INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS
VOLUME XII, ISSUE IX, NOV. 2019, WWW.IJCEA.COM, ISSN 2321-3469

# STUDY OF MACHINE LEARNING CLASSIFIERS FOR SENTIMENT PREDICTION

**Y. Sri Lalitha, Y. Gayathri, Althaf Hussain Basha Sk**

[1,2] Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India
[3]Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, Andhra Pradesh, India

**ABSTRACT:**

***Product Review Analysis has developed into a crucial application for all businesses. This will give the company the chance to examine customer product reviews and learn what the market thinks of their goods. It necessitates a comprehensive computational analysis of the behaviour of discrete entities with regard to consumer purchasing similarity and the extraction of the customer's perspective on the business entity. Customer satisfaction is the constant yardstick by which corporate performance is judged. In this newly emerging era of e-commerce and social networking, the introduction of a new product requires a thorough examination of consumer opinions on current products and their needs in the product. Since so many reviews are being produced from different sources, it is becoming more and more challenging. The issue of categorizing reviews into positive and negative opinion is addressed in this study. The work presented here used Naive Bayes, Stochastic Gradient Decent, Random Forest, Multinomial, and Logistic Regression techniques to analyze the product reviews.***

## [1] INTRODUCTION

Opinion Prediction is a subset of data mining that uses natural language processing (NLP), computational linguistics, and text analysis to collect and analyze subjective data from the Web, primarily from social media and related sources, to gauge the propensity of people's opinions. The studied data quantifies public opinions or responses to particular services, individuals, or concepts and reveals the contextual polarity of the information. Sentiment Analysis is another name for opinion mining.

Today’s classifier-based Opinion analysis systems can reliably handle massive amounts of end user opinions, consistently and accurately. When used in conjunction with text analytics, sentiment analysis shows the user's viewpoint on variety of subjects, including your goods and services to your locality, your marketing, and even your rivals. In addition to polarity of product, opinion mining can extract the information about the user, product and users opinion from the text.

## [2] LITERATURE SURVEY

The main problem with opinion analysis is opinion polarity or categorisation. The challenge is deciding whether to classify a review as favourable, negative, or neutral based on its sentiment. There are three measures of opinion polarity differentiation, depending on the extent of the review: the level of documentation, the level of the phrase, and the entity and aspect level[1,3-6]. The concern of the document level is the overall analysis of a piece of writing to determine whether it communicates negative or positive sentiment; meanwhile, the sentence level deals with the sentiment coding of each individual sentence. Finding out precisely what people are into or not from their opinions is the key concern of the entity level. In terms of grading, opinion prediction is fundamentally an issue. Opinion prediction calls for features that involve perceptions to be recognised or identified prior to classification[11,13]. Sentiment categorization is fundamentally a grading problem, where elements that involve perceptions must first be detected or identified.

The field of Opinion study is widely used in Recommendation Systems[7-9]. For being a direct business use-case this is an area of interest for researchers with ever demanding optimized models. Lot of researchers are contributing to this field. In [12] KNN, Apriori methods were employed and evaluated to detect Music Polarity in people. In [5] Instinctively categorizing the themes based on NaiveBayes, ML Models and HMM classifiers are applied on tweet data.

Figure 1 depict the Process of this work.

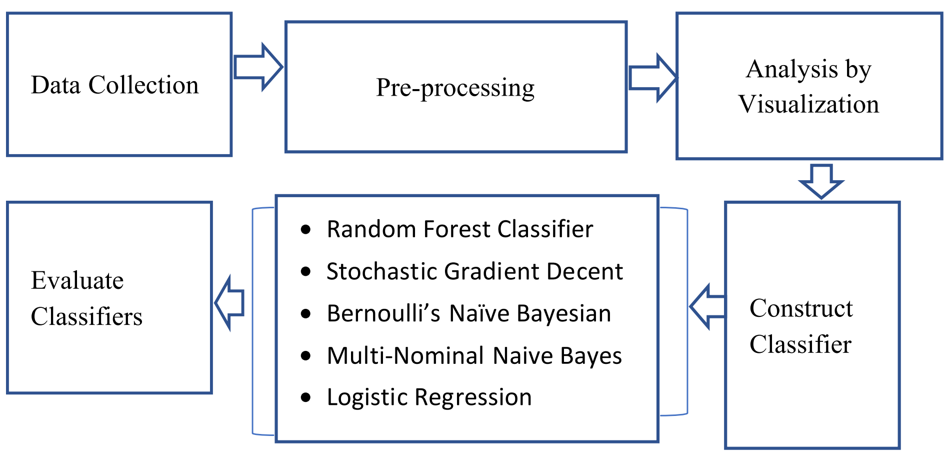



*Figure 1 : Product Opinion Prediction Process*

**Dataset Collection:** The Amazon review dataset [2] is taken for this study. The dataset includes 142.8 million reviews total, dating from May 1996 – July 2014. The dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). Sample dataset is depicted in Figure 2. There are various product categories such as: Books, Electronics, Movies and TV, CDs and Toys and Games, Video Games etc. The dataset is thoroughly studied to identify missing value records, outliers, incorrect data and identified that "Toys and Games" dataset has lesser outliers and more consistent than the other datasets. So review related to "Toys and Games" dataset in studied in this work.'

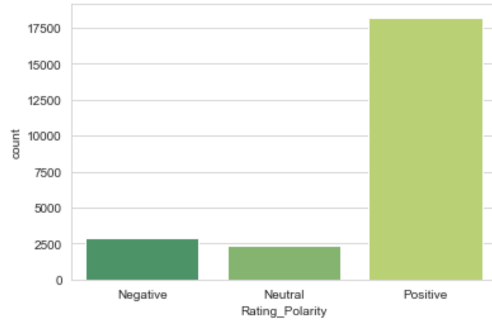
Below are the features that are found in the dataset:

- ReviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- helpful - helpfulness rating of the review, e.g. 2/3
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

| reviewerI | asin | reviewerN | helpful | reviewTex | overall | summary | unixReviev | reviewTim | Division N | Department Name |
|---|---|---|---|---|---|---|---|---|---|---|
| A1YJEY40Y | 7.81E+09 | Andrea | [3,4] | Very oily a | 1 | Don't wast | 1.39E+09 | 01 30, 201 | Initmate | Electronic |
| A6OXNB87 | 7.81E+09 | Jessica H. | [1,1] | This palett | 3 | OK Palette | 1.4E+09 | 04 18, 201 | General | Shoes |
| A3G6XNM | 7.81E+09 | Karen | [0,1] | The textur | 4 | great quali | 1.38E+09 | 09 6, 2013 | General | Dresses |
| A1PQFP6S. | 7.81E+09 | Norah | [2,2] | I really car | 2 | Do not wo | 1.39E+09 | 12 8, 2013 | General Pe | Bottoms |
| A38FVHZT | 7.81E+09 | Nova Amo | [0,0] | It was a lit | 3 | It's okay. | 1.38E+09 | 10 19, 201 | General | Tops |
| A38TN14H | 7.81E+09 | S. M. Ranc | [1,2] | I was very | 5 | Very nice p | 1.37E+09 | 04 15, 201 | General | Dresses |
| A1Z59RFKI | 7.81E+09 | tasha "luv | [1,3] | PLEASE DC | 1 | smh!!! | 1.38E+09 | 08 16, 201 | General Pe | Tops |
| AWUO9P6 | 7.81E+09 | TreMagnif | [0,1] | Chalky,No | 2 | Chalky, Nc | 1.38E+09 | 09 4, 2013 | General Pe | Tops |
| A3LMILRN | 9.76E+09 | | [0,0] | Did nothin | 2 | no Lighten | 1.41E+09 | 07 13, 201 | General | Dresses |
| A3OIP88QI | 9.76E+09 | Amina Bint | [0,0] | I bought th | 3 | Its alright | 1.39E+09 | 12 27, 201 | General | Dresses |
| APBQH4BS | 9.76E+09 | Charmmy | [0,0] | I have mix | 3 | Mixed feel | 1.4E+09 | 05 20, 201 | General | Dresses |
| A3FE8W8L | 9.76E+09 | Culture C S | [0,0] | Did nothin | 1 | Nothing | 1.39E+09 | 02 18, 201 | General Pe | Dresses |
| A1EVGDO | 9.76E+09 | Jessica "Ar | [0,1] | I bought th | 5 | This works | 1.39E+09 | 01 23, 201 | General Pe | Dresses |
| APSWTCM | 9.76E+09 | Layla B | [0,0] | This gell di | 1 | Does noth | 1.39E+09 | 01 11, 201 | Initmates | Intimate |
| A21IM16P | 9.76E+09 | mdub9922 | [0,1] | i got this tc | 5 | it works | 1.39E+09 | 02 18, 201 | General | Dresses |
| A1TLDR1V | 9.76E+09 | Mickey O f | [0,0] | I used it fo | 2 | burns | 1.4E+09 | 04 6, 2014 | General | Bottoms |
| A6F8KHOJ1 | 9.76E+09 | SanBen | [2,4] | I order this | 5 | Did work f | 1.38E+09 | 09 14, 201 | General | Bottoms |
| AXPKZA7U | 9.76E+09 | Shirleyyy | [2,4] | Good proc | 4 | excellent | 1.38E+09 | 10 18, 201 | General | Tops |
| A2SIAYDKG | 9.76E+09 | theredtran | [0,1] | I didn't use | 3 | weird sme | 1.38E+09 | 11 1, 2013 | General | Jackets |
| A1QVSIH6 | 9.79E+09 | armygirl | [24,24] | I haven't b | 5 | Love the s | 1.32E+09 | 09 19, 201 | General | Dresses |
| A3UQXHI8 | 9.79E+09 | D. Greene | [0,0] | We gave tl | 5 | Happy | 1.38E+09 | 08 10, 201 | General | Tops |
| A2EK2CJN | 9.79E+09 | Nikki | [1,1] | This is the | 5 | Very good | 1.32E+09 | 11 28, 201 | General | Dresses |
| A2GWNGC | 9.79E+09 | Pholuke "L | [2,4] | So I got thi | 5 | Lurrrrrrrv. | 1.34E+09 | 05 27, 201 | General | Dresses |
| ABV67T13 | 9.79E+09 | Sandra | [0,0] | This produ | 5 | Great Scer | 1.36E+09 | 02 2, 2013 | General | Dresses |
| A2FQZKL2I | 9.79E+09 | Ellie B. | [1,1] | I'm very pi | 5 | Spring Gar | 1.39E+09 | 03 11, 201 | General | Tops |

Figure 2 : Sample Dataset

**Feature Extraction**: It is an important phase, in model building process. It is important to convert the text data into a feature vector so as to process text in an efficient manner[10]. We dropped the records with null value columns. Then preprocessing techniques like special characters, punctuations, numbers, extra spaces from the review text. Performed text tokenization of the review text, removed stop words, identified $\sum Positive$, $\sum$ Negative and $\sum$ Neutral tokens in the text. The inherent polarity of words in the text is shown in below fig3.




**Construct Classifier:** Several models developed to study the performance of different classifiers on Text reviews.

**NaiveBayes** is selected for its simplistic approach, a fast classifier that can be applied for binary, multi-class classification problems, most widely used in real-time applications, for dynamic data changes. **Logistic Regression:** A classifier, an extension of Linear regression applicable for categorical class labels. It is a simple and efficient method, with low variance and provides probability score for an observation. As more and more relevant data comes in, the algorithm betters the prediction performance.

**Ensemble***:* These methods uses multiple learning algorithms to obtain better prediction than obtained by any single learning model. We employed Random forest in this work.

**Random forest** method is a decision forest method applied to Classification and Regression tasks. The method constructs multiple Decision trees at learning stage and outputs a model that is accurate many a times. With random forest approach, overfitting is reduced by which the prediction accuracy improves. Multiple trees reduce the chance of stumbling across a classifier that doesn't perform well. In case of unbalanced datasets, random forest has balancing error in class population. It has capabilities to compute similarities in the data and identify outliers. Thus, it can be extended to unlabeled data, leading to unsupervised learning, data views and outlier detection.

**Stochastic gradient descent** is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. It's an inexact but powerful technique. Stochastic gradient descent is widely used in machine learning applications. Combined with backpropagation, it's dominant in neural network training applications.

**Evaluate Classifier :**
Accuracy : Area Under ROC Curve are considered to evaluate the classifier.
Confusion Matrix: Firstly, Confusion matrix is computed, followed by Precision, Recall, Accuracy and Area Under ROC curve were calculated.

| | | Predicted | |
|---|---|---|---|
| | | **Negative** | **Positive** |
| **Actual** | **Negative** | True Negative | False Positive |
| | **Positive** | False Negative | True Positive |

**Accuracy :** Can be defined as the number of correct predictions made to the total number of predictions made. Precision, Recall measure can be obtained from confusion Matrix. Precision is a metric to know the correct positive predictions out of all the positive predictions. High precision indicates low false positive rate.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

**Recall:** Recall is the ratio of correctly predicted positive values to the actual positive values. It helps us to know the number of positive predictions that are made out of all actual positives.

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$

**Area Under ROC curve** : It is a performance metrics which is used to represent a model's ability to discriminate between positive and negative classes.
The above measure are used to evaluate classifiers.

## [4] EXPERIMENTAL RESULTS

The figure 3 depicts the word cloud of the product reviews. We can notice that high score words are great, love, book so on and so forth. In word cloud the size of words varies with the frequency of occurrences. When compared to figure 4, we can see that disappointed word is the most frequent words in the reviews of average scored words. We get a overview of the review dataset from figures 3, 4.

Figure 3 : High Scored Words in Reviews

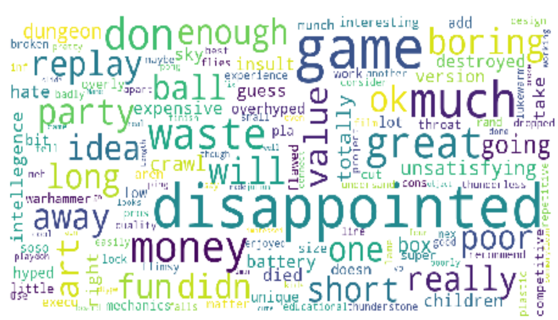


Figure 4: Average Scored Words in Reviews

After comparing the accuracy scores of all the models, We concluded that the model generated using Logistic Regression is better and has an accuracy score of 99.5%. The Table1 depicts the various classifiers implemented in this work.

Table 1 : Comparison of the classifiers

| Sl.No. | Classifier | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 1. | Logistic | 99.9 | 99.9 | 100 | 99.5 |
| 2. | Naïve bayes | 93 | 93 | 93 | 92.7 |
| 3. | Multinomial Naïve Bayes | 97 | 97 | 97 | 97.1 |
| 4. | Random Forest | 86 | 93 | 89 | 92 |
| 5. | Stochastic Gradient Decent | 95 | 95 | 94 | 95 |

Below graph represents the Roc curve of each classifier model and we can see that logistic regression has an AUC (Area under curve) of 0.95 which is greater than other models.

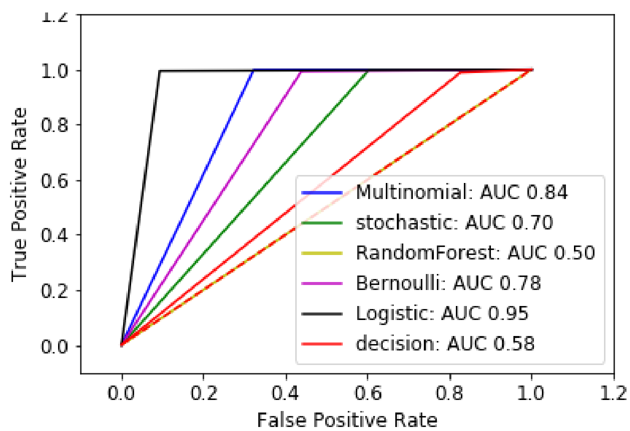



Figure 5 Area under ROC Curve

## [5] CONCLUSIONS AND FUTURE WORK

In this work different classifiers are studied to understand the sentiment of a product by end users given in the form of text review. The area of Opinion prediction is gaining lot of research interest, and deemed to grow further in future. Since it is a direct implication of a business use case, lot of research is encouraged as well as observed by business giants. Hence from mere likes and reviews, the businesses are expecting the customer expectations of the products and develop such products for better business. The models developed in this work will enable the businesses to get to know the sentiment of the products by drill through thousands of reviews at a single stretch. By using the models, business can understand how the end users feel about different areas of the business. Researchers and Businesses are interested in understanding the thoughts of people and how they respond to everything happening around them. AI based product promotions are evolving using the sentiment analysis applications. Hence in our future work we want to explore this study to consider multi-model inputs and study the behaviour of customer and their product ratings. We work with Deep Learning methods to deal with huge product review data and come up with a much efficient model.

## References

[1] Chetviorkin, P. Braslavskiy and N. Loukachevich, "Sentiment Analysis Track “ at ROMIP 2011.
[2] Link: http://jmcauley.ucsd.edu/data/amazon/
[3] Sri Lalitha Y., Govardhan A, “Semantic Framework for Text Clustering with Neighbors” in Intelligent Systems and Computing 249, © Springer International Publishing Switzerland December 2013 pp.261-271.

[4] S. Kiritchenko, X. Zhu, C. Cherry and S. Mohammad, "NRC-Canada-2014: Detecting aspects and sentiment in customer reviews", *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, pp. 437-442, 2014

[5] L. McClendon and N. Meghanathan, "Using machine learning algorithms to analyze crime data," *Mach. Learn. and Appl.: an Intl. J. (MLAIJ)*, vol.2, no.1, Mar. 2015.

[6] Alessia et. al. (2015), "Approaches, Tools and Applications for Sentiment Analysis Implementation" at International Journal of Computer Applications (0975 – 8887), Volume 125 – No.3, September 2015.

[7] Rincy J, Varghese S Cl, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach", 2016, International Conference on Data Mining and Advanced Computing (SAPIENCE)

[8] Jagbir Kaur, and Meenakshi Bansal, "Multi-layered sentiment analytical model for product review mining", In Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 415-420, 2016.

[9] Singla Z, Randhawa S, Jain S. Statistical and sentiment analysis of consumer product reviews. In 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6, 2017.

[10] Sri Lalitha Y., N.V. Ganapathi Raju, et.al "Analysis of Parts of Speech Tagging in Text Clustering", International Journal of Innovative Technology and Exploring Engineering, Jun 2019, pp : 2287-2291.

[11] Chaturvedi S, Mishra V, Mishra N. Sentiment analysis using machine learning for business intelligence. In: IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI); 2017. Pp: 2162-2166.

[12] Lucia M. and Maria N.C, "Applying Data Mining for Sentiment Analysis, Trends in Cyber-Physical Multi-Agent Systems" at PAAMS 2017 (pp.198-205)

[13] Sri Lalitha Y, et.al "Semantic Framework for Text Clustering with Neighbors" ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI, Volume II, Advances in Intelligent Systems and Computing 249, © Springer International Publishing Switzerland 2013 December 2013 pp.261-271, ISBN: 978-3-319-03095-1.