

Sentiment Detection for Amazon Product Review

¹Prof. Satish Manje, ²Mr. Tejas Vishe, ³Mr. Kunal Vishe ⁴Mr. Rohan Patil.

¹Asst.Professor, ^{2,3,4}UG Student, ^{1,2,3,4}Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

¹satishmanje93@gmail.com, ²vishetejas13@gmail.com, ³kunal.vishe6@gmail.com, ⁴rohanpatil3715@gmail.com

Abstract - In this paper, the evaluation of the sentiments in the present technological age over the reviews of online products is performed, online products are used by the majority of people. They provide their feedback and then products are recommended for purchase and sale on that factor too. 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. It may use machine learning methods to perform this experiment. Sentiment Analysis is research in which consumers are conscious of a product reaction. A Kaggle of amazon product reviews gathers the data collection used. It uses various Logistic Regression, Naive Bayes, and Random Forest methodology for classifying feedback and achieving the best of precision. Among all the algorithms used it finds that the Random forest machine learning algorithm to be the most accurate.[1]

Keywords- Machine Learning Algorithm, Logistic Regression, Naïve bayes.

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 on these platforms by different brands. Thus, it will be quite difficult to choose a product that is useful and reliable. To get a useful product, a user goes through the reviews of the product, to understand the product and to decide whether to purchase it or not. 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 buys or cancel a product only based on reviews. 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.[1]

II. AIMS AND OBJECTIVE

a) Aim

Healthcare is an unavoidable assignment to be done in human life. Cardiovascular sickness a classification for a variety of infections that affect the heart and veins. Early methods for evaluating cardiovascular diseases helped in making decisions about the progressions that should have occurred in high-chance patients, which decreased their risks.

b) Objective

- The offered study has taken into account informative gathering from Kaggle and does not call for information pre-handling systems like the expulsion of noise data, evacuation of lacking information, or filling default esteems.
- Classification, accuracy, sensitivity, and specificity analyses are used to measure the performance of the diagnosis model.

III. LITERATURE SURVEY

Paper 1: Collaborative attention neural network for multi-domain sentiment classification

Multi-domain sentiment classification is a challenging topic in natural language processing, where data from multiple domains are applied to improve the performance of classification. Recently, it has been demonstrated that attention neural networks exhibit powerful performance in this task. In the present study, This paper propose a collaborative attention neural network (CAN). A self-attention module and domain attention module work together in our approach, where the hidden states generated in the self-attention module are fed into both the domain sub-module and sentiment sub-module in the domain attention module. Compared with other attention neural networks, It uses two types of attention modules to conduct the auxiliary and main sentiment classification tasks. The experimental results showed that CAN outperforms other state-of-the-art sentiment classification approaches in terms

of the overall accuracy based on both English (Amazon) and Chinese (JD) multi-domain sentiment analysis data sets.[3]

Paper 2: Sentiment Analysis on Amazon Product Reviews with Stacked Neural Networks

The project was intended to work on Multi-sense word embedding by sentiment specific bagging of words. Stacked LSTM-GRU along with individual LSTM and GRU was implemented on Amazon Product Reviews (2018) on Products sold under Electronics Category to understand the performance of all the 3 ML algorithms in the task of sentiment classification and analysis of review text.[4]

Paper 3: Feature Selection Based Twin-Support Vector Machine for the Diagnosis of Parkinson’s Disease.With growing number of ageing population, Parkinson's disease has become a serious problem to huge fraction of people above 60. There is no cure for the disease which makes it more difficult when the disease is diagnosed later. For early stages of Parkinson's disease, there are some medications to improve the symptoms. There are certain symptoms like slurred speech, problems in utterances, etc. which are seen earlier. These symptoms can be leveraged to diagnose the disease in its earlier stages. Recently, computers and machine learning algorithms have been widely used in diagnosis of various diseases. Speech

attributes can be analyzed using machine learning algorithms to build predictive models for detection of Parkinson's disease. In this paper, twin-support vector machine (TSVM) based on feature selection technique has been discussed along with other ML techniques for the early diagnosis of Parkinson's disease. . In this paper, twin-support vector machine (TSVM) based on feature selection technique has been discussed along with other ML techniques for the early diagnosis of Parkinson's disease.[5]

IV. EXISTING SYSTEM

People rely on online products from food to cloth and from home to electronics, rather than going outside. Several products are available on these platforms by different brands. Thus, it will be quite difficult to choose a product that is useful and reliable. To get a useful product, a user goes through the reviews of the product, to understand the product and to decide whether to purchase it or not.

In existing system they are using LSTM method to detect and identify sentiments which is given by the reviewer. Due to LSTM the speed of analyzing and detection of sentiments is sometimes complex. It is difficult for accessing the data from huge amount of dataset . That data set used to train the models is derived from Kaggle repositories.

V. COMPARATIVE STUDY

Sr No.	Author	Project Title	Publication	Technology Used	Purpose
1.	Chunyi Yue Hangiang Cao,Guoping Xu & Youli Dong	Collaborative attention neural network for multi domain sentiment	IEEE, 2020	Data Mining,, Decision Tree,, Logistic Model Tree	The purpose of this study was to examine how various decision tree effect the classification accuracies of the data.
2.	Apoorva Mysore Suresha	Sentiment analysis on amazon product reviews with stacked neural networks	IEEE,2020	SVM Wrapper Method	The objective of this study is to predict the life expectancy for patients with hepatitis based on a hepatitis data
3.	Surendrabikram Thapa, Surabhi Adhikari, Awishkar Ghimire	Feature Selection based Twin Support Vector Machine for the Diagnosis of Parkinson’s Disease	IEEE,2020	CART Algorithm	This paper purpose to classify dataset and develop decision support system
4.	Supratim kundu, Swapnajit Chakraborti	A Comparative study of online consumer reviews of apple iphone across Amazon, Twitter and Mouth Shut Platforms	IEEE,2020	Decision Tree, Logistic Regression; Random Forest	This Paper purposes appropriate performance in predicting HBsAg Seroclance.
5.	Afsan Ejaz	Opinion mining approaches on Amazon product reviews :A comparative study	IEEE,2017	Decision tree, genetic algorithm, particle swarm optimization, and multilinear regression models	The purpose of this study was to categorized the dataset and for disease risk prediction

VI. PROBLEM STATEMENT

Customer reviews or ratings aim to define the attitude of the writer towards the product. It may be positive, negative, or neutral. Some people give a product four or five stars and express their final satisfaction with it, and others give a product one or two stars and express their final dissatisfaction with it. This does not present any difficulty in sentiment analysis. However, other people give three stars, although obviously expressing their final satisfaction with it. This leads to confusing other customers, as well as companies, who want to know their actual opinion.[6]

VII. PROPOSED SYSTEM

Here's where Machine learning comes into play. Machine learning and AI has revolutionized everything .The use cases of machine learning have been well explored in fields like healthcare analytics, business, sentiment analysis and so on . 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.

ALGORITHM

1.Logistic Regression

```
# Fitting Logistic Regression
from sklearn.linear_model import LogisticRegression
lg_model = LogisticRegression()
lg_model.fit(X_train, y_train)
# Predicting the Test set Results
y_pred = lg_model.predict(X_test)
```

2.Random Forest

```
# Fitting RandomForest
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier()
rf_model.fit(X_train, y_train)
# Predicting the Test set Results
y_pred = rf_model.predict(X_test)
```

3.Naive Bayes

```
from sklearn.naive_bayes import BernoulliNB
bnb_model = BernoulliNB()
bnb_model.fit(X_train, y_train)
# Predicting the Test set Results
y_pred = bnb_model.predict(X_test)
```

VIII. MATHEMATICAL MODEL

1. Logistic Regression

Logistic Regression builds on the concepts of Linear Regression, where the model produces a linear equation relating the input features(X) to the target variable (Y)[8].

The two major differentiating features of the Logistic Regression algorithm are

1. The target variable is a discrete value (0 or 1) unlike a continuous value, as in the case of Linear Regression, which adds an additional step after calculating output from the linear equation, to get discrete values.

2. The equation built by the model focuses of separating the various discrete values of target — trying to identify a line such that all 1's fall on one side of the line and all 0's on the other.

Consider the following data, with two input features — X1, X2, and one binary (0/1) target feature — Y

Logistic Regression will try to find the optimum values for the parameters w1, w2, and b, such that

$$z = w_1 * x_1 + w_2 * x_2 + b$$

$$y^{\wedge} = H(z)$$

Here, the function H, also known as the activation function, converts the continuous output values of y to a discrete value. This will ensure that the equation is able to output a 1 or 0 similar to the input data. The algorithm finds these optimum values using the following steps:

1. Assign random values to w1, w2, and b.
2. Pick one instance of the data and calculate the continuous output (z)
3. Calculate the discrete output (ŷ) using the activation function H().
4. Calculate loss — Did our assumptions lead us close to a 1, when the actual target was 1?
5. Calculate the gradient for w1, w2 and b — How should it change the parameters to move closer to the actual output?
6. Update w1, w2 and b.
7. Repeat steps 2–6 until convergence.

2.3.2 Support Vector Machine (SVM)

SVM or support vector machine is the classifier that maximizes the margin. The goal of a classifier in our example below is to find a line or (n-1) dimension hyper-plane that separates the two classes present in the n-dimensional space.

Breaking Curse Of Dimensionality:

In 1970, mathematicians Vapnik and Chervonenkis gave the concept of VC dimension where they estimated future testing error(R(α)) as a function of training error and some function of VC dimension (monotonically increasing function).

$$R = (\alpha) = R_{\text{train}}(\alpha) + \sqrt{\frac{f(h)}{N}}$$

$$F(h) = h + h \log(2N) - h \log(h) - c$$

Margin = p

$$\text{Relative Margin} = \frac{p}{D}$$

$$h \leq \min(\{d, \left\lceil \frac{D^2}{p^2} \right\rceil\}) + 1$$

The VC dimension, h, was written as a minimum of the inverse of relative margin square and the dimensions of the data. Hence, if we could maximize the relative margin, we would be minimizing its inverse square, and if that goes below dimensions of data, h will become independent of dimension.

$$g(x) = W^T x + b$$

Maximize k such that :

- $w^T x + b \geq k$ for $d_i = 1$
- $w^T x + b \leq k$ for $d_i = -1$

Value of $g(x)$ depends upon $\| w \|$:

- 1) Keep $\| w \| = 1$ and maximize $g(x)$ or ,
- 2) $g(x) \geq 1$ and minimize $\| w \|^2$

We use the approach 2 and formulate the problem

$$\Phi(w) = \frac{1}{2} w^T w - \text{minimize}$$

Subject to $d_i (w^T x + b) \geq 1 \forall i$

Integrating the constants in Lagrangian form we get:

$$\text{Minimize : } J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i d_i (w^T x_i + b) + \sum_{i=1}^N \alpha_i$$

Subject to : $\alpha_i \geq 0 \forall i$

Method of Lagrange multipliers states that J is minimized for w and b as before, but it has to be maximized for α . The point that J represents is called a saddle point.

The function J currently is represented in its primal form we can convert it into its dual form for the solution.

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i d_i (w^T x_i + b) + \sum_{i=1}^N \alpha_i$$

At the optimum $\frac{\partial J}{\partial w} = 0$ and $\frac{\partial J}{\partial b} = 0$

$$\Rightarrow \omega_0 = \sum_{i=1}^N \alpha_i d_i x_i \text{ and } \sum_{i=1}^N \alpha_i d_i = 0$$

Also, from the KKT condition of the language Multipliers, we can say that all the terms corresponding to the Lagrange Multipliers in the J function should go to 0 .

$$\Rightarrow \alpha_i [d_i (w_0^T x_i + b_0) - 1] = 0$$

$$\Rightarrow \text{either } \alpha_i = 0 \text{ or } d_i (w_0^T x_i + b_0) = 1$$

It implies that non zero Lagrangian coefficients corresponds to the support vector data point using the above equations, we can write j as:

$$J(w, b, \alpha) = \sum_{i=1}^N \alpha_i + \frac{1}{2} w^T w - w^T \sum_{i=1}^N \alpha_i d_i x_i -$$

$$b \sum_{i=1}^N \alpha_i d_i$$

$Q(\alpha)$ represents the dual form J which is only dependent on α as rest are all known scalars.

Decision Tree(DT)

Decision Tree are a non-parametric (fixed number of parameters) Method of supervised learning used for regression and categorization. The objective is to learn straightforward decision rules inferred from the data attributes in order to build a model that predicts the label of a target variable.

The two fundamental components of a decision tree are nodes, which divide the data, and leaves.

,where we got outcome.

We have the following two types of decision trees:-

- Classification decision trees
- Regression decision trees(Continuous data types)

Classification

A given set of data is categorised into classes through the process of classification. Both structured and unstructured data can be used to conduct it. Predicting the class of the

provided data points is the first step in the procedure. Common names for the classes include target, label, and categories..[2]

Decision Tree Classification

In this kind of decision trees, the decision variable is categorical. The above decision tree is an example of classification decision tree. Such a tree is created using a technique called binary recursive partitioning. The data is being partitioned by an iterative method.

Tree Classification:

You shouldn't use a classification algorithm like SVM, K-means, or Naive Bayes for a dataset with random distribution.

SYSTEM ARCHITECTURE

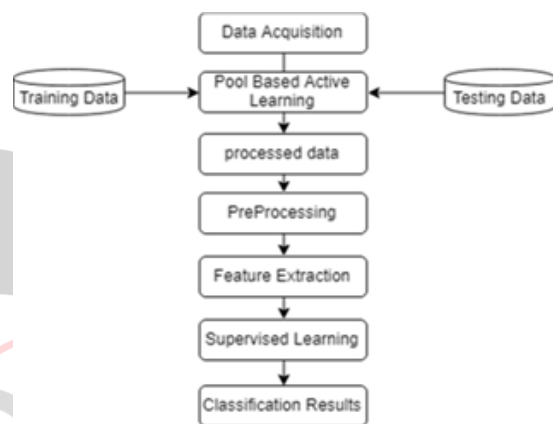


Fig.1: System Architecture

- **Data Collection:** The system collects textual data from various sources, such as social media, customer reviews, or other relevant text data sets.
- **Text Preprocessing:** The collected text data is preprocessed to prepare it for analysis. This step may include tasks such as tokenization, stopword removal and stemming or lemmatization.
- **Feature Extraction:** In this step, relevant features or representations are extracted from the preprocessed text data.
- **Sentiment Classification:** The extracted features are then used as input to a machine learning model, such as a supervised classifier (e.g., logistic regression, support vector machines), to classify the sentiment of the text data into different categories (e.g., positive, negative, neutral).
- **Model Training:** The machine learning model is trained using labeled data, which consists of text data with pre-assigned sentiment labels. The model learns from this data to make predictions on new, unseen text data.
- **Model Evaluation:** The trained model is evaluated using a validation or test set of labeled data to assess its accuracy, precision, recall, F1-score, or other

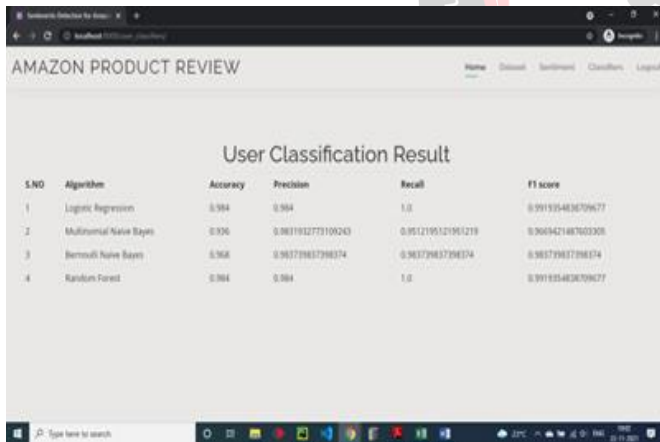
performance metrics. Model performance is iteratively improved by tuning hyperparameters, feature representations, or model architecture.

- **Model Deployment:** Once the sentiment detection model is trained and evaluated, it can be deployed in a production environment to process real-time or batch text data. This may involve integrating the model into a web application, API.
- **Monitoring and Maintenance:** The deployed model may need ongoing monitoring and maintenance to ensure its performance remains accurate and relevant. This may involve retraining the model periodically with updated data or making updates to the system architecture as needed.

IX. ADVANTAGES

- Helps solve complex real-world problems with several constraints.
- Tackle problems like having little or almost no labeled data availability.
- Ease of transferring knowledge from one model to another based on domains and tasks.
- Provides a path towards achieving Artificial General Intelligence some day in the future.
- It automatically detects the important features without any human supervision.
 - In less quantity of data, It can achieve more accuracy.

X. DESIGN DETAILS

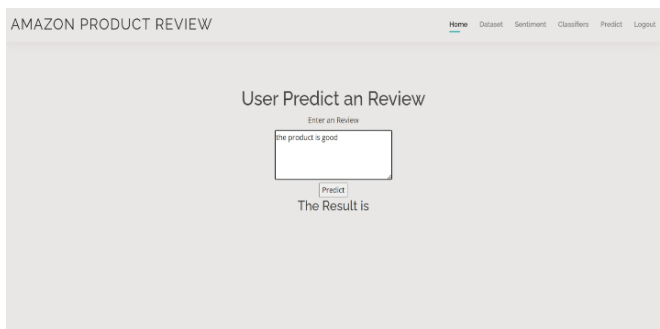


AMAZON PRODUCT REVIEW

User Classification Result

SNO	Algorithm	Accuracy	Precision	Recall	F1 score
1	Logistic Regression	0.964	0.964	1.0	0.9919354838709677
2	Multinomial Naive Bayes	0.936	0.981193277109243	0.9512195121951219	0.9669421487603306
3	Bernoulli Naive Bayes	0.968	0.9837983798374	0.9837983798374	0.9837983798374
4	Random Forest	0.964	0.964	1.0	0.9919354838709677

Fig 2: Classification Result



AMAZON PRODUCT REVIEW

User Predict an Review

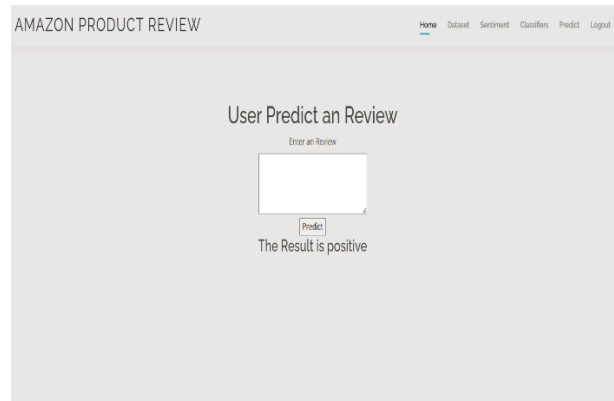
Enter an Review

the product is good

Predict

The Result is

Fig 3. Input



AMAZON PRODUCT REVIEW

User Predict an Review

Enter an Review

Predict

The Result is positive

Fig.3 Result

XIV.CONCLUSION

Thus we have tried to implement “Sentiment detection for amazon product review” by “Bickey sha”, “Amar Jaiswal”. ICCV,2021 and the conclusion is as follows. 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.

XIV.REFERENCE

[1]Sentiment detection for amazon product review by Bickey shah, Amar Jaiswal.ICCC,2021

[2] Prof Vishal R Shinde, “Text Classification on Twitter Data” in IJREAM, ISSN : 2454-9150, Volume07,Issue 02,Special Issue, MAY 2021

[3]Collaborative attention neural network for multi-domain sentiment classification by Chunyi Yue Hangiang Cao,Guoping Xu & Youli Dong. IEEE,2020

[4]Sentiment analysis on amazon product reviews with stacked neural networks by Apoorva Mysore Suresha.IEEE,2020

[5]Feature Selection based Twin Support Vector Machine for the Diagnosis of Parkinson’s Disease by Surendra bikram Thapa, Surabhi Adhikari, Awishkar Ghimire. IEEE,2020.

[6] Prof Vishal R Shinde, “Detection of Suicide Related Posts in Twitter Data Stream” in IJREAM,ISSN:2454-9150, Vol-06, Special Issue, June 2020.

[7] Prof Vishal R Shinde, “Determining Fake Statement Made Public Figures by Means Artificial Intelligences” in IJREAM, ISSN : 2454-9150, Vol-06, Special Issue, June 2020.

[8]Kaur, P., & Jain, K. (2019). Sentiment Analysis of Amazon Product Reviews using Machine Learning Techniques. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-5.