Machine Learning Centric Product Endorsement on Flipkart Database

M. Sri Lakshmi, S Prem Kumar, M. Janardhan

Abstract: ***-The growing need for individual mining information from text leads to analyzing sentiment and viewpoints. Retailers, E-commerce companies, product companies, media houses, real estate firms, and whom all have recognized that sentiment analysis is the key to success. They perform sentiment analysis necessary to get customer information related to feelings; attitudes, reactions, and opinions of existing and potential buyers towards their product or services. In this context, evaluating an individual's viewpoint or humor from a piece of text is challenging. In recent years the need for this analysis has increased due to the benefits obtained from it. In this paper, we conduct sentiment analysis on Flipkart product reviews using machine learning techniques to address the above challenge.***

Keywords: Sentiment analysis, Machine Learning, Flipkart,Product Analysis.

I. INTRODUCTION

Nowadays, people are spending more time on e-commerce websites for purchasing items, and the online platform almost entirely covers the global business site. For this reason, it is also common to read and understand the reviews for the products before purchasing them. Apparently, customers are more likely to buy a product if it has been given with positive reviews; therefore, analyzing these customer review data is essential to make them more dynamic. The major attention of this paper is to classify the positive and negative sentiments of the consumers over different products and articulate a supervised learning model to polarize a large number of reviews [8]. A study on online e-commerce sales increases due to its product reviews shown as per the survey, in the figure 1.

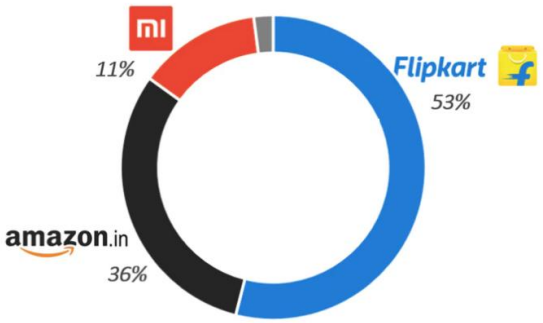

Figure 1. Online share in India over eCommerce business

In E-commerce market, customer reviews testimonials, or even social media posts which influence the growth of the company's business potential when company gains customer trust. Figure 2 demonstrates how effective is the customer content at increasing the conversion from product visits to purchases.

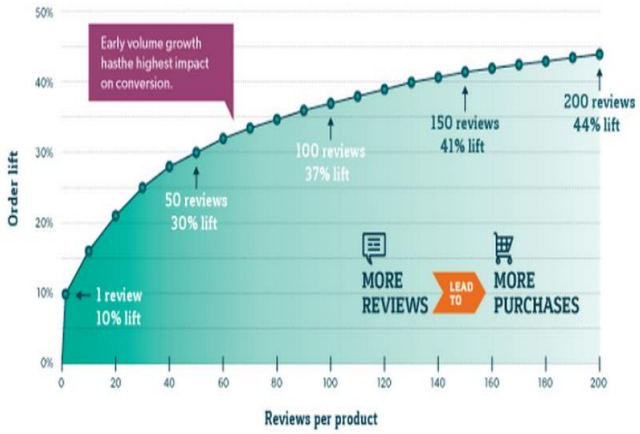

Figure 2. Increasing sales based on Reviews

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems which try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.

Polarity: a positive or negative opinion that the speaker expresses,

Subject: the thing that is being talked about,

Opinion holder: the person, or entity that expresses the opinion.

The Impact of Reviews: In this digital era, people are overwhelmingly using the internet to search and research online.

Revised Manuscript Received on October 20, 2019.

***** Correspondence Author

M. Sri Lakshmi *, Assistant Professor, Department of CSE, G.Pullaiah College of Engineering and Technology, Kurnool.

Dr. S Prem Kumar, Professor & Dean , Department of CSE, G.Pullaiah College of Engineering and Technology, Kurnool.

M. Janardhan, Associate Professor, Department of CSE, G.Pullaiah College of Engineering and Technology, Kurnool.

Daily, purchase decisions are being made, while pre-purchase research is being done mostly online. Reports and studies show that 9 out of 10 consumers conducted online research via search engines before making a purchase [7].
More importantly, a large portion of that research comes from browsing reviews. Negative reviews have become quite influential in undermining a business' reputation, leading to consequences:

- **Reputational risk**: Negative reviews cause potential customers to trust a company, lesser.
- **Hard to fix**: Having an abundance of negative reviews makes it difficult to regain trust and rebrand.

On the other hand, positive reviews provide a business with a positive reinforcement loop:

- **Improve reputation:** Consumers has been trusting the company over lower-rated competitors. Positive feedback from past customers increases the likelihood of a prospect to choose the company.

Overall negative reviews can have a significant impact on a business. Online reputation management is becoming essential in every company's management toolbox [9].
Online reviews are a natural way for consumers to share their experiences while their purchase the product of a brand. It's essential to the business that we monitor and manage positive and negative reviews. It's the best way to leverage their effectiveness and ensure the specific brand is accurately represented.

II. RELATED WORKS

So far numerous research articles have been published on product reviews, sentiment analysis, and opinion balloting. In [1], Ellie, Maria and Yi-Fan have collected opponents' views and analysed the results to form an economic model. They claim that the usage of tools revealed that they predominantly offer high precision. The use of commercial analysis has made their decision more appropriate. They also worked to identify the emotions gained from the examination by identifying false comments based on gender, by word. The most commonly used programming languages are Python and R. The classification techniques are mainly Naïve Bayesian Multinomial Classification (MNB) and Support Vector Machine (SVM). In the article [2], the existing supervised learning algorithms were used by the author to estimate the number of revisions using only the text. They have mutual validation using 70% data as training data and 30% as test data. In this article, the author cast-off different classifications to determine the exact values and recall them. The author [3] applied the work undertaken in the field of natural language processing and sentiment analysis to data from Amazon review datasets. Naïve Bayesian and decision list classifiers were used to tag a given review as positive or negative.
Extracting features from user opinion information is an emerging task.

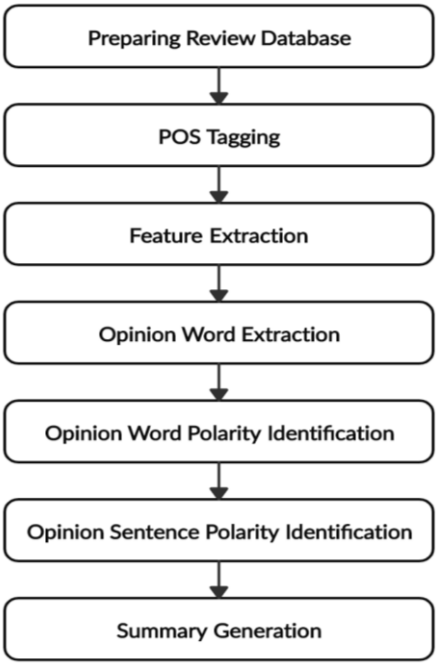

Figure 3. A generic model of feature extraction from opinion information

For the above process, large dataset has been used to produce the effective results and which could make it as better decision using machine learning approaches [6]. Furthermore, the system setup with several feature extraction methods which concludes with a higher accuracy than the existing research works.

III. PROBLEM STATEMENT

Customer Experience (CX) is the key to business success. Nowadays, more than ever, it's key for companies to pay close attention to Voice of Customer (VoC) to improve the customer experience. By analyzing and producing insights from customer feedback, companies have attained better information to make strategic decisions, which further creates an accurate understanding of what the customer actually wants and, as a result, a better experience for everyone. Sentiment analysis is the automated process of understanding the sentiment or opinion of a given text. This machine learning approaches can provide insights by automatically analyzing product reviews and separating them into tags: Positive, Neutral, Negative. By using sentiment analysis to structure product reviews, the strategists can:

1. Understand what the customers like and dislike about the product.
2. Compare product reviews with those of other competitors.
3. Get the latest product insights in real-time.
4. Save hundreds of hours of manual data processing.

Objective of the paper

- Scrapping product reviews on various websites featuring various products specifically amazon.com.

- Analyze and categorize review data.
- Analyze sentiment on dataset from document level (review level) using machine learning approach.
- Categorization or classification of opinion sentiment into:
  - Positive
  - Negative

IV. METHODOLOGY

1. Sentiment-Analysis-of-Flipkart-review-data

The dataset from the Flipkart Reviews Kaggle competition was used for the purpose of the achieving the objective in this paper. The goal is to perform sentiment analysis to determine whether a review is positive or negative using a classifier in Python for sentiment analysis on Amazon reviews. Sentiment analysis is frequently used to develop the emotion/opinion uttered in a text [4]. The main aim is to conduct the sentiment analysis on Flipkart product reviews by means of machine learning techniques. The trained model was to be used to predict users' sentiment based on their online reviews.

Part 1. Data Exploration

The dataset consisted of 400 thousand reviews of products from flipkart.com. The data set has the following fields: Text – The review data Label – Binary label (positive/negative). Below are some summary statistics about the data:

- Total number of reviews: 412529
- Number of positive reviews: 209932
- Number of negative and reviews: 202597

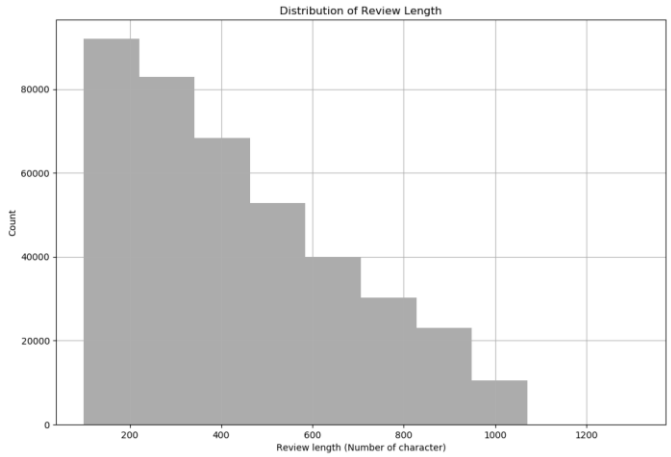

Figure 4. Distribution of review length.

Part 2. Data Preparation

Step 1: Load Product review data as input for the system, which consist of positive, negative and neutral reviews on products.

Step 2: Stop word filtering: "Stop words" are the most common words in a language like "the", "a", "on", "is", "all". These words do not carry important meaning and are usually removed from texts. It is possible to remove stop words using Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing.

from nltk.corpus import stopwords
stopset = set(stopwords.words('english'))
def stopword_filtered_word_feats(words):
return dict([(word, True) for word in words if word not in stopset])
evaluate_classifier(stopword_filtered_word_feats)

Step 3: Stemming words:

Stemming is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look). The main two algorithms are Porter stemming algorithm (removes common morphological and in flexional endings from words) and Lancaster stemming algorithm (a more aggressive stemming algorithm). In the "Stemming" sheet of the table some stemmers are described. Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words "chocolates", "chocolatey", "choco" to the root word, "chocolate" and "retrieval", "retrieved", "retrieves" reduce to the stem "retrieve".

Some more example of stemming for root word "like" include:
->"likes"
->"liked"
->"likely"
->"liking"

Stemming using NLTK:

Code:

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
stemmer= PorterStemmer()
input_str= "There are several types of stemming
algorithms."
input_str=word_tokenize(input_str)
for word in input_str:
   print(stemmer.stem(word))
```

Part 3. Machine Learning algorithm

Models for evaluation of Multinomial Naive Bayes Neural Networks Decision Tree were created using the following implementation methods.

Algorithm Approach:

Multinomial Naive Bayes:

Step:1 Input Text
Step 2: Discrete Count for number of occurrences in given input
Step 3: Multinomial Distribution
Step:4 *Calculate*

$$P(X_1 = x_1, \ldots\ldots X_k = x_k) = \frac{n!}{x_1! \ldots\ldots x_k!} p_1^{x_1} \ldots\ldots p_k^{x_k}$$

Step 5: Count the number of word occurrence in step 4
Step 6: Classify based on classifiers
Step 7: Predict the Result as output.

V. IMPLEMENTATION

Fit feature vectors to a supervised learning algorithm using Multinomial Naïve bias, Neural Networks, and Decision Trees in learn Load pre-trained model and predict the sentiment of the new data.

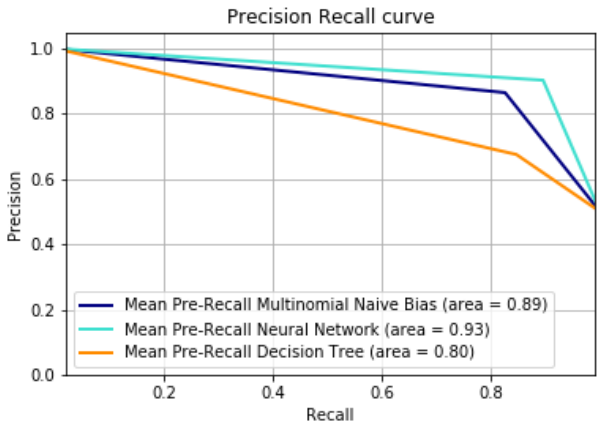

Figure 5. Multinomial Naïve Bayes algorithm

Naive Bayes is a simple, and due to its simplicity, this algorithm might outperform more complex models when the data set isn't large enough, and the categories are kept simple [10]. The training documents are allowed to calculate the class and evidence of the text document where classification is applied on test documents and choosing the class with the maximum probability. We articulate this by applying Bayes' rule: *P(cj |di;ˆθ) = P(cj |ˆθ)P(di|cj; ˆθj) P(di|ˆθ)*.

Table 1. Classification report

Method	Precision	Recall	Accuracy
Naïve Bayes	0.85	0.85	0.8483
Neural Network	0.90	0.90	0.9002
Decision Tree	0.75	0.74	0.7424

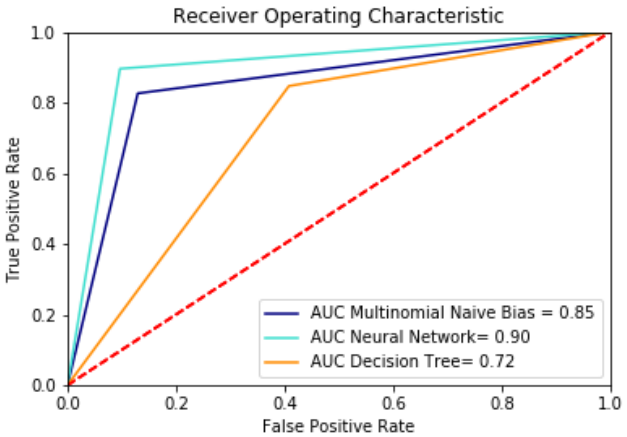

Figure 6. Receive Operating Characteristic

Part 4. Visualization

For the purposed of developing the visualization of the results and some analysis of the data, plotty tool was used.

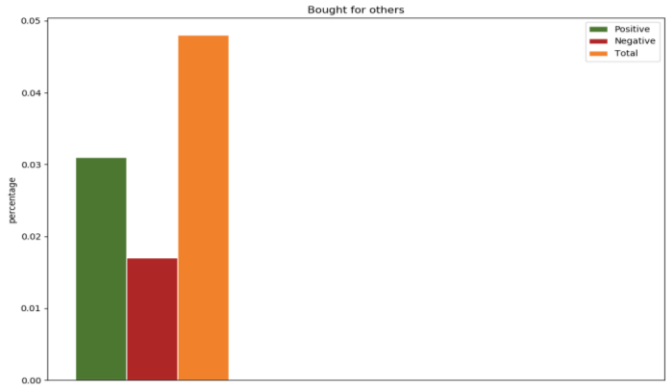

Figure 7. Analysis of items bought by customers for others

From the above analysis, we conclude that sentiment analysis has been applied on Flipkart product review data set using machine learning algorithm and compared with three different classifiers and evaluated the result in terms of precision, recall, and accuracy.

VI. CONCLUSION

This paper validates the machine learning classifiers, how they make use for sentiment analysis on text-based Flipkart Product review contents. Likewise, evaluation was performed also for the performance of sentiment analysis of Product review dataset using a machine learning approach. Subsequently, best classification performer would be enabled for a product review based its performance measuring parameters.

REFERENCES

1. Elli, Maria Soledad, and Yi-Fan Wang, "Amazon Reviews, business analytics with sentiment analysis." 2016.
2. Xu, Yun, Xinhui Wu, and Qinxia Wang, "Sentiment Analysis of Yelp's Ratings Based on Text Reviews." (2015).
3. Rain, Callen. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning. "Swarthmore College (2013).
4. N. Nodarakis, S. Sioutas, A. Tsakalidis, and G. Tzimas, "Large-Scale Sentiment Analysis On Twitter with Spark". Mar 15, 2016.
5. Enock Kanyesigye , Sumitra Menerea, "Sentiment Analysis Of Reviews Using Hadoop", 2016.
6. J. McAuley, R. Pandey, J. Leskovec, "Knowledge Discovery and Data Mining", 2015.
7. J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015
8. Joachims, Thorsten, "Text categorization with support vector machines: Learning with many relevant features", Springer Berlin Heidelberg, 1998.
9. Pravesh Kumar Singh, Mohd Shahid Husain, "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques", February 2014.
10. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", 2008.