3rd National Level Students' Research Conference on "Innovative Ideas and Inventions in Computer Science & IT with Its Sustainability"
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Sentimental Analysis of Customer Reviews: By using Data Analysis

Vikas Shukla, Prof. Sachin Bhoite
School of Computer Science, MIT-World Peace University, Pune, India

ABSTRACT

Online shopping and e-commerce are an area that has experienced considerable growth over the last 5-6 Years. Sentiment analysis - this area of research attempts to determine the feelings, opinions, emotions, among other things, of people on something or someone. To work on this, we use new techniques NLP (Natural Language Processing) and machine learning algorithms. Reviews on product by customer make a big impact on sellers profit margin However, negative reviews can reduce the profits for companies as people become transparent in expressing their' opinions and post them online without considering the consequences the companies would face. To help the business owner we are going to do this project, in which we will be taking E-commerce website FLIPKART for our dataset and the product as headphones. Reading one by one reviews takes a lot of time, so what we will do is summarize the whole reviews into 3 points. For this we will be using Sentiment intensity analyzer algorithm. It is more efficient than any other algo like visualization or data mining.

Keywords— Data Science, Sentiment analyze, opinion mining, reviews, e-commerce, natural language processing, semantic analyze

I. INTRODUCTION

Now a days everyone buys items online on different e-commerce websites like flipkart, amazon, snapdeal and many more. In growing country like INDIA e commerce, play a big role and their business depends upon the reviews given by the customer. The sentiment is a feeling that expresses judgement, attitude or thought. Sentiment analysis, also known as opinion mining, studies people's sentiments towards certain entities.

In today's world reviews dropped by users on a certain product/item makes a big impact on the owner's business. It is practically not possible for seller to keep track and analyse all these feedbacks.

So basically, the main objective of this project is to analyse the sentiment of customers by reviews given by them on the review section.

We will use different libraries which are already defined in Python like PANDAS, SKLEARN, NLTK etc. The model we are going to use is in NLTK – sentiment intensity analyser. Usually, business owners read the reviews and takes a lot of time ten minutes to read 7-8 reviews which is not suitable but using this model we can read up to 15 thousand reviews in seconds and make a summary out of it. The motive of this study is to use Python and

to predict the sentiment of reviews and how this could deliver valuable information to the manager, how can you make the product better or what are the problems customer facing.

II. RELATED WORK

Various research and students have published related work in national and international research papers, thesis to understand the objective, types of algorithms they have used and various techniques for pre-processing, Feature.

Hanan Alasmari have used Tableau, Python & Knowledge discovery and data mining (KDD) is a common methodology, which refers to the overall process of detecting useful insights from a collection of data they said "It is true that it is possible to understand customers opinions towards products through the massive scale of unstructured text online that are also informal" [1]

Shweta Rana and Archana Singh used SVM, and Naïve Bytes techniques and they said "The future scope of the work is that we can explore our data to a wider genre of different products on social networking sites or e-commerce as day by day the user is moving online and they prefer buying stuff online so we can identify the accuracy rates of the products like books, games etc." [2]

Zied Kechaou, Mohamed Ben Ammar and Adel.M Alimi "Improving e-learning with sentiment analysis of users' opinions" Applying a sentiment analysis to examine the nature and the structure of web forums and e-learning blogs turns out to be an important endeavour; however, the current accuracy is promising for effective analysis of forum conversation sentiments. Such analysis can help provide a better understanding of users' opinions regarding the e-learning system for the sake of its improvement. [3]

"Sentiment Analysis and Opinion Mining: A Survey" G.Vinodhini, RM.Chandrasekaran. "The main challenging aspects exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features/attributes, complexity of sentence/ document, handling of implicit product features , etc.". [4]

"Thumbs up? Sentiment Classification using Machine Learning Techniques" Bo Pang and Lillian Lee they said "The classification accuracies resulting from using only unigrams as features are shown in line (1) of Figure 3. As a whole, the machine learning algorithms clearly surpass the random-choice baseline of 50%." [5]

III. DATA SET DESCRIPTION

For this analysis, we will be using Flipkart Review dataset present on Kaggle. The dataset contains the reviews given by the users on a particular product named "BoAt Rockerz 235v2 with ASAP Charging".

IV. RESEARCH METHODS

Our methodology comprises of three sections –Data collection, Data pre-processing and VADER Sentimental Analyzer.

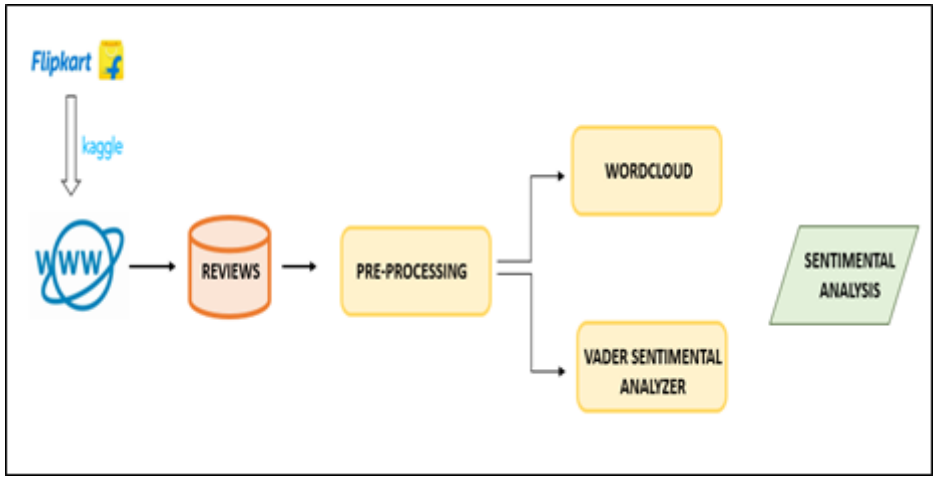

Fig. 1. Architecture of methodology to process reviews

A. Data Collection

In this section, we discuss the architecture framework and programming libraries we used in our experiment to pull and process the unstructured reviews from the Flipkart product page. To extract reviews from the products page we have used third party website Kaggle database which is a web application that exports stores different data to a comma-separated file (CSV).

Libraries used and a brief about them:

1. **NumPy**-NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. [7]
2. **Pandas**-Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series [7]
3. **Matplotlib**-Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. [7]
4. **Seaborn**-Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. [7]
5. **NLTK**-The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language. [7]
6. **WordCloud**-Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analysing data from social network websites. [7]

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

Reviewdata = pd.read_csv("flipk.csv")
Reviewdata.head(10)
```

	product_id	product_title	rating	summ
0	ACCFZGAQJGYCYDCM	BoAt Rockerz 235v2 with ASAP charging Version ...	5	Terrific purch
1	ACCFZGAQJGYCYDCM	BoAt Rockerz 235v2 with ASAP charging Version ...	5	Terrific purch
2	ACCFZGAQJGYCYDCM	BoAt Rockerz 235v2 with ASAP charging Version ...	5	Sup
3	ACCFZGAQJGYCYDCM	BoAt Rockerz 235v2 with ASAP charging Version ...	5	Sup

Fig. 2. The Dataset overviews

B. Pre-Processing

- There are 1293 Attributes with null values, were dropped from location columns and were replaced with their upper attributes.

```
print(Reviewdata.isnull().sum())
```

```
product_id          0
product_title       0
rating              0
summary             0
review              0
location         1293
date                0
upvotes             0
downvotes           0
dtype: int64
```

Fig. 3. Null Values

- Cleaned the data which is not useful for us. Dropped the following columns: product_id, product_title, location, upvotes, downvotes.
- Removed all the symbols like ' "", . ! @' and converted all the text into lowercase.

```
Reviewdata['cleaned_summary']=pd.DataFrame(Reviewdata.summary.apply(cleaned1))
Reviewdata['cleaned_review'] = pd.DataFrame(Reviewdata.review.apply(cleaned1))
Reviewdata.head(10)
```

	rating	summary	review	cleaned_summary
0	5	Terrific purchase	1-more flexible2-bass is very high3-sound clar...	terrific purchase
1	5	Terrific purchase	Super sound and good looking I like that prize	terrific purchase
2	5	Super!	Very much satisfied with the device at this pr...	super
3	5	Super!	Nice headphone, bass was very good and sound i...	super
4	5	Terrific purchase	Sound quality super battery backup super quali...	terrific purchase
5	5	Wonderful	Wowwww it's amezing bluetooth nice look, nice ...	wonderful
6	4	Pretty good	Awesome colour! Amazing experience .. but only...	pretty good
7	5	Terrific purchase	For the first time, I am posting a review, jus...	terrific purchase

Fig. 4. Removing Symbols and converting to lowercase

C. Feature Selection

Let's start by deleting the unnecessary or redundant columns. For data analysis, we do not need the :

id , title , location , upvotes , downvotes :

Some of these columns may looks like they are important but all of them are not usable in the project . The columns being used are:

Ratings, summary , review and date .

```
Reviewdata['cleaned_summary']=pd.DataFrame(Reviewdata.summary.apply(cleaned1))
Reviewdata['cleaned_review'] = pd.DataFrame(Reviewdata.review.apply(cleaned1))
Reviewdata.head(10)
```

Fig. 5.

D. VADER(Algorithm)

We applied Valence Aware Dictionary and sentiment Reasoner or VADER library in python programming language to calculate the semantic score of each comment. The positive sentiment, negative sentiment and neutral were extracted from each sentence using NLTK in Python 3.5.1. This was completed by utilizing Naïve Bayes Analyzer module "it finds the probability of an event given the probability of another event that has already occurred.

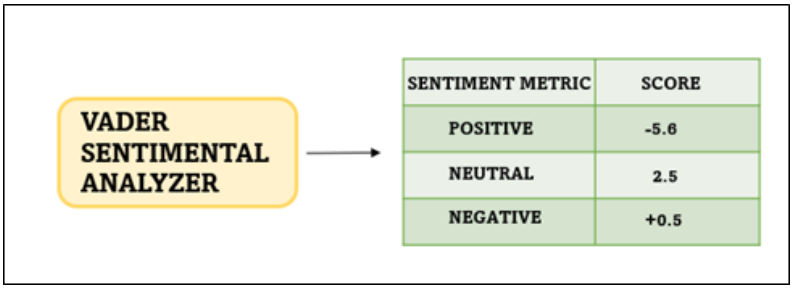

Fig. 6. Sentimental module uses VADER to calculate sentimental score .

V. EXPOLATORY DATA ANALYSIS

We predominantly used Python Libraries. Specifically, we used the Pandas Python Library, which is a data manipulation script . We used Pyplot for data visualization.

The Rating column of the data contains the ratings given by every reviewer. So let's have a look at how most of the people rate the products they buy from Flipkart:

A. Rating By Percentage

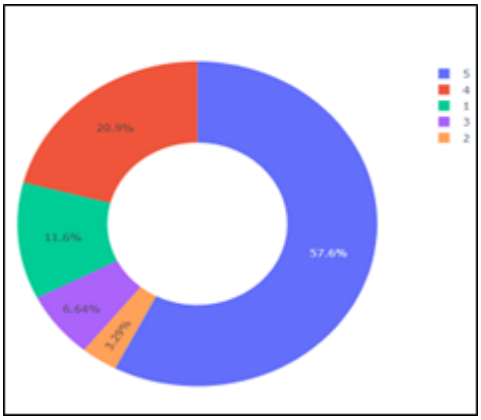

Fig. 7. Ratings in percentage

So 57.6% of the reviewers have given 5 out of 5 ratings to the products they buy from Flipkart.
Now let's have a look at the kind of reviews people leave. For this, I will use a word cloud to visualize the most used words in the review's column.

B. Word Cloud from Reviews

Fig. 8. WordCloud of the reviews written by user.

The bigger the size of word is visible the most it is used in the reviews.

C. Count of Ratings

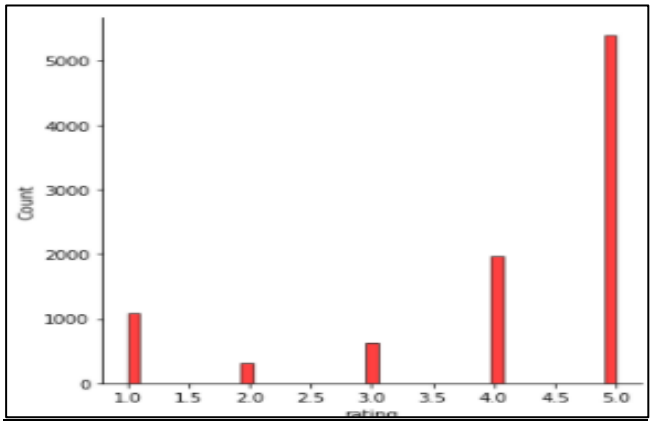

Fig. 9. Count of the rating.

As we can see in the above graph 5 star rating is given by more than 5000 customers and 1 star rating is given by 1000 plus customers.

D. Adding Column in CSV

Now I will analyse the sentiments of Flipkart reviews by adding three columns in this dataset as Positive, Negative, and Neutral by calculating the sentiment scores of the reviews:

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
```

	review	Positive	Negative	Neutral
0	1-more flexible2-bass is very high3-sound clar...	0.167	0.017	0.815
1	Super sound and good looking I like that prize	0.759	0.000	0.241
2	Very much satisfied with the device at this pr...	0.278	0.000	0.722
3	Nice headphone, bass was very good and sound i...	0.286	0.082	0.632
4	Sound quality super battery backup super quali...	0.600	0.000	0.400

Fig. 10.Positive , Negative and Neutral scores adding .

E. Word Cloud by One-star Reviews

Fig. 11. 1star rating reviews

Here we can see that words like ‘QUALITY’,’SOUND QUALITY’, ‘VOICE’ , ‘WORST’ and many more by this company can work on their products and analyse what’s the major problem customer is facing.

F. Word Cloud by five-star rating

Fig. 12. 5star rating reviews

VI. KEY FINDINGS

Now let's see how most of the reviewers think about the products and services of Flipkart:

```
x = sum(Reviewdata["Positive"])
y = sum(Reviewdata["Negative"])
z = sum(Reviewdata["Neutral"])

def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print("Positive 😊 ")
    elif (b>a) and (b>c):
        print("Negative 😠 ")
    else:
        print("Neutral 🙂 ")
sentiment_score(x, y, z)
```

```
Neutral 🙂
```

Fig. 13.The overall .

So most of the reviews are neutral. Let's have a look at the total of Positive, Negative, and Neutral sentiment scores to find a result about Flipkart reviews:

VII. RESULTS AND DISCUSSION

In this research we have considered a particular e-commerce website and a product. This will help business personals to open a new restaurant there. Such analysis is essential part of business. Having a prior idea about the product and how customer are feeling with that. So, most people give Neutral reviews, and a small proportion of people give Negative reviews. So, we can say that people are satisfied with Flipkart products and services. You can refer the word cloud –[Fig.11&12] to check what keywords are mostly used by customer and what are the defaults in the product. So, in future you can develop a more enhanced product

Various methods have been used to measure the performance. From the performance achieved by these methods it is difficult to judge the best choice of classification method, since each method uses a variety of resources for training and different collections of documents for testing, various feature selection methods and different text granularity.

For our research we have used Boat headphone as product and the scores from the reviews are like this :

```
print("Positive: ", x)
print("Negative: ", y)
print("Neutral: ", z)
```

```
Positive:  3439.9740000000415
Negative:  532.5680000000009
Neutral:  5401.457999999954
```

Fig. 14.Positive , Negative and Neutral scores.

VIII. CONCLUSION

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. There are likely to be many other applications that is not discussed. Based on customer reviews. We concluded that the product is neutral. And for many customers it is positive also. By this a company can make their product more efficient. For future studies, this study could be extended to focus on integration of the sentiments found from the reviews with the tokens in each cluster. It is also recommended to use the sentiment analysis with the regression line as a dependent variable to investigate their causal relationship. The outcomes of this study could support Flipkart business managers to develop the usefulness of the product and enhance the way of marketing this product.

In future, more work needed in this field for further improving the performance measures . Sentiment analysis can be applied for new applications. Although the techniques and algorithms used for sentiment analysis are advancing fast, however, a lot of problems in this field of study remain unsolved.

IX. REFERENCES

[1]. "Sentimental Visualization: Semantic Analysis of Online Product Reviews Using Python and Tableau" Hanan Alasmari, IEEE ON BIG DATA VO., XX, NO., X, DECEMBER 2020.

[2]. "Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques" Shweta Rana and Archana Singh, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016)

[3]. Zied Kechaou, Mohamed Ben Ammar and Adel.M Alimi " Improving e-learning with sentiment analysis of users' opinions "2011 IEEE Global Engineering Education Conference (EDUCON)

[4]. "Sentiment Analysis and Opinion Mining: A Survey" G.Vinodhini, RM.Chandrasekaran Volume 2, Issue 6, June 2012(ijarcsse)

[5]. "Thumbs up? Sentiment Classification using Machine Learning Techniques" Bo Pang and Lillian Lee

[6]. Flipkart Product Review (Naushad Shukoor)

[7]. Google Search (Definition)

[8]. Chafale, Dhanashri, and Amit Pimpalkar. "Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic. "International Journal of Computer Sciences and Engineering (IJCSE) 2 (2014): 14-18

[9]. J.Ortigosa-Hernández,J.D. Rodríguez, L. Alzate, M. Lucania, I. Inza and J.A. Lozano, Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers, Neurocomputing 92 (2012)

[10]. B. Keith, E. Fuentes and C. Meneses, A hybrid approach for sentiment analysis applied to paper reviews, 2017.

[11]. E. Boiy and M.-F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12(5) (Oct. 2009), 526–558.

[12]. Tsur, D. Davidov, and A. Rappoport , " A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews". In Proceeding of ICWSM. of Context Dependent Opinions,2010.