Sentiment Analysis of Customer Reviews on Laptop Products for Flipkart

Janhavi N L¹, Santhosh Kumar K L², Jharna Majumdar³

¹M.Tech Student, Dept. of M.Tech Computer Science & Engineering, NMIT, Bangalore, Karnataka
²Assistant Professor, Dept. of M.Tech Computer Science & Engineering, NMIT, Bangalore, Karnataka
³Dean R&D, Professor & Head, Dept. of M.Tech Computer Science & Engineering, NMIT, Bangalore, Karnataka
---***---

Abstract - *E-commerce is rising rapidly now a days, purchasing items on online has grown to be more and more fashionable outstanding of more options like lower in price, better supply system, therefore buyers plan to do Online shopping. User's comments are useful information to estimate product quality. This paper has a tendency to analyze the fundamentals of opinion mining. It consist different approaches including Extraction, Clustering and Classification. Extracting reviews from the website using flipkart product API, using product API we can easily fetch the brand name, reviews, rating and other related things for product, clustering using ROCK and using CART algorithm to classify reviews as positive and negative words from the comments and finally they come to know which product having more percentage of positive reviews.*

Key Words: **Product API, Data Mining techniques, Machine Learning algorithm, ROCK, CART.**

1. INTRODUCTION

Online shopping is a way of buying products and services from the vendors over the internet by different browsers and apps. User can search their wishing product in a different websites by looking around the different webpage of e-commerce of various vendors of same product, its availability and price. There will be difference in price, specification of product and availability in different vendors. Users consider reviews of other buyers while buying also concentrate on few recommendations and interact with different search engines.

"The process of finding user opinion about the topic or product or problem is called as opinion mining." Or "It can also be defined as the process of automatic extraction of knowledge by means of opinions expressed by the user who is currently using the product is called as opinion mining." The motivation of opinion mining is toward build system to identify and convey sentiments in reviews.

Sentiment analysis is valuable in several ways. It helps vendors measure the occurrence of fresh product initiate, verify the range of a products or services is on demanding and identify the statistics of like or dislike particular product features. Sentiment analysis is a way of finding users opinion about exacting matter or a problem or product and aim is to determine expressed reviews are positive, negative or neutral. Opinion mining comes under web content mining.

In this work, we are extracting Laptop product reviews from Flipkart online shopping website, extracted data is store in CSV file. Clustering done on extracted data that is pre-processed laptop products Reviews (CSV) file. Using classification technique, the reviews are classified as positive and negative and performance measured.

2. Literature Survey

The ongoing research work related to the Opinion mining and Sentiment Analysis is given in this section. In [1], focuses tools and techniques used in opinion mining. The process of opinion summarization has three main steps, such as "Opinion Retrieval, Opinion Classification and Opinion Summarization." User comments are retrieved from review websites. These comments contain subjective information and they are classified as positive or negative review. Depending upon the frequency of occurrences of features opinion summary is created.

In [2], focuses on review mining and sentiment analysis on Amazon website. Users of the online shopping site Amazon are encouraged to post reviews of the products that they purchase. Amazon employs a 1-to-5 scale for all products, regardless of their category, and it becomes challenging to determine the advantages and disadvantages to different parts of a product. In [3], focuses only on the reviews taken from Amazon website analyze the results using three different algorithms such as Naïve Bayes, Logistic Regression and SentiWordNet.

In [4], Decision Making and Analysis on Customer Reviews using Sentiment Analysis Dictionary (DMA) Tools developed for both customer and manufacturer or seller to check whether the selected product is a good or bad and same has been tested on the Humanoid robot for the Human-Robot Interaction.

In [5], focuses on analyzing reviews from different E-Shopping Websites. The main focus of the system is to analyze reviews for online shopping services. The reviews are classified according to positive, negative and neutral. These results help to select particular site for e-shopping, based on maximum number of positive reviews and rating. Firstly collect the E-shopping websites dataset which

contains review related to the services of particular websites. Then apply some preprocessing methods on datasets for removing undesirable things and arranging data in proper manner. After that we use POS tagger for assigning tags to each word according to its role. It uses "sentiwordnet dictionary" for finding score of each words. Then sentiments are classified as positive, negative and neutral. The analysis of the services according to positive and negative reviews can be shown in the graphical format.

In [6], focuses on Sentimental Analysis of Flipkart reviews using algorithms Naïve Bayes and Decision Tree. Using users reviews about product and review about retailers from Flipkart as dataset and classifies review by subjectivity/objectivity and negative/positive opinion of buyer. Such reviews are helpful to some extent, promising both the shoppers and products makers. It presents an empirical study of efficacy of classifying product review by semantic meaning. Classifying comments employs hybrid algorithm combining Decision Trees and Naïve Bayes algorithm.

The procedure of Web mining has 4 stages, those are: Data collection, Data preprocessing, Pattern discovery, Pattern analysis as shown in the figure 1.
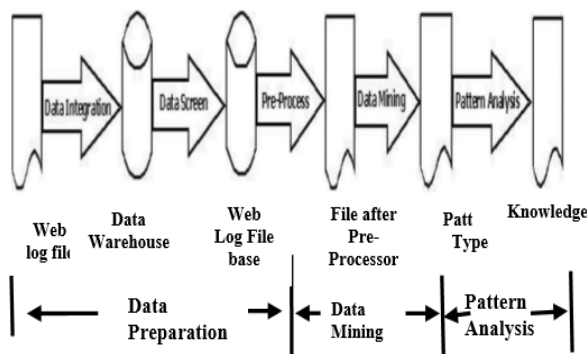

Fig 1: Process of web data mining

In [7], concentrate on Different Clustering Algorithms and methodology on Amazon Reviews. Customers of online shopping site like Amazon, showing thousands of customer comments earlier than purchasing product may be a challenging assignment. Using unsupervised machine learning methods are used for analyze the preprocessed data from online shopping websites to supply consumers with enhanced customers knowledge earlier than buying product. K-means and Peak-searching clustering algorithms are to perform clustering of product on Amazon reviews.

In [8] "ROCK: A Robust Clustering Algorithm for Categorical Attributes", ROCK clustering algorithm is a hierarchical clustering technique and it relates to group of agglomerative hierarchical clustering algorithms. Rock algorithm randomly takes sample dataset retrieved by database. Hierarchical clustering technique uses links to the

points, clusters are requires sampled points; those points are utilized to allocate left over data points on database to proper clusters.

In [9], survey concentrates on decision tree algorithm techniques for classification in data mining. The most popular classification technique is decision tree methods. The basic learning strategy divide and conquer technique is used in decision tree. A decision trees are similar to tree structure, every implicit node represents attribute, every branch represents an output of the test, and class label is represented by each leaf node. Algorithms of different types of trees are ID3, C4.5 and CART. [10]

ID3 (Iterative Dichotomiser 3) Algorithm: ID3 algorithm is to build the decision tree by utilizing a top-down, greedy search via the known sets to test data every attribute tree node. This method uses information gain to regulate reasonable belongings for every node of decision tree.

C4.5 Algorithm: C4.5 is used for classification technique and also C4.5 is referring as statistical classifier and this algorithm uses information gain. Categorical or numerical values both are accepted in this algorithm. In order to hold continues data values is produced by threshold and after that divide attributes based on threshold value and these values can lower or equivalent to the threshold value. Handling missing values is easy by C4.5 algorithm, and lost attributes are not used in information gain.

CART Algorithm: CART acronym is Classification And Regression Trees. CART was invented in 1984 by Breiman. In CART algorithm there are two things classification and regression. Classification is for binary splitting of the attributes and it also uses Gini index for splitting calculate in choosing the splitting attribute. Regression is to predict a reliant variable known as group of variables among certain amount of given time. CART algorithm uses continuous and nominal attribute value and has average speed of processing.

3. Methodology

The basic purpose of analyzing the review is to help buyer to decide whether product is good or what. Usually if we want to buy anything online we will first go to the website and see how many people have recommended the product? How many people have rated the product as 5*? What are the pros and cons of the product? Etc. Nowadays websites also provide compare options, which will allow the user to compare two or more products and take decision if the buyer is in confusion. Current work uses data mining algorithm, such as ROCK for clustering and for classification CART algorithm along with various mechanisms.

The need of this application is to show an analysis of customer opinions on decision making. It will help the customer to decide whether the product is good or bad and which product is best and customer to choose among variety of product by seeing reviews of the products. Detailed work

flow diagram of proposed system is shown in figure 2. The system architecture of proposed method is shown in figure 3.
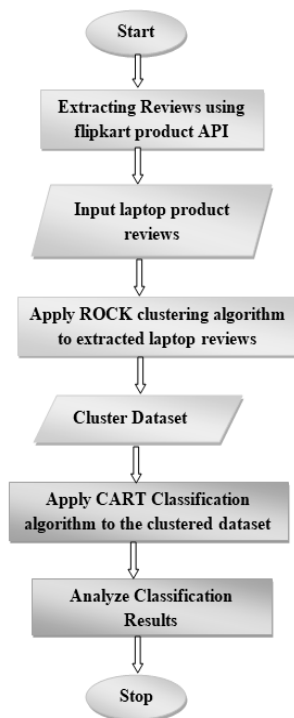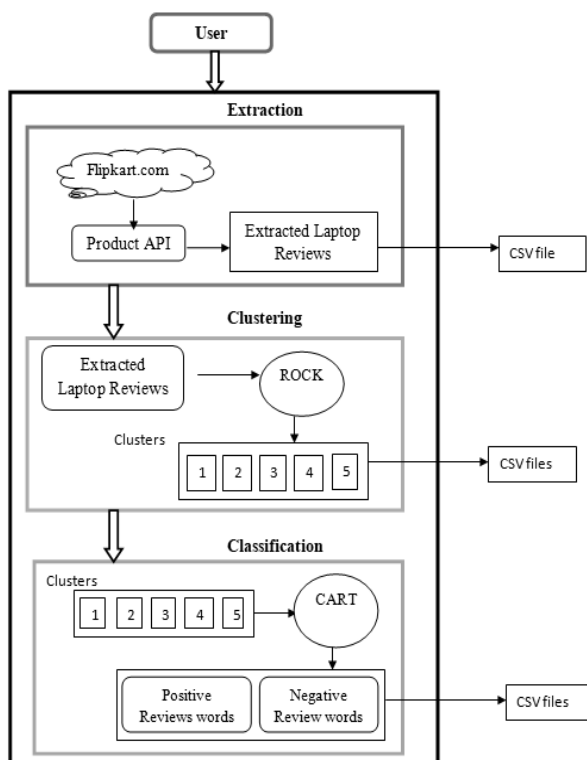

Fig. 2: Flowchart of the proposed system

Fig. 3: System Architecture of the Proposed Method

3.1 Extraction

Input: flipkart product API.

An *API product* is a collection resource in order to provide a specific level of access and functionality for client application developers using Json and Jsoup. In order to fetch the laptop product details like brand name, link, rating, stars and other details of product.

Output: Fetched reviews from website are stored in CSV file.

3.2 Clustering

On the extracted data ROCK algorithm is applied. ROCK algorithm along with the steps carried out in performing it is as below:

3.2.1 ROCK

Clustering technique can be used for the information among Boolean and categorical attributes. Clustering is a method of collecting data points; these data points are in a cluster or group having same characteristics and while data points in separate clusters are different. For instance, assume the database of market basket consist one transaction per consumer, every transaction consist a group of product consumer purchased.

ROCK clustering technique used link for data concerning links among two points and while making decision on those points to be combined into one cluster.

For example, link (i,j) can be the listing familiar neighbors among i and j. From the meaning of links, it follow that if link (i,j) is greater, then it is further probable that i and j feel right to the similar cluster.

This approach deals with overall difficulty of clustering. It catches the world wide facts of surrounding data points into connection among pair of points. Consequently, ROCK algorithm uses the data concerning links among points focuses when decision creation on those points to be combined into one cluster, it is exceptionally robust.

ROCK algorithm is divided into three general parts:

1. Obtaining a random sample of the data.

2. Performing clustering on the data with the link agglomerative method. A goodness measure is use to establish which pair off of points is combined at every step.

3. Using clusters the left data on disk are assigned to them.

ROCK Algorithm Steps

1. ROCK Algorithm takes input as dataset *DS* of num data points could be clustered, along with the group of preferred clusters *cl*.

   Rock_clus_alg(DS,cl)

2. Algorithm process starts with performing links among pair of points.

   li = compute_link(DS)

3. Initially every data point is individual cluster.

4. Every separate cluster *i*, construct *local_heap c[i]*. *c[i]* having each cluster like *li[pi,pj]* is non zero value and j cluster with *c[i]* arranged in decreasing manner.

   for i to DS   c[i]=build_lheap(li,i)

5. Including *local_heaps c[i]* for every cluster *i*, and also maintain another global heap *Qi* having all cluster.

   Qi= build_lheap (DS, c)

6. Later on, clusters in *Qi* are sorted *g1(j1,max(q[j1]))* for arranging clusters *j1* in *Qi*, when *emax(q[j1])*, the maximum element in *q[j1]*, is the finest cluster to combine with cluster *j1*.

   while Size(Qi)>cl

   ui=emax(q[j1])

   vi= max(q[ui])

7. Every iteration in loop, maximum cluster *j1* within *Qi* and the maximum clusters *q[j1]* are the good pair of cluster could be combined and find the best cluster to merge.

   Wi=mergec(ui,vi)

8. Clusters *ui* and *vi* clusters combined; data in *ui* and *vi* are removed if it is not necessary from *Qi* and Clusters *ui* and *vi* are combined to generate a cluster *Wi* having *ui+vi* point. Update *Qi* in every *x*.

   for x to q[ui] & q[vi] li[x,Wi]=li[x,ui]+li[x,vi]

9. Evaluate one time clusters *ui* and *vi* are combined, primary if for each clusters to having *ui* or *vi* is in *local_heap*, *ui* and *vi* to be changing by means of the latest combined cluster *w* and the *local_heap* wants to be restructured and a local heap *w* get generated. List of links among clusters *x* and *w* is addition number of links among *x* & *u* and *x* & v.

   Insert_c(q[x],Wi,gi(x,w))

   Update_c(Qi,x,q[x])

   Insert_(Qi,Wi,q[Wi])

   Deallocate(q[ui])

Input: extracted data that is pre-processed laptop products Reviews file (CSV file).

Output: Each Cluster data stored in the different CSV/Excel formats.

3.3 Classification

Classification is a data mining technique that assigns dataset in a group to object categories or classes.

Classification and Regression Tree (CART) is a type of the decision tree algorithms. CART algorithm is used for building both Classification and Regression Decision Trees, CART classification algorithm is applied for classifying the reviews either positive or negative words with the help of dictionary words. The impurity (or purity) measure used in building decision tree in CART is Gini Index.

Classification Tree

Classification tree is used, when decision or target variable is categorical. Example: Predicting the food choices of the customers (nominal variable) using set of independent variable.

Regression Tree

Regression decision tree is used, when the decision or target variable is continuous variable. Example: Predicting house prices using attributes of houses such as size, type and others. These variables can be continuous and categorical.

Generally in recursive partitioning approach and CART, split each of the input nodes into two child nodes, in other term CART is a Binary Decision Tree

CART Algorithm Steps

This involves few simple steps as follows:

1. **Input dataset**: Target Variable and a list of Independent Variables

2. **Best Split**: Find Best Split for each of the independent variables

3. **Best Variable**: Select the Best Variable for the split

4. **Split the input data** into Left and Right Nodes

5. **Continue step 2-4** on each of the nodes until meet stopping criteria

6. **Decision Tree Pruning** : Steps to prune Decision Tree built

4. Experimental Results

Online connection is established to the webpage of flipkart.com for laptop details using flipkart product API using Json.

The Laptop details are fetched using product API. The class name of element is specified to get the required Laptop details. The extracted data of laptop details is stored in CSV file format. It contains six attributes such as Star Rating, Reviews, Brand, Link and Review Text. ROCK algorithm is applied on the extracted CSV files. It forms five clusters based on the algorithm. Each cluster stores in separate CSV files. The graphical representation of clusters is shown in figure 4:
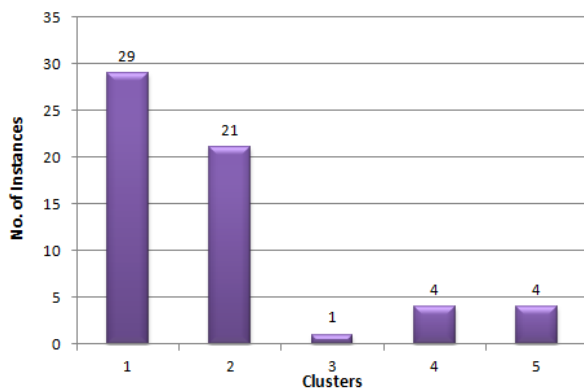

Fig 4: Visualization of Clusters

In laptop dataset there are 59 instances, the instances are clustered into five clusters. The table 1, shows the number of instances and its result in percentage in each cluster.

Clusters	1	2	3	4	5
Instances	29	21	1	4	4

Table 1: Result of Clusters

On the clustered data CART algorithm is applied, it compares the dictionary words with the dataset of clustered data and then classifies the review text into their positive and negative words. Table 2 shows the result in positive opinion percentage of nine different brands.

Laptop Brands	1	2	3	4	5	6	7	8	9
Instances	90	67	77	60	90	93	100	60	85

Table 2: Result of positive percentage of laptop brands

Word Cloud class shows the prominence of the individually classified word by highlighting words in diverse font styles. Figure 5 and 6 shows the word cloud for positive and negative words respectively.

Fig 5: Word Cloud display of positive words

Fig 6: Word Cloud display of negative words

CONCLUSION AND FUTURE WORK:

In this paper, the Classification and Analysis of laptop product reviews for the flipkart website is achieved by using ROCK and CART algorithms. The Laptop details are extracted by using flipkart product API. This work classifies positive and negative words from reviews and it calculates percentage of positive and negative words. Therefore the result analysis of review percentage it helps the user to conclude based on the positive review percentage of the product. Future work can be concentrated on mining of reviews from multiple website and multiple products etc. The same work can be extended to incorporate many more classification algorithms which will help us to decide or to choose the best classifier for opinion mining and sentiment analysis.

ACKNOWLEDGEMENT:

REFERENCES

[1] G.Angulakshmi, R.ManickaChezian, "An Analysis on Opinion Mining: Techniques and Tools." International

Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014

[2] Callen Rain, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning", Swarthmore College Compter Society, November 2013

[3] Santhosh Kumar K L, Jayanti Desai, Jharna Majumdar "Opinion Mining and Sentiment Analysis on Online Customer Review", 2016 IEEE International Conference on Computational Intelligence and Computing Research at Agni College of Technology, Chennai during December 15th to 17th 2016

[4] Gurunath H Naragund, Santhosh Kumar K L, Jharna Majumdar "Development of Decision Making and Analysis on Customer Reviews using Sentiment Dictionary for Human-Robot Interaction", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Volume 4, Issue 8, August 2015

[5] U Ravi Babu, Narsimha Reddy, "Sentiment Analysis of Reviews for E-Shopping Websites", International Journal of Engineering And Computer Science, ISSN: 2319-7242 Volume 6 Issue 1 January. 2017

[6] Gurneet Kaur, Abhinash Singla, "Sentimental Analysis of Flipkart reviews using NaïveBayes and Decision Tree algorithm", International Journal of Advanced Researching Computer Engineering & Technology (IJARCET) Volume5, January 2016

[7] Chantal Fry, Sukanya Manna, "Can we Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms", IEEE Tenth International Conference on Semantic Computing 2016

[8] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", International Journal of Science, Engineering and Technology Research (IJSETR), 2015

[9] Delveen Luqman Abd AL-Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)", Computer Engineering and Intelligent Systems(IISTE) Vol.4 , 2013

[10] Himani Sharma, Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR) 2015