An Aspect of Sentimental Analysis: Sentimental analysis on the Reviews of the Amazon Fine Food

Pradeep Angadageri  
Data Analytics  
 Dublin Business School   
 Dublin2, Dublin, Ireland

**ABSTRACT**

The increasing Numbers of the new technologies and digital data, the business trend dependent on the dicisions and operations made by the computer technologies. Sentimental analysis or opinoin mining is a machine learning study of people’s opin**io**n, attitude, sentiments and émotions expressed in the communication[1].

Social media given a ample of Opportunity for the consumer in terms of understanding the product quality by reading and examining the reviews that are expressed by the other consumers regarding the perticular product in the online shopping platform[2]. The shopping platforms like AMAZON.COM provides an option that even makes easier to choose the products, they provide a lable as ‘helpful’ if they find content of the reivew is valuable. This helps both consumers and manufacturer in effecient of the products. This study deals with building the automated text classifiaction system to predict the helpfullness of the product irrespective of the time that the review is posted. Sometimes the reivews which are uploaded in recent times which gets lower rating goes down the list and higher voted reviews come to radars first. The data is collected from the Amazon.com the consisting the reviews of the fine food. In this classifiers used for the text-classification are Support Vector Machine, Logistic Regression and Naïves Bayes are compared with the Decision tree(Extremely Random Tree)[3]. The content based features like reivew polarity, review subjectivity, review charecters, word count and summary charecter count are signifiant predictors of the review helpfullness.

**Key Words : Machine Learning, Data Processing, Sentimental Analysis**

**INTRODUCTION**

A great revolution has began in last few years with developments of the new technologies that are helps in improving the market value of their company through their products values[2]. The new computers using for market analysis and decision making by using the data analytics. There are many ways to get a opinion from the user like from the social media as like Facebook, Twitter, LinkedIn, apart from the Amazon we can still get the reviews from the user from the social media. Social media also helps consumers to receive the better accessibility to the product information. Studies infer that the customers shows more trust towards the online consumers reviews than the vendors.

Consumers can judge the products by checking the online reviews given by the other consumers. In earlier these reviews are not organized, so sometimes it leads customers to an uneven situations. To overcome this situation Amazon comes up with ‘helpful’ field, that along with the reviews it lets the user to give the rating regarding the information available on the product is helpful or not.

**Problem Definition**

The goal of this paper is to develop accurate text classification that can predicts the accurate helpfulness of the reviews of the consumers. The problem is to perform the binary classification by using the combination of the text features and machine learning algorithms. The binary class will be defined as ‘1’ will be helpful and ‘0’ will be not helpful. The helpful is defined with the number of the user who voted the reviews are helpful. The algorithms used in classification are Support vector machine, logistic regression, Multinomial naïve bayes, Extremely randomized tree.

**Literature Survey**

Several algorithms are applied om sentimental analysis in the fast few years. Peter D. Turney proposed an unsupervised approach of SO-PMI-IR to categorize reviews as thumbs-up (positive) or thumbs-down (negative). The observed accuracy varied from 66% for movie reviews to 84% for automobile reviews. V K Singh also used SO-PMI-IR to mine the students’ opinion regarding different subjects by collecting feedback from them in textual format.

Aggarawal C.C. [4] used the sentiWordNet lexical resources on a set of the 300 dataset. This technique provides the 61% accuracy. P.V. Rajeev et al. [5] and Esha Tyagi et al. [6] used the naïve bayes and support vector machine for the feature extraction of the reviews on the online reviews of smart phones products as a database. The proposed system gave the classification[6] based on the star based rating helps in picking the right products. The use of the support vector machine that give the 89.98% accuracy. The expected model works only for one product at a time.

H. Alasmari [7] proposed the sentimental analysis on the online products on the Amazon by using the tableau and Python. C.Fry [8] focused on the clustering, an unsupervised machine learning technique, then group them with similar topics. They compared with the different algorithms K-means and peak-searching.

Marium Nafees [9].Santhosh Kumar KL [10], he used both naïve bayes, SVM and Logistic regression on reviews from twitter and Amazon respectively. The technology used are good to conclude Sentimental analysis, but SVM is the best approach to attain utmost precision of the polarity division.

It has been estimated that the ‘Helpfulness Voting’ system brings in Amazon.com about $2.7 billion in additional revenue[14]. It is imperative for any business organization to perceive which factors determine the helpfulness of the online reviews. This can help the online business managers increase the revenue by designing and implementing an automated system for classification of the vast quantity of the online consumer reviews data.

**METHODOLOGY**

**(i).Machine Learning**

Training the machine is basic requirement for the sentimental analysis. Along the lexicons the machine learning most used the technology for the machine learning[10]. There are many algorithms are used in sentimental analysis, some are supervised and some are unsupervised machine learning.

**(ii). Data Acquisition**

The Internet is good source of the digital data, this is why it is becomes easy to collect the data for the analysis. Most of the user opinion found in the social media.[10] In this study Amazon fine food dataset is used. The dataset data span period of more than 10 years which includes the more than 500,000 reviews up to October 2012. The reviews include product and user information. It also include reviews from all other Amazon categories.

Dataset Details:

Number of reviews: 568,454

Number of users: 256,059

Number of products: 74,258

Timespan: Oct 1999 - Oct 2012

Number of Attributes/Columns in data: 10

**(iii).Data Preprocessing**

Data preprocessing is necessary for the data analysis. We create the dataset sometimes that contain many garbage data[12]. So we need to clean the data before applying any machine learning models. Preprocessing the data includes removing the numbers, punctuations, images, lemmatizing or stemming the words.

**Diagram

Description automatically generated**

**Index for Methodology :**

Collection and Loading of data

Exploratory data analysis

-Deduplication

-General analysis

Preprocessing

-Preprocessing review text

Vectorization

Applying algorithm with below steps

**1)Collection and Loading of data :-**

Before starting the actual machine learning implementation of any algorithm, we need to get the dataset. For my project that is sentimental analysis of amazon fine food dataset i have collected dataset from the popular site Kaggle. The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon. After loading the dataset found below observation from the dataset. In order to load the data, We have used the SQLITE dataset as it is easier to query the data and visualize the data efficiently.

**Objective :**

Given a review, determine whether the review is positive (rating of 4 or 5) or negative (rating of 1or 2).

**2)Exploratory data analysis**

Data Cleaning - Deduplication

It is observed (as shown in the table below) that the reviews data had many duplicate entries[15]. Hence it was necessary to remove duplicates in order to get unbiased results for the analysis of the data. For example if same person has reviewed a same product for multiple times, we need to remove it.

-Remove all the entries where Helpfulness Numerator is greater than Helpfulness Denominator because Helpfulness Numerator cannot be greater than Helpfulness Denominator.

**General Analysis :**

In this section you can do general analysis of the dataset like we can find number of rows, number of columns in the dataset. We can also find unique values in each of the attribute, We can analyze meaning of each of the attribute Preprocessing - Preprocessing review text Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags

2. Remove any punctuations or limited set of special characters like , or . or # etc.

3. Check if the word is made up of English letters and is not alpha-numeric.

4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)

5. Convert the word to lowercase

6. Remove Stop words

7. Finally Snowball Stemming the word (it was observed to be better than Porter Stemming)

**Vectorization :-**

-Here we will be using four different techniques for text vectorization and we will be applying machine learning algorithm on each of the preprocessed text format

-Here our main task is

-Conversion of review text into vector using bag of word technique

-Conversion of review text into vector using TFIDF

-Conversion of review text into vector using word to VEC

-Conversion of review text into vector using TFIDF word to VECT

**Applying Machine Learning Algorithm :-**

-Here we will be using two versions of SVM algorithm

-Linear kernal

-RBF kernal

When you are working with linear kernel, use SGD Classifier’ with hinge loss because it is computationally less expensive. When you are working with ‘SGD Classifier’ with hinge loss and trying to find the AUC score, you would have to use Calibrated ClassifierCV. when you are working with RBF kernel It is better to reduce the number of dimensions.

**Hyper Parameter tuning :**

In the SVM algorithm we have two variables for which we can perform hyper parameter tuning that is alpha value and penalty value. here our task is to find the best alpha value and penalty value which will give the maximum AUC value. We can use k-fold cross validation or simple cross validation of data for finding best value for the variables.

**Feature importance :**

It is important to print the best features or top important feature. Here for the demo purpose we can show top 10 important features for bow and TFIDF techniques.

**Feature Engineering :-**

To increase the performance of your model, you can also experiment with feature engineering like: Taking length of reviews as another feature, Considering some features from review summary as well.

**Representation of results**

It is important to represent the result in the correct format with proper confusion matrix using seaborn heatmaps. Also we can include ROC curves and AUC curves.

**CONCLUSION**

As per study on the sentimental analysis, it can be seen that variety of the tasks are done till now[13]. In this research we have represented the helpfulness in the reviews that are given by the user that helps for the other users in buying the product, or else which products got ratings that product can be checked efficiency.

ACKNOWLEDGMENTS

Sentiment analysis is the process of detecting positive or negative sentiment in text. In this paper we are detecting sentiments in the reviews of the amazon fine food that which are helpful for the other customers[13]. We are applying the more than 2 machine learning algorithms to get the accuracy and compare between them with highest accuracy we will deploy the model. In this project we will take some survey or a poll to get to know more about how the customer react to the product which are bad or good quality.

**REFERENCES**

[1] K. S. Madhu, B. C. Reddy, C. Damarukanadhan, M. Polireddy and N. Ravinder, "Real Time Sentimental Analysis on Twitter," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1030-1034, doi: 10.1109/ICICT50816.2021.9358772.

[2] Salehan, M. and Kim, D.J., 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems, 81, pp.30-40.

[3] V. D. Kaur, "Sentimental Analysis of Book Reviews using Unsupervised Semantic Orientation and Supervised Machine Learning Approaches," 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), 2018, pp. 519-524, doi: 10.1109/ICGCIoT.2018.8753089.

[4] Aggarwal, C.C. and Zhai, C., 2012. A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.

[5] P. Venkata Rajeev and V. SmrithiRekha, "Recommending Products to Customers using Opinion Mining of Online Product Reviews and Features", International Conference on Circuit Power and computing technologies, 2015.

[6] Esha Tyagi and Arvind Kumar Sharma, Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm, vol. 10, no. 35, September 2017, [online] Available: 10.17485/ijst/2017/v10i35/118965.

[7] H. Alasmari, "Sentimental Visualization: Semantic Analysis of Online Product Reviews Using Python and Tableau," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 1-3, doi: 10.1109/BigData50022.2020.9391769.

[8] C. Fry and S. Manna, "Can We Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms," 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), 2016, pp. 374-377, doi: 10.1109/ICSC.2016.71.

[9] Marium Nafees et al., Sentiment Analysis of Polarity in Product Reviews In Social Media 10.1109/ICET.2018.8603585.

[10] K L Santhosh Kumar, Jayanti Desai and Jharna Majumdar, "Opinion Mining and Sentiment Analysis on Online Customer Review", 978-1-5090-0612-0/16/\$31.00 ©2016 IEEE.

[11] C. Nanda, M. Dua and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, pp. 1069-1072, doi: 10.1109/ICCSP.2018.8524223.

[12] P. K. Chaki, I. Hossain, P. R. Chanda and S. Anirban, "An Aspect of Sentiment Analysis: Sentimental Noun with Dual Sentimental Words Analysis," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), 2017, pp. 1242-1246, doi: 10.1109/CTCEEC.2017.8455159.

[13] S. K. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016, pp. 566-569, doi: 10.1109/ICRITO.2016.7785019.

[14]Cao, Q., Duan, W. and Gan, Q., 2011. Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. Decision Support Systems, 50(2), pp.511-521.

[15] G. Satyanarayana, J. Bhuvana and M. Balamurugan, "Sentimental Analysis on voice using AWS Comprehend," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-4, doi: 10.1109/ICCCI48352.2020.9104105.