International Journal of Mechanical Engineering

Comparative Analysis on Machine Learning Techniques: A case study on Amazon Product Reviews

Shubhangi V. Urkude
Faculty of Operations & IT, ICFAI Business School (IBS), Hyderabad, The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956), Hyderabad-India.
ushubhu@gmail.com

Hasanuzzaman
Faculty of Operations & IT, ICFAI Business School (IBS), Hyderabad, The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956), Hyderabad-India.
hasantext@gmail.com

Vijaykumar R. Urkude
Department of ECE, Vignan’s Institute of Management and Technology for Women, Ghatkesar, Hyderabad, India.
cool.viju1721@gmil.com

C. Srinivasa Kumar
Department of CSE, Vignan’s Institute of Management and Technology for Women, Ghatkesar, Hyderabad, India.
drcskumar41@gmail.com.

Abstract:

With the digitization of the entire world, e-commerce and online shopping become more popular among the customers. Drastic change in shopping style develops the need of e-commerce sites. This proliferation of customer’s interest to check the particular product review before buying the products leads to the sentiment analysis. Sentiment analysis is the analysis of customer’s opinion using natural language processing. Customer reviews should be analysed properly to provide correct suggestions to the customers. This paper aims to classify Amazon product reviews for electronics parts into two categories as positive and negative by using different machine learning algorithms such as Support Vector Machine (SVM), Naïve Bays (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Stochastic Gradient Descent (SGD). The analysis shows the logistic regression having highest accuracy of 83.89% and decision tree has lowest accuracy of 73.3%.

Keywords: E-commerce, Customer Reviews, Exponential, Sentiment analysis, Natural Language Processing

1. Introduction

Nowadays, more and more people are using the internet and turning towards online shopping. Due to that many e-commerce websites are developed such as Amazon, Flip cart, Myntra and so on. All these sites provide the facility of giving feedback on the product. Customers can share their experience about the product in terms of reviews and ratings. This feedback is utilized by the customers and businesses in order to understand their customer choice and business. Depending on those different strategies are developed to increase the sale and get more profit. These customer’s feedback or reviews are used by the other customers to buy the particular product. So that the customer reviews should be analyzed properly to get insight from it, this process is called as sentiment analysis. Sentiment analysis or opinion mining play a vital role in text mining using natural language processing.

As the businesses are completely switched to online mode, customers are exchanging items through various business sites. Therefore, evaluating items before purchasing is a critical situation. So, breaking down the information from those text reviews and extract the useful data to get benefit is a fundamental field these days. It is enrapturing to comprehend the importance of all the purchasers in the world. Many papers are available on sentimental analysis to separate the customer reviews on several products and fabricate an administered learning model to enrapture enormous measure of reviews. Sentimental analysis or opinion mining is coming under natural language processing. As per survey conducted in 2019 on Amazon more than one third of e-commerce users trust reviews to buy the product. More positive responses on a particular product gives the genuineness of the product and produce more impact on the other buyers. We utilized both manual and dynamic learning approach to mark our datasets better. In the active learning process different classifiers are considered to give precision until coming to palatable level. Different classifiers are used in this area to analyse the customer reviews, such as NB, SVM, DT, LR, SGD, RF, deep learning algorithms and neural networks etc.

The main objective of this paper is to classify Amazon product reviews for electronics products into two categories as positive and negative. This review analysis will help the customer to buy new product. Rest of the paper is organized as follows: the literature review in the sentiment analysis is explained in section 2. All the machine learning algorithms used in the paper are discussed in the section 3. The data collection; preprocessing and model architecture is described in section 4. Whereas result analysis is discussed in section 5 and finally conclusion and future scope is discussed in the last section.

2. Literature Review

Many researchers worked on the sentiment analysis, opinion mining and text mining using different approaches. They have used many techniques to pre-process the data and extract the necessary features to get better results. The sentiment analysis on various product reviews has been extensively used for business analysis and also useful to the customers for online shopping. Tanjim Ul Haque et al. [1], proposed a supervised machine learning model to a large unlabeled review dataset using different approaches of feature extraction. They created a hybrid method by combining different approaches and got good accuracy as compared to the other classifiers. Shangdi Sun et al. [2] proposed a novel architecture for short text product reviews classification. It consists of deep convolutional features from convolution neural network with SVM classifier. Taysir Hassan A. Soliman et. al [3] proposed opinion mining approach to mine ungrammatical and unstructured customer reviews by using feature classification and polarity classification. They got the precision of 93% approximately.

Sanjay Dey et al. [4] compared naive bayes with SVM classifier for Amazon product reviews giving good accuracy related to other models. Ahlam Alrehili [5] proposed sentiment classifiers to differentiate negative and positive feedback using the method of ensemble machine learning. In the ensemble technique they had considered five weak classifiers as naive bayes, support vector machines, random forest, bagging and boosting. Jahanzeb Jabber et al. [6] used SVM machine learning technique to perform opinion analysis on musical instruments and beauty products from Amazon e-commerce website. To tackle the basic problem, sentiment analysis on sentence level categorization and review level categorization was performed. Xing Fang et al. [7] proposed opinion polarity categorization on online reviews from Amazon. They performed detailed analysis to categorize the review data. Pankaj, [8] proposed the sentiment polarity categorization and analysis on the reviews on different smartphones by partitioning them into negative, neutral and positive behaviors. Wanliang Tan et al. [9] proposed, customer review analysis using traditional algorithms like naive bayes, SVM and K-nearest neighbors (KNN) and long short-term memory (LSTM). As per their study LSTM is given more accuracy on test data for one type of feature. Various algorithms used in the product review analysis are discussed in the following section.

3. Supervised Machine Learning Algorithms (SMLAS)

This section described about the various machine learning algorithms used is this paper. SMLA are basically of two types: (i) classification (ii) regression. Regression analysis is performed to give the future predictions such as market trends, time series analysis, weather forecasting and so on. When excepted output is a number the regression analysis will be used. To get the categorical output or your expected output is categorical in nature then we have to go for classification task. Various classification algorithms are available that will divide the given data into binary or multiple classes depending on the application such as customer will take credit card or not, student will pass or fail, dog vs cat and so on. The algorithms used in this paper are discussed below.

3.1. Naive Bayes

Naïve Bayes is a supervised learning algorithm built on the bayes theorem and frequently used for classification task. NB will make simple naïve assumptions that each pair of features is independent of each other and all are equally important, which is never true in real time datasets. There are many variants of NB available in the market in that the Gaussian NB is the most popular one. Bayes theorem find the probability of the future event depending on the probability of the past event. Daniel Jurafsky et al. [10] stated the applications of NB in text categorization and spam detection. They take single observation, extracted the useful features and finally classified into some predefined classes. NB is a generative classifier, based on the input given it will classify to the predefined classes. We have used NB because it is the simplest algorithms and can be trained on small dataset so computation is fast. Bayes theorem is used for decision making and future predictions [11]. Bayes theorem is stated as below:

$P(Y \mid X) = [P(X \mid Y) * P(Y)] / P(X)$

Where:

P(Y|Xi): is the conditional probability/ posterior probability that event Y occurs, if X has occurred.
P(X) and P(Y): probability of X and Y independent of each other.
P(X|Y): the conditional probability that event X occurs, given that Y has occurred.

3.2. Support Vector Machine

SVM is mostly preferred by many researchers for its dual task as regressor and classifier. SVM will also give high accuracy and less computation power. SVM categories the input data points into multiple classes by using hyperplane.

3.3. Stochastic Gradient Descent

SGD take one random input like gradient descent during weight change which is consider as input for entire training data. It is very efficient and fast while handling large data set as compared to gradient descent. Its time complexity is O(pn + kn) where p : dimension of each input, n: number of records, k:number of responses (outputs).

3.4. Decision Tree

A decision tree is often used to visually and explicitly represent decisions. It is looking like a tree in which features/variables are represented as non-leaf nodes and output is represented by leaf node. The feature having highest information gain is decided as root node and shows the maximum similarity between the features.

3.5. Logistic Regression

Logistic Regression is supervised machine learning (ML) method used for classification purpose. LR is preferred for binary classification and it is based on the probability value. The probability above 0.5 is consider as positive and below 0.5 is considered as negative. It is mostly used in financial sector to detect the defaulters.

3.6. Random Forest

Random forest is a supervised learning algorithm. It is depending on concept of bagging, in which input data is divided into random samples and each sample dataset is given to separate learners. It improves the overall performance of the model. It is mostly used due to its simplicity and variety.

4. Methodology

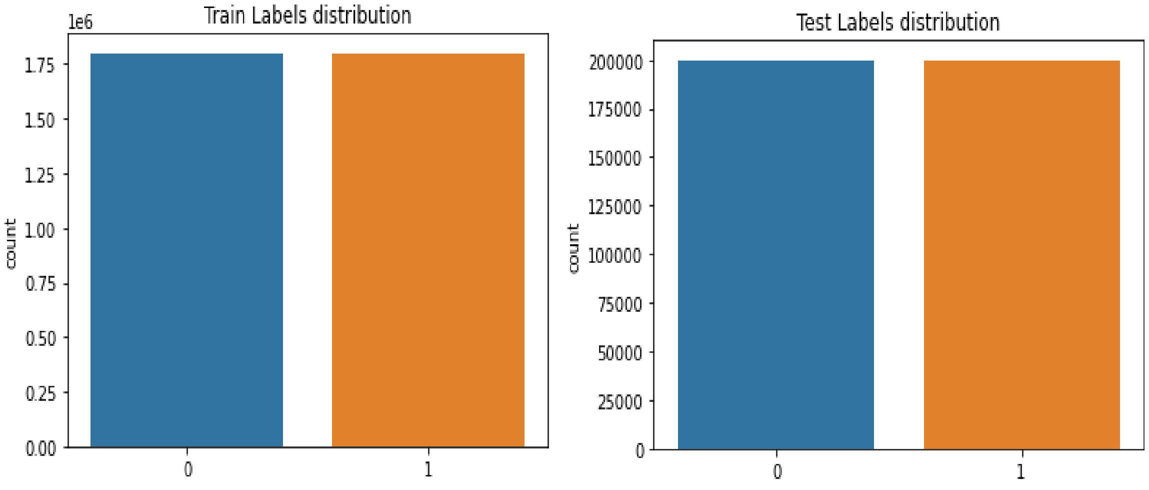
E-commerce websites are very popular among the customers. Depending on the quality of product reviews and ratings will change. This huge amount of reviews should be analyzed properly to increase business profit. The data analyst plays important role in the business to analyze all the reviews, give suggestions, ideas to attract the customers and helps in the business growth.

4.1. Data Collection

The dataset used for opinion analysis is collected from open platform Kaggle.com and consist of customer's reviews of all electronics parts. This is the labelled data consist of two JSON files training and testing dataset separately. Each file is having two columns, the first column is output label and second column is the review text. The training data set consist of 3600000 and testing data set contains 400000 product reviews. The data set is having product ratings from 1 to 5, among that 1 is considered as lowest, 5 is the highest and 3 is taken as neutral rating.

Figure 1(a) and 1(b) shows balanced dataset used in classification. The figure shows that the dataset used is exactly balanced one that means positive and negative reviews are present in 50% of each in training and testing dataset. The dataset is said to be balanced when all the labels are present in almost equal proportion. Balanced dataset will give unbiased result in the analysis.

Figure.1 Balanced dataset (a) For Training label distribution (b) For Testing label distribution



a

b

4.2. Data Pre-processing

In sentiment analysis the output is depend on the various preprocessing techniques applied. We have applied following techniques to preprocess the data.

(A) Lowercasing: Lower casing is the process of converting all uppercase letters into small case for easy processing.
(B) Tokenization: It is very important process of separating the string of alphabets into individual phrases, symbols,

keywords called as tokens. In tokenization unwanted characters are discarded and tokens are given as input to the parsing and text mining process.

(C) Stemming and lemmatization: Stemming and Lemmatization both are similar methods and produce root word after applying some standard techniques. In stemming the unwanted characters are removed that is called as stem and generate base word. Lemmatization is removing unwanted characters to produce meaningful word.

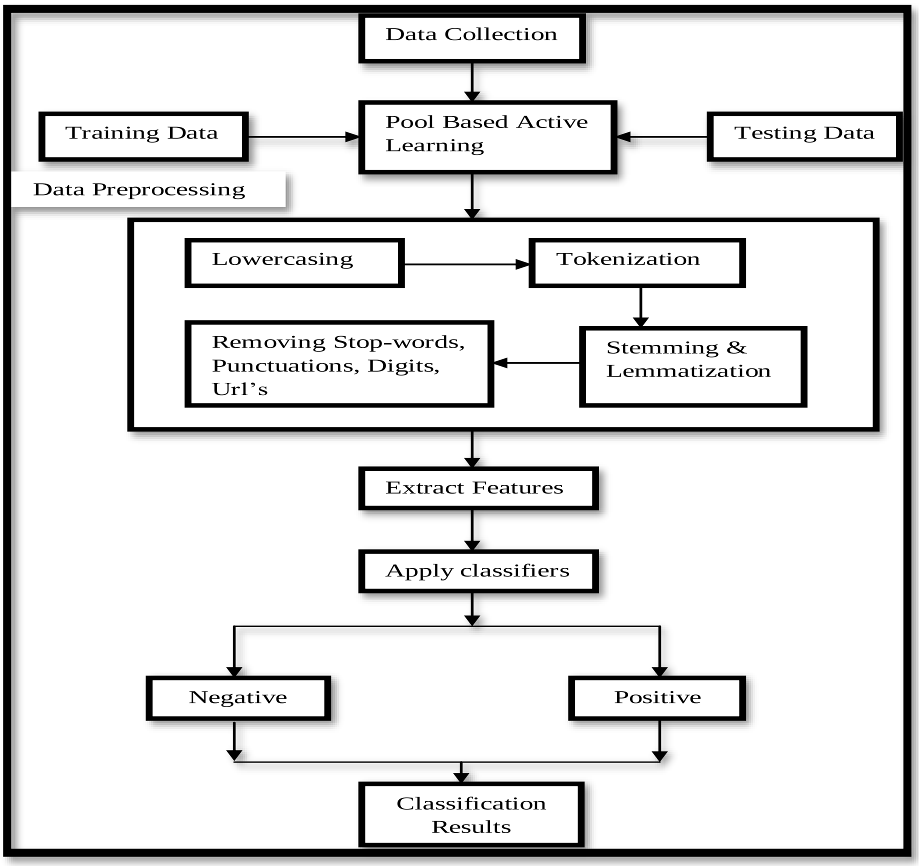
(D) Removing Stop words: Stop words are the objects like preposition, conjunction and common words appears in the sentences. These stop words reduce the accuracy of the model so it has been removed in the data preprocessing.

(E) Removing Punctuations, Digits, URL's: Punctuations symbols, number and the URL are not necessary for mining process. These symbols are used to make the sentence more readable so can be removed in the data preprocessing.

4.3. Work Flow of Model

Figure 2 shows the work flow of model. The first step in the model is data collection after that collected data is divided into training as well as testing data and given to pool based active learning. Pool based active learning (PBAL) is kind of supervised machine learning. It is used in different condition such as (i) when the size of data set is too big or too small (ii) having limited processing power (iii) annotation is time and cost consuming for large data set. PBAL is preferred to enhance the performance of classifier by identifying which label is more beneficial to learn from it. So those inputs are selected actively to train the model. Active learning is used to handle the problem of data labeling by creating expert instance. This unlabeled data is labeled by active learning using oracle instance and that data will be run on some classifier to find the accuracy. If the accuracy is more than 90% then converted data is combined with pre-labeled data and model is tested on the new data. This data is then given to data preprocessing block, in which all preprocessing steps are applied. After extracting the features different classifiers such as, SVM, LG, SGD, RF, DT and NB are applied to classify the positive and negative reviews.

Figure.2 Work flow of model



4.4. Evaluation Measures

Evaluation measures are very important for any classifier. Accuracy is one of the popular evaluation measures. Accuracy of the classifier is defined as the number of inputs classified correctly. Finding the accuracy, we should know some of the general terms that are:

TP (True Positive): gives correctly classified input data

FP (False Positive): gives correctly misclassified input data

FN (False Negative): gives number of incorrect data classified as correct

TN (True Negative): gives correct data classified as incorrect.

(A) Precision**:** Precision denotes number of inputs classified correctly among the total number of inputs.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (1)$$

(B) Recall: Recall is referred as sensitivity of a classifier. Recall is the ratio of total correct classification to the total correct predictions.

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

(C) F1 score: F1 score is used to keep balance between precision and recall. It is the weighted harmonic mean of precision and recall. The harmonic mean is calculated as number of inputs divided by sum of its reciprocals and it will penalize the model.

$$F1\ Score = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}} \qquad (3)$$

(D) Accuracy**:** Accuracy is the most important and widely used measure for classification model. It tells whether the model is trained properly or not.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad (4)$$

5. Result and Discussion

Table.1. Analysis result of Electronics product reviews for all algorithms

Classifier	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Naïve Bayes	82.9	0.8299	0.8291	0.8289
Support Vector Machine	82.6	0.827	0.827	0.8269
Stochastic Gradient Descent	78.1	0.7811	0.781	0.7809
Decision Tree	73.3	0.7337	0.7328	0.7327
Logistic Regression	**83.89**	**0.8391**	**0.8389**	**0.8389**
Random Forest	83.3	0.8336	0.8329	0.8328

Interpretation of table-1.

Table 1 shows the analysis result of electronics product reviews for all the algorithms. According to the analysis done the logistic regression is giving highest accuracy of 83.89%. The other classifier such as random forest is showing 83.3% of accuracy, naïve bayes and support vector machine are having approximately 82.9% & 82.6% accuracy. The decision tree is giving lowest accuracy 73.3% among all.

Figure.3 Accuracy for all algorithms

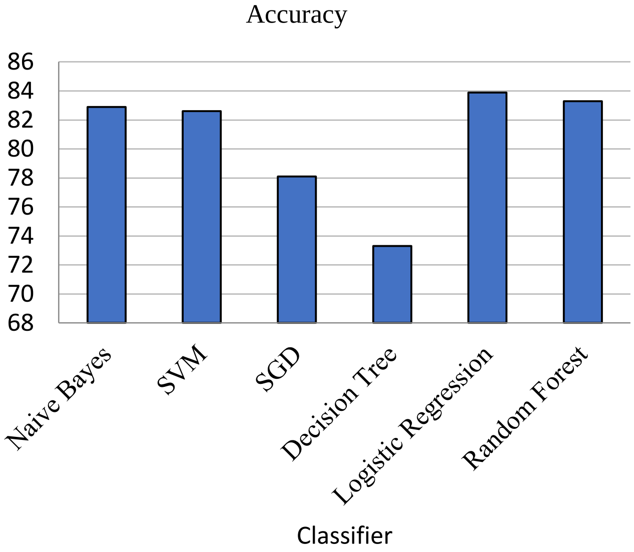
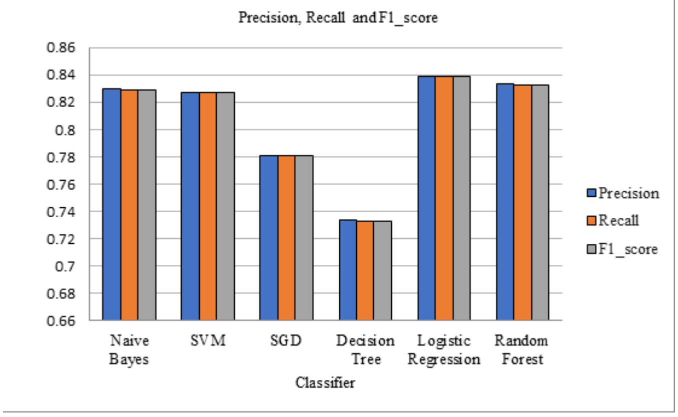


Figure 3 shows the accuracy of all the algorithms, in which x-axis represents all the machine learning algorithms and y-axis represents accuracy in percentages that is found in the analysis. The figure 4 shows the other evaluation metric like precision, recall & f1_score for all the analysed algorithms. In this figure x-axis represents all the machine learning algorithms and y-axis represents precision, recall and f1_score achieved by the entire algorithm between 0 and 1.

Figure.4 Evaluation metric for all algorithms



6. Conclusion and Future Scope

In this paper we proposed classification of electronics product reviews using machine learning. We have taken customer reviews from Amazon website and review dataset is analysed after doing pre-processing. The processed reviews are categories into binary classes as positive and negative with the help of six machine learning algorithms. This analysis shows logistic regression has highest accuracy of 83.89%, random forest has 83.3%, naive bayes has 82.9%, support vector machine has 82.6%, stochastic gradient descent has 78.1% and decision tree has 73.3% accuracy which is the lowest among all the algorithms.

In future this model can be improved further by using neural network and deep learning methodology. It can be used for sentiment analysis of any product. Customers can take benefit of this proposed model to evaluate the product and save their time and money in selecting the correct product. We can also create dynamic web application to classify the review data based on the ratings and find the customer satisfaction about the product.

References

1] Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews", International Conference on Innovative Research and Development, 2018.

2] Shangdi Sun and Xiaodong Gu, "Support Vector Machine Equipped with Deep Convolutional Features for Product Reviews", 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 130-135, 2017.

3] Taysir Hassan A. Soliman, Mostafa A. Elmasry, Abdel Rahman Hedar and M. M. Doss, "Utilizing Support Vector Machines in Mining Online Customer Reviews", 22nd International Conference on Computer Theory and Applications, 192-196, 2012.

4] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey, "A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews", International Conference on Contemporary Computing and Applications, 217-220, 2020.

5] Ahlam Alrehili and Kholood Albalawi, "Sentiment Analysis of Customer Reviews Using Ensemble Method", IEEE International Conference on Computer and Information Sciences, 2019.

6] Jahanzeb Jabber, Iqra Urooj, Wu JunSheng and Naqash Azeem, "Real-time Sentiment Analysis on E-Commerce Application", 16th International Conference on Networking, Sensing and Control, IEEE, 391-396, 2019.

7] Xing Fang and Justin Zhan, "Sentiment analysis using product review data", Journal of Big data, Vol 2, Issue 5, 2015.

8] Pankaj, Prashant Pandey, Muskan and Nitasha Soni, "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, 320-322, 2019.

9] Wanliang Tan, Xinyu Wang and Xinyu Xu, "Sentiment Analysis for Amazon Reviews", 2018.

10] Daniel Jurafsky and James H. Martin, "Speech and Language Processing", September 21, 2021.

11] J K Sharma, "Business statistics solutions and problems", Pearson Education, 2010.