Labelling the Learning data
- Splitting the data
- Fitting the training data to classifier
- Classifying the evaluation data
- Comparison of the classifiers

Labelling the Learning data

The labelled learning data is important for training the classifier on the basis of the known content. The learning data is labelled based on the sentiment polarity of the content using the textblob package.

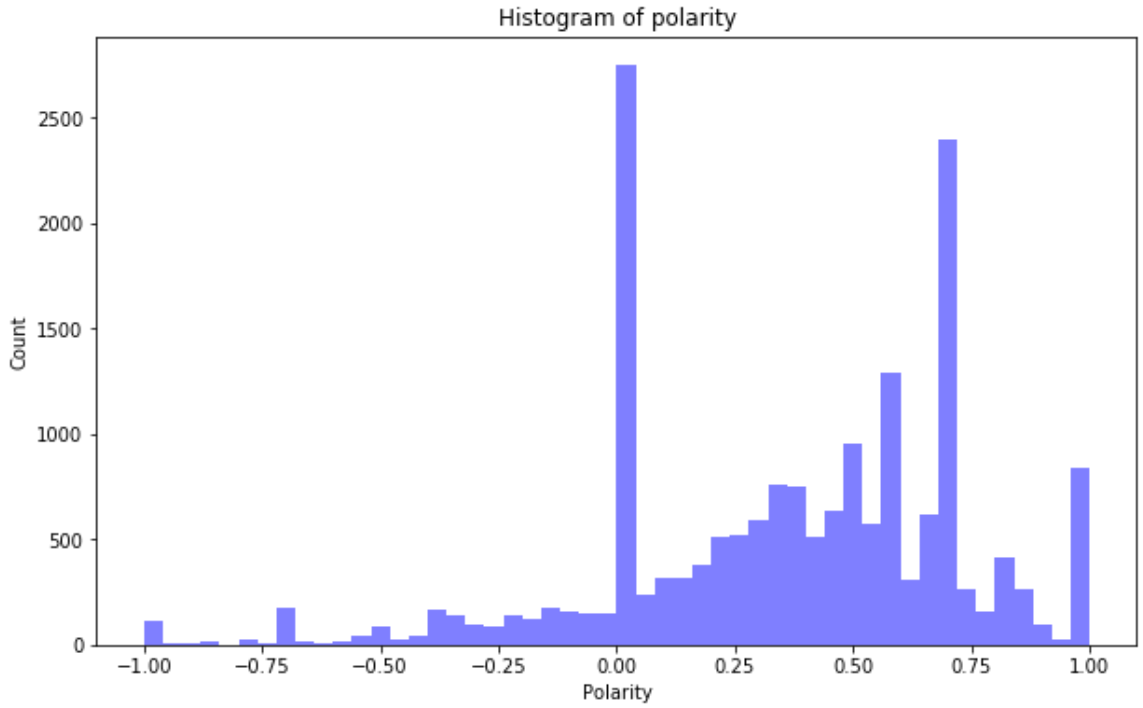


Fig 7. Histogram of Polarity

The graph shows the histogram of the polarity of the data.

Based on the polarity the reviews are labelled as Positive, Negative and Neutral.

If the polarity is greater than zero then the reviews are **"Positive"**.

If it is lesser than zero then the reviews are **"Negative".**

If it is equal to zero then the reviews are **"Neutral"**.

Splitting the data:

To evaluate the performance of the classifiers, the labelled data is classified into training data and evaluation data in the ratio of 75:25.

The no. of training dataset used are 13821.

The training data is used for training the classifier and the evaluation data is used for evaluating the performance of the classifier.

Fitting the training data to classifier:

The training data is fitted in the classifier to create a model for classifying the reviews.

Classifying the evaluation data:

Based on the knowledge from the training data, the evaluation data is classified into Positive, Negative and Neutral.

Comparison of the classifiers:

The classifiers that are used are

- ✓ K-Nearest Neighbor (K-NN) Classifier
- ✓ Decision Tree Classifier
- ✓ Multinomial Naïve Bayes Classifier
- ✓ Support Vector machines

Other than labelling the data other steps uses the sklearn package for splitting the data, fitting the model, predictions based on the model and the classification report for the model.

The Classification report is important for evaluating the classifiers based on the accuracy, precision, recall and etc..

The confusion matrix and classification report of the classifiers are given below:

K-Nearest Neighbor (K-NN) Classifier:

Confusion matrix

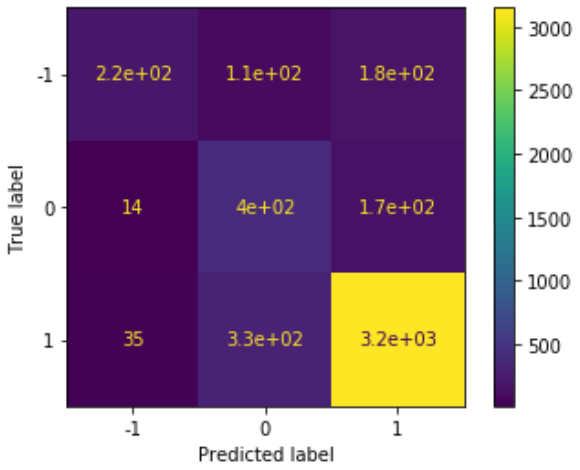


Fig 8. Confusion matrix of K-NN

CM: K-NN	Predicted Label			
True Label		NEG	NEU	POS
	NEG	218	110	176
	NEU	14	403	169
	POS	35	332	3151

Table 4. Confusion matrix of K-NN

Classification Report:

K-NN	Precision	Recall	F1-Score
NEG	0.82	0.43	0.57
NEU	0.48	0.69	0.56
POS	0.90	0.90	0.90
Macro Avg.	0.73	0.67	0.68
Weighted Avg.	0.84	0.82	0.82
Accuracy:	0.82		

Table 5. Classification Report of K-NN

Decision Tree Classifier:

Confusion matrix

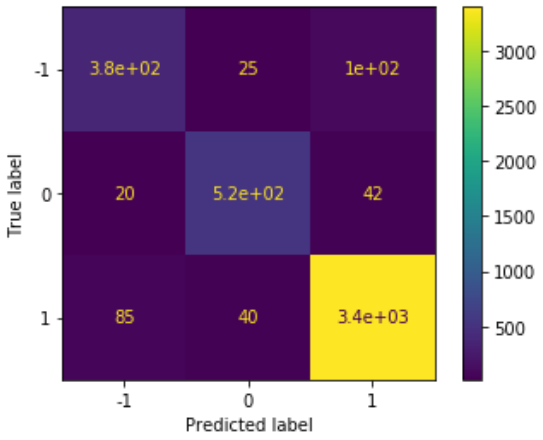


Fig 9. Confusion matrix of decision tree

CM: DT	Predicted Label		
		NEG	**NEU**	**POS**
True Label	**NEG**	376	25	103
	NEU	20	524	42
	POS	85	40	3393

Table 6. Confusion matrix of Decision tree

Classification Report:

DT	Precision	Recall	F1-Score
NEG	0.78	0.75	0.76
NEU	0.89	0.89	0.89
POS	0.96	0.96	0.96
Macro Avg.	0.88	0.87	0.87
Weighted Avg.	0.93	0.93	0.93
Accuracy:		0.93	

Table 7. Classification report of Decision tree

Multinomial Naïve Bayes Classifier:

Confusion matrix

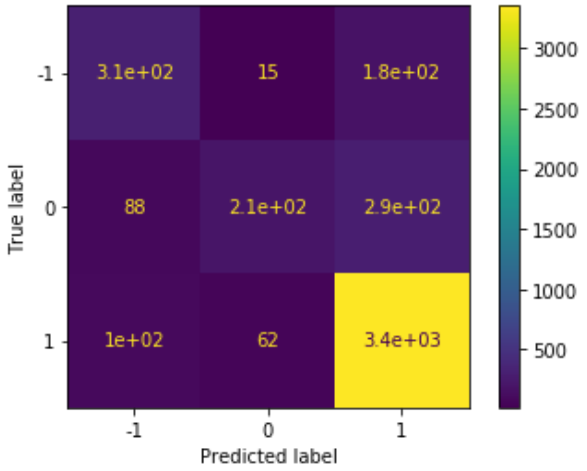


Fig 10. Confusion matrix of Multinomial Naïve Bayes

CM: M-NB	Predicted Label		
		NEG	**NEU**	**POS**
True Label	**NEG**	309	15	180
	NEU	88	206	292
	POS	101	62	3355

Table 8. Confusion matrix of Multinomial Naïve Bayes

Classification Report:

M-NB	Precision	Recall	F1-Score
NEG	0.62	0.61	0.62
NEU	0.73	0.35	0.47
POS	0.88	0.95	0.91
Macro Avg.	0.74	0.64	0.67
Weighted Avg.	0.83	0.84	0.83
Accuracy:	0.84		

Table 9. Classification report of Multinomial Naïve Bayes

Multinomial Naïve Bayes Classifier (TF-IDF):

Confusion matrix

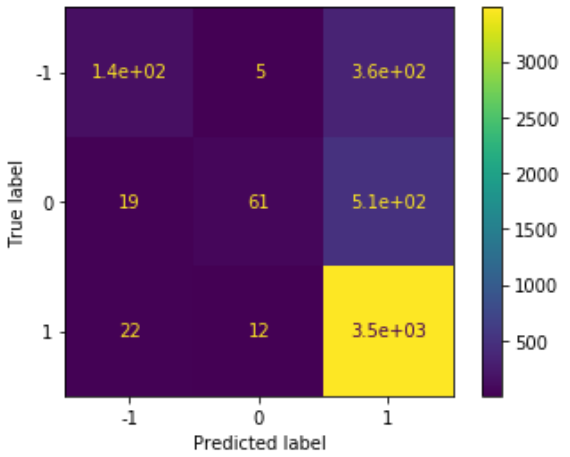


Fig 11. Confusion matrix of Multinomial Naïve Bayes 2

CM: M-NB(TFIDF)		Predicted Label		
		NEG	**NEU**	**POS**
True Label	**NEG**	143	5	356
	NEU	19	61	506
	POS	22	12	3484

Table 10. Confusion matrix of Multinomial Naïve Bayes 2

Classification Report:

M-NB(TF-IDF)	Precision	Recall	F1-Score
NEG	0.78	0.28	0.42
NEU	0.78	0.10	0.18
POS	0.80	0.99	0.89
Macro Avg.	0.79	0.46	0.50
Weighted Avg.	0.80	0.80	0.75
Accuracy:	0.80		

Table 11. Classification report of Multinomial Naïve Bayes

Support Vector machines:

Confusion matrix

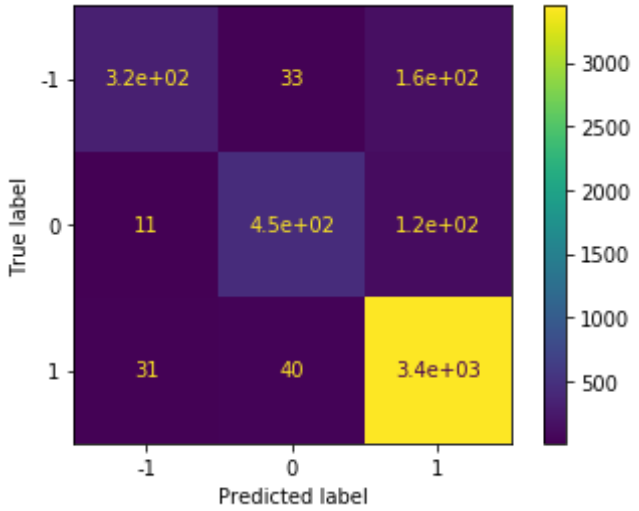


Fig 12. Confusion matrix of Support vector Machines

CM: SVM	Predicted Label			
		NEG	**NEU**	**POS**
True Label	**NEG**	316	33	155
	NEU	11	452	123
	POS	31	40	3447

Table 12. Confusion matrix of Support vector Machines

Classification Report:

SVM	Precision	Recall	F1-Score
NEG	0.88	0.63	0.73
NEU	0.86	0.77	0.81
POS	0.93	0.98	0.95
Macro Avg.	0.89	0.79	0.83
Weighted Avg.	0.91	0.91	0.91
Accuracy:	0.91		

Table 13. Classification report of Support Vector Machines

Comparison of Precision:

Classifier	Neg	Neu	Pos
K-NN	0.82	0.48	0.90
Decision Tree	0.78	0.89	0.96
Multinomial Naïve Bayes	0.62	0.73	0.88
Multinomial Naïve Bayes(TF-IDF)	0.78	0.78	0.80
SVM	0.88	0.86	0.93

Table 14. Comparison of classifier's Precision

From the viewpoint of precision, the negative category has highest precision in SVM, second highest in neutral category and third in the positive category. As a whole considering all the categories SVM has the best precision value.

Comparison of Recall:

Classifier	Neg	Neu	Pos
K-NN	0.43	0.69	0.90
Decision Tree	0.75	0.89	0.96
Multinomial Naïve Bayes	0.61	0.35	0.95
Multinomial Naïve Bayes(TF-IDF)	0.28	0.10	0.99
SVM	0.63	0.77	0.98

Table 15. Comparison of classifier's Recall

From the viewpoint of Recall, as a whole considering all there categories SVM is the best classifier among the others.

Comparison of F1-Score:

Classifier	Neg	Neu	Pos
K-NN	0.57	0.56	0.90
Decision Tree	0.76	0.89	0.96
Multinomial Naïve Bayes	0.62	0.47	0.91
Multinomial Naïve Bayes(TF-IDF)	0.42	0.18	0.89
SVM	0.73	0.81	0.95

Table 16. Comparison of classifier's F1-score

From the viewpoint of F1-score, DT is the best classifier as it has the highest F1-score among all categories.

Comparison of Accuracy:

Classifier	Accuracy
K-NN	82%
Decision Tree	93%
Multinomial Naïve Bayes	84%
Multinomial Naïve Bayes(TF-IDF)	80%
SVM	91%

Table 17. Comparison of classifier's Accuracy

The DT has the highest accuracy followed by SVM. Considering Precision, Recall, F1-score and Accuracy of all the categories, SVM will be a most suitable classifier. Also SVM is the best and powerful classifier[20].

4.1.6 Polarity of the reviews

Using the trained SVM classifier the polarity (Positive, Negative, Neutral) of the reviews are labeled.

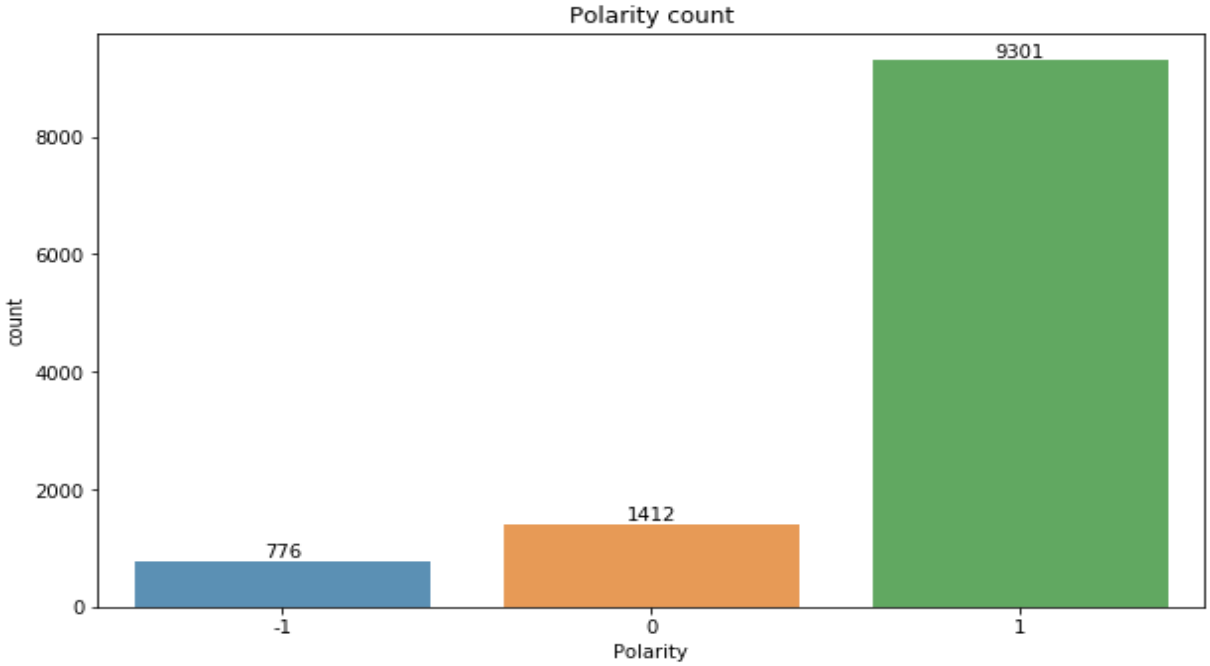


Fig 13. Polarity Count of the test data

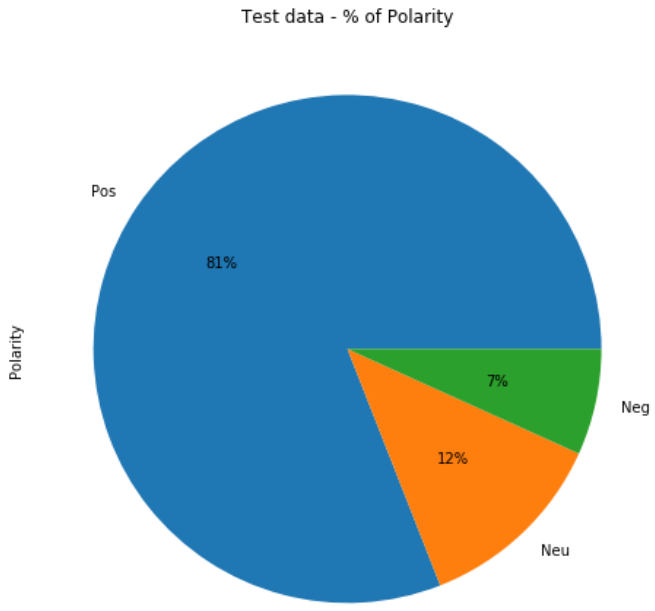


Fig 14. % of Polarity in test data

4.1.7 Product score

The score is generated for all reviews based on the Ratings, Subjectivity, Polarity, Age and Count of words of the reviews.

Score = Ratings + Subjectivity + Polarity + Age + Count

The mean of all scores for the particular product is used as the product score.

The Intelligent Investor - 117.47

Nikon D3500 - 17.65

Apple iphone 11 Max pro - 3.37

Echo Input Portable Smart Speaker - 2.20

Vim dishwash gel - 72.49

Fastrack Reflex 2.0 Activity Tracker - 54.96

Samsung Galaxy M31 - 42.67

4.1.8 Descriptive Statistics of Reviews

The graph gives the distribution of the reviews based on the ratings. The 5-star ratings has the highest amount of reviews. From this we can say that the 80% of the reviews are positive considering the 4-star and 5-star ratings.

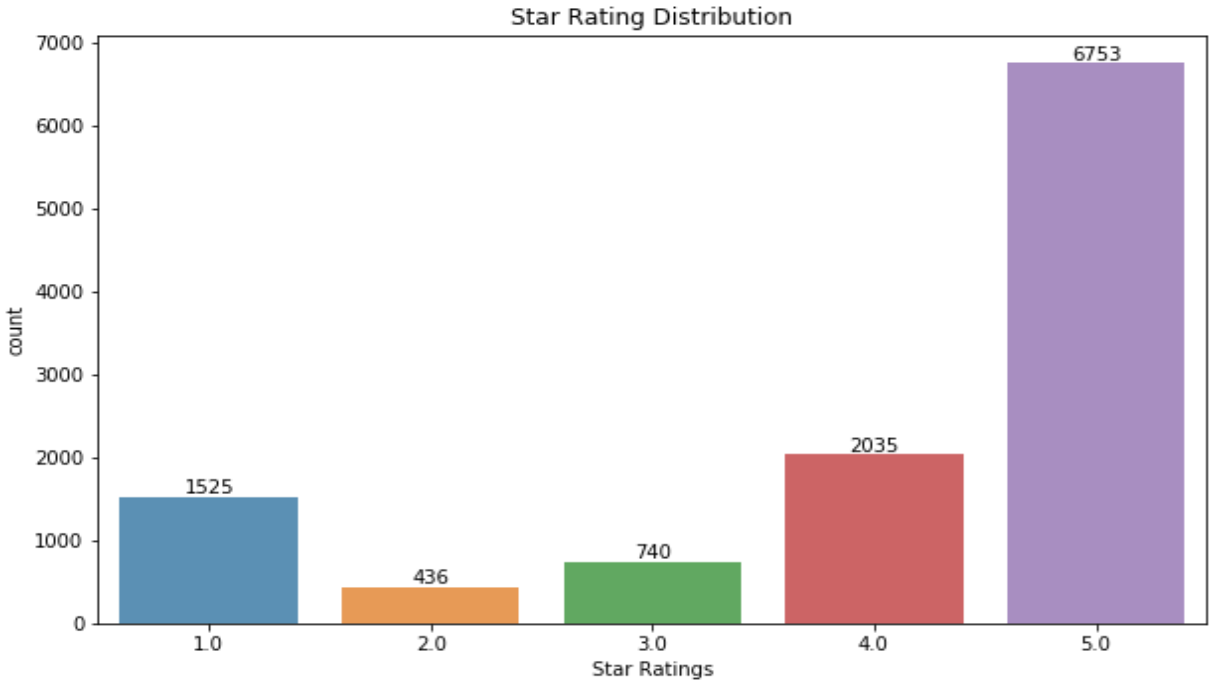


Fig 15. Distribution of Ratings

The graph gives the distribution of the reviews based on the ratings. Most of the reviews present in the dataset are reviewed in the year 2018, 2019 and 2020. Even though the year 2020 has only 5 months of data, it has the huge amount of dataset.

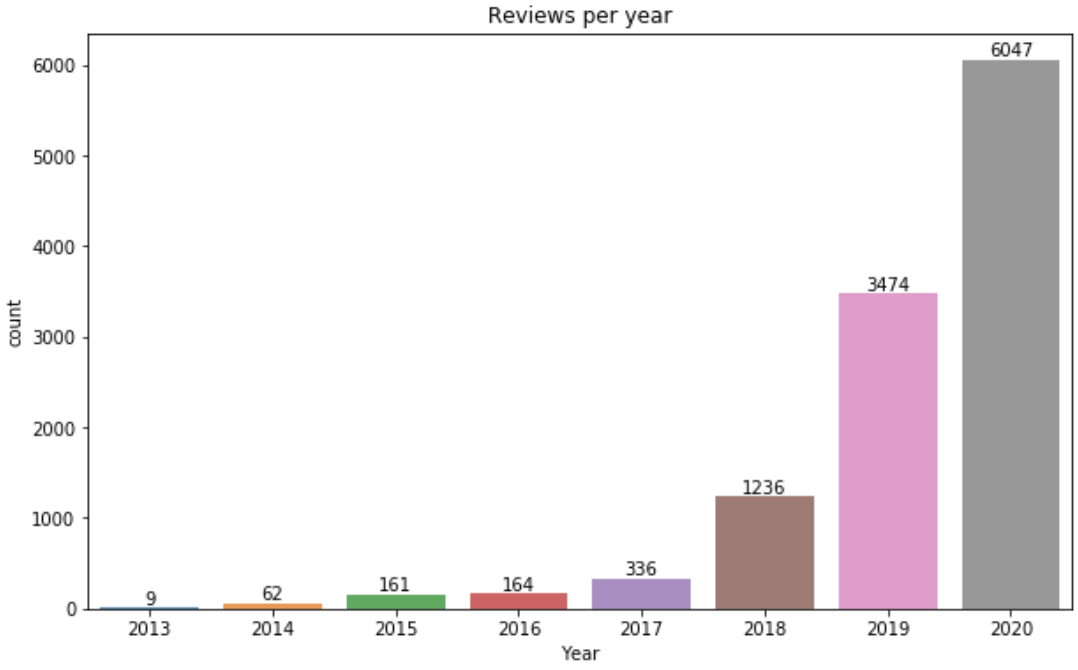


Fig 16. Distribution of Reviews per year

The graph shows the wordcloud of the dataset.

Fig 17. Wordcloud of the dataset

TF-IDF

Based on the TF-IDF ranking the 10 smallest and the largest feature names are

Smallest	Largest
Worry	Supper
Soc	Help
Bezel	Great
Asphalt	Grip
960fp	Grt
2x	Gud
Conclusion	Handwash
Models	Happy
Drops	Happi
Qualcomm	Hai

Table 18. Term Frequency and Inverse Document Frequency

The wordcloud for top 50 words

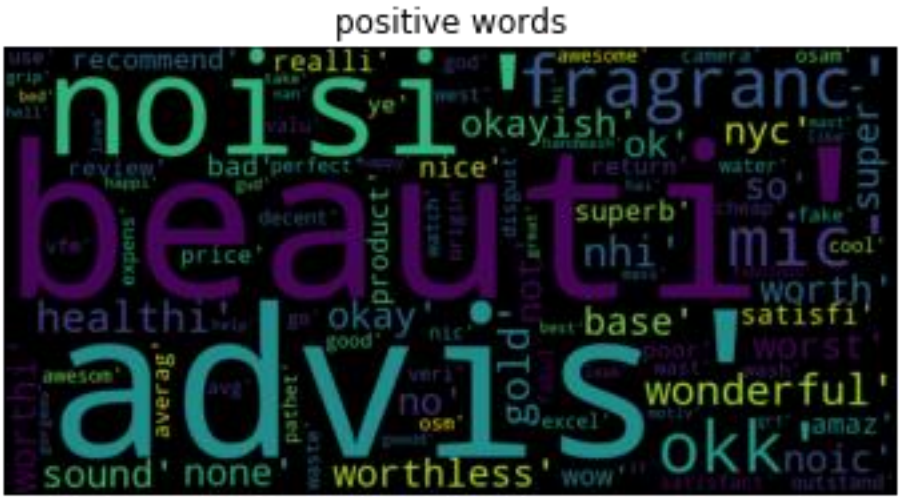


Fig 18. Wordcloud of the positive words

The wordcloud for the least 50 words

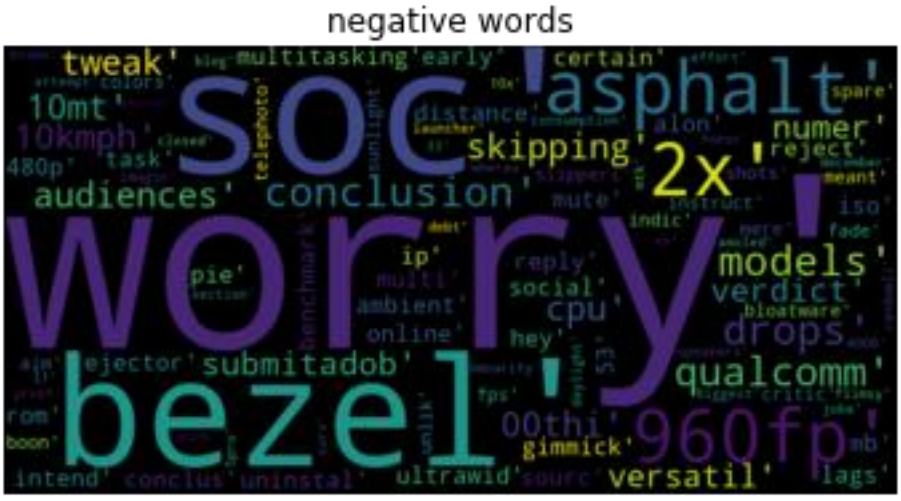


Fig 19. Wordcloud of the negative words

Summary

The table gives the Mean, Standard deviation, Maximum and Minimum values

	Ratings	Year	Age	Subjectivity	Count	Polarity	Product Score
Count	11489	11489	11489	11489	11489	11489	11489
Mean	4.05	2019	285.09	0.49	20.59	0.74	310.98
Std. Dev.	1.41	1.10	363.74	0.25	27.48	0.57	359.40
Min.	1.00	2013	2.00	0.00	0.00	-1.00	11.00
25th quantile	4.00	2019	71.00	0.38	5.00	1.00	100.57
50th quantile	5.00	2020	119.00	0.54	14.00	1.00	162.00
75th quantile	5.00	2020	376.00	0.60	26.00	1.00	394.00
Max	5.00	2020	2514.00	1.00	625.00	1.00	2543.19
Skewness	3.03						
Kurtosis	15.450						
Jarque-Bera	91872.77						
Total Reviews	11500						

Table 19. Summary of the dataset

The graph gives the count of Positive, Negative and Neutral reviews based on the products.

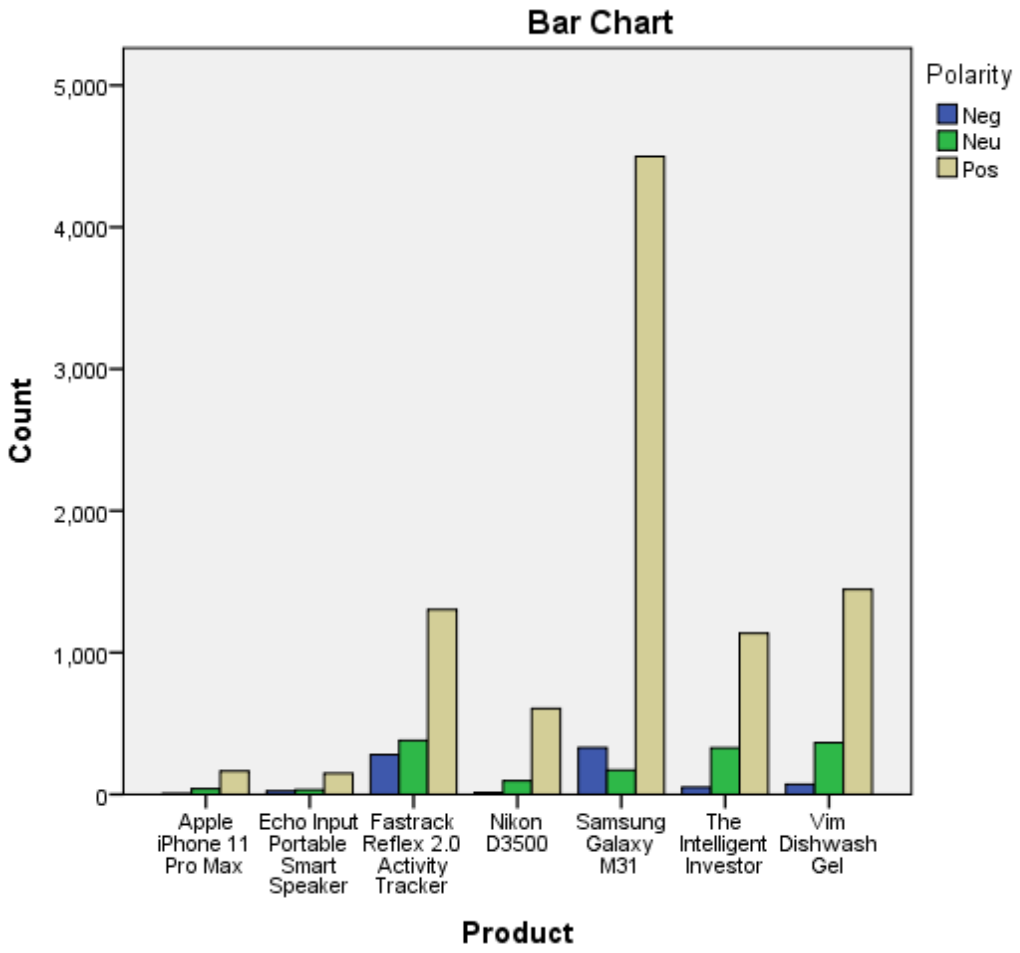


Fig 20. Bar chart for Products

Product & Polarity Crosstabulation

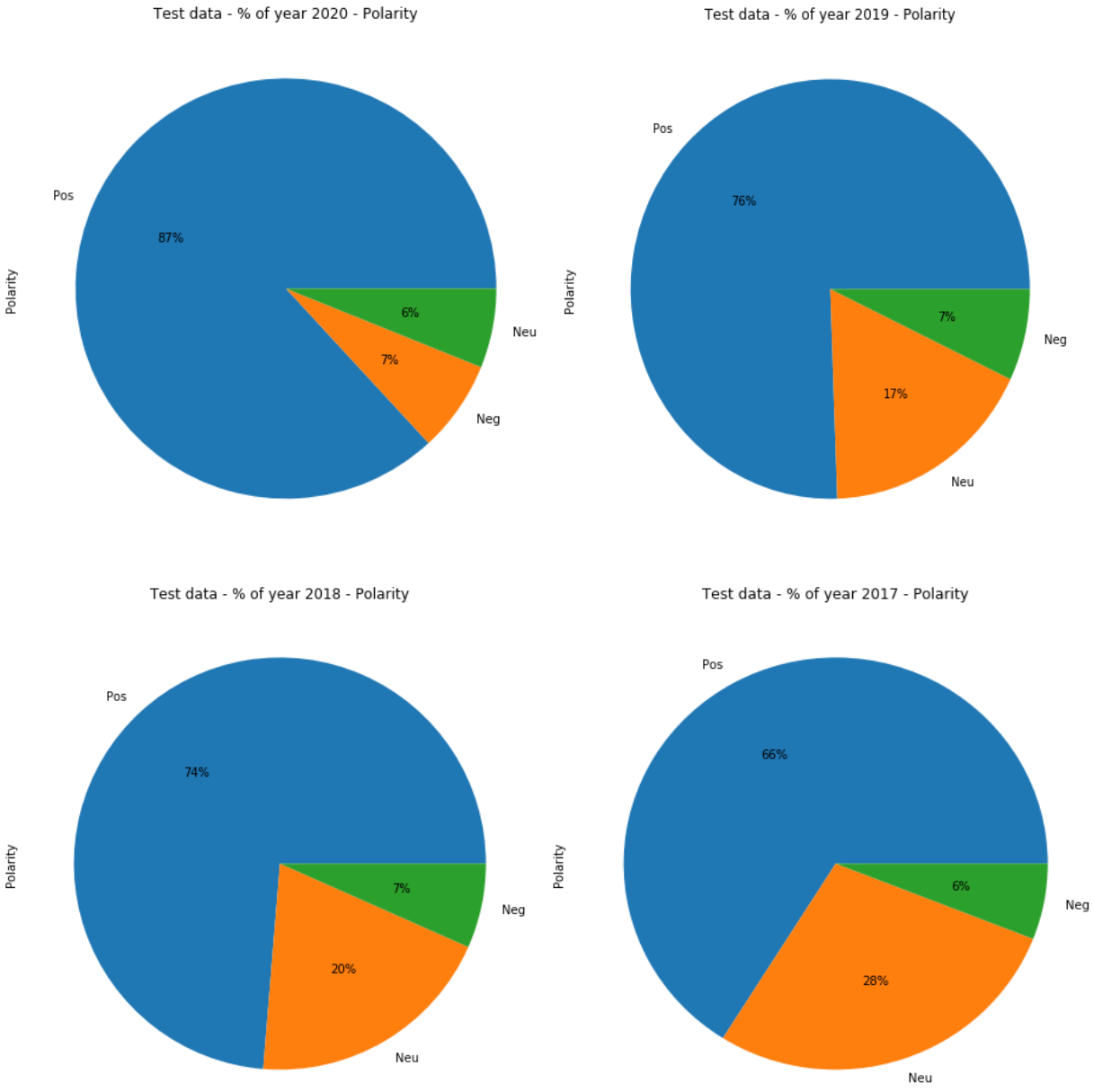
			Polarity			Total
			Neg	Neu	Pos	
Product	Apple iPhone 11 Pro Max	Count	7	39	165	211
		% within Product	3.3%	18.5%	78.2%	100.0%
		% within Polarity	0.9%	2.8%	1.8%	1.8%
		% of Total	0.1%	0.3%	1.4%	1.8%
	Echo Input Portable Smart Speaker	Count	23	33	148	204
		% within Product	11.3%	16.2%	72.5%	100.0%
		% within Polarity	3.0%	2.3%	1.6%	1.8%
		% of Total	0.2%	0.3%	1.3%	1.8%
	Fastrack Reflex 2.0 Activity Tracker	Count	281	381	1303	1965
		% within Product	14.3%	19.4%	66.3%	100.0%
		% within Polarity	36.2%	27.0%	14.0%	17.1%
		% of Total	2.4%	3.3%	11.3%	17.1%
	Nikon D3500	Count	13	95	605	713
		% within Product	1.8%	13.3%	84.9%	100.0%
		% within Polarity	1.7%	6.7%	6.5%	6.2%
		% of Total	0.1%	0.8%	5.3%	6.2%

	Samsung Galaxy M31	**Count**	330	172	4498	5000
		% within Product	6.6%	3.4%	90.0%	100.0%
		% within Polarity	42.5%	12.2%	48.4%	43.5%
		% of Total	2.9%	1.5%	39.2%	43.5%
	The Intelligent Investor	**Count**	50	327	1136	1513
		% within Product	3.3%	21.6%	75.1%	100.0%
		% within Polarity	6.4%	23.2%	12.2%	13.2%
		% of Total	0.4%	2.8%	9.9%	13.2%
	Vim Dishwash Gel	**Count**	72	365	1446	1883
		% within Product	3.8%	19.4%	76.8%	100.0%
		% within Polarity	9.3%	25.8%	15.5%	16.4%
		% of Total	0.6%	3.2%	12.6%	16.4%
Total		**Count**	776	1412	9301	11489
		% within Product	6.8%	12.3%	81.0%	100.0%
		% within Polarity	100.0%	100.0%	100.0%	100.0%
		% of Total	6.8%	12.3%	81.0%	100.0%

Table 20. Product & Polarity Cross tabulation

Year wise polarity

The below pie-charts shows the percentage of polarity of the reviews in their respective year.



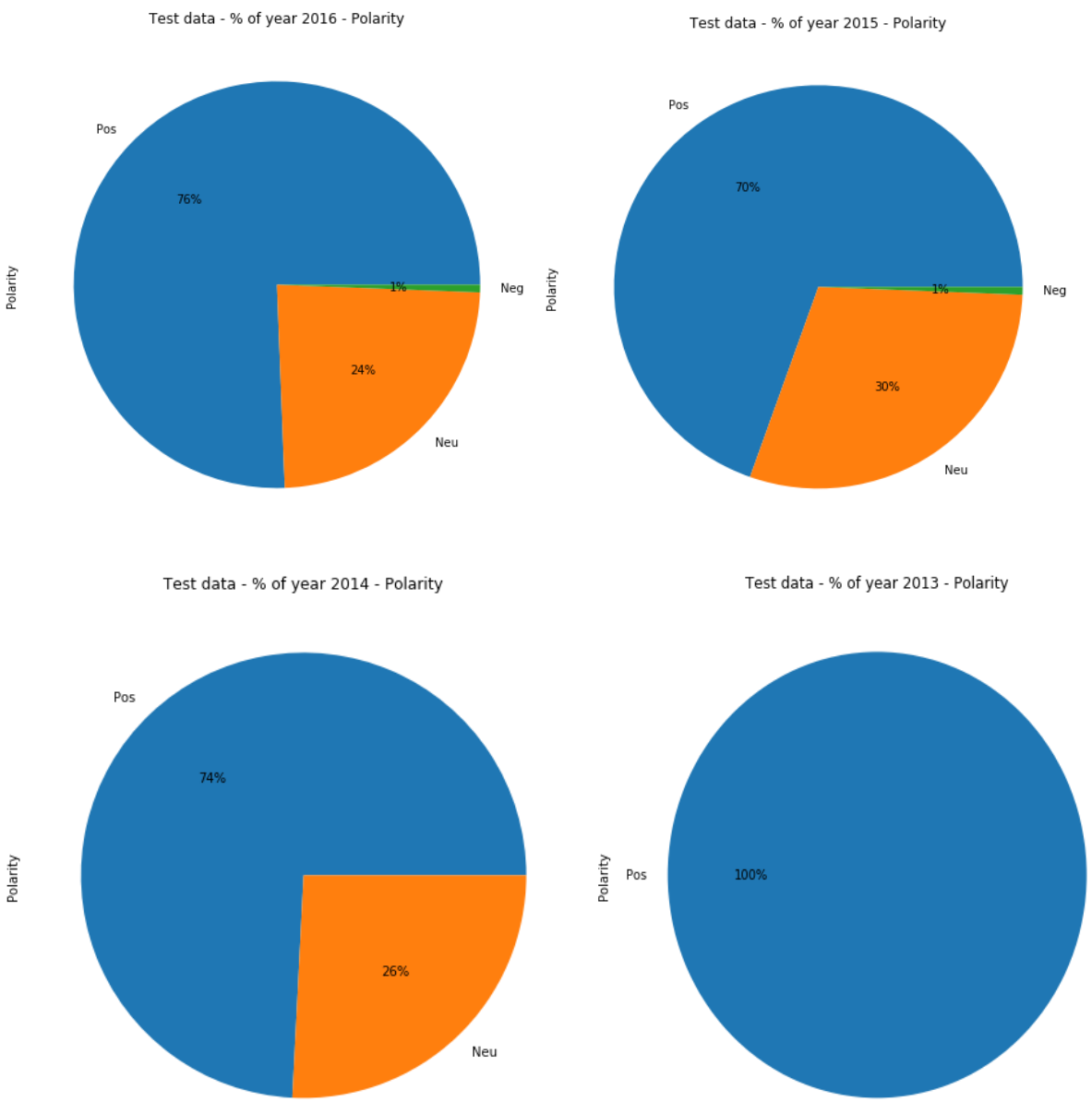


Fig 21. % of polarity year wise summary

Month wise summary

The March month has the highest percentage of reviews. The first four months (Jan, Feb, Mar and Apr) amounts for the 50 percentage of reviews.

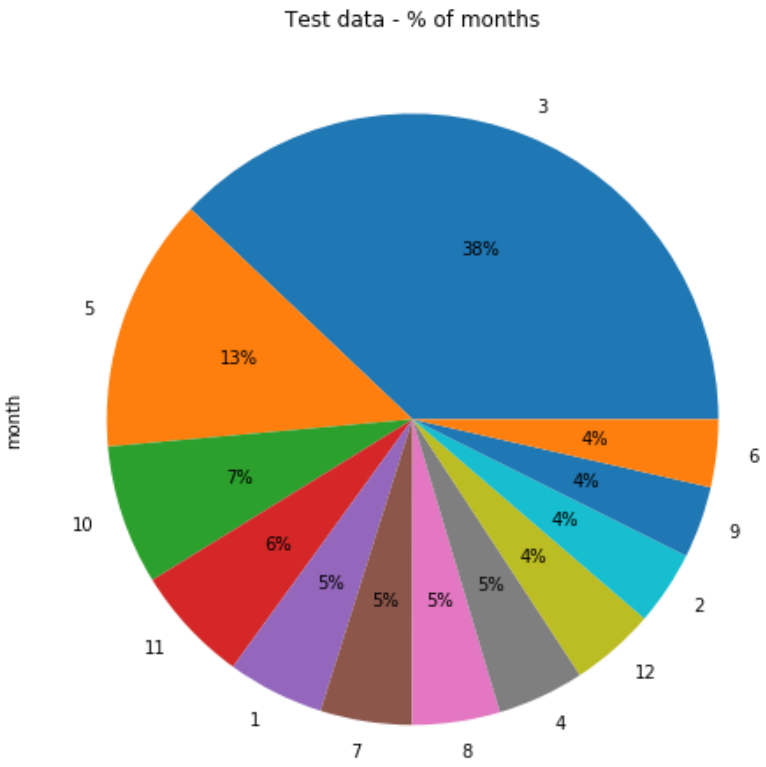


Fig 22. % of Month wise summary

2020

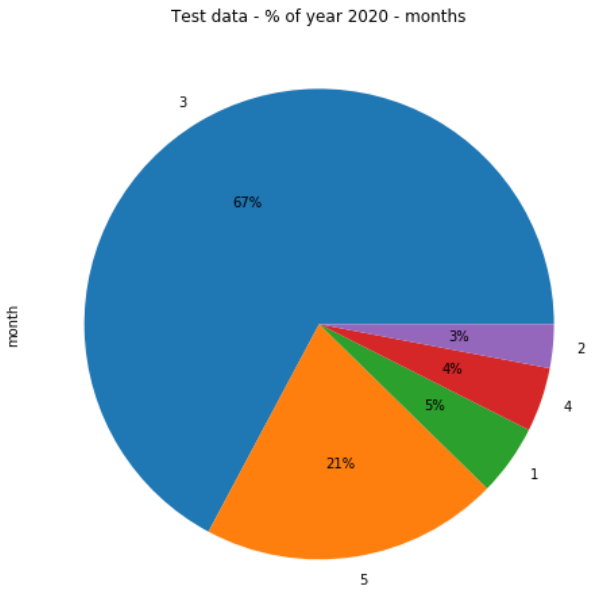


Fig 23. % of 2020's Month wise summary

2019

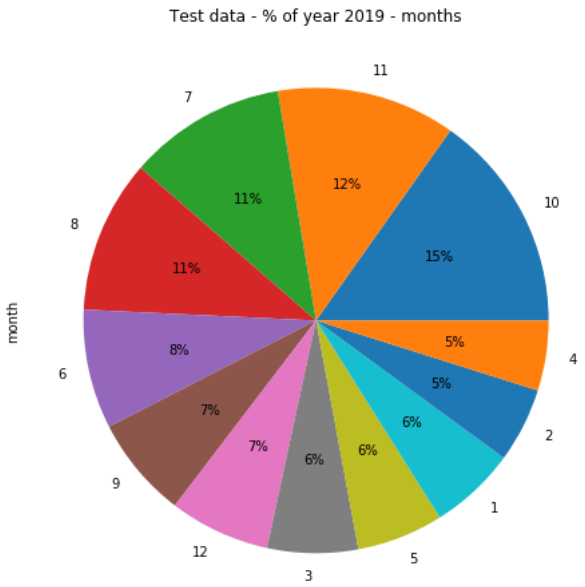


Fig 24. % of 2019's Month wise summary

2018

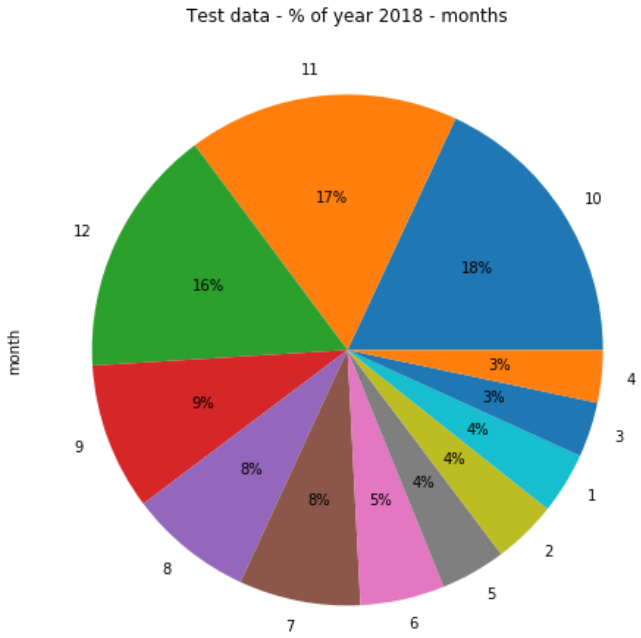


Fig 25. % of 2018's Month wise summary

4.1.9 Hypothesis testing

H1: The Customer Sentiment Polarity has a positive impact on the Customer Ratings.

H0: There is no relation between Customer Sentiment Polarity and Customer Ratings.

The hypothesis variables are categorical data. So the test for hypothesis is Chi-Square test. The test of normality confirms that the data is normal distributed or not.

Test of Normality:

Polarity		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	Df	Sig.	Statistic	df	Sig.
Ratings	NEG	.430	776	.000	.566	776	.000
	NEU	.284	1412	.000	.744	1412	.000
	POS	.372	9301	.000			

Table 21. Test of Normality

From the Shapiro-Wilk test, the dataset does not follow a normal distribution as the p-value (7.40959e-071) is less than α.

Rating score & Polarity Cross tabulation

			Polarity			Total
			Neg	Neu	Pos	
Rating score	1	Count	573	404	548	1525
		% within Rating score	37.6%	26.5%	35.9%	100.0%
		% within Polarity	73.8%	28.6%	5.9%	13.3%
		% of Total	5.0%	3.5%	4.8%	13.3%
	2	Count	99	69	268	436
		% within Rating score	22.7%	15.8%	61.5%	100.0%
		% within Polarity	12.8%	4.9%	2.9%	3.8%
		% of Total	0.9%	0.6%	2.3%	3.8%

	3	**Count**	55	113	572	740
		% within Rating score	7.4%	15.3%	77.3%	100.0%
		% within Polarity	7.1%	8.0%	6.1%	6.4%
		% of Total	0.5%	1.0%	5.0%	6.4%
	4	**Count**	26	168	1841	2035
		% within Rating score	1.3%	8.3%	90.5%	100.0%
		% within Polarity	3.4%	11.9%	19.8%	17.7%
		% of Total	0.2%	1.5%	16.0%	17.7%
	5	**Count**	23	658	6072	6753
		% within Rating score	0.3%	9.7%	89.9%	100.0%
		% within Polarity	3.0%	46.6%	65.3%	58.8%
		% of Total	0.2%	5.7%	52.9%	58.8%
Total		**Count**	776	1412	9301	11489
		% within Rating score	6.8%	12.3%	81.0%	100.0%
		% within Polarity	100.0%	100.0%	100.0%	100.0%
		% of Total	6.8%	12.3%	81.0%	100.0%

Table 22. Rating score and Polarity cross tabulation

Chi-Square test

	Value	**Df**	**Asymptotic significance(2-sided)**
Pearson Chi-Square	3626.186	8	.000
Likelihood ratio	2851.817	8	.000
N of valid cases	11489		

Table 23. Chi Square results

The *p*-value indicates that these variables are not independent of each other and that there is a statistically significant relationship between the categorical variables.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	**Phi**	.562	.000
	Cramer's V	.397	.000
N of Valid Cases		11489	

Table 24. Phi and Cramer's V values

The Phi and Cramer's V value gives the strength between the two variables. Both the customer sentiment polarity and customer ratings has a strong relation between them.

4.1.10 Relationship between Customer Ratings and Customer Sentiments

OLS

The linear regression gives the relationship between the Customer ratings and the Customer Sentiment Polarity.

The dependent variable is Customer ratings.

The independent variables is Customer sentiment Polarity and subjectivity of the reviews.

Ratings = 3.101 + 1.29 (Polarity).

The overall model is significant.

The Polarity is significant.

R – Square : 27.33%

VIF

Variables	VIF
Year	2393.414
Month	139.758
Days	48980.418
Sub	1.108
Count	526.109
PD score	90302.270
Polarity	2.305

Table 25. VIF for variables

There exists a multicollinearity between the variables Year, Month, Days, Count and PD score. As the product score is calculated based on Polarity, Age, Subjectivity, Count and ratings, we can say that the collinearity can be known easily without the VIF. The VIF proves that there exists an multicollinearity.

Because of the multicollinearity the model is changed to

Ratings = 3.101 + 1.29 (Polarity).

Correlation coefficients

	Category	Product	Ratings	Year	Month	Days	Subjectivity	Count	Polarity	Score
Category	1.0000	0.4673	-0.1684	0.1151	0.1631	-0.1694	0.0430	-0.1845	-0.1316	-0.1864
Product		1.0000	-0.1008	0.6530	-0.3532	-0.6309	0.0609	0.2606	0.0328	-0.6189
Ratings			1.0000	-0.0489	0.0050	0.0524	0.1156	-0.0086	0.5228	0.0572
Year				1.0000	-0.5064	-0.9730	0.0750	0.2342	0.0744	-0.9669
Month					1.0000	0.2949	-0.0116	-0.2430	-0.0746	0.2797
Days						1.0000	-0.0799	-0.1960	-0.0631	0.9972
Subjectivity							1.0000	-0.0056	0.2303	-0.0798
Count								1.0000	0.1141	-0.1217
Polarity									1.0000	-0.1526
Score										1.0000

Table 26. Correlation coefficients of variables

4.1.11 Recommendation of the product

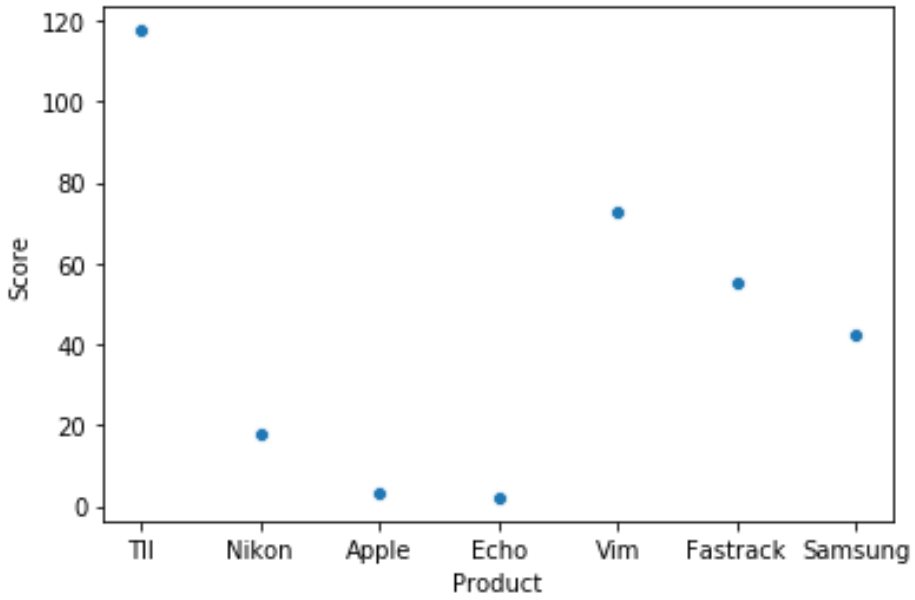


Fig 26. Product score for each products

Based on the product score of the products the most recommended products are The intelligent investor, Vim dishwash gel, Fastrack activity tracker. The least recommended products are Nikon D3500, Apple iphone 11 Max Pro, and Echo smart speaker.

Including the percentage of the product reviews in the dataset,

Products	Total reviews	Negative %	Neutral %	Positive %
The intelligent investor	1513	3.3%	21.6%	75.1%
Nikon D3500	713	1.8%	13.3%	84.9%
Apple iphone 11 Max Pro	211	3.3%	18.5%	78.2%
Fastrack activity tracker	1965	14.3%	19.4%	66.3%
Vim dishwash gel	1883	3.8%	19.4%	76.8%
Echo smart speaker	204	11.3%	16.2%	72.5%
Samsung Galaxy M31	5000	6.6%	3.4%	90%
Total	11489	6.8%	12.3%	81%

Table 27. Product wise Polarity of the reviews

Based on the table, the most recommended products are Samsung Galaxy M31 and Nikon D3500.

The least recommended product is Fastrack activity tracker.

4.2 Results

From the findings, the research question can be answered

Research Question 1: What are the impacts of customer sentiment on the customer satisfaction?

The impacts of the customer sentiment on the customer satisfaction is analyzed through our hypothesis and its testing. The hypothesis is based on the relationship between the customer sentiment polarity and customer ratings. The chi-square significance value proved our alternate hypothesis that there is an relationship between the variables and they are dependent of each other.

The phi and Cramer's v value shows the strength between the two variables are strong. The correlation coefficient shows that there is a positive correlation between the two. The OLS model is

Ratings = 3.101 + 1.29 (Polarity).

Hence the sentiment of the customers will have an impact on the customer satisfaction which will be measured through the customer ratings.

Research Question 2: Which classification algorithm will be best suited for the prediction of polarity of the reviews?

The ML classifier trains the model and classifies the evaluating data with its respective classifier. Based on the Precision, Recall, F1-score and Accuracy, the SVM classifier is the best classifier among NB, DT and K-NN. SVM is an powerful classifier for text classification.

Research Question 3: What is the opinion of the customers towards the products?

The opinion of the customers towards the products can be either satisfied or dissatisfied. The satisfaction of the customer can be understand through the sentiment analysis of the reviews. The polarity of the reviews are classified into Positive, Negative and Neutral. From the product cross tabulation table, we can see that the 81% of the dataset is positive. From the polarity of the sentiment based on the product, all the product has higher positive sentiment and low negative sentiment. The highest negative polarity among the product is 14.3%. The opinion of the customer towards the product can vary based on the product dataset.

Research Question 4: Which products can be recommended to the customers?

The product score is calculated based on the age, polarity, ratings and length of the review. Considering the product score and the product polarity percentage, the product can be recommended. The product with highest positive polarity and product score can be recommended.

The most recommended products are The intelligent investor, Vim dishwash gel, Fastrack activity tracker. The least recommended products are Nikon D3500, Apple iphone 11 Max Pro, and Echo smart speaker.

CHAPTER V: SUGGESTIONS AND CONCLUSION

5.1 Conclusion

The main objective of this study is to find the relation between the customer satisfaction towards the products in online platforms. The E-commerce platforms, digital population and mobile users growth has impacted the consumer buying behavior through online stores. The reviews are the only way to know whether a customer has satisfied by the product and service of the store. The review and ratings are essential in measuring the customer satisfaction. If the customer is satisfied, it leads to customer trust, loyalty, repurchase and etc.. The main objective of the study is proved by the hypothesis testing that there is an strong relationship between the customers sentiment and customers satisfaction. The sentiment which is measured by the polarity of the reviews have helped to understand the customer satisfaction through customer ratings.

The secondary objectives paves way for the primary objective. The sentiment analysis helps to find the polarity of the reviews and the best ML algorithm for the classification of polarity is evaluated based on the accuracy measures. The product score and the products polarity helps the customers on which product to prefer.

The managerial implications of the study are the understanding the customer needs without any efforts. The traditional approach includes survey and a long process to know the customer needs. It is less time consuming and easier to process large sets of data. It also benefits the customer decision making process.

5.2 Limitations and Scope for future research

The limitations in this study is the optimization of ML algorithms. The methodology does not focus on that because optimization of classifiers is not the objective of the study. This is one of the future scopes for the research, finding the optimized classifier for text classification. More number of product can be compared across different categories. The same product sentiment polarity analysis across different E-commerce sites can gives us some valuable insights about the sites and the products.

REFERENCES

[1] Drus, Z., & Khalid, H. (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. Procedia Computer Science, 161, 707-714.

[2] Endo, S., Yang, J., & Park, J. (2012). The investigation on dimensions of e-satisfaction for online shoes retailing. Journal of Retailing and Consumer Services, 19(4), 398-405.

[3] Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. Procedia Computer Science, 161, 765-772.

[4] Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. Tourism Management, 61, 43-54.

[5] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. Morgan Kaufmann.

[6] Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.

[7] Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. CIRP Journal of Manufacturing Science and Technology, 23, 128-144.

[8] Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. Computational Intelligence, 22(2), 100-109.

[9] Li, J., Lowe, D., Wayment, L., & Huang, Q. (2020). Text mining datasets of β-hydroxybutyrate (BHB) supplement products' consumer online reviews. Data in Brief, 105385.

[10] Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. Information & Management, 56(2), 172-184.

[11] Liu, Z., Lv, X., Liu, K., & Shi, S. (2010, March). Study on SVM compared with the other text classification methods. In 2010 Second international workshop on education technology and computer science (Vol. 1, pp. 219-222). IEEE.

[12] Mansour, S. (2018). Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. Procedia Computer Science, 140, 95-103.

[13] Mirtalaie, M. A., Hussain, O. K., Chang, E., & Hussain, F. K. (2018). Extracting sentiment knowledge from pros/cons product reviews: Discovering features along with the polarity strength of their associated opinions. Expert Systems with Applications, 114, 267-288.

[14] Mulajati, M., & Hakim, R. F. (2017). Sentiment Analysis on online reviews using Naïve Bayes Classifier Method and Text Association (Case Study: Garuda Indonesia Airlines Passengers Reviews on TripAdvisor Site). Indian Journal of Scientific Research, 274-281.

[15] Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative study of machine learning approaches for Amazon reviews. Procedia computer science, 132, 1552-1561.

[16] Rajeev, P. V., & Rekha, V. S. (2015, March). Recommending products to customers using opinion mining of online product reviews and features. In 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015] (pp. 1-5). IEEE.

[17] Rita, P., Oliveira, T., & Farisa, A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. Heliyon, 5(10), e02690.

[18] Sun, Q., Niu, J., Yao, Z., & Yan, H. (2019). Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level. Engineering Applications of Artificial Intelligence, 81, 68-78.

[19] Tsao, W. C., & Hsieh, M. T. (2012). Exploring how relationship quality influences positive eWOM: the importance of customer commitment. Total Quality Management & Business Excellence, 23(7-8), 821-835.

[20] Vechtomova, O. (2009). Introduction to Information Retrieval Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (Stanford University, Yahoo! Research, and University of Stuttgart) Cambridge: Cambridge University Press, 2008, xxi+ 482 pp; hardbound, ISBN 978-0-521-86571-5.

[21] Xiong, S., Wang, K., Ji, D., & Wang, B. (2018). A short text sentiment-topic model for product reviews. Neurocomputing, 297, 94-102.

[22] Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, 111-121.

APPENDICES

APPENDIX I

Web Scraper

Web Scraper is an extension in google chrome which is used to extract the data from the sites. It is used for extracting the product review info from the amazon website.

Link: https://webscraper.io/

APPENDIX II

SPSS

SPSS is short for Statistical Package for the Social Sciences, and it's used for complex statistical data analysis.

The version used for the study is IBM SPSS Statistics 23.

APPENDIX III

Anaconda – Jupyter notebook

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

The version of Anaconda used for the study is Anaconda Navigator (Anaconda3)

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

The Jupyter Notebook uses the Anaconda environment.

Link: https://jupyter.org/

Link: https://www.anaconda.com/

APPENDIX IV

Python Packages

The python packages used for the data cleaning, transformation, statistical modelling, data visualization, machine learning and much more are Pandas, NumPy, matplotlib, seaborn, nltk and sklearn

APPENDIX IV

Tableau

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format.

The version used for visualization is Tableau Public

The link contains the dashboard visualization of the dataset.

Link:https://public.tableau.com/views/SIP_15965675852920/Story1?:language=en-GB&:display_count=y&:origin=viz_share_link

TSM	FEEDBACK FORM	FOR/QSP/IT
		Rev. No / Date :01/09.09.2016
		Page 1 of 1
		Approved By: DIRECTOR

Name of the Student : KARTHICK RAJ S

Title of Summer Project : A TEXT MINING APPROACH TO ANALYZE THE CUSTOMER's SATISFACTION FROM THE ONLINE PRODUCT REVIEWS (OPRs)

Name of Faculty Guide : V. SENTHIL

1-Poor 2-Fair 3-Good 4-Very Good 5-Excellent

	1	2	3	4	5
1. Motivation to learn	☐	☐	☐	☑	☐
2. Domain Knowledge	☐	☐	☐	☑	☐
3. Sense of responsibility	☐	☐	☐	☑	☐
4. Analytical ability	☐	☐	☐	☑	☐
5. Ability to take initiative	☐	☐	☑	☐	☐
6. Communication skills	☐	☐	☑	☐	☐
7. Creativity	☐	☐	☐	☑	☐
8. Effective utilisation of time	☐	☐	☐	☑	☐
9. Interpersonal relations	☐	☐	☑	☐	☐
10. Ability to finish the work before deadline	☐	☐	☐	☑	☐

Would you like to nominate the student for summer project competition ☐ Yes ☑ No

ONLINE COURSE CERTIFICATES

Stanford | ONLINE

COURSE CERTIFICATE

06/25/2020

Karthick Raj

has successfully completed

Machine Learning

an online non-credit course authorized by Stanford University and offered through Coursera

EDUCATION FOR EVERYONE · coursera · COURSE CERTIFICATE

Associate Professor Andrew Ng
Computer Science Department
Stanford University

SOME ONLINE COURSES MAY DRAW ON MATERIAL FROM COURSES TAUGHT ON-CAMPUS BUT THEY ARE NOT EQUIVALENT TO ON-CAMPUS COURSES. THIS STATEMENT DOES NOT AFFIRM THAT THIS PARTICIPANT WAS ENROLLED AS A STUDENT AT STANFORD UNIVERSITY IN ANY WAY. IT DOES NOT CONFER A STANFORD UNIVERSITY GRADE, COURSE CREDIT OR DEGREE, AND IT DOES NOT VERIFY THE IDENTITY OF THE PARTICIPANT.

Verify at coursera.org/verify/GV3WKQGGQ9Y9

Coursera has confirmed the identity of this individual and their participation in the course.

ILLINOIS

COURSE CERTIFICATE

05/18/2020

Karthick Raj

has successfully completed

Text Mining and Analytics

an online non-credit course authorized by University of Illinois at Urbana-Champaign and offered through Coursera

ChengXiang Zhai
Professor and Willett Faculty Scholar
Department of Computer Science

Verify at coursera.org/verify/KTGQEJKFCBD9

Coursera has confirmed the identity of this individual and their participation in the course.

ORIGINALITY REPORT

chapters

ORIGINALITY REPORT

17%	10%	12%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Monireh Alsadat Mirtalaie, Omar Khadeer Hussain, Elizabeth Chang, Farookh Khadeer Hussain. "Extracting sentiment knowledge from pros/cons product reviews: Discovering features along with the polarity strength of their associated opinions", Expert Systems with Applications, 2018 Publication	1%
2	Saurabh Tewari, U. D. Dwivedi. "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies", Journal of Petroleum Exploration and Production Technology, 2020 Publication	1%
3	Doaa Mohey El-Din Mohamed Hussein. "A survey on sentiment analysis challenges", Journal of King Saud University - Engineering Sciences, 2016 Publication	1%
4	www.ijariit.com Internet Source	1%