# A Review Paper on Machine Learning Techniques and (Product) Reviews from Social Network Services

**Shreya M Dholaria[1] Sangani Vivekkumar D[2] Madhuri Thakar[3] Daxa V Vekariya[4]**
[1,3]PG Student [2,4]Assistant Professor
[1,3,4]Department of Computer Engineering [2]Department of Electrical Engineering
[1,3,4]Noble Group of Institution, Bhesan Road, Via Vadal, Nr Bamangam, Junagadh, Gujarat, India
[2]Dr.Subhash technical campus, Khamdhrol Road, Nr Khodiyar temple, Junagadh, Gujarat, India

*Abstract*— Sentiment analysis is developing area of research to extract the subjective information by applying Natural Language processing and Machine learning techniques. The field of opinion mining also called as sentiment analysis. Social media helps in connecting themselves with social networking sites. E-Commerce business is developed rapidly and larger number of products is sold online. Products are now available on the Internet. The challenges area defines extract reviews from large number of reviews, because it is impossible to read all the reviews. Reviews can be given based on features of products. This paper presents the Machine learning techniques which including decision tree, Support vector Machine, naïve Bayes classifier, Maximum Entropy classifier and Evaluation formulas of opinion, Extract Product Rating reviews.

***Key words:*** Machine Learning, Opinion Mining, Sentiment Analysis, Natural Language Processing, Feature Based Opinion, Classification, Product Review, Support Vector Machine, Native Based Classifier, Maximum Entropy Classifier

## I. INTRODUCTION

Data mining (DM) is a widely used and popular knowledge extraction method for knowledge discovery also called as KDD process. The opinion is used by manufacturers, politicians, news groups, and some organization to know the opinions of customer, people, End user and Websites. If users get clear idea about product reviews and service huge amount of users are using the Internet to post the opinions about the Services or products. Reviews or Comment will be easier to help for user to take the right decision. Today, we are in 21st century and people do not find time to come & interact with each other to make decision to purchase effective product or services. Social media services helps in connecting themselves with social networking sites like flipkart, amazon, Snapdeal, Jabong through which now users can stay far and yet remain connected. Recently, there has been a lot of interest in the continuum between completely supervised and unsupervised learning Classification is used for social network opinion mining.[1] The social media, blogs, forums, e-commerce web sites, etc. encourages people to share their reviews, emotions and feelings publically. Opinions as feedback from the customers can improve the quality of their product or services. It also very useful for decision makers or policy makers and organization. In presence the usage of the Internet increases across a worldwide various of fields. A number of machine learning approaches are used to distinguish the Reviews given by customers for their product. These techniques are Support Vector Machines (SVM), Naïve Bayes (NB), and Maximum Entropy .Machine learning approaches starts from collecting training dataset, then to train a classifier on the training data. Once a supervised classification technique is selected, then an important step: decision to make is feature selection[2].The important product aspects are identified based on two observations: 1) the important aspects are usually commented on by a large number of consumers and 2) consumer opinions on the important aspects greatly influence their overall opinions on the product.To perform sentiment analysis the most and common source of data set are web pages, social web site like face book , twitter, YouTube *etc*.[2]

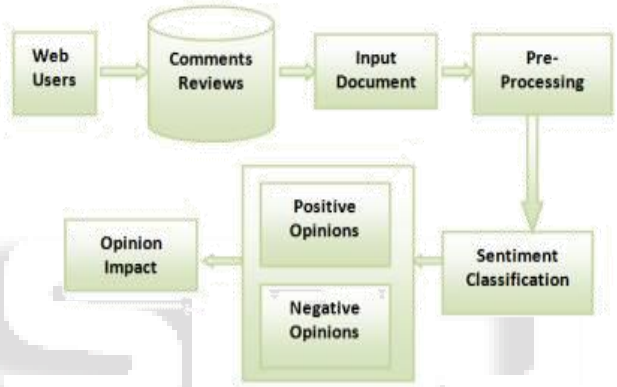


Fig. 1: Workflow of Opinion Mining [3]

## II. BASIC REQUIREMENT FOR OPINION MINING

### A. *Sentiment Analysis of Data Set Domains:*

- Reviews - product, movie and music reviews.
- Web discourse – Social Website (face book, twitter, YouTube).
- News articles – Online articles and web pages.

### B. *Data Size and Data Source:-*

Data size means the number of sentence/ expression/feedback/review on which techniques are applied for sentiment analysis. Data source mean that from which place (e.g. website, movie review, web pages, journals, customer feedback) data sets have been taken.

### C. *Accuracy, Precision and Recall*

We can see the formula to compute the accuracy, precision and recall values
Accuracy: Accuracy or Accuracy rate (or percent correct), is defined as the number of correct cases divided by the total number of cases. [5]

 

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$

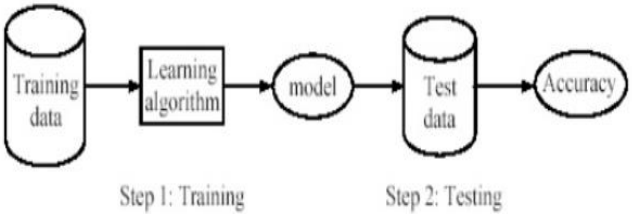


Fig. 2: Training And Testing of Data

Precision: Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant or it is the percentage of selected items that are correct

Recall: Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved or it is the percentage of correct items that are selected [5] we can calculate the above mentioned measures by using the formulas discussed below [5]

Accuracy = (tp + tn) / (tp + fp + fn + tn)

Precision = tp / (tp + fp)

Recall = tp / (tp + fn)

- False positives (FP) - number of reviews incorrectly labeled as belonging to particular class.
- False negatives (FN) - number of reviews were not labeled as belonging to the particular class but should have been labeled.
- True positives (TP) - number of reviews correctly labeled as belonging to particular class (positive/negative).

| | Actual Class (Observation) | |
|---|---|---|
| Predicted class (Expectation) | $t_p$ (true positive Correct result) | $f_p$ (False positive unexpected result) |
| | $f_n$ ( False Negative Missing Result) | $t_n$ (True negative Correct Absent of Result) |

Table 1: Discription of symbols used for formulas[5].

## III. SENTIMENT ANALYSIS OF MACHINE LEARNING CLASSIFICATION TECHNIQUES

(Learning = Representation + Evaluation + Optimization)

Classification is most frequently used popular data mining technique [4]. The aim of Machine Learning is to develop an algorithm so as to optimize the performance of the system using example data or past experience. The Machine Learning provides a solution to the classification problem that involves two steps: 1) Learning the model from a corpus of training data 2) Classifying the unseen data based on the trained model.[7].The classification phase of the process finds the actual mapping between patterns and labels (or targets) Active learning, a kind of machine learning is a promising way for sentiment classification to reduce the annotation cost [8].

1) Supervised Learning Approach
2) Unsupervised Learning Approach

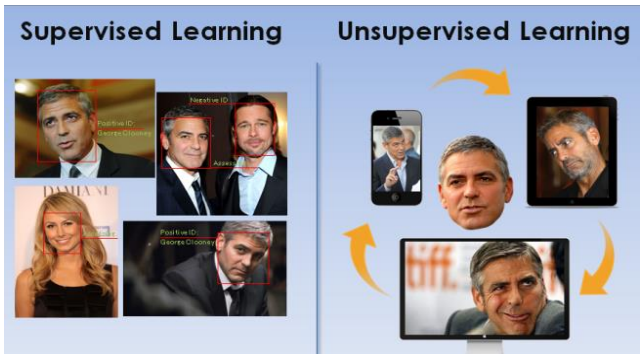


**Fig. 3: Machine learning techniques of SA.**

Supervised Learning Approach

1) Support Vector Machine,
2) Naïve Bayes
3) Maximum Entropy
4) Decision Tree.

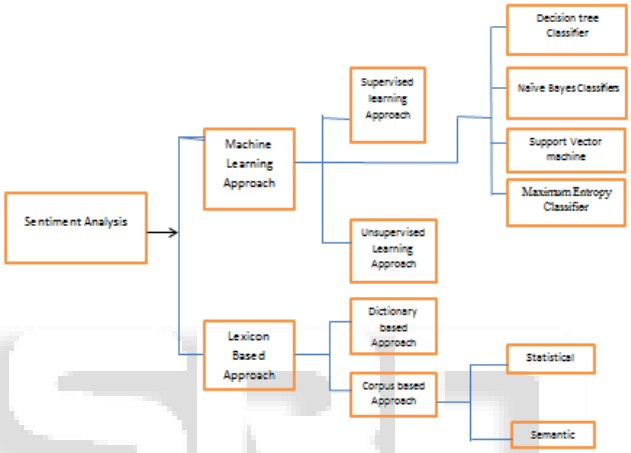


Fig. 4: Classifications of Sentiment Analysis

The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a "teacher" gives the classes (supervision).

- Test data are classified into these classes too.
- These methods are usually fast and accurate.
- Supervised learning is the Data mining task of inferring a function from labeled training data.

### *A. Machine Learning Structure:*

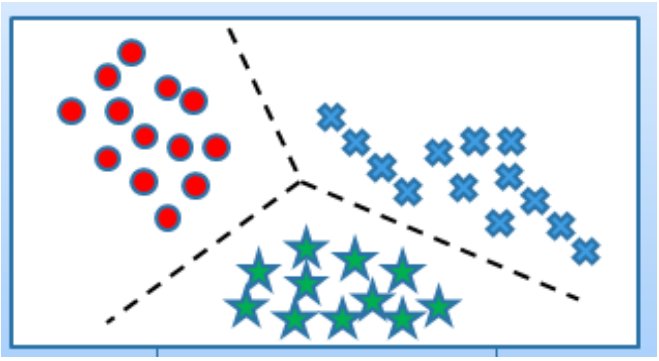

Fig. 5: Supervised learning structure

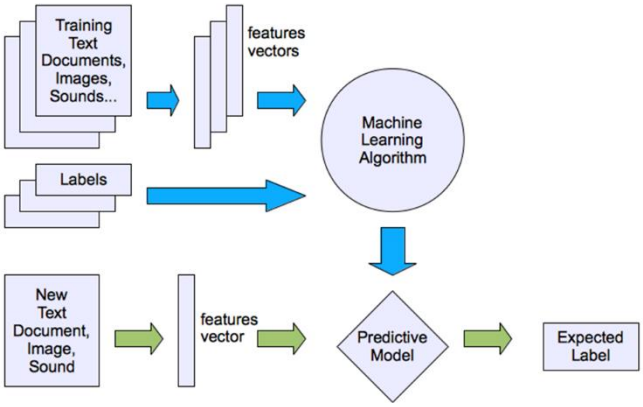


Fig. 6: Supervised learning Algorithms

*B. Support Vector Machine:*

Support Vector Machines (SVMs) are a set of supervised learning methods which have been used for Classification, regression and outlier's detection. Basically Support Vector Machine is a binary classifier that classifies the samples into either true class or in false class. There are number of benefits for using SVM such as: i) It is versatile because holds different kernel functions can be specified for the decision function. ii) It is effective is high dimensional space, iii) Uses a subset of training points in the decision function so it is also memory efficient, Common kernels are provided, but it is also possible to specify custom kernels. The support vector machine method is a particular way to learn linear classification model. Using appropriate kernels, these methods can thus actually learn non-linear models and also handle naturally the complex data types that are often met in the context of computational biology. Compared with other methods, SVMs usually provide state-of-the-art results but like neural networks, they are essentially black-box models and not easy to use for non-specialists.

Advantages of SVMs:

- High accuracy,
- Nice theoretical guarantees regarding over fitting
- Memory-intensive and hard to interpret

*C. Native Bayes Classifiction:*

Naive Bayes is a probabilistic learning method that assumes terms occur independently. In order to incorporate unlabeled data, the foundation Naïve Bayes was build. The task of learning of a generative model is to estimate the parameters using labeled training data only. The estimated parameters are used by the algorithm to classify new documents by calculating which class the generated the given document belongs to The naive Bayesian classifier works as follows:

1) Consider a training set of samples, each with the class labels T. There are k classes, C1, C2, . . . ,Ck. Every sample consists of an n-dimensional vector, X = { x1, x2, . . . ,xn}, representing n measured values of the n attributes, A1,A2, . ,An, respectively.
2) The classifier will classify the given sample X such that it belongs to the class having the highest posterior probability. That is X is predicted to belong to the class Ci if and only P(Ci |X) > P(Cj |X) for 1= j = m, j≠ i.

Thus we find the class that maximizes P(Ci |X). The maximized value of P(Ci |X) for class Ci is called the Maximum posterior hypothesis.

By Bayes Theorem

A Survey Of Classification Methods…

P (A|B) = (1)

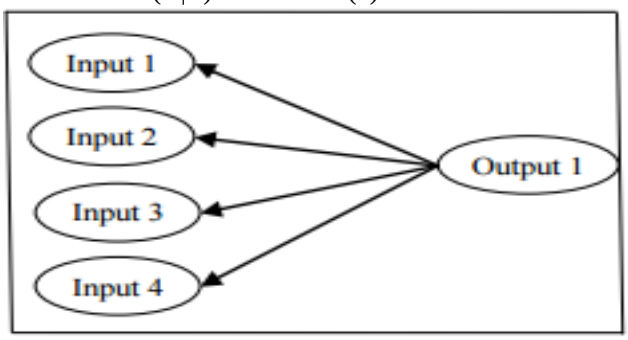


Fig. 7: Naive Bayes model

The simplicity of the naïve bayes theorem is very useful when it comes to document classification The main idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The simplicity of the Naïve Bayes algorithm makes this process efficient has proposed an improved version of the Naïve Bayes algorithm and a unigrams + bigrams was used as the feature, the gap between the positive accuracy and the negative accuracy was narrowed to 3.6% compared to when the original Naïve Bayes was used, and that the 28.5% gap was able to be narrowed compared to when SVM was used. [9] As shown in the below figure, the links in a Naive Bayes model are directed from output to input, which gives the model its simplicity, as there are no interactions between the inputs, except indirectly via the output[10].

Advantages of Naive Bayes:

- Super simple technique
- Logistic regression
- Less training data Required

*D. Maximum Entropy Classifier:*

It is also called as Conditional Exponential Classifier. It uses encoding technique to convert labeled feature set to vectors and then weights are calculated for each feature to determine most relevant label for a feature [11].

*E. Decision Tree:*

This algorithm is prediction based model. The knowledge source used for the decision tree is another classification method[12]. Decision Trees are trees that classify instances by sorting them based on feature values, where each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume [10]. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves . In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

 

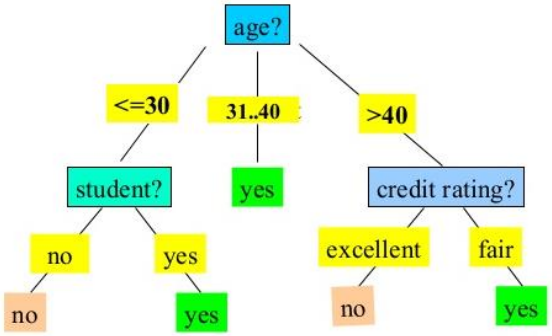


Fig. 8: Decision Trees for "buy_computer"

Advantages of Decision Trees

- Produce intensive results
- Easy to understand
- Easy to interpret and explain

*F. UNSUPERVISED LEARNING (Clustering):*

Class labels of the data are unknown. Given a set of data, the task is to establish the existence of classes or clusters in the data. The model is not provided with the correct results during the training. Clustering comes under unsupervised learning.

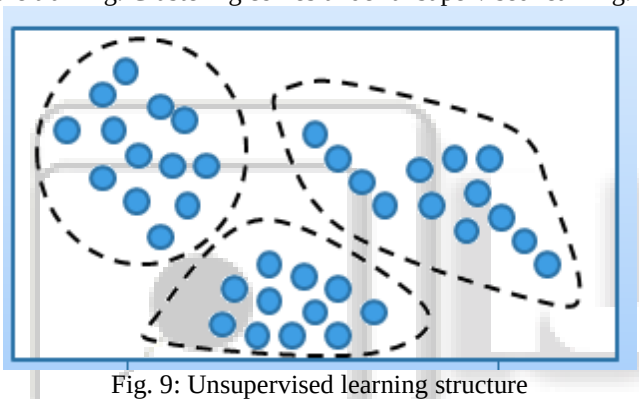

Fig. 9: Unsupervised learning structure

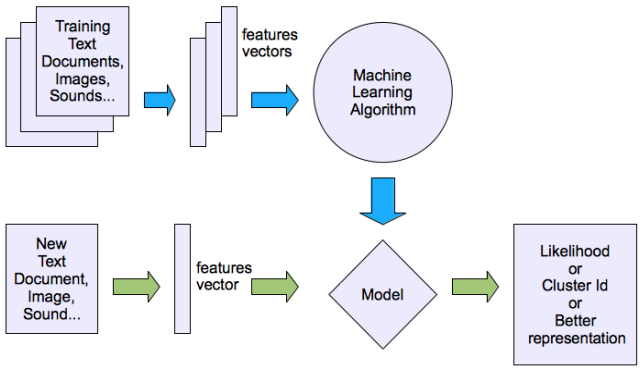


Fig. 10: Unsupervised learning Algorithms

*G. Classification of Opinions from Social Networks:*

Machine Learning approaches commonly used for Sentiment Classification. [7] The opinions are gathered from social networks. They are in the form of sentences. We gathered data from various social networking sites like, shopping sites, Blogs, Twitter, and Facebook. Classification follows these steps:

Step1: Preprocessing of opinions
Step2: Extraction of feature-phrases of products
Step 3: Classification of opinions.

*1) Preprocessing of opinion:*

In this process noise or common words are removed by comparing with list of common words. Product features are usually noun or noun phases.

*2) Extraction of feature-phrases of products:*

In this process we extract feature phrase pattern. Some of the positive and negative phrases are used in the customer Opinions [6]. Holds well-organized knowledge structure[14].

## IV. APPLICATIONS AREAS OF OPINION MINING AND SENTIMENT ANALYSIS

Opinion based or feedback based application are more helpful to user for right decision to purchase product, now a days, the natural language processing are shows much interest in Sentiment Analysis and Opinion Mining system. Customers' reviews and experience are very useful Attributes in decision making process. The major applications of Opinion mining and sentiment analysis are the following [14]. Purchasing Product or Service Quality Improvement in Product or service Marketing research Recommendation Systems Detection of "flame" Opinion spam detection Policy Making Decision Making

## V. CONCLUSION

In This review paper gives the brief ideas about different techniques of machine learning field to mining opinion from various shopping site for their sold product based on reviews given by customers. Also this literature survey paper it is seen that sentiment analysis/opinion mining play vital role to make decision about product or services. To mine the opinions of the people opinion mining and sentiment analysis is the best approach. From this survey, it can be concluded that supervised techniques provide better accuracy compared to dictionary based approach. More future research works could be committed to rating product based on features and give rank.

## REFERENCES

[1] Mr. Nitin N. Pise Dr. Parag Kulkarni "A Survey of Semi-Supervised Learning Methods" International Conference on Computational Intelligence and Security 2008

[2] Shailendra Kumar Singh, Sanchita Paul and Dhananjay Kumar "Sentiment Analysis Approaches on Different Data set Domain: Survey"International Journal of Database Theory and Application Vol.7, No.5 (2014).

[3] Akanksha Dubey, Nikhil Chitre and Vasundhara Ghate "A Survey on Opinion Mining" Vol. 4, Issue 6, June 2016 International Journal of Innovative Research in Computerand Communication Engineering

[4] Rashid A, Anwer N, Iqbal M, Sher M; "A Survey Paper: Areas, Techniques and Challenges of Opinion Mining",2013;10(6) IJCSI International Journal of Computer Science Issues

[5] K S Kushwanth Ram1 , Sachin Araballi2 ,Shambhavi and Shobha G "Sentiment Analysis Of Twitter Data" Volume 3 Issue 12, December 2014 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)

[6] Dr. S. Sagar Imambi, Y. Chandana, G. Raphi "Analysing Customer Reviews Using Opinion Mining" November-2015, Imambi et al., International Journal of Advanced

 

Research in Computer Science and Software Engineering 5(11),
[7] S. ChandraKala and C. Sindhu " opinion mining and sentiment classification a survey" OCT 2012, Volume: 03, ICTACT Journal on soft computing
[8] Shoushan Li, ShengfengJu, Guodong Zhou and Xiaojun Li, "Active Learning for Imbalanced Sentiment Classification", Proceedings of the International Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 139–148, 2012
[9] M.Govindarajan and Romina M "A Survey of Classification Methods and Applications for Sentiment Analysis" Volume 2 2013 The International Journal Of Engineering And Science (IJES)
[10] Iqbal Muhammad1 and Zhu Yan "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY" APRIL 2015, VOLUME: 05, ICTACT JOURNAL ON SOFT COMPUTING.
[11] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques."Vol 7 No 6 Dec 2015-Jan 2016 International Journal of Engineering and Technology (IJET)
[12] International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013 Copyright to IJARCCE www.ijarcce.com 4667 A Survey on Supervised Learning for Word Sense Disambiguation Abhishek Fulmari, Manoj B. Chandak
[13] Yen-Liang Chen and Lucas Tzu-Hsuan Hung, "Using decision trees to summarize associative classification rules", Expert Systems with Applications, Vol. 36, No. 2, Part 1, 2009
[14] K S Kushwanth Ram and Shobha G " Opinion Mining and Sentiment Analysis - Challenges and Applications" Volume 3, Issue 5, May 2014 International Journal of Application or Innovation in Engineering & Management (IJAIEM)

 