



# **iJRASET**

International Journal For Research in  
Applied Science and Engineering Technology



---

# **INTERNATIONAL JOURNAL FOR RESEARCH**

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume: 9      Issue: XII      Month of publication: December 2021**

**DOI: <https://doi.org/10.22214/ijraset.2021.39725>**

**[www.ijraset.com](http://www.ijraset.com)**

**Call:  08813907089**

**E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)**

# Data Analysis and Sentiment Analysis on Amazon Reviews

Raj Sinha

Department of Information Technology, BIT Sindri, Dhanbad

**Abstract:** *In the present scenario, a person wants ease in their lives, so E-commerce has become a great and admirable involvement in providing the availability of any product at the doorsteps. But how a person can know the efficiency and originality of the product just by looking at the pictures and the details of the product on the websites. To overcome these issues the E-commerce websites have introduced the concept of the Reviews. Reviews are written by the customers who have already purchased it. Studies show that Product reviews are one of the most important points one considers during the purchasing from E-commerce websites like Flipkart, Snapdeal, Amazon and so on.*

*This paper proposes a model that detects whether the given review is positive, negative, or neutral using the method of sentiment analysis. And using Data Analysis we can find the extension of this paper, we are planning to use a type of sentiment analysis, Opinion Mining which is the research field that predominantly makes automatic systems that will find opinion from the text written in human language. Using opinion mining, we can find whether the given reviews are fake or not. In this paper we have used Amazon food reviews data and based on the rating given by the user we are classifying reviews as positive, negative, or neutral. For positive review ratings given were 4 and 5. For negative review ratings given were 1 and 2. For neutral, rating given was 3. Based on these ratings, we are performing sentiment analysis using Scikit Learn and finding the accuracies of various classification algorithms. We are using Jupyter Notebook for visualization of documents and live coding.*

**Keywords:** *Data analysis, classification algorithms, data visualization, machine learning*

## I. INTRODUCTION

In today's world of online shopping, product reviews play an important role in consumers' online shopping activities. Most people go through these reviews before buying any product online. These reviews can be positive or negative or neutral. Therefore, these product reviews can affect any business and they also have the potential to bring along financial losses or profits. Every day a lot of reviews are posted by the customers to put forward their views regarding the product they have bought. Generally, reviews appear in e-commerce web sites and applications like Amazon, Flipkart, etc. Most of the product reviews are posted by real consumers to express their views and share their shopping experience with other people, but fake reviews also appear in the e-business web sites because of financial reasons. This has affected influenced habits everywhere in the world. Data used in this study are online product reviews collected from Kaggle.com. For example, if a customer writes very negative reviews for a mobile phone say I-phone 11 on an apple review website due to its bad service. This review will present a bad impression of the product to its potential customers and damage its business. After the arrival of internet, people have now started buying products online, thanks to the increasing popularity of the World Wide Web. With more and more people becoming comfortable and finding ease in using the internet, an increasing number of people post their reviews, ratings, and comments online. This posting of reviews is also resulting to be highly beneficial for other users who use online platforms to buy products. Hence, this is resulting in an increased number of reviews and thus its impact on the purchase of goods is also increasing. Popular products get hundreds of reviews that make or break their sales and thus hold large importance. Our application, thus, proposes to separate the reviews as positive and negative in order to guarantee more accuracy in the existing system and enhancing the sales of genuine products based on real reviews. We are doing here is to compare the supervised learning algorithms:

- 1) Multinomial Naive Bayes
- 2) Gaussian Naive Bayes
- 3) Support Vector Machine
- 4) Logistic Regression for classifying the review using a dataset of online review from various E-commerce websites.

Classification in sentiment analysis is done between two classes:

Namely positive or negative.

The sentiments are extracted from the reviews and then classified based on polarity.

Mentions the classification of sentimental analysis in three different levels namely aspect level, sentence level and the document level.

The basis of aspect level is the fact that every review has its own sentiment, and the sentiment analysis distinguishes the sentiment based on specific entities of the aspects.

Similarly, the basis of the sentence level is to put forward the opinion in every sentence. Lastly the document level, finds the sentiment from a document and then categorize that into the class of positive and negative opinions.

This paper aims to classify the database of online reviews based on sentiment using the different approaches for classification. This is the secondary aim (for future work), will be the to categorize these reviews and label them as fake v/s real by the help of all these algorithms and check for the most accurate results. Sample review system using star rating is given below.



Fig. 1 Review / Rating Comparison

## II. BACKGROUND

### A. Statistical Data

Online reviews have certainly taken the consumers to the level of satisfaction they demand while the choice, from hotels, bars, restaurants, online products and many more. In 2015, The market researchers of GFK carried out research with 3,729 Britons for the Competition and Markets Authority (CMA). The results mentioned that over 54 percent of the middle-aged population relied on the online reviews and influence for the annual consumption of about 2133.48 billion Indian rupees.

But the accessibility to these reviews has opened the path for proliferation. There can be positive reviews that are mentioned in order to promote business whereas the negative ones can be there to ruin them written by the competitors. This makes the reviewing platform susceptible to abuse. According to the Best SEO Companies data.

### B. Literature Survey

In this section we are presenting the study of the previous and related work to the field that has used sentiment analysis or any other statistical methods to achieve similar results.

Mentions the matters that are caused due to sentiment analysis. The journal specifically mentions the two issues as the viewpoint and the second one as a way of expressing one's sentiment. The viewpoint signifies that what may be a negative comment to someone it may be a positive comment to others.

The second point signifies that a minor change doesn't change the meaning of a review or comment.

Various websites is depending on the overall rating than the individual review submitted by the purchaser. But there are certain web pages that even allows the text reviews and an elaborated description for the object purchased.

This proposed that the textual reviews are to be categorized on the grounds of their topic or aims. The classification distinguishes the negative versus positive reviews. In general, there are two phases of the Supervised learning method, the first is selecting the review and then the second one is extracting the relevant reviews.

Star Level	General Meaning
★	I hate it.
★★	I don't like it.
★★★	It's okay.
★★★★	I like it.
★★★★★	I love it.

Fig. 2 Meaning of Star Ratings in review system.



### C. Dataset Statistics

The dataset we are using is provided on this link: <https://kaggle.com/snap/amazon-fine-food-reviews/data>

1.	Number of reviews	5,68,454
2.	Number of users	2,56,059
3.	Number of products	74,258
4.	User with > 50 reviews	260
5.	Median Number of words per review	56

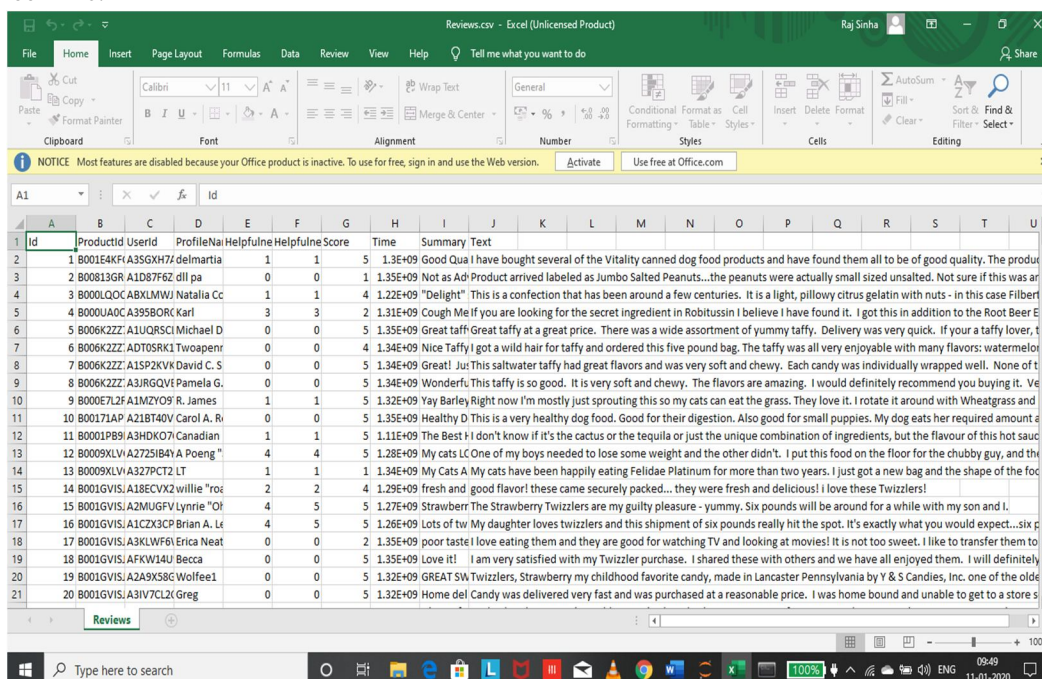
### D. Data Fields Explanation

The Amazon Fine Food Reviews dataset consists of 568,454 food reviews. This dataset consists of a single CSV file, Reviews.csv.

The columns in the table are:

- 1) Id - Unique row number
- 2) Productid - unique identifier for the product
- 3) User Id - unique identifier for the user
- 4) Profile Name
- 5) Helpfulness Numerator - number of users who found the review helpful
- 6) Helpfulness Denominator - number of users who indicated whether they found the review helpful
- 7) Score - rating between 1 and 5
- 8) Time - timestamp for the review
- 9) Summary - brief summary of the review
- 10) Text - text of the review

Sample Records look like:



Id	Productid	Userid	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
1	B001E4KF	A35GH77	delmartia	1	1	5	1.3E+09	Good Qual	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The produ...
2	B00813GR	A1D87F6Z	dll pa	0	0	1	1.35E+09	Not as Adi	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was ar...
3	B000LQOC	A8XLMWJ	Natalia Cc	1	1	4	1.22E+09	"Delight"	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filbert...
4	B000UAOC	A395B0RC	Karl	3	3	2	1.31E+09	Cough Me	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer E...
5	B006K2ZZ	A1U0RSC1	Michael D	0	0	5	1.35E+09	Great taff	Great taff at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, t...
6	B006K2ZZ	ADT0SRK1	Twoapent	0	0	4	1.34E+09	Nice Taffy	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon...
7	B006K2ZZ	A1SP2KV6	David C. S	0	0	5	1.34E+09	Great!	Jus This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of t...
8	B006K2ZZ	A3JRGQV6	Pamela G.	0	0	5	1.34E+09	Wonderfu	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Ve...
9	B000E7L2F	A1MZY09	R. James	1	1	5	1.32E+09	Yay Barley	Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and...
10	B00171AP	A21B740V	Carol A. Ri	0	0	5	1.35E+09	Healthy D	This is a very healthy dog food. Good for their digestion. Also good for small puppies. My dog eats her required amount a...
11	B0001PB9I	A3HDK07	Canadian	1	1	5	1.11E+09	The Best I	I don't know if it's the cactus or the tequila or just the unique combination of ingredients, but the flavour of this hot sauc...
12	B0009XLV4	A2725B4Y	A Poeng	4	4	5	1.28E+09	My cats LC	One of my boys needed to lose some weight and the other didn't. I put this food on the floor for the chubby guy, and the...
13	B0009XLV4	A3Z7PCT2	LT	1	1	1	1.34E+09	My Cats A	My cats have been happily eating Felidae Platinum for more than two years. I just got a new bag and the shape of the foc...
14	B001GVIS	A18ECVX2	willie "roi	2	2	4	1.29E+09	fresh and	good flavor! these came securely packed... they were fresh and delicious! I love these Twizzlers!
15	B001GVIS	A12MUGV9	lynnie "Of	4	5	5	1.27E+09	Strawberri	The Strawberry Twizzlers are my guilty pleasure - yummy. Six pounds will be around for a while with my son and I.
16	B001GVIS	A1C3XCPB	Brian A. L4	4	5	5	1.26E+09	Lots of tw	My daughter loves twizzlers and this shipment of six pounds really hit the spot. It's exactly what you would expect...six p...
17	B001GVIS	A3KLWF61	Erica Neat	0	0	2	1.35E+09	poor taste	I love eating them and they are good for watching TV and looking at movies! It is not too sweet. I like to transfer them to...
18	B001GVIS	AFKW14U	Becca	0	0	5	1.35E+09	Love it!	I am very satisfied with my Twizzler purchase. I shared these with others and we have all enjoyed them. I will definitely...
19	B001GVIS	A2A9X58G	Wolffee1	0	0	5	1.32E+09	GREAT SW	Twizzlers, Strawberry my childhood favorite candy, made in Lancaster Pennsylvania by Y & S Candies, Inc. one of the olde...
20	B001GVIS	A3IV7CL2	Greg	0	0	5	1.32E+09	Home del	Candy was delivered very fast and was purchased at a reasonable price. I was home bound and unable to get to a store s...

Fig 2: Dataset sample records view in MS Excel

### III. IMPLEMENTATION

#### A. Problem Description

Motto of the e-commerce websites of introducing review system is to further improve customer satisfaction and online shopping experience. Thus, these websites allow their customer to put forward their reviews on the products listed or purchased by them. With more internet users, a huge number of people are coming forward to write the reviews and post them on the website which is becoming beneficial for other customers. It also decides the profit or loss for any e-commerce merchant. Thus, it has become very important for the seller to get positive reviews for their products. Due to these things, number of reviews is continuously surging. Now any customer can write any opinion text or review, which can draw the individual's attention before buying that product.

#### B. Methodology

The aim of this paper is to analyse the review data from various websites or e-commerce stores. We will be using weka tool for text classification

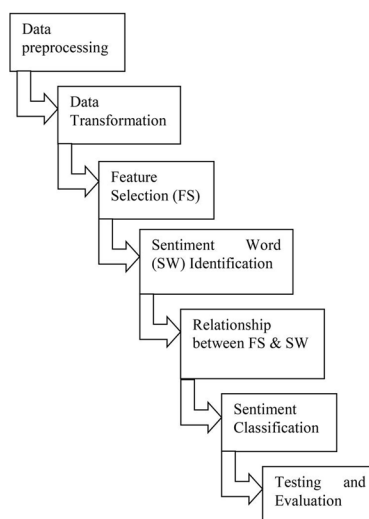


Fig.3 Techniques of sentiment analysis

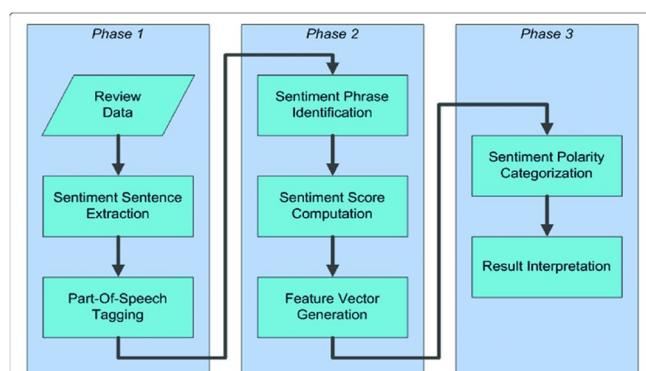


Fig. 4 Sentiment Polarity Categorization Process

We are following a series of steps to achieve the aim of the paper we are submitting: The steps we are involving is diagrammatically mentioned in the figure 3. The textual description for the steps is as follows:

- 1) *Collection of the data*: To analyse the sentiment value of the review posted by consumers on the e-commerce websites we need a standard dataset. So, in this paper we will be relying on the reviews originally collected from website like Amazon.
- 2) *Pre-processing of the Data*: Exploring the figure 1 we can see the two steps for pre-processing the data and converting it into an appropriate sentiment analysis task. Without this process of pre-processing all we will have an inconsistent set of text that would be impossible to analyse. Mentioning the steps within the pre-processing:

- a) *String To Word Vector*: We are using a filter of Word to Vector for preparing the dataset. To perform this, we will be using a text analysis tool called Weka. As the name suggests the String Value to Vector will convert the reviews which are currently in the form of the text to the Positive or Negative for all single words, depending on whether the word appears in the document or not. This filtration process is used for configuring the different steps of the term extraction. This filtration process is again subdivided in two phases namely Configuration of the tokenizer and specifying the list of stop words. The first step converts the content as a set of features and thus making the document classifiable. The second sub-process contains the list of words to filtrate before sending it to the classifier for training it. Some of those words are commonly used (e.g., "a," "the," "of," "I," "you," "it", "and") but has no significance in terms of meaning or labelling, these types of words are known as Stop Words. This on the other hand creates confusion for the classifier. Removing such words also is beneficial in terms of less utilization of the memory.
- b) *Selection of Attribute*: All the attribute of the dataset does not add significance to the result. Some of them has minimal implication in the accuracy so it is better to remove it than to use it. As we cannot risk the performance of the model therefore we need to use an attribute selection scheme.
- 3) *Selection of relevant Features*: As mentioned, not all the attributes of feature are relevant to the performance of the model, And why waste time and space on these irrelevant features. So, instead we take a subset of these features that are more likely to be related to the goals of the model we intend to make. In sentiment analysis five features are mostly used in the task of classifying. With different subsets of features used the results can be different from the other features used.
- 4) *Algorithm of Sentiment Classification*: Sentiment analysis classification have various algorithms which can be applied in various domains like biology, commerce, retails and many more. There are several techniques for classification. Some of them are Naive Bayes, Decision tree, KNN, SVM and genetic algorithm.
- a) *Naïve Bayes (NB)*: Naive Bayes algorithm is a classification algorithm based on Bayes' Theorem. It is a group of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The fundamental Naive Bayes assumption is that each feature makes an:
  - Independent
  - Equal contribution to the outcome.
- b) *Support Vector Machine (SVM)*: A Support Vector Machine (SVM) is a classification algorithm defined by a separating hyperplane or, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
- c) *K-Nearest Neighbor (K-NN)*: K-Nearest Neighbors is also a classification algorithm in Machine Learning. It is a part of the supervised learning domain (works on labelled data) and is used in intense application in pattern recognition, data mining and intrusion detection. It is used in real-life scenarios as it is non-parametric, i.e. it does not make any underlying assumptions about the distribution of data.
- d) *K Star (K\*)*: K\* is an instance-based classifier, i.e. the class of a test instance is based on the class of those training instances similar to it, and determined by the similarity function. It differs from other instance-based learners because it uses an entropy-based distance function.
- e) *Decision Tree (DT-J48)*: It is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is used to display an algorithm that only contains conditional control statements. It is a flowchart like tree structure, here each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. In this paper we are using few of these techniques 1. Multinomial Naive Bayes, 2. Gaussian Naive Bayes and 3. Logistic Regression for identification and achieving the project goal.
- 5) *Taking the Final Decision*: Now we have trained our model by using the above algorithms, so we need to predict the output of it on the test data set. The results that we obtain has following attributes:
  - a) *True Positive*: The review classed as Positive (P) and are originally true as well comes under this category.
  - b) *False Positive*: The review that classifier thinks to be positive (P) but in reality, are not true that is incorrect classification by the classifier.
  - 5.3) *True Negative*: The review classed as negative(N)and are originally true as well comes under this category.
  - c) *False Negative*: The review that classifier thinks to be negative(N) but, are not true that is incorrect classification by the classifier.

- Fake Positive Reviews Rate =  $FP/FP+TN$
- Fake Negative Reviews Rate =  $FN/TP+FN$
- Real Positive Reviews Rate =  $TP/TP+FN$
- Real Negative Reviews Rate =  $TN/TN+FP$
- Accuracy =  $TP+TN/TP+TN+FN+FP$
- Precision =  $TP/TP+FP$

	REAL	FAKE
REAL	True Negative Reviews (TN)	False Positive Reviews (FP)
FAKE	False Negative Reviews (FN)	True Positive Reviews (TP)

[illegible]

2205





## 2206





10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)