# Sentiment Analysis Using Product Based Reviews

Disha Wadhe[1], Rutuja Jangamwar[2], Shweta Narwade[3], Vedika Gawande[4]
Students, Department of Information Technology[1,2,3,4]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India

**Abstract:** *Sentiment analysis is opinion mining in which it uses natural language processing and extracts the reviews in positive, negative and neutral categories. This helps users to identify the emotional tone behind the body of a text. Sentiment analysis computes the user opinion, attitudes towards the product, and the emotions to that product. Some machine learning techniques are used to identify the sentiment for the product. This model tests the reviews using various machine learning algorithms. Logistic regression algorithm has given the highest accuracy as compared to other algorithms.*

**Keywords:** Machine Learning, Sentiment Analysis, Logistic Regression

## I. Introduction

Due to the rapid growth and market of e-marketing technology, progressively buyers prefer to buy on different e-marketing websites. Contrast with the method of offline purchasing in real shops, buyers can shop at anytime and anywhere, in addition to saving time and effort by not having to wait for holidays to go shopping. In addition, the products of the e-marketing websites are rich in variations and also in type, so customers can purchase the things without going anywhere. While electronic shopping enhances customer convenience through the virtuality of e-commerce platforms, there are numerous difficulties with products put on sale on platform like merchandise. Therefore, it is worth nothing to perform sentiment analysis to effectively evaluate the products purchased on the e-commerce platform.

Feedbacks are given by consumers after they utilize the things which give us the merits and demerits of this thing and help further consumers choose the product correctly. Rating helps further customers' experience of the product completely. Reviews with content specifying the merits, demerits, characteristics, responsibilities, etc. Review changes from product class for example, electronic gadget products will supply a variety of knowledge while non electronic gadgets will supply a variety of knowledge.

Sentiment analysis is the most ordinary text categorization tool that analyses arriving information and tells whether the basic sentiment is positive, negative or neutral. Considering consumers' feelings is crucial for businesses since consumers are able to convey their ideas and emotions more freely than ever before. It is difficult for a human to go through each and every review and identify whether a particular review is good or bad. Now with machines, we can impulsively analyze consumers' responses, from survey feedback to social media discussion. Brands are able to listen carefully to their consumers, and customize results and resources to meet their requirements.

Sentiment analysis drops into the Natural Language Processing (NLP) subclass of artificial intelligence (AI). All the feelings are recorded using NLP. It is your responsibility to enable your computer to understand speech at the human level. After all, when it comes to natural language, computers do not have the same natural ability as humans. However the advent of machine learning has helped fill the gap significantly. Nowadays, the amount of information is huge and unstructured, so NLP is becoming more popular. You can use natural language processing to find out trends, popularity etc.

Sentiment analysis has a key function in e-commerce. Almost all websites offer alternatives for consumers to post their report and rating on different aspects of their profession. This feedback from previous customers can be presented in different ways, for instance textual comments under the product being viewed, 5-star scale rating based on the average number of recommendations, graphic interpretation, and product overview. Current buyers can review past comments and make decisions. In this paper, we perform aspect-level sentiment analysis for one sector of such e-commerce sites.

## II. Literature Survey

In this section we will discuss the analysis of methods which are used to perform more accurate sentiment analysis. Nowadays sentiment analysis is bringing more attraction due to complexity in human language. In order to give a better presentation we proposed various combinations of methods. Following methods are operated to classify the sentiment into three groups positive, neutral and negative.

DOI: 10.48175/IJARSCT-3201

Huyen Trang Phan, Van Cuong Tran, Ngoc Thanh Nguyen proposed the methods like creating tweet embeddings this is done using various steps like lexical vector which is drived on syntactic n-grams , it is the extension of n – gram model. Then the next is word-type vector which is based on the POS tag of the word in the tweet. Next is the polarity sentiment vector that is associated with the information such as contrary words, basic emotional words, and fuzzy semantic words. Fourth is the semantic vector it is based on the word implantation. The last is the position vector; it is formed from the position vector; this information is useful for the convolution encoders. The analysis of SENTIMENT OF TWEETS CONTAINING FUZZY SENTIMENT it is done by using CNN model tweet embedding layer, max layer and convolution layer are used to complete the model.

Alaa Noor and Mohrima Islam carried the sentiment analysis using various machine learning algorithms. The primary motive of this research is to find how much correctness we get susing various algorithms. First they performed various processes on the collected dataset like pre-processing, feature extraction, and attribute selection after completing all this processes they had used five different algorithms like Naïve Categorize the analysis as useful or unuseful, the system uses two machine learning algorithms KNN and Naïve Bayes classification algorithms and to stem the review porter stemmer algorithm is used and to compute new rating system uses rule-based extraction method. K Nearest Neighbour will choose the nearest neighbour class to the test analysis and categorize the analysis into two types that is either class = good or class = bad, whereas the Naïve Bayes algorithm uses a probabilistic approach to classify the product into good or bad by selecting the highest probability class label layers, AdaBoost, JRip, J48 algorithm and Sequential Minimal Optimization (SMO) under Support Vector Machines (SVM). Among all the classifiers it is seen that SMO has the highest accuracy among the all five classifiers at 80.87 % and J48 has the lowest accuracy at 71.25 %.

Shihab Elbagir and Jing Yang examined the sentiments using various machine learning algorithms to get the proper sentiment using twitter data. They also gave details of the suggested approach for sentiment analysis. The collected data and then pre-processed the tweets using. In further process they perform the feature extraction using the count vectorizer and term frequency inverse document frequency (TF–IDF). Now the model is trained using various machine learning classifiers and those classifiers are Support Vector Regression, Decision Tree, Random Forest. After all the results show that Decision Tree and Random Forest achieve better correctness as compared to support vector regression. Decision tree gives the highest accuracy that is 91.81%.

Fattesingh Rane, Gaurish Kauthankar , Akhil Naik, Sulaxan Gawas proposed the sentiment model to categorize the analysis into good or bad, the system uses two machine learning algorithms KNN and Naïve Bayes classification algorithms and to stem the review porter stemmer algorithm is used and to calculate new rating system uses rule-based extraction method. K Nearest Neighbour will choose the nearest neighbour class to the test review and categorize the analysis into two classes that is either good or bad, whereas the Naïve Bayes algorithm uses a probabilistic approach to categorize the product into good or bad by choosing the highest probability class label.

Li Yang, (Member, IEEE), Ying Li, Jin Wang (Senior Member, IEEE), and R. Simon Sherratt proposed the model called as SLCABG . They differentiate the sentiment analysis result of the SLCABG model with the ordinary sentiment analysis models on the dataset. Using Naïve Bayes Algorithm, Support vector Machine, CNN and BiGRU. They added the attention mechanism based on the deep learning model can improve the categorization performance of the model. m. First, they used the sentiment lexicon to increase the sentiment features in the analysis. Then the CNN and GRU networks are used to withdraw the leading sentimental and contextual attribute of the analysis, and attention mechanism is used to weight them.

## III. Dataset and Features

We are using Amazon Dataset for our project which is collected from Kaggle.com. It contains item reviews from Amazon. This dataset contains more than 10,000 reviews of instruments in CSV format.

The files contain attributes ‘reviewers ID’, ‘ASIN’, ‘Reviewers Name’, ‘Reviewers Text’, ‘Helpful’, ‘Summary’, ‘Rating’, and ‘Review time’. It demonstrates that there are 5 group rankings from 1 to 5. In comparison, these five classes are potentially imbalanced as class 1 and class2 have limited quantities of data while class 5 has over 6000 ratings.

## IV. Methodology

Sentiment analysis is widely used in Python, an open source tool for detailed mathematical study. Python carries out main tasks of sentiment analysis and gives an optical description of that study. There are five steps to analyze the sentiment data. It involves Data Collection, Data Pre-processing, Visualization, Extracting features from Clean Reviews and Model Building.

DOI: 10.48175/IJARSCT-3201

### 4.1 Data Collection

Data Collection is one of the most important and crucial aspects of the sentiment analysis application. Due to the wide adoption of the machine learning models, simply having a large dataset on a domain specific task does not ensure superior performance. The performance of the model depends on the quality of dataset and labeling/annotation. As ML models learn from the data they are trained with, automatic predictions are likely to mirror the human disagreement identified during annotation. As a result, having a proper guideline to annotate data is also more important.

There are different ways to collect the data some of them are Using API provided by social media platform which allows to collect data in streaming fashion, Using web scrapers that crawl up data and collect specific information, Using a web browser plugin with which users can extract information from any public website with which users can extract information from any public website using HTML and export the data to the desired file format, Using existing open-source repositories of data that are cleaned and compiled which can be used directly. In this project we are going to use the dataset which is downloaded from the Kaggle.com website.

### 4.2. Data Pre-Processing

Data pre- processing involves the transformation of the raw dataset into an understandable format. Pre-processing of the data is a fundamental stage in data mining to improve data efficiency. The data preprocessing method directly affect the outcomes of any analytic algorithm.

### 4.2.1 Handling Missing Values

First step of data pre-processing is data cleaning. Most of the data that we work today are not clean and requires substantial amount of data cleaning. Some have missing values and some have junk data in it. If these missing values and inconsistencies are not handled properly then our model would not give accurate results. Missing values is very crucial when it comes to building a model. It can break your complete model by predicting inaccurately if not handled properly.

Before handling null values or missing values we must know how many null values are there in our dataset. So that we used isnull() function to know it. After that we got there are 27 missing values in reviewerName and 7 missing values in reviewText. To handle this problem, we replaced these null values with “Missing” word.

### 4.2.2 Review Text - Punctuation Cleaning

There are many operations that are being developed using Natural Language Processing. The text is the main input of any type of model like classification, sentiment analysis and many more. So the text contains different symbol and words which does not convey meaning to the model while training. So we will remove them before providing to the model in an efficient way. Model accuracy depends on how well the text is cleaned before training any model.

To remove the punctuation firstly we convert the review to the lower case, then we remove all the punctuations along with the words containing numbers and links that are present in the review.

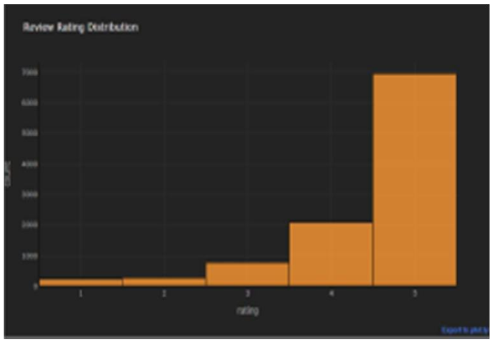
### 4.2.3 Review Text Stop-Words

Stop words are words which does not add too much meaning to the sentence. These words are separated before or after the pre-processing phase of the text because they can generate a lot of noise when applying machine learning to text data. Therefore, take out these unrelated terms from the analysis. Stop words are observed as the crash in the contents. Stop words mainly mention to the bulk of familiar words such as “and”, “the”, “a”, “y”, “any” etc in language. There is no single worldwide list of stop words. The stop word list is subject to change due to issues. For stop words, there is the NLTK toolkit with a predetermined list of stop words corelated to the most similar words.

### 4.3. Visualization

Data visualization is a pictorial presentation of details and data. By utilizing the visual elements such as histogram, bar chart and plan, data measurement gadget give an approachable way to spot and recognize shifts, outliers and design. Today we have a lots of data in our hand that is in the trend of big data, data measurement gadget and technologies are crucial to examine vast quantity of data and make data-based conclusion.

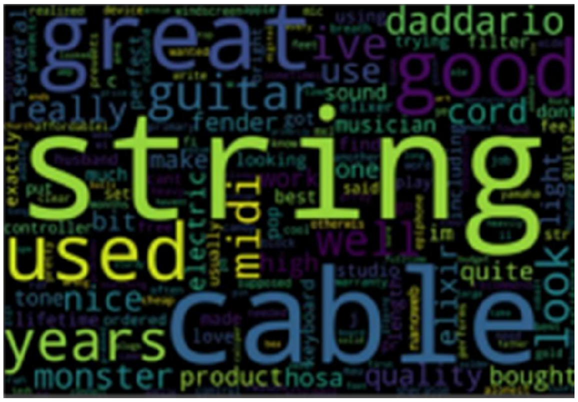
DOI: 10.48175/IJARSCT-3201



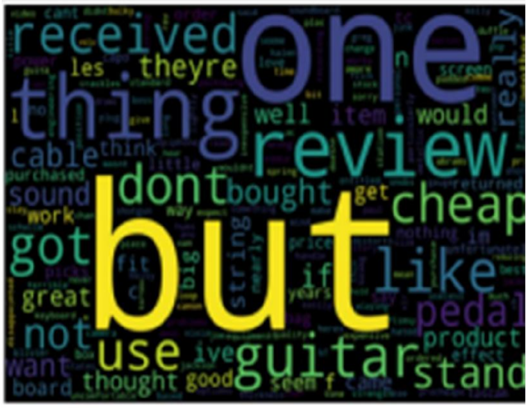
### 4.3.1 Review Rating Distribution

Here, we used histograms for data visualization. A histogram is pictorial presentation of the arrangement of a data. Even though its aspects is alike to that of a quality histogram, rather of creating contrast in the middle of non-identical thing or classification or exhibiting shifts over time, a bar graph is a plan that allow you to express the fundamental frequency distribution or the probability distribution of single continuous numerical variable. Below fig 4.3.1 shows review rating distribution.

Word cloud is a data visualization technique that displays terms in a specific matter on the head chart. Some features of this technique related to this graph are that common or important words are display in large bold and uncommon or unimportant words are display in small or bright typeface. This data visualization technique is very useful in NLP tasks to be analyzed. Fig 4.3.2 show a cloud of positive review words, Fig 4.3.3 show a cloud of negative review words, and Fig 4.3.4 show a cloud of neutral review words.

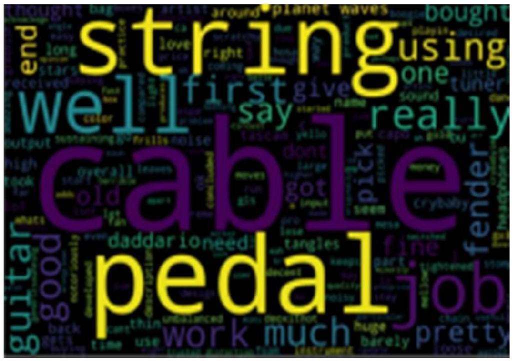


### 4.3.2 Word Cloud - Positive Reviews



DOI: 10.48175/IJARSCT-3201

### 4.3.3 Word Cloud – Negative Reviews



### 4.3.4 Word Cloud – Neutral Reviews

### 4.4. Extracting Features From Cleaned Reviews

Feature extraction is the procedure of modifying original data into digit attributes that can be prepared while keep up the data in the real data set. The feature extraction process is helpful at the moment that you desire to reduce the quantity of resources required for clearing without release crucial or relevant knowledge. Feature extraction can also decrease the quantity of unessential words for a given examine. As well, the depletion of the text and the machine's work in structuring varying combinations resources the rate of studying and initiated steps in machine learning process.

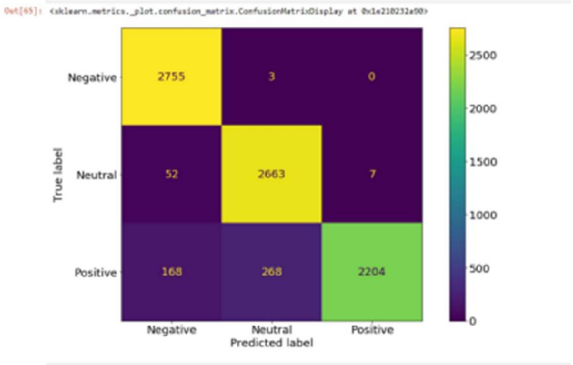
Before developing the model for sentiment analysis, computer cannot understand emotion like words and human so we need to transform the evaluation text into vector format. Here, we used TF-IDF method to convert the text into vector formation. In encoding target variable sentiment, we encode our target variable with label encoder in that 0 indicate negative, 1 indicate neutral and 2 represent positive

Another method is stemming reviews, this is approach of achieve root word from the twisted words. Here we take out the review and change the words in reviews to its main word. For example word going will be convert to go. TF-IDF is a short form for Term Frequency - Inverse Document Frequency. The present method is generally utilized approach in knowledge recapturing and text mining. We found that our objective function has plenty of positive emotions contrast to negative and neutral emotions. Therefore it is necessary to balance the classes in such situation. Now, we used SMOTE to balance out the imbalanced dataset problem where, SMOTE is an abbreviation for Synthetic Minority Oversampling Technique.

### 4.5 Model Building

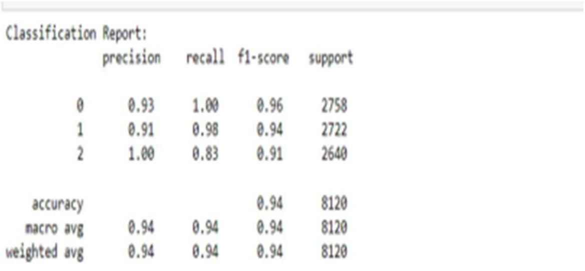
Machine learning models are built by learning from training dataset and generalizing it, applying the knowledge gained to new things making prediction, and fulfilling their goals. A machine learning model is a box that has been instruct to recognize particular type of pattern. It trains the model on the dataset and provides algorithm for the model to use and train.

Here, we train our dataset on various algorithms that are Logistic Regression, KNN, Decision Tree, Support Vector Classifier, Naive Byes. We got highest accuracy using logistic regression i.e. 88%. After Testing logistic regression model on testing dataset we got the accuracy 93%. To know in better way we use classification matrix which gives us the clear idea.



```
Out[65]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1e210232a90>
```

### 4.5.1 Classification Matrix



```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      2758
           1       0.91      0.98      0.94      2722
           2       1.00      0.83      0.91      2640

    accuracy                           0.94      8120
   macro avg       0.94      0.94      0.94      8120
weighted avg       0.94      0.94      0.94      8120
```

### 4.5.2 Classification Report

Fig 4.5.1 shows that sum of diagonal element number are reviews are correctly predicated and other reviews are not predicted correctly.

## V. Conclusion

With the expeditious growth of E-marketing platforming modern time, the sentiment analysis automation of product reviews has obtained more attention. In this paper we develop a processes that can classify reviews into positive, negative and neutral form easily. By make use of our model to examine user reviews, we can help merchants on e-marketing platform to obtain user feedback in time to improve their service quality and attract more customer to patronize.

## References

**[1].** Huizhi Liang, Thomas Thorne "A Dynamic Bayesian Network Approach for Analysis topic-sentiment Evolution" IEEE Access Special Section on Advance Data Mining Methods For Social Computing, volume 8, March 2020.

**[2].** Huyen Trang Phan, Dosam Hwang "Improving the Performance of sentiment analysis of Tweets Contaning Fuzzy Sentiments Using the feature Ensemble Model" IEEE Access, Volume 8, Feb 2020.

**[3].** Sayyed Johar, Samara Mubeen "Sentiment Analysis on Large Scale Amazon Product Review" International Journal of Scientific Research in Computer Science and Engineering Volume8, Issue 1, pp 07-15, February 2020.

**[4].** Li Yang, Jin Wang "Sentiment Analysis for E-Commerce Product Review based on Lexicon and Deep Learning" IEEE Access, Volume 8, Feb 2020.

**[5].** Shihab Elbagir, Jing Yang "Twitter Sentiment Analysis Based on Ordinal Regression" IEEE Access, Volume 7, November 2019.

**[6].** Alaa Noor, Mohrima Islam "Sentiment Analysis for Women's E-Commerce Review using Machine Learning Algorithms" IEEE, Volume 10, July 2019.

**[7].** Fattesingh Rane, Akhil Naik "Online Product Review Classification" International Conference on Advance in Information Technology, 1-7281-3241, 2019.

**[8].** V. Uma Devi, Dr.Vallinayagi V "Product Reviews Sentiment Analysis- A survey" Journal of Emerging Technology and Innovative Research (JETIR) Volume 6, Issue 2, February 2019.

**[9].** Dipak R. Kawade, Dr. Kavita s. oza "Sentiment Analysis: Machine Learning Approach" International Journal of Engineering and Technology volume 9 No 3 June 2017.

**[10].** Sobia Wassan, Xi Chen ,Tian Shen "Amazon product Sentiment Analysis using Machine Learning Techniques" Revista Argentina De ClinicaPsicologica , N.1, 695-703.

**[11].** Xing Fang ,Justin Zhan "Sentiment Analysis Using Product Review Data" Journal of Big Data , December 2015.

DOI: 10.48175/IJARSCT-3201