IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis Of Customer Sentiment Through Computational Study Of Product Reviews Using Machine Learning Approach

Preetha D'Souza, Ashitha Naik, Bhargavi Pai, Meghana R, MisikaProfessor, Student, Student, Student
Department of Electronics and Communication, St. Joseph Engineering College, Mangalore, India

Abstract: With the increasing trend of online shopping, it is necessary for companies to know what customers think about their products through reviews and feedback in order to make improvements in the future. If there are a large number of reviews or feedback, handling these manually can become a challenging task to perform, which can be simplified by using sentiment analysis. Opinions can be widely grouped into three categories positive, negative and neutral. In this paper we have considered set of textual English reviews on product from Amazon website. Text present in the review is unstructured in nature, hence we perform pre-processing techniques. Then features are extracted from the pre-processed data using TF-IDF feature extraction method using CountVectorizer library. Sentiment analysis is done using three classification algorithms that is Support Vector Machine, Naive Bayes and Decision Tree considering Accuracy, Precision, Recall and F-Score performance parameters. Upon comparison using parameters, Naïve Bayes was found to be the better classification algorithm for sentiment analysis.

Keywords—Sentiment Analysis, TF-IDF, Support Vector Machine, Naive Bayes, Decision Tree, Accuracy, Precision, Recall,F-Score and CountVectorizer library.

I. INTRODUCTION

In today's era, buyers intend to do online shopping and post comments or review regarding on their buying experiences. Such reviews are important resources for both future decision making customers and retailers to improve their products/goods and services. E-commerce giants like Amazon, Flipkart, etc. provide a platform to consumers to share their experience and provide real insights about the performance of the product to future buyers. To extract valuable insights from a large set of reviews, classification of reviews into positive and negative sentiment is required. Hence there is a need to automate the process of sentiment analysis to ease the tasks of determining public's opinions without having to read millions of reviews manually. This process of analyzing and summarizing the opinions expressed in these huge, opinionated user generated data is usually called Sentiment Analysis**.**

There are several approaches for sentiment analysis: Machine learning based approach (ML) uses several machine learning algorithms (supervised or unsupervised algorithms) to classify data. In this paper, unstructured data of reviews have been extracted from amazon websites. The data has been pre-processed to evaluate the sentiment of reviews using supervised learning, and noisy data has been filtered out. Machine learning classification models like Naïve Bayes, Support Vector Machine (SVM), and Decision Tree have been used to classify the reviews and Finally, the performance is compared in terms of precision, recall, f-score and accuracy.

II. METHODOLOGY

The overall view of our proposed system is portrayed in Fig. 1. This proposed work is done based on Supervised learning and have used Support Vector Machine(SVM), Naive Bayes and Decision Tree algorithms. A Supervised learning is a ML techniques in which the model is trained by initially providing the input and output data sets. Then the model is tested by new set of data called the testset and make it predict the output.

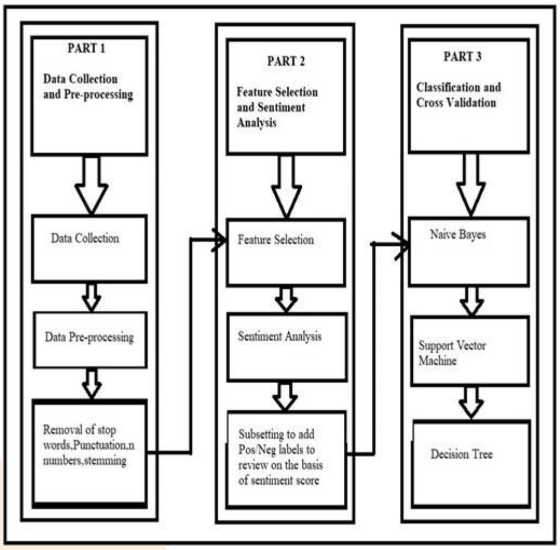

PART 1
Data Collection and Pre-processing
Data Collection
Data Pre-processing
Removal of stop words,Punctuation,n numbers,stemming
PART 2
Feature Selection and Sentiment Analysis
Feature Selection
Sentiment Analysis
Subsetting to add Pos/Neg labels to review on the basis of sentiment score
PART 3
Classification and Cross Validation
Naive Bayes
Support Vector Machine
Decision Tree

Fig. 1 Block Diagram

The whole project was divided into 3 segments: Data collection and Pre-Processing, Feature Extraction and Classification and testingof the algorithm.

3.1 Data Collection and Pre-processing

Here, the data set of 1043 reviews were taken from Kaggle website of the product HP wireless mouse and divided into train set and test set. We have used 835 reviews for train set and 209 reviews for test set. In Pre-processing, involves transforming raw data into a format that is suitable for analysis and modeling. This process involves Case conversions, Stop-words removal, Punctuation removaland Lemmatization.

- Case Conversion: All words are converted either into lower case or upper case to remove the difference between “Text” and“text” for further processing.
- Stop-words Removal: The commonly used words like a, an, the, has, have etc. which carry no meaning i.e., do not help indetermining the sentiment of text while analyzing should be removed from the input text.
- Punctuation Removal: Punctuation marks such as comma or colon often carry no meaning for the textual analysis hence, theycan be removed from input text.
- Lemmatization: Deals with removal of inflectional endings only and to return the base or dictionary form of a word, whichis known as the lemma.

3.2 Features extraction

After the 1st sub process the output of Pre-processed data is taken and feature extraction is done. We also convert textual data into anumerical format that can be used by the machine learning algorithms.

This is done using Count Vectorizer. It transforms a given text into a vector on the basis of the frequency i.e., count of each word that occurs in the entire text. Here, relevant features that are required to classify the review into positive, negative and neutral are found and considered. This is done using Term Frequency - Inverse Document Frequency(TF-IDF).The basic idea behind TF-IDF is to give each term in a document a weight that reflects its importance in the document and across a collection of documents. The weight is calculated by combining two factors:

- Term frequency (TF): This measures how often a term appears in a document. The more often a term appears in a document, the more important it is likely to be.
- Inverse document frequency (IDF): This measures how common or rare a term is across all documents in a collection. Therarer a term is, the more important it is likely to be.
- The TF-IDF score for a term in a document is calculated by multiplying its TF by its IDF. The resulting score reflects howimportant the term is in the document relative to its importance across the collection of documents.

$w(\mathrm{i},\mathrm{j}) = \mathrm{tf}(\mathrm{i},\mathrm{j}) * \log(\quad N$
) (1)

- Where, tf (i, j) = number of occurrence of i in j.df (i) = number of document containing i.
  N= total number of document. $df(i)$

3.3 Classification using different algorithms

1) Support Vector Machine (SVM) : It is one of the popular machine learning algorithm. These are useful in NLPs tasks as they can handle high-dimensional data. The SVM model is trained to classify text to positive, negative and neutral by finding the hyperplane that maximizes the margin between two classes. The hyperplane can be then used to classify new reviews. Consider the pre-processed data, create two files : pos.txt and neg.txt which will contain all the positive and negative words which exist in English dictionary. Find senti-score value .According to this value the reviews are classified as positive or negative or neutral.

2) Naive Bayes : It is a probabilistic model that calculates the probability of a given text belonging to a particular sentiment class such as positive, negative or neutral. In Naive Bayes, each word in the text is treated as a feature, and the model learns the probability of each word occurring in each sentiment class. The probability of the entire text belonging to a particular class is then calculated by multiplying the probabilities of each word occurring in that class. The class with the highest probability is then assigned to the text. Here, it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
It's based in the Bayes's Theorem.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \qquad (2)$$

Where,
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

3) Decision Tree : In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

D. Word Cloud

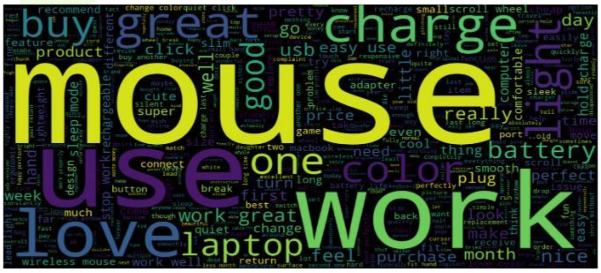


Fig. 2 Word cloud for positive reviews

The above figure i.e. Fig.2 shows the word cloud for positive reviews obtained from the cleaned reviews. Word Cloud is a pictorial representation of commonly used words in a particular dataset. We provided our dataset of cleaned reviews to the model to generate this word cloud. The entire word cloud represents the most frequently used words. The words with a larger font occur more commonly than the words with a smaller font. Similar word cloud can be obtained for negative and neutral reviews.

3.4 Performance Measures

Once a classifier for sentiment analysis is selected, the trained model classifier must be validated using cross fold validation. For a classifier, the possible assessments are true positive (TP), true negative (TN), false positive (FP), false negative (FN) and the performance of the model can be determined using the following measures:

Accuracy: It is the ratio of correctly predicted observations to all the observations. The formula follows as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision: Ratio of predicted positive observations to the total number of positive observation is known as precision. The formulafollows as:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall: Ratio of correctly predicted positive observations to all observations in actual class is known as recall. The formula follows as:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F-Score: Weighted average of recall and precision is called F-Score.The formula follows as:

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

III. RESULTS AND DISCUSSION

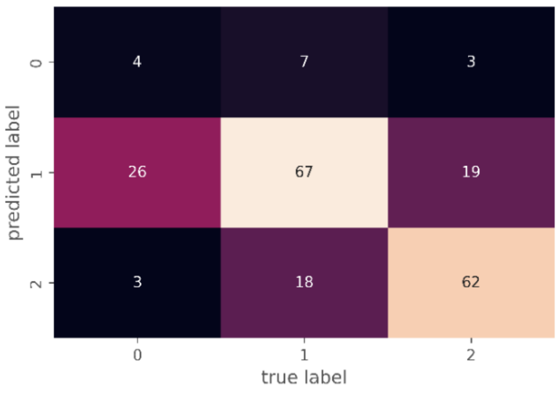
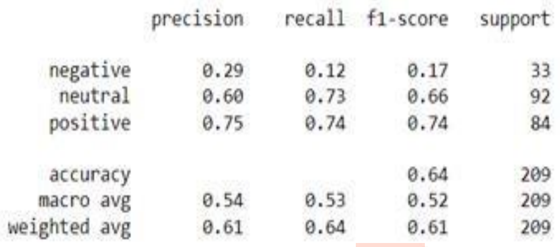


Fig. 3 Confusion Matrix for SVM

The above Fig. 3 depicts the confusion matrix for SVM using TF- IDF Vectorizer. The true label mentions the actual sentiment of the review and the predicted label is the predicted sentiment of the review. The confusion matrix implies that 4+67+62 =133 reviews were predicted with the same sentiment as their true label. It also implies that 7+3+26+19+3+18 =76 reviews were predicted with awrong sentiment.



```
              precision    recall  f1-score   support

    negative       0.29      0.12      0.17        33
     neutral       0.60      0.73      0.66        92
    positive       0.75      0.74      0.74        84

    accuracy                           0.64       209
   macro avg       0.54      0.53      0.52       209
weighted avg       0.61      0.64      0.61       209
```

Fig. 4 SVM Result obtained using four parameters

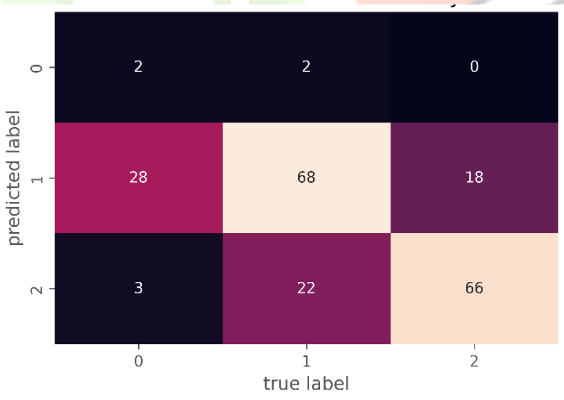
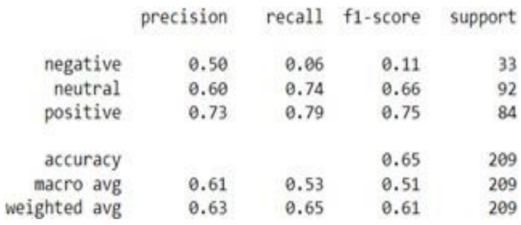


Fig. 5 Confusion Matrix for Naïve Bayes

The above Fig. 5 depicts the confusion matrix for Naïve Bayes using TF- IDF Vectorizer. The confusion matrix implies that 2+68+66
=136 reviews were predicted with the same sentiment as their true label. It also implies that 2+28+18+3+22 =73 reviews were predicted with a wrong sentiment.



```
              precision    recall  f1-score   support

    negative       0.50      0.06      0.11        33
     neutral       0.60      0.74      0.66        92
    positive       0.73      0.79      0.75        84

    accuracy                           0.65       209
   macro avg       0.61      0.53      0.51       209
weighted avg       0.63      0.65      0.61       209
```

Fig. 6 Naïve Bayes Result obtained using four parameters

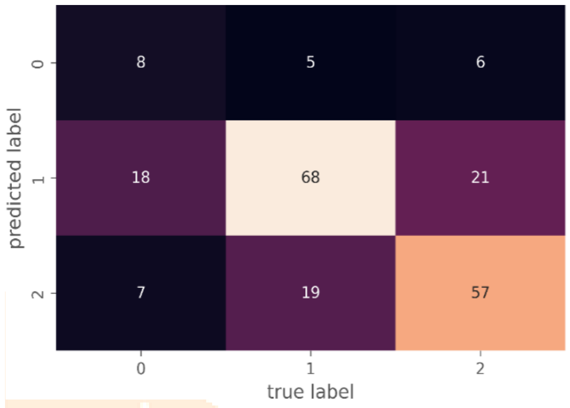
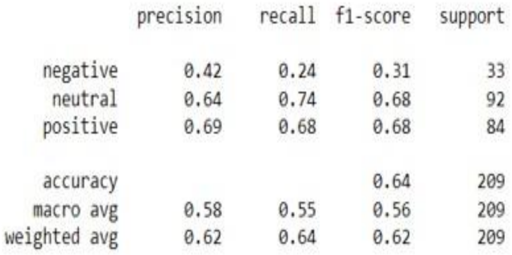


Fig. 7 Confusion Matrix for Decision Tree

The above Fig. 7 depicts the confusion matrix for Decision Tree using TF- IDF Vectorizer. The confusion matrix implies that8+68+57=133 reviews were predicted with the same sentiment as their true label. It also implies that 5+6+18+21+7+19 =76 reviews were predicted with a wrong sentiment.



```
              precision    recall  f1-score   support

    negative       0.42      0.24      0.31        33
     neutral       0.64      0.74      0.68        92
    positive       0.69      0.68      0.68        84

    accuracy                           0.64       209
   macro avg       0.58      0.55      0.56       209
weighted avg       0.62      0.64      0.62       209
```

Fig. 8 Decision Tree Result obtained using four parameters

IV.CONCLUSION AND FUTURE SCOPE

In this project, a total of 1043 textual English reviews were classified into three sentiment categories: positive, negative, and neutral. The pre-processing of the textual English reviews was carried out using Google Colab and Jupyter Notebook. Threepopular machine learning algorithms, SVM, Naïve Bayes and Decision Tree were written and executed for sentiment analysis and their performance was evaluated using key metrics such as accuracy, recall, precision, and F-score.

Comparing accuracy between all the 3 model Naïve Bayes is having 65% accuracy while SVM and decision tree is having 64% accuracy and when we compare precision values we can see that Naïve Bayes is having greater than or equal to 50% precision value for positive, negative and neutral sentiment while decision tree and SVM have less than 50% precision value for negative sentiment hence we can say that Naïve Bayes is having better precision and accuracy value.

Identifying and predicting market trends and make decisions based on market sentiment. It also helps in keeping an eye on the brand’s image. Data from customer feedback can be used to identify areas for improvement. Sentiment analysis can help to extract value and insights from customer feedback data, as well as develop effective customer satisfaction strategies. Observing and analyzing conversations on social media this can help any company plan its future strategies much more effectively.

V. ACKNOWLEDGMENT

We would like to express our sincere gratitude and profound indebtedness to my guide Ms. Preetha D'Souza, Assistant Professor, Department of Electronics and Communication Engineering for her valuable guidance and encouragement in pursuing this project. We are grateful to Dr Dayakshini, Head of the Department, Electronics and Communication Engineering, for her support and encouragement.We wish to express our sincere gratitude to all the Faculty and Technical staff of the Department of Electronics & Communication Engineering, SJEC Mangaluru, for their valuable guidance, help and support.

REFERENCES

[1] R. M. N. and N. , "Sentiment Analysis for Predicting Customer Reviews using a Hybrid Approach," no. IEEE, 2020.

[2] B. S. B. N. K. Reddy and . K. B. Naidu, "Deep Learning for Sentiment Analysis Based on Customer Reviews," no. IEEE,2020.

[3] P. P. Pandey, M. N. S. and M. Rachna, "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews,"no. International Conference on Machine Learning ,Big Data,Cloud and Parallel Computing, 2019.

[4] R. Ahuja, A. C. S. K. S. G. and P. A. , "The Impact of Features Extraction on the Sentiment Analysis," Procedia Computer Science, vol. 152, pp. 341-348, 2019.

[5] S. R. D'Douza and K. Sonawane, "Sentiment Analysis based on multiple reviews by using Machine Learning ," ICCMC,2019.

[6] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets using Machine LearningApproach," no. IC3, 2018.

[7] Z. Singla, S. Randhawa and S. Jain, "Sentiment Analysis of Customer Reviews using Machine Learning," no. I2C2,2017.