**PROJECT ON RECOMMENDATION OF DIFFERENT PRODUCTS BASED ON REVIEWS**

Submitted towards partial fulfilment of the criteria

for award of PGPBABI by Great Lakes Institute of Management

Submitted By

Group No. 1 [Batch: 2019]

Group Members

1. Abhilash R
2. Angela Susan Mathews
3. Ashwin. S
4. Deepthi H
5. Sajin Madhavan S
6. Sneha Devadass

Research Supervisor

Ms Anjana Agrawal



**Great Lakes Institute of Management**

# 

# 

# 

# 

# **Abstract**

The customer's review plays an important role in deciding the purchasing behavior of both online shopping as well offline, as a customer prefers to get the opinion of other customers by observing their opinion through online products’ reviews, blogs and social networking sites, etc. A subjective conclusion from the user's review could help the customers in deciding whether to purchase that product of particular brand or leave it. Sentimental analysis of the reviews of the products were performed and classified the reviews as "Positive “or "Negative" using various classifier algorithms. This project aims at building a model that classifies various products as "good" or "bad" based on the reviews available from various platforms. Also this paper describes the techniques used for data collection, sentimental analysis performed and various modelling methods that have been used.

|  |  |
| --- | --- |
| Nature | Report |
| Version | 1.0 |
| Keyword List | Sentiment Analysis, Naive Bayes, Text Mining,  Polarity, Random Forest, R, Tableau, Data Scraping, Product Reviews, Ecommerce,social media reviews,  [Classification algorithms](https://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:Classification%20algorithms&newsearch=true),Text Analytics,Python |

# 

# **Table of Contents**

[**Abstract**](#_3s9modjsr75t) **3**

[**Table of Contents**](#_380znri7vw6l) **4**

[**List of Tables and Figures**](#_l94z2ljfd3zj) **6**

[**Abbreviations**](#_5jqnh72fjocn) **6**

[**Chapter 1**](#_sqz4awh2czru) **7**

[**1.1 Introduction**](#_qjek82niksc6) **7**

[**Chapter 2**](#_puw6rsv6prb) **8**

[**2.1 Literature Review**](#_upaqqywzfik9) **8**

[2.1.1 Techniques used in Text Mining](#_5j2idjrj5qlx) 8

[2.1.2 Sentiment analysis](#_i3gi0ce9ceo9) 9

[**CHAPTER 3**](#_kvjhi27uxzpt) **10**

[3.1 DATA COLLECTION](#_rv3wspwudki9) 10

[3.1.1 Purpose](#_tsjlg0fhrk16) 10

[3.1.2 Sources Of Data](#_2g7u8l4h993v) 10

[3.2 Techniques used to scrape data](#_a6556nhcdhi1) 10

[3.2.1 .Web Scraper browser extension (mouthshut.com, official websites,gadgets.ndtv.com)](#_nsqc2vx4wa6g) 10

[3.2.2 API (For Twitter)](#_hulmy01ofbsm) 11

[3.2.3 Inspect elements —> curl command —> python script (to scrape facebook data)](#_tts98v73vapu) 11

[Methods used to overcome limitations :](#_67d79dxj50c) 11

[3.3 Data Cleaning](#_gwsvaf2hsbpy) 11

[3.3.1 Cleaning Data using Excel](#_c3pxmhohae0m) 11

[3.3.2 Data Conditioning using Text Mining Package:](#_x7puhr9mv580) 12

[Chapter 4](#_uvupaexfism6) **12**

[4.1 Exploratory Data Analysis](#_obu7m4mt8mpq) 13

[4.1.1 Summary of dataset](#_n3lkavfksfbk) 13

[4.2 Feature extraction & analysis](#_bm5xjqxe48me) 13

[4.2.1 Most Frequently Occurring Words](#_yzp6agky7ov) 14

[4.2.2 Top 20 frequently occurring unigrams extracted](#_3yorik550i3d) 14

[4.2.3 Top 20 frequently occurring sequence of bigrams](#_8ahoiz6h22np) 15

[Interesting Inferences](#_d56ruhmfnbjb) 15

[**Chapter 5**](#_ep0x9h9vg5q4) **15**

[5.1 Sentiment Analysis](#_3b40ggz7tfjk) 15

[5..1.1 Summary of Sentiment Scores obtained](#_vlbrs6mbp7kl) 16

[5.1.2 Histogram - Visualize the text polarity sentiment scores](#_sb142ujmk9zj) 16

[5.1.3 Visualize the reviews and its polarity with respect to categories and brand](#_n03vzwppp66r) 17

[**Chapter 6**](#_uelnvt567xk4) **18**

[**Modelling**](#_jsdjq1b44wpx) **18**

[6.1 Model 1: Naive Bayes](#_b5xfqfmjy5gx) 18

[6.1.1 Implementation of Algorithm](#_d04fo7ycf497) 18

[6.2 Model 2 : Logistic Regression](#_j4i37ym6mxvp) 19

[6.2.1 Implementation of Algorithm](#_bj46mosl0ood) 19

[6.2.2 Simple N fold logistic regression model](#_5jriw69fyq8j) 20

[6.2.3 Model with Pruned Vocabulary](#_hkn350flt7le) 21

[6.2.4 Model using N-grams](#_eev0la65zh0q) 22

[6.2.6 Accuracy Metrics of 4 Models](#_1bnncj122r4y) 24

[6.2.7 Inference](#_dswoc3mww9lz) 24

[**Chapter 7**](#_uj6f86h53q2k) **24**

[7.1 Further Refinement](#_11hgiepc3guz) 24

[**Chapter 8**](#_o75w4wxx4f19) **25**

[8.1 How to use these predicted values(1,0) in classifying a product as good or bad](#_qkb5s8yccilv) 25

[**Chapter 9**](#_ao4p507e4qhp) **26**

[9.1 Modelling Techniques Performed And Its Accuracy Metrics](#_lk7fu6xg7uy9) 26

[**Conclusion and Recommendation**](#_3lzf1djx8vg9) **26**

[**Annexure**](#_5vii5lpwqpo4) **28**

# **List of Tables and Figures**

|  |  |  |
| --- | --- | --- |
| Table 8.1A | Modelling Techniques Performed and Its Accuracy Metrics | 27 |
| Figure 3.2.1A | Web Scraper browser extension | 11 |
| Figure 4.1.1A | Summary of dataset | 13 |
| Figure 4.2.1A | Most Frequently Occurring Words | 14 |
| Figure 4.2.2A | Top 20 frequently occurring unigrams extracted | 15 |
| Figure 4.2.3A | Top 20 frequently occurring sequence of bigrams | 16 |
| Figure 5.1.1A | Summary of Sentiment Scores obtained | 17 |
| Figure 5.1.2A | Histogram - Visualize the text polarity sentiment scores | 17 |
| Figure 5.1.3A | Visualize the reviews and its polarity with respect to categories and brand | 18 |

# **Abbreviations**

|  |  |
| --- | --- |
| TM | Text Mining |
| NLP | Natural Language Processing |
| API | Application Programming Interfaces |
| DTM | Document-Term Matrix |
| AUC | Area Under Curve |

# **Chapter 1**

# **1.1 Introduction**

For a business to succeed, customer experience and satisfaction is a key factor. Internet penetration among the youth, internet penetration in the urban and rural geography, customers can interact with a brand and express their sentiments towards a brand/business. Multitudes of reviews are available on various platforms on the internet which expresses their sentiments and emotions for a business/brand. If the businesses can analyze these reviews and categorize them into negative and positive reviews, they will be able to provide a better customer experience.

New purchases by existing or new customers are influenced by online reviews. So it is important for a brand or a business to classify the review as positive or negative ones.

The project aims to perform a behavioral analysis of web and social media-based reviews of products. Reviews for various products (brand and category wise) were extracted from different social media platforms like Twitter, YouTube, Mouthshut, Croma, etc. were extracted using web scraping techniques. Brand, Category of the products, Dates on which the reviews were written and the reviews for the products were the variables that were extracted in the dataset. A prediction algorithm for the dataset using Natural Language Processing was modelled. The objective of the project was to classify the products as “recommended” or “not recommended” based on their reviews. This was achieved by doing a sentimental analysis of the reviews collected and classifying the reviews as “Positive”, “Negative” or “Neutral”. If the proportion of the “Positive” and “Neutral” reviews were more than “Negative”, we classify the product as “recommended”.

The classification steps involved statistical models like Naïve Bayes, Logistic Regression, Support Vector Machines, Random Forest.

* Naïve Bayes is a family or probabilistic algorithm that uses Bayes’ theorem to predict the category of a text under the assumption that each pair of features being classified is independent of each other.
* Logistic Regression: In natural language processing, logistic regression is the baseline supervised machine learning algorithm for classification, and also has a very close relationship with neural networks. Logistic regression has a number of advantages over Naive Bayes. Naive Bayes has overly strong conditional independence assumptions. Logistic regression is much robust to correlated features. Thus, when there are many correlated features, logistic regression will assign a more accurate probability than Naive Bayes.
* Support Vector Machines is a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they’re mapped to.
* Random Forest classifier is an ensemble algorithm (i.e., it uses more than one algorithm of same or different kind for classifying objects). Random forest classifier creates a set of decision trees from a randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

# **Chapter 2**

# **2.1 Literature Review**

With the advancement of technology, more and more data is available in digital form. But most of the data is in an unstructured textual form. Text mining is a process of analyzing text to extract information that is useful for a specific purpose. Text mining identifies facts, relationships and assertions which remain hidden in the huge amount of text data that is available. The information that is extracted is converted into a structured form that can be further analyzed. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP). The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.

There are a series of activities to be performed to mine the information that we require in the text mining process.

**1.Pre-preprocessing:** It involves a series of steps as below.

* Text collecting: This can be done from any websites, database or can be from survey that we give out
* Text Clean up: removing any unnecessary or unwanted information
* Text Transformation: A text document is represented by the words it contains and their occurrences. Bag of words and Vector space are the two main approaches to document representation.

Domain knowledge should be applied while doing text clean up and transformation.

**2. Text mining operations:** On the processed data, various text mining operations can be done. Classification techniques, clustering, topic modelling, similarity and distance computing and transforming the data into more reduced data can be done.

**3. Post-preprocessing:** Once the text mining operations are done, domain knowledge can be applied for pattern interpretations and evaluation which can be then validated using experts and data triangulation.

## **2.1.1 Techniques used in Text Mining**

The five basic techniques used in text mining is discussed below:

**1. Information extraction:** This technique is used to analyze the unstructured text by finding out the important words and finding the relationships between them using pattern matching. This gives a structure to the unstructured data.

**2. Categorization**: The text document is classified into one or more categories. This process includes pre-processing, indexing, dimensional reduction and classification. For categorization of text data, techniques like Naïve Bayesian classifier, Decision tree, Nearest Neighbor classifier and Support Vendor Machines can be used.

**3. Clustering:** This is used to group text documents which have similar contents. There are different partitions called clusters and each cluster will have several documents with similar content. A commonly used clustering technique is K-means, which compares each cluster and finds how well the data is connected to each other. This technique can be used to create a database with thousands of similar documents.

**4. Visualization:** To simplify the process of finding relevant information, this technique is used. Visualization techniques help to display textual information in a more attractive way.

**5. Summarization:** This is used to reduce the length of the document and summarize the details of the documents in brief. It summarizes large text documents easily and quickly.

Text mining helps to track opinions over time, summarize a document, to extract concepts from the documents and present in a simpler way.

Many predictive modelling methods can be done on the data that is obtained from text mining. This enables better and smart decision making, solving knowledge discovery problems in different areas of business. Text mining helps to extract patterns from large amounts of unstructured data.

## **2.1.2 Sentiment analysis**

It is an automated process capable of understanding the feelings or opinions that underlie a text. It is one of the most interesting subfields of NLP( Natural language processing), a branch of Artificial Intelligence (AI) that focuses on how machines process human language and studies the subjective information in an expression, that is, the opinions, appraisals, emotions, or attitudes towards a topic, person or entity. Expressions can be classified as positive, negative, or neutral. Sentiment analysis systems can extract other relevant attributes like the subject (the topic, entity, person or event that the opinion is referring to) and the opinion holder (the person who is expressing the opinion). Sentiment analysis is a powerful tool that can have a great impact in many business areas, like social media management, marketing, product, and customer support.

Some of the examples of sentimental analysis are:

* Social media monitoring
* Brand monitoring
* Customer support
* Customer feedback
* Market research

As mentioned earlier most of the data available to us is in unstructured form. Sentimental analysis can help make sense of the unstructured data by automatically tagging it. Huge amounts of data can be processed in an efficient and cost-effective way.

Sentimental analysis is modelled as a classifier technique where for a text, a category, positive, negative or neutral is assigned. In a training process model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training. The feature extractor transfers the text input into a feature vector. Pairs of feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning algorithm to generate a model. In the prediction process, the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, positive, negative, or neutral).

# **CHAPTER 3**

## **3.1 DATA COLLECTION**

Data scraping is commonly manifest in Web scraping (or screen scraping) is a way to get data from any website. With so much information now online, getting that data can often prove the difference between success and stagnation

### **3.1.1 Purpose**

Collected product review comments from social media’s and official websites to perform behavioral analysis on the products reviews

### **3.1.2 Sources of Data**

1. Twitter
2. Facebook
3. Mouthshut.com
4. Official websites of the respective product and brand
5. [gadgets.ndtv.com](https://gadgets.ndtv.com/)

## **3.2 Techniques used to scrape data**

## **3.2.1. Web Scraper browser extension (mouthshut.com, official websites,gadgets.ndtv.com)**

By using a web scraping tool, sometimes called a website scraper, you’re able to extract lots of data through an automated process. The tool works by sending a query to the requested pages, then combing through the HTML for specific items.

1. Downloaded and installed the Google’s web scraper browser extension.
2. Create sitemaps

* Specify the URL of a product for which we will be scraping the data.
* Created the sitemap where we have specified the details of variables like review comments and date. Also, we have taken care of pagination.

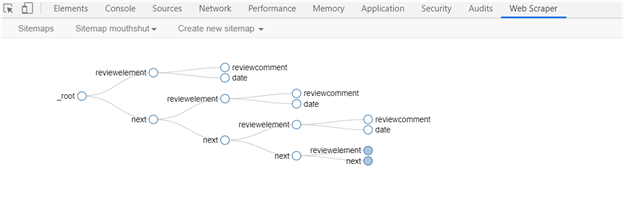


Figure 3.2.1A

### **3.2.2 API (For Twitter)**

Used R language and twitter API to scrape data from twitter on products.

### **3.2.3 Inspect elements —> curl command —> python script (to scrape Facebook data)**

**Limitations:**

● Multiple requests from the same IP.

● Web scraping bots fetch data very fast which makes websites become unresponsive.

#### **Methods used to overcome limitations:**

● Rotating IPs using VPN.

● Added sleep calls and some delays after crawling

## **3.3 Data Cleaning**

### **3.3.1 Cleaning Data using Excel**

**Step 1: Removing null rows**

Skimming through the dataset, though filter, a lot of null values (in the reviews column) were found and removed from that particular row.

**Step 2: Formatting “dateofreview” column to one single format**

The dates extracted from different websites were of different formats, so the format has been made unique to all the rows

Due to lots of null values, dateofreview column has been dropped from the dataset.

**Step 3: Clustering the products into 3 common “categories”**

For each brand we have decided to collect different products and its reviews.

For example: For the brand “samsung”, we have collected the reviews for mobiles, televisions, washing machines. Here we have grouped mobile phones into “mobiles”, “televisions, washing machines” into “Home Appliances” and “fitband, watches” into “electronic gadgets” product categories.

So we have clustered various products into 3 categories which are “Mobiles”,”Home Appliances”, “Electronic Gadgets”.

After performing these basic cleaning of data in Excel, the dataset was loaded into R to perform further cleaning of the data and analysis

### **3.3.2 Data Conditioning using Text Mining Package:**

Text mining framework provided by the tm package contains methods for data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices. Few mining techniques used on dataset are,

Convert dataset to Corpus

Convert document to lowercase

Remove Punctuations (comma, hyphen, period):

Remove Stopwords which are extremely common words like “and”, “or”, “not”, “in”,etc.

Remove Numbers

Remove Extra whitespaces

Stem the document where word is replaced with its most basic conjugate form.

# **Chapter 4**

## **4.1 Exploratory Data Analysis**

The major tasks of this analysis are: 1. create a basic script of summary statistics about the data. 2. use 1-grams, 2-grams, and 3-grams to understand frequencies of words and word pairs. 3. discuss plans for creating a prediction algorithm.

### **4.1.1 Summary of dataset**

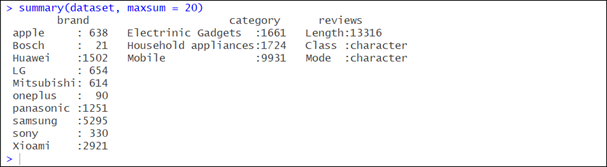


Figure 4.1.1A

**Brand wise inference -** There is no sufficient data for “Bosch” and “oneplus” to analyse, and build a model.

**Category wise inference -** Data is biased as there are more data for mobiles but not to other categories.

**Action**

We have removed those brands from the dataset to start performing sentiment analysis and apply modelling techniques on the updated dataset.

When coming to categories, collecting data for other categories would be the best solution but as there are no dependencies between the categories, having a good number of data in each category would make the model better is what the call taken.

## **4.2 Feature extraction & analysis**

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing.

### **4.2.1 Most Frequently Occurring Words**

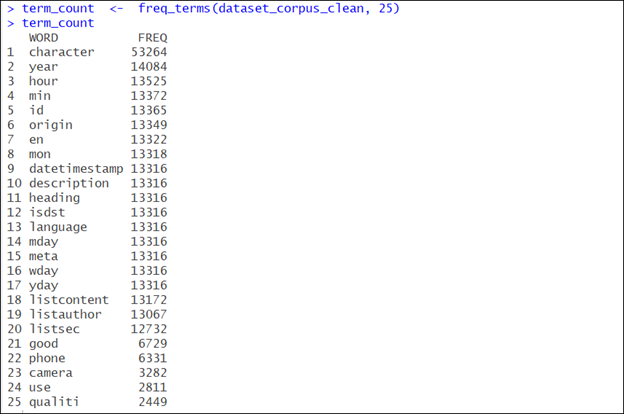
****

Figure 4.2.1A

### **4.2.2 Top 20 frequently occurring unigrams extracted**

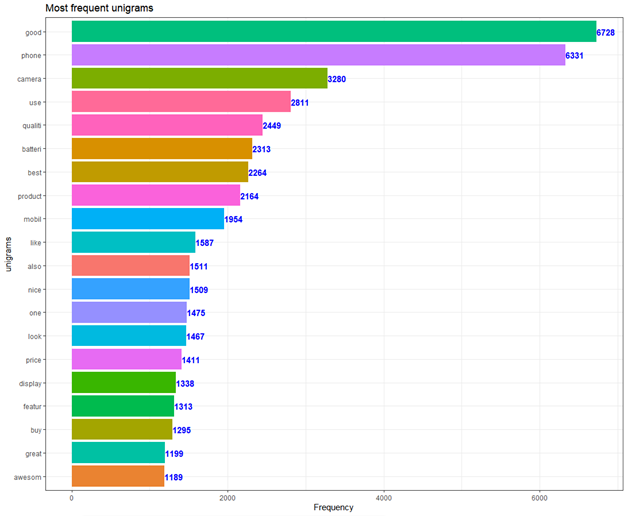
****

Figure 4.2.2A

### **4.2.3 Top 20 frequently occurring sequence of bigrams**

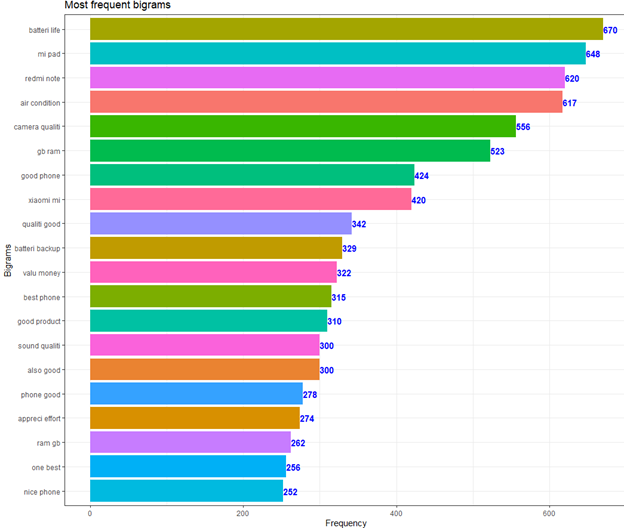
****

Figure 4.2.3A

### **Interesting Inferences**

* First, in the cleaning process, we have to remove the stopwords such as “the”, “and”,“are”, etc. These kinds of words don’t provide useful information, but they occur very frequently. Initially, we did not remove them
* Second, we realize that it is very important to look at the data with two-word phrases or three-word phrases.This is also very important for prediction of negative words or to capture negative reviews.
* Finally analysis on the extracted feature helped to calculate term frequencies and review the top tokens and also to examine and compare the positive and negative comments from the reviews

# **Chapter 5**

## **5.1 Sentiment Analysis**

The reviews are bought in to a clean format to perform sentiment analysis.

In order to perform sentiment analysis, and extract the sentiment scores for each reviews from the dataset, “Sentimentr” package is installed.

Sentiment() function is applied on the character vector of reviews, which in turn returns the text polarity sentiment in a sentence level**.**

### **5..1.1 Summary of Sentiment Scores obtained**

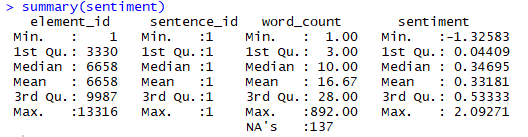
****

Figure 5.1.1A

The sentiment scores are now grouped into 3 categories - Positive, Negative and Neutral comments for better analysis. The categorization parameter is taken based on the mean of the sentiment scores (0.34).Hence, Positive group includes all values greater than 0.34,Negative group consists of all values less than 0, and Neutral group consists of the values under the cap of 0 and 0.34.The sentiment scores extracted are now merged to the final dataset data frame to proceed with further analysis**.**

### **5.1.2 Histogram - Visualize the text polarity sentiment scores**

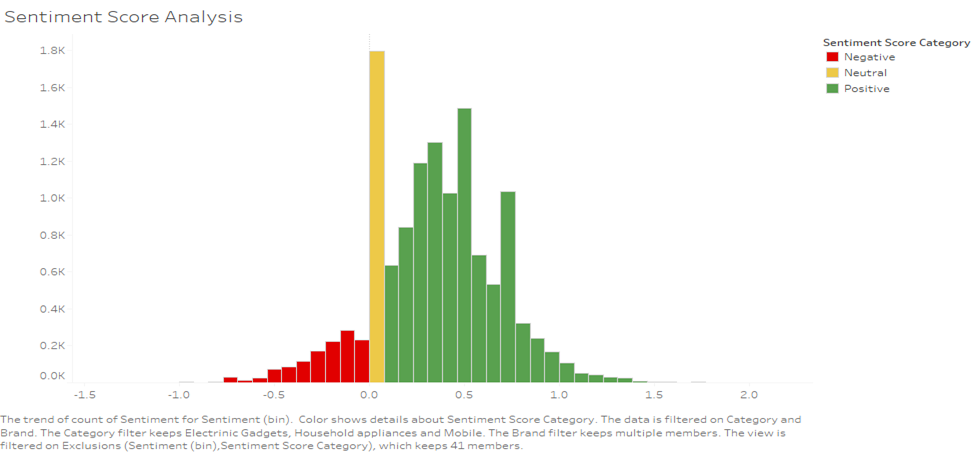
****

Figure 5.1.2A

**Inference**

Scores get overpopulated in the center of the histogram and fall under the 0-1 range of values which means neutral and positive reviews are much in the dataset.

Initially 3 class (target variable) models have been built which have been an over fit model with 99.9% accuracy due to biased data.

**Action**

Two class (1,0 - positive, negative) models have been decided to build.

### **5.1.3 Visualize the reviews and its polarity with respect to categories and brand**

This gives the spread of data collected

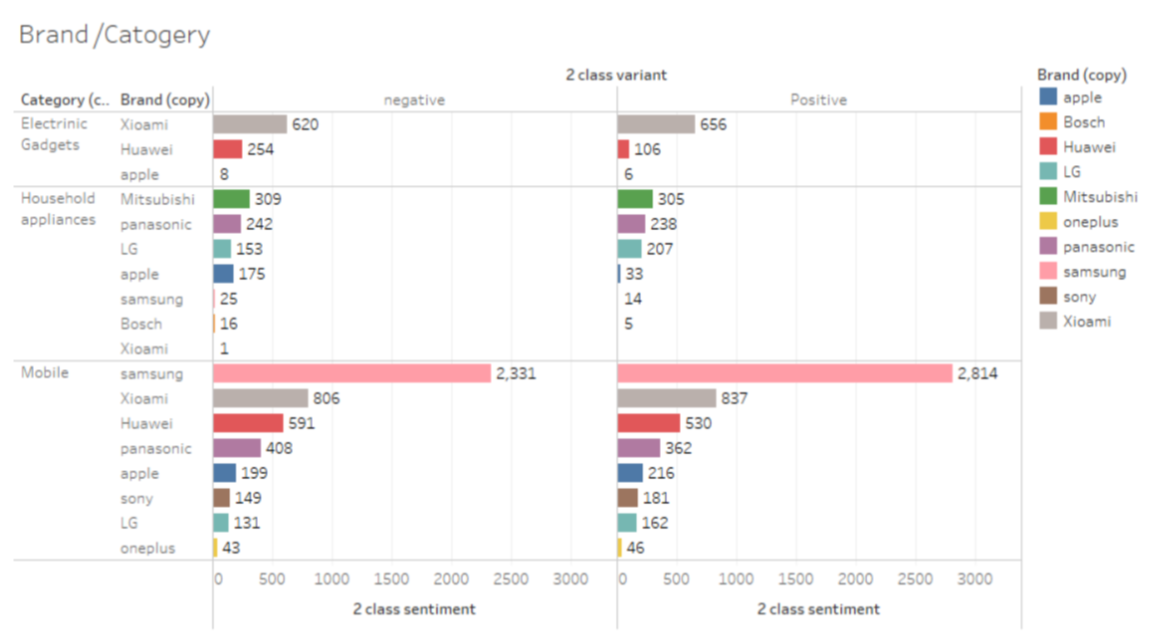


Figure 5.1.3A

# **Chapter 6**

# **Modelling**

## **6.1 Model 1: Naive Bayes**

Naive Bayes classifiers are based on popular Bayes’ probability theorem, that are known for creating simple yet well performing models, especially in the fields of document classification. As it performs well when we have multiple classes and working with text classification, Naive Bayes have been chosen at first.

Using Naive Bayes text classifiers might be a really good idea, especially if there’s not much training data available and computational resources are scarce. Usually, results are pretty competitive in terms of performance if features are well engineered.

It is a method to approach the following segments of the review which the customer are providing such that the review are consisting of several unwanted words such that classifier are being in search of the tag words such that the defined tag words in the vectors which are positive and negative such that the there are several condition to calculate the following probability of calculating the points to determines about the classification of the major review where there are several special symbol which must be removed by the classifier.

Thus the navies’ Bayes classifier is an method and an approach of statistics to get through the vectors contain the word and several library package to determine that the particular tag word must be counted in order to give there result about an accuracy of data to which we can determine whether the user can be able to identify the review rating about user reviews. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

### **6.1.1 Implementation of Algorithm**

Fetching all product review comments.

Conversions of the unwanted format of comment data to a wanted understandable format document

The process of determining the tokens that are to be keywords.

The implementation and processing of unwanted stop character and the required tagging of word by the determined tagged word.

Performing sentiment analysis by generating sentiment scores for each review.

Binning sentiment scores into 3 bins (-1,0,1) which are profiled as negative reviews, neutral reviews and positive reviews.

With the 3-class variable, the model has been trained and tested which has given an accuracy of 67% which is where 2 class classification has come into picture.

Now binned sentiment scores into 2 bins (0,1) which are profiled as positive and negative reviews as the data collected are highly biased when classified into 3 class variables.

Now with 2 class variables, the model has been trained and tested with an accuracy of 77% which is better than 3 class variables.

## **6.2 Model 2: Logistic Regression**

To predict the response variable, here defined as the Sentiment scores grouped into +1, 0 and -1 which represents outcomes such as positive, neutral and negative, a Logistic regression Model is built and applied on the refined dataset for Text Classification. This model estimates the probability of the categorical response based on predictor variables here defined as “Brand”, “Category”, “Reviews”.

Dependent variable here has more than 2 outcome categories which is to be analyzed by Multinomial logistic regression. Target variable has 3 classes which implies that the model will use (3-1) independent binary logistic classifier model. One target group is first considered as a reference class and builds the rest (3-1) models which compares each of the remaining classes to the picked reference class.

Since the target variable is not ordinal (1,2, 3,..) in nature, we omit the possibility of using the Ordinal Logistic Regression model for our classification.

### **6.2.1 Implementation of Algorithm**

Logistic Regression Model was applied on the data set with 3 groups as the target variable. A very high accuracy nearing 100% was witnessed.

The target variable after grouped into 2, Positive (+1) and Negative (-1) is now balanced in the following manner.

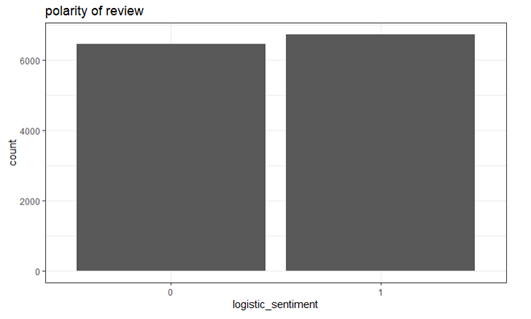


FIGURE 6.2.1A

Before the data is fed to the model, using package “text2vec” we perform certain preprocessing to receive expected results.

Review column is to be tokenized word by word and marked with an ID to be iterated, further which a Vocabulary of the comments is built.

For the model to recognize the text data, a document term matrix is created by vectorizing the vocabulary and then the logistic regression model is fit into the DTM.

### **6.2.2 Simple N fold logistic regression model**

Glmnet with k fold cross validation is applied which by default uses 10-fold and here is adjusted to 7 folds using NFOLDS. Fold value when given as 7 provided the sensible accuracy whereas any other value in the range would provide bizarre outputs.

Alpha () the elastic net mixing parameter is given as default Lasso value 1. The performance of the model is interpreted using the AUC curve.

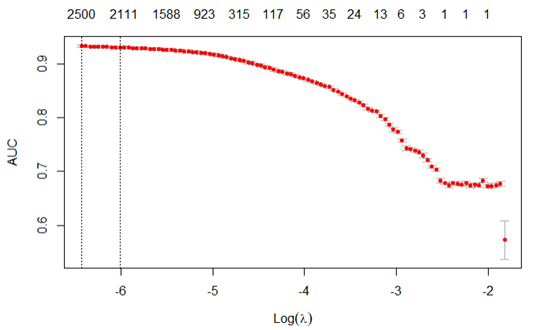


FIGURE 6.2.2A

AUC here gives the rate of successful classification by the logistic model. The model when fed to train data received an accuracy of 93.29%.

Once the test data is converted to a DTM, the trained model is fed to the test data and is made to check the model's performance. The accuracy obtained for prediction on test data is 93.42%.

### **6.2.3 Model with Pruned Vocabulary**

To ensure if a better accuracy is possible or not, pruning of the vocabulary built is done which in turn filters very frequent and very infrequent terms. These vocabulary elements are mapped to indices and made a vectorizer using vocab\_vectorizer ().

Again, train and test DTMs are created in multiple threads(itoken) with the pruned vocabulary vectorizer.

The train pruned vocabulary DTM is fed to the model with 7 fold cross validation and Alpha () the elastic net mixing parameter is again given as default Lasso value 1. Fold value when given as 7 provided the sensible accuracy compared to other values in the range.

The performance of the model on train data is interpreted using the AUC curve. An accuracy of 93.19% is obtained using pruned vocabulary vectorizer.

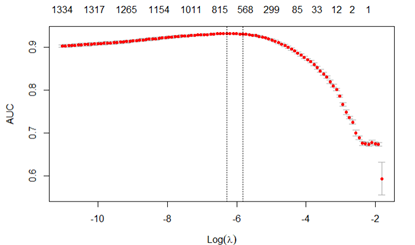


FIGURE 6.2.3A

### **6.2.4 Model using N-grams**

To ensure if the accuracy can/cannot be varied much, a model is built using 2-grams instead of individual words. Here text is vectorized by creating a map from words or n-grams to a vector space. The vocabulary is created using bigrams, pruned and finally converted to a vectorizer.

Train and test DTMs are created in multiple threads(itoken) with the bigram vectorizer.

As trained before, the train bigram DTM is fed to the model with 7 fold cross validation and Alpha () the elastic net mixing parameter is again given as default Lasso value 1. Fold value when given as 7 provided the sensible accuracy compared to other values in the range.

The performance of the train data model is interpreted using the AUC curve. An accuracy of 93.07% is obtained using N-gram vectorizer.

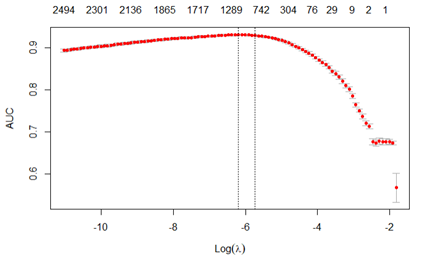


FIGURE 6.2.4A

**6.2.5 Model using Feature Hashing**

Feature Hashing is another effective way to model data and uses less memory and address locations instead of actual data.

Here, the list of tokens is converted to vectors by mapping words to indices. This is done by applying a hash function to the data and the hash values are used as indices.

Hash size, number of buckets for hashing is given as 2^14 and the n-gram, lower and upper bounds for “n” in n gram is given as 1,2 for feature hashing using hash\_vectorizer ().

Train and test DTMs are created in multiple threads(itoken) with the hashed vectorizer.

The train feature hashed DTM is fed to the model with 7 fold cross validation and Alpha () the elastic net mixing parameter is again given as default Lasso value 1. Fold value when given as 7 provided the sensible accuracy compared to other values in the range.

The performance of the train data model is interpreted using the AUC curve. An accuracy of 90.91% is obtained using feature hashing technique.

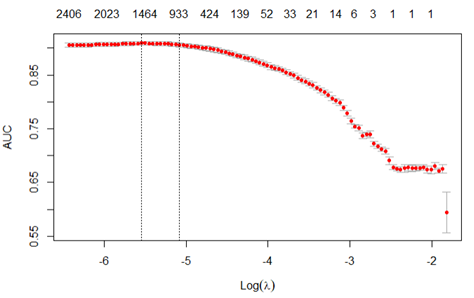


FIGURE 6.2.5A

From the above 4 modelling techniques applied, the output accuracies obtained when applied on the test set are summarized as below.

### **6.2.6 Accuracy Metrics of 4 Models**

AUC for **basic model** in Test dataset= 0.93

AUC for **pruned\_model** in Test dataset= 0.94

AUC for **bigram\_model** in Test dataset= 0.934

AUC for **feature\_hashing\_model** in Test dataset= 0.91

### **6.2.7 Inference**

The conclusion inferred from the Logistic regression models is that when a model with pruned vocabulary vectorizer is fed to a new data, the results procured are as expected and of high reliability.

# **Chapter 7**

## **7.1 Further Refinement**

In terms of refining the modeling techniques applied, 3 class grouping is to be considered. In an ideal situation, neutral reviews are as important as positive and negative ones. The existence of neutral reviews can put a significant influence in shaping consumers' motivation and ability of perceiving positive and negative reviews. Thereby enhancement of the current model could be done by creating subsets of the combinations of data(eg: positive,neutral and neutral, negative and negative, positive). On this data further SMOTE could be applied to get a well-balanced spread of neutral reviews and then feed them into the modelling techniques like SVM, Random Forest to achieve accurate predictions.

# **Chapter 8**

## **8.1 How to use these predicted values(1,0) in classifying a product as good or bad**

Here the predicted values are grouped by brand and category to classify a product as good or bad product

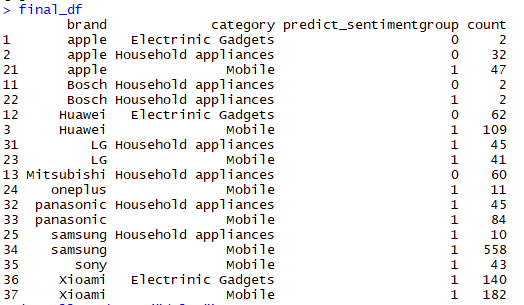


FIGURE 8.1A

After modelling the dataset, with the predicted sentiment group values on the holdout dataset- a recommendation is done to the end users based on the number of corresponding review counts for each group (as in positive, negative.

For eg, The predicted values from Naive Bayes Model is considered here. As seen in figure : a conclusive table is formed with the highest count of reviews.

Category “Household Appliances” under Brand “Panasonic” has a higher count of positive reviews when compared to the negative and neutral reviews received. Whereas, Category “Household Appliances” under Brand “Apple” has higher neutral count of reviews compared to positive and negative reviews. From this, it is inferred that when the Household Appliances of the 2 Brands “Panasonic” and “Apple” are compared, highly recommended Brand to user/consumer would be of “Panasonic” under household appliance products.

# **Chapter 9**

## **9.1 Modelling Techniques Performed And Its Accuracy Metrics**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Machine Learning Model** | **Accuracy** |
| 1. | Naive Bayes  3 groups classification  2 groups classification | 67%  76% |
| 2.  2.1  2.2  2.3  2.4 | Logistic Regression (2 group classification)  Simple Model  Using Pruned Vocabulary  Using N-Gram  Using Hashing Feature | 93.42%  93.66%  93.37%  91.33% |

TABLE 9.1A

# **Conclusion and Recommendation**

Categorization of reviews assist the customers to make an informed choice on whether to buy a product or not based on it's True Positive,False Positive,Truenegative and Falsenegative points by reducing the time that they would have spent reading through a loads of reviews from different social media.The proposed approach in this paper tries to predict sentiments from reviews posted by users on social media like facebook, twitter and other official websites of that particular brand. The approach used a lexicon based approach. It tries to evaluate each opinion bearing word on the basis of its intensity. It assigns a valence to words found in the tweets/paragraph using AFINN lexicon dataset. The logistic regression and Naïve Bayes classifiers are used for training and testing the various product review dataset.

From a business aspect, this project built would help understand the product acceptance of customers from various social media platforms irrespective of where the purchase was made. Model with further analysis will help businesses recognize product features which are less liked by users and improve on them for its success. Similarly, highlighting of features/products which are much liked could be done and thereby showcase the product overall success and increase visibility.

In future, we can enlarge the number of users by providing an environment(app) to share their reviews no matter how they purchase the product(though e-commerce websites or offline) which also helps to collect the data directly from users along with other sources. Also, there should be some automatic method that can recommend products based on features. One can use some mining technique or soft computing methods over these features to recommend products for users.

**Bibliography**

1. Taboada, Maite, et al. "Lexicon-based methods for sentiment analysis." Computational linguistics.
2. McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization.
3. Yousef, Ahmed Hassan, Walaa Medhat, and Hoda Korashy Mohamed. "Sentiment Analysis Algorithms and Applications: A Survey.
4. Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources.
5. Dobra, Alin. "Decision Tree Classification." Encyclopedia of Database Systems.
6. Tableau, „Mission Tableau,“ Tableau, [Online]. Available: http://mission.tableau.com.
7. Tableau, „Add-in Reshaping Data Excel,“ 05 December 2014. [Online]. Available:http://kb.tableau.com/articles/knowledgebase/addin-reshaping-data-excel.
8. Microsoft, „Microsoft Excel Data Types,“ Microsoft, [Online]. Available: https://msdn.microsoft.com/en-us/library/ms712640%28v.aspx
9. Tableau, „Quick Start Guides,“ Tableau, [Online]. Available: http://www.tableau.com/support/manuals/quickstart.
10. Tableau, „Knowledge Base,“ Tableau, [Online]. Available: kb.tableau.com.
11. iosrjournals.org/iosr-jce/papers/Conf.17015-2017/Volume-1/15.%2069-73.pdf?i
12. R. Shermann, Business Intelligence Guidebook, Morgan Kaufmann,2014
13. Few, „Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message,“ perceptual edge.
14. Guide, Developer. "Amazon Elastic MapReduce."
15. Bollen, Johan, Huina Mao, and Xiaojun Zeokng. "Twitter mood predicts the stock market." Journal of Computational Science.
16. Logistic Regression. url: http : / / scikit - learn . org / stable / modules/generated/sklearn.linear\_model.LogisticRegression. Html
17. Trevor Hastie, Jonathan Taylor, Robert Tibshirani, Guenther Walther, et al. “Forward stagewise regression and the monotone lasso.” In: Electronic Journal of Statistic
18. Rune Haubo B Christensen. A Tutorial on fitting Cumulative Link Models with the ordinal Package. 2015. url: https : / / cran . r - project.org/web/packages/ordinal/vignettes/clm\_tutorial. Pdf
19. Janez Demšar. “Statistical comparisons of classifiers over multiple data sets.” In: Journal of Machine learning research
20. Bartlett P. and Shawe-Taylor J., “Generalization performance of support vector machine and other pattern classifiers”, In C.~Burges B.~Scholkopf, editor, “Advances in Kernel Methods--Support Vector Learning”.
21. Osuna E., Freund R., and Girosi F., “Support Vector Machines: Training and Applications”, A.I. Memo, Artificial Intelligence Laboratory
22. Trafalis T., "Primal-dual optimization methods in neural networks and support vector machines training",
23. Vapnik V., ”Statistical Learning Theory”, Wiley, New York.

# **Annexure**

Here is our entire code which demonstrates other EDA’S and models that are tried.

