A Study Of Sentiment Classification On Amazon Product Reviews

Akhilesh Mahajan
Department of Computer Science And Engineering
Delhi Technological University
Delhi, India
akhileshmahajan_2k18co040@dtu.ac.in

Massoud Massoudi
Software Engineering Department
Assistant Professor
Parwan University
Charikar,Afghanistan
Ph.D Candidate at DTU, Delhi,India
massoud.massoudi@hotmail.com

Akash Kumar
Dept. Of Computer Science And Engineering
Delhi Technological University
Delhi,India
akashkumar_2k18co038@dtu.ac.in

***Abstract*— In this present technological age and Digitalisation, online products are used for the majority of people. Therefore, it is most important to check a product. The large e-commerce platforms such as Flipkart, Myntra, Amazon, and many others enable their users to review the Products. To buy a commodity, the consumer will examine to have a better quality understanding of the product And product work. The interpretation will be a really simple product polarized into positive, neutral, and negative Product checks. We may use machine learning methods to do this. Sentiment Analysis is research in which consumers are conscious of a product reaction. A kaggle of amazon product reviews gathers the data collection used. We use various Logistic Regression, Naive Bayes, and Random Forest methodology for classifying feedback and achieving the best of precision.**

Keywords—Machine learning, Classification, Linear regression, Naïve Bayes, Random forest, reviews.

I. INTRODUCTION

In this present age of technology and digitalization, everything is going online. People rely on online products from food to cloth and from home to electronics, rather than going outside. Thus e-commerce platforms have raised a lot. Several products are available at these platforms by different brands[1]. Thus, it will be quite difficult to choose a product that is useful and reliable. Though, to get a useful product, a user goes through the reviews of the product, to understand the product and to decide whether to buy it or not [2]. Whenever a person is going online shopping, one of the prior things a user will check is reviews about the product. A user trust more on other people experience and their views. Most of the time, a person buy or cancel a product only based on reviews [3]. Thus, it is clear to show the importance of reviews. Although, it will be quite difficult to go through thousands of reviews whenever a person thinks of buying a product. Thus it will be good to scratch out some useful info from these reviews. Here's where Machine learning comes into play. Sentiment Analysis is a computational technique through which one gets to know about the sentiment or views of a person about a product a thing [4].In principle, it is a classification process that emphasizes that a specific review expresses a neutral, negative, and positive feeling. In this new age of internet evolution, this area of study has become more popular. But these large numbers of feedback have certain drawbacks. First, even, all these online reviews do not guarantee the original product or have a guarantee of it [5].This is because of the reason that fake users can also make fake comments and gives fake opinions. The second limitation is that most of the online reviews and reviewers are unavailable. This sometimes is responsible for the loss of these shopping platforms [6].

We can use sentiment analysis to review the structural product: Understand what customers like and dislike, the contrast between opinion of our products and the opinions of our adversary products, to perceive information about up to date products, and to conserve several hours of physical conversion of hours.

The main motive is to tag the reviews into positive and negative sentiments of the user so that the user can make a quick decision whether to buy a product or not. Yet, several approaches have been made to classify the reviews. Some of the approaches we will discuss in this paper [9].

II. RELATED WORK

In the publication cited in 2002, Pang, Lee, and Vaithyanathan suggested a sentiment categorization using machine learning techniques in the dataset of film reviews. They studied the sentiment analysis on monographs and data bigrams based on the Naïve Bayes, Max Entropy, and Help Vector Machine models. SVM, combined with unigram function extraction, yielded the best results in their experiment. They recorded an accuracy of 82.9% [3]. The Sentiment Classification of Jewelry and Footwear shoe Product Criticism was completed in the 2004 publication of Mulle and Collier. They contrasted the hybrid Support vector machine, Naïve Bayes, Logistic regression, and decision tree methods with the Lemmas and Osgood theory-based feature extraction methods. The support vector machine achieves the foremost results with 86.6 percent accuracy in its study [4]. Elmurngi and Gherbi proposed to detect false movie reviews in the 2017 publication. SMV's, decision book, and knit performance for a corpus with stop words and a corpus without stop words were studied and compared to Naive Bayes. SVM appears highest, with 81.75% and 81.35% precision in both different cases [5].

The experiment with SVM, TF-IDF, combining the Next Word Negation with an Amazon Product Opinion dataset and produced a precision of 89% was performed in 2018 by

Bijoyan Das and Sarit Chakraborty [6]. A relative study of the different machine techniques, morphological-based approaches, and sentimental analysis in film reviews has been carried out by Bhavitha, Rodrigues, and Chiplunkar in 2018. They have produced 74 percent for the Senti-WordNet method and 86.40 percent for the SVM method [7]. A supervised learning technique was proposed to polarise several untagged product opinion datasets for the 2018 IEEE publication, including Tanjim, nudrat, and Faisal. It is a monitored method of learning and uses a combination of two kinds of extractor method. With F1 measurement, precision, and a recovery of 90%, they were able to achieve accuracy of more than 90% [8]. A sentimental analysis of mobile phone reviews classified into positive or negative sentiments in the 2017 publication, Ceenia Singla, Sukhchandan Randhawa, and Sushma Jain. Naïve Bayes, Support Vector Machine, and Decision Tree, three classifiers were used. SVM's predictive accuracy is found to be the best with 81% accuracy [9].

In this paper, we will compare different methodologies to implement Sentiment Analysis and get their accuracies.

III. Procedures in sentiment analysis

There are two main approaches to sentiment analysis. It includes a machine learning technique and a word-book-based technique. The former is relying on mainstream symbolic methods that are used by machine learning techniques for text classification and lexicon-based approaches. Text learning can be categorized based on certain learning techniques, strategies such as learning by realizing cases, root learning, and analogy [2].
It is possible to roughly divide the machine learning approach to supervised learning and unsupervised learning techniques.

A. Machine Learning approach

Machine learning is interpreted as an important bureau of Artificial intelligence that takes action with the use of practicing a code that allows the system to comprehend. This technique makes use of morphological and expressive features. It examines Sentiment analysis as an issue of periodic text categorization in which we have several documentations for priming and categorization. The prototype is instructed to forecast the class mark for the latest example [2].
Some widely used classifiers include the decision tree, the neural tree network, Naïve Bayes, logistic regression, and Support Vector Machine. We carry out these classifiers using supervised and unsupervised learning.

1) Supervised learning method

Supervised learning is one of the techniques of machine learning that uses a labeled data set. These tutoring records consist of, along with expected output, a few input data. Then, using machine learning classifiers, new instances are classified. There are many different types of methodologies, some of which are described in section, that have been introduced and documented.

2) Unsupervised learning method

Supervised learning needs a set of tagged training that is not in every case that boosts the use of unsupervised learning, it is available. Such learning strategies do not require any labeled information. Lowly supervised learning utilizes a large percentage of untagged learning data and a very small division of data that is tagged. Non-monitored learning, however, includes input training devices and no predicted yield values are communicating to them. Some examples of unsupervised learning include collection analysis and expectation-maximization algorithms.

B. Lexicon based approach

The Lexicon Based approach turns on detecting a point of view. Lexicon incorporates a multitude of words of opinions that are known and precompiled. To analyze the text, this viewpoint lexicon can be used. The lexicon-based approach mostly includes three methods, including the dictionary-based method, the corpus-based method, and the manual approach to opinion [2]. This is preferably used with the rest of the two methodologies in combination.

Table1: Different Methodologies used**.**

Approach	Methodologies used
Machine Learning	1)Support Vector Machine (SVM) 2)Decision Trees 3)Neural Networks 4)Naïve Bayes 5)Logistic Regression
Wordbook Approach (lexicon)	1)Dictionary-based approach 2)Manual opinion approach 3) Corpus-based approach

IV. Framework

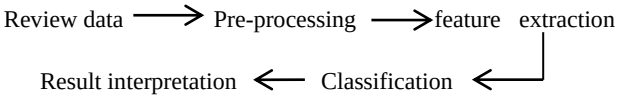
A. Dataset and features

Amazon being one of the finest Electronic commerce websites and we can see a large number of reviews there. We used a dataset from kaggle which has a record of consumer opinions for Amazon products. The dataset includes product ID, rating, review text, and basic product

information. All datasets are in CSV format. We have also used two other datasets of amazon product reviews to balance our dataset as our previous dataset was heavily biased toward positive reviews.

B. Accession

The access is succeeded by the sub-structure put forwarded. To start with, the exploratory data is gathered from kaggle-Amazon product reviews of an e-commerce website. Each record is in the file setup of Comma Separated Values (CSV). In the next step, data is pre-treated, we have removed the null values and we find out that our dataset is strongly biased towards positive reviews using data visualization technique, so we add some more neutral and negative reviews to balance the dataset. Records to detatch stop words, diacritical marks, whitespaces, figures, and special tokens are pre-processed in the next step. In the lower case, the review text is converted, stop words removed, and stemmed. We performed feature extraction using TF-IDF vectorization in the next step. After that, we split the data set into priming and experimenting data to train the model. At last, we classify the priming and experimenting data using Logistic Regression, Naïve Bayes, and Random Forest Classifier. We get the accuracy and classification report by different algorithms.



Framework for sentiment analysis

V. METHODOLOGY

A. Data-collection and Visualisation

First, we collected the dataset from Kaggle - Amazon product reviews which has a record of consumer reviews for Amazon products. We extracted four important features that are relevant for us from the dataset that is (ID, Review text, Review rating). Then we visualize the data and found out the data is heavily biased towards positive reviews so we nixed some neutral and negative reviews to make our dataset balance. Then we did exploratory data analysis to found out the most common words in the text of the reviews.

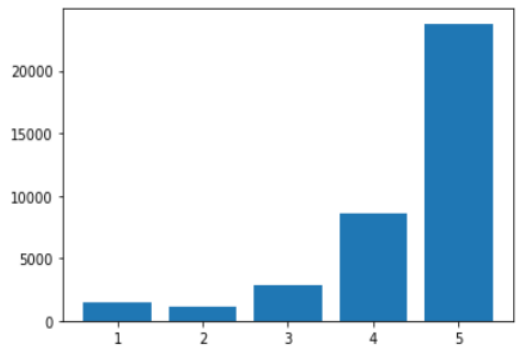


Figure 1: Number of Reviews Vs Rating Visualization.

B. Preprocessing

Tokenization: Tokenization is designated as the technique by which a succession of words is separated into individual words, such as names, keywords, expression, and tokens. Tokens could be phrases, even full sentences of single or individual words. We discard certain characters, like punctuation marks, in the tokenization process. For distinct course of action such as parsing and text mining, the tokens act as input values.

Lower case: Word to lower case (TREE -> tree) conversion.

Cleaning Stop Words:- Stop words are terms in an expression that are not required or needed in text analytics in any field. So, to enhance the efficiency of the study, we mostly disregard these terms. Depending on the country and their culture and language etc., there are various kinds of stop words in different formats. There are some stop words in the English format, so we have to delete them.

Stemming: It is the method of normalizing the word into its root form.

Assigning sentiment scores: We assigned sentiment scores to different ratings. We assigned a score of 0 to reviews having ratings 1 and 2, score 1 to reviews having ratings 3, and score 2 to reviews having ratings 4 and 5 respectively.
Score 0: negative sentiment
Score 1: neutral sentiment
Score 2: positive sentiment

C. Feature Extraction

TF-IDF: It is an analytical metric that evaluates how foremost a term is to a document or record. This is computed by two metrics: how many times a word appears and inverse document frequency of word across the document. Every word or expression has a TF and IDF score of its own. We can accordingly optimize that the unlikely the word, and the other way round, the maximum is the TF*IDF measure. The TF of a term or an expression is the number of times in a text that a term or expression appears. The IDF of a term is the course of action of the value in the corpus of that name. It will always be encompassed in the head search results when terms have a large TF*IDF measure inexpression/name/term/figure, so everyone can:

1. Avoid thinking about stop-words being used,
2. Find terms that are searched for the maximum number of times and lower competition successfully.

We extracted the function that we got from preprocessing from the cleaned text.

VI. CATEGORIZATION

Categorization is the process of analyzing opinions based on their emotions into three divisions: Positive, Neutral, and Negative. 80% of data is used for priming the model and 20% for experimenting.

A. *Classifiers*

We used several classifiers
Logistic regression.
Naive Bayes.
Random forest classifier.

Logistic regression: The probability of a categorical subordinate variable is anticipated by Calculated Relapse. The subordinate encompasses a parallel variable coded as yes or no. Calculated relapse works superior to the large test estimate. The calculated work may be a sigmoid work (as seen in figure 2), which takes any genuine input x and yields esteem between zero and one.

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

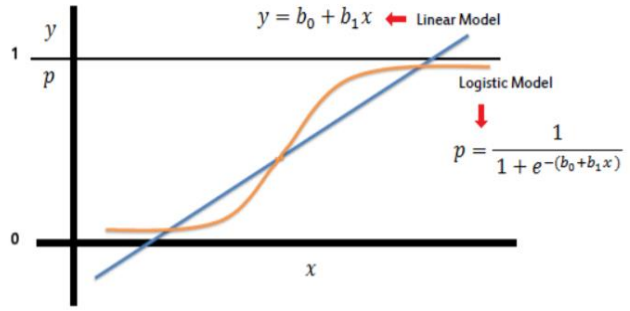


Figure 2: Logistic Regression Graphical Representation.

Naïve Bayes Method: The probabilistic classifier technique is Naïve Bayes. It depends on Bayes theorem There are two different naïve bayes variants for text. Multi-nomial naïve Bayes and benerouolli naïve Bayes. In Multinomial naïve Bayes method data simply follows a multinomial distribution and here each feature value is count. Bernouolli naïve bayes data follows a multivariate distribution and each feature is binary.
Since proof Y is calculated by the Bayes rule by the Finding emotion, the conditional likelihood of event X occurs.

$$P(X/Y) = [P(X)\ P(Y/X)] / P(Y).$$

Random forest: Having supremacy on the top of a single decision tree concerning reliability and efficiency, the random forest classifier was chosen. It is a method of ensemble which is based on bulging. The categorizer works in this way: (as shown in figure 3) given D, the disposer first generates k bootstrap D specimens, with Di denoting each of the specimens. A Di has almost the same proportion of D-rows that are selected with D-substitution. It demonstrates that a few authentic D rows may not be encompassed in Di, whereas other tuples may occur more than once, by sampling with replacement. Depending on each Di, the classifier will then create a decision tree. Consequently, a "forest" is generated consisting of k decision trees. Each tree returns its genre prognosis enumerating as single ballot to identify an unknown tuple, X.

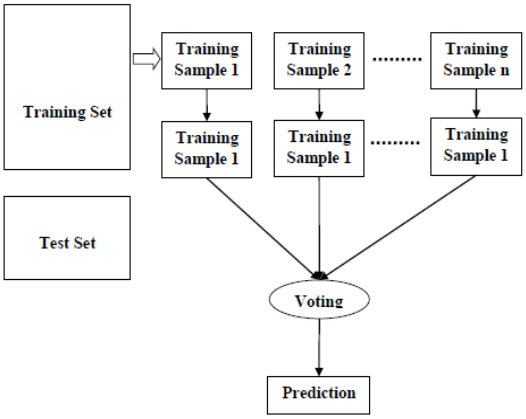


Figure 3: Working of Random Forest.

VII. EVALUATING METRICS

In assessing the classification efficiency of a model, evaluating metrics plays an important role. The most widely used technique for this purpose is precision measurement. The efficiency of a sequencer on a given experimenting dataset is the proportion of those datasets that are properly grouped by the sequencer, and the accuracy measure of the text mining method is often not sufficient to provide proper decision or result, so we have to take some other metrics to assess the output of the classifier. Three other significant indicators widely used are memory, accuracy, and F-measurement. There are some words we need to get acquainted with before addressing various steps.

TP (True Positive) refers to the sample of the proportion of positive instances predicted accurately.
FP (False Positive) refers to the sample of the proportion of positive instances classified inaccurately.
FN (False Negative) refers to the sample of the proportion of negative instances classified inaccurately.
TN (True Negative) applies to the sample of proportion negative instances accurately classified.

Precision: The efficiency of the classifier tests how many numbers of the return records are accurate. Excessive accuracy means low false positives, while nether accuracy means high false positives. It (P) is the proportion of flawlessly classified instance numbers from the total categorized positive instances. It may be evaluated as

$$TP / TP+FP = P$$

Recall: It measures a classifier's sensitivity; how much optimistic data it returns. Excessive remembrance suggests fewer incorrect negatives. A recall is the ratio of the

correctly classified amount of instances to the total amount of expected instances. This can be shown as

$$R = TP / TP+FN$$

F-Measure: Single metrics known as F-measure are generated by combining precision and recall, and that is evaluated as the measured harmonic mean of accuracy and recall. It may be explicated as

$$F = 2P*R / P+R$$

Accuracy: Accuracy or efficiency evaluates how much the proper prediction is made by the classifier. Accuracy is the ratio of the samples of predictions that are accurate to the total samples of predictions.

Accuracy = correct prediction/total data points

Table 2-Classification report of Logistic regression

Sentiment score	precision	recall	F1 score
0	0.79	0.63	0.70
1	0.62	0.35	0.45
2	0.93	0.98	0.96

Table 3-Classification report of Multinomial Naive Bayes

Sentiment score	precision	recall	F1 score
0	0.81	0.25	0.38
1	0.54	0.01	0.02
2	0.87	1.00	0.93

Table 4-Classification report of Bernoulli Naive Bayes

Sentiment score	precision	recall	F1 score
0	0.61	0.48	0.54
1	0.37	0.23	0.28
2	0.91	0.95	0.93

Table 5-Classification report of Random forest

Sentiment score	precision	recall	F1 score
0	0.89	0.68	0.77
1	0.90	0.46	0.60
2	0.94	1.00	0.96

Observations

After experimenting and priming the model with the dataset, their expected efficiency is decided and compared to expose which methodology is best categorizing reviews. As shown using table 5,

The Random Forest model undergoes the finest predictive validity admist the three models and Naïve Bayes has the inferior anticipating precision.

Table 5; CALCULATED ACCURACY

Method	Accuracy
Logistic regression	90.88
Multinomial Naive Bayes	87.09
Bernoulli Naive Bayes	86.46
Random Forest	93.17

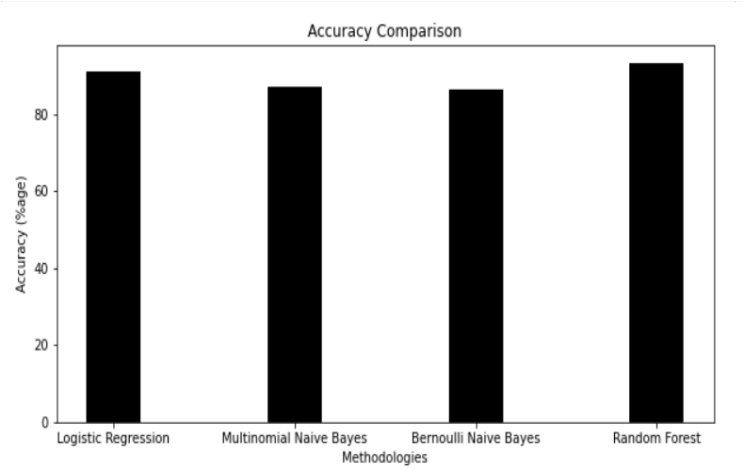


Figure 5: Comparison of accuracies of different methods used.

VII Conclusion and Future Work

A radical change from virtual platforms to digitalized platforms can be seen in a new age. The dependence of clients and consumers on online feedback has increased especially. Digital opinions have enhanced a forum for raising belief and shaping the trends of customer purchasing. By performing an opinion analysis of Amazon product checks and categorizing the opinions into optimistic, neutral, and negative feelings, our project aims to accomplish this. Four classification models were used to identify reviews after combining the data with some neutral and negative opinions. Out of these classifiers, i.e., Multinomial and Bernoulli Naïve Bayes, Logistic Regression, and Random Forest, with 93.17 percent accuracy, the predictive accuracy of Random Forest is found to be the highest.

The work may be extended in the future to estimate the grades of a product from the opinion. This would allow consumers with a dependable grading because the grades

sustained by the product and the emotion of the review often do not fit each one. By combining our dataset more with equal numbers of optimistic, negative, and neutral feedback, we will also get more consistency. The preferred augmentation of work will be very helpful for electronic commerce assiduity as it will improve customer loyalty and confidence.

REFRENCES

[1] Chhaya Chauhan, Smriti Sehgal, "SENTIMENT ANALYSIS ON PRODUCT REVIEWS", International Conference on Computing Communication and Automation(ICCCA2017).

[2] A Study on Sentiment Analysis on Product Reviews, February 2018 DOI:10.1109/ICSNS.2018.8573681 Conference: 2018 International Conference on Soft-computing and Network Security(ICSNS) Anil Singh Parihar, Bhagyanidhi.

[3] Bo Pang and Lillian Lee "Thumbs up? Sentiment Classification using Machine Learning Techniques" In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[4] Tony Mullen and Nigel Collier, Sentiment Analysis using Support Vector Machines with Diverse Information Sources, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004.

[5] Elshrif Elmurngi, Abdelouahed Gherbi, "Detecting Fake reviews through sentiment analysis using machine learning techniques" DATA ANALYTICS 2017: The Sixth International Conference on Data Analytics

[6] Bijoyan Das Sarit Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation" arXiv: 1806.06407, 17 June 2018.

[7] B. Bhavitha, A. P. Rodrigues, N. Chiplunkar "Comparative study of machine learning techniques in sentiment analysis" 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT).

[8] Tanjim Ul Haque Nudrat Nawal Saber Faisal Muhammad Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD) 978-1-5386-5283-1/18/$31.00 ©2018 IEEE.

[9] Zeenia Singla, Sukhchandan Randhawa, Sushma Jain, "Sentiment Analysis of Customer Product Reviews Using Machine Learning" 2017 International Conference on Intelligent Computing and Control (I2C2).

[10] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, pp. 347–354.