I J R B A T, Issue (IX), Vol. II, May 2021: 63-69

A Double-Blind Peer Reviewed & Refereed Journal

OPEN ACCES

e-ISSN 2347 – 517X

Original Article

Reg. No : MH-NGPX00011

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY

© www.ijrbat.in

ASPECT EXTRACTION AND COMPUTATION OF LEXICON FROM NYKAA PRODUCT REVIEWS

T Sai Sravani[1], K Maneesha Reddy[2], S Nirupama[3]

[1] Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India
[2] Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India
[3] Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India
e-mail: saisravani0427@gmail.com, maneeshakomirelly@gmail.com ,
nirupama.sheelwant@gmail.com

Communicated: 15.03.21	Revision : 21.04.2021 Accepted: 11.05.2021	Published: 30.05.2021

ABSTRACT:

Lexicon based approach is one of the basic methods used to judge a customer view or how a consumer interprets the product depending on the words in the review. This process is considered Opinion Mining. The reviews from various websites can be utilized to fetch descriptive statistics to make a comparative analysis. The opinion of the consumer is polarized (positive, negative, neutral). This paper gives a brief idea of a product from Nykaa.com categorized in terms of brands. The result helps the customer in understanding the characteristics of each brand.

Keywords: *Opinion Mining, Descriptive and Comparative Analysis, Polarity, Lexicon*

INTRODUCTION:

The measure of data over the years is recorded as 2.5 quintillions (2.5 billion gigabytes) per day. It is estimated that the value might reach up to 1.7 megabytes per second for each person in the world. All the unstructured data can be utilized to perform analysis and the result can be brought into play to make effective decisions. To make use of this unorganized data, it should undergo preliminary steps.

Opinion Mining is also known as Sentiment Analysis which is to portray the insights, convictions, and views of the patron. This concept of the computer to fetch the emotion from the review given by the consumer is known as Natural Language Processing. NLP uses the approach of POS taggers. It is the process of identifying the POS of the words and their importance in the sentence to comprehend the meaning it is depicting. Different POS tagging techniques include

1)Lexicon based method

2)Rule-based method

3)Probabilistic methods

4)Deep Learning method

The lexicon-based method focuses on the corpus data-driven approach, which means that we identify the context of the sentiment word from the sentence. This outlook informs the algorithm with the context of conflicting opinion words, hence reduces the ambiguity of the POS tagger. The algorithm which is represented in this paper uses the NLTK (Natural Language Tool Kit) package to analyze the structure of the reviews and compute the word polarity. This helps in analyzing the unstructured data from the web and quantify the reaction of the people towards a certain product. This also assists the organization to make decisions regarding the enhancement of the product or services.

The organization of the paper is as follows, Section I contains the introduction of Opinion Mining, Section II contain the related work of Sentiment Analysis of various websites, Section III contain the Configuration with a flow chart, Section IV contains some methods for Descriptive Statistics and Lexicon based Opinion Mining, Section V describes the Results, section VI concludes the research Results and Section VII consists of future directions.

I. LITERATURE REVIEW

Sentiment depicts an emotion of a human which can be positive, negative, or neutral and the reviews are a collection of data that is sentiment rich will aid the product manufacturer to make satisfactory decisions for the growth of the company. Websites like Amazon and Trivago analyze the reviews based on Sentiment Analysis to build a large customer base. This result or outcome is used by the customer to make the right decisions while purchasing the product based on different feedback.Most of the data on online websites is unstructured and it is further cleaned and analyzed to develop or identify the opinion based on polarity. Author Zeenia Singhla has stated in the paper that algorithms like Support Vector Machine are used for processing. A combination of Product reviews is collected based on volume, variety, and velocity for the understanding of People's perspective to detect the mindset and behavior of them on a particular product.The author Meena Belwal proposes the idea of collaborative data for the same product available on different websites and can be utilized by the manufacturer for the enrichment of Business Intelligence and proper decision making. The results are visualized using graphs for better understanding and get a quick idea of the entire feedback.

II. CONFIGURATION

Flow Chart: a) Data Assortment: Unstructured data in the form of reviews have been collected from Nykaa.com.

b) Data Preprocessing: Preprocessing steps such as cleaning, Integrating, Transformation, Reduction are performed.

c)Descriptive Statistics: Feature, Selection, Feature extraction, and Visualization of the outcome are done.

d)Lexical Based Opinion Mining: It is performed to generate the polarity of different brands of lipsticks and then visualize.

***Descriptive Statistics*:** Descriptive statistics focus on the gathering of unorganized data to decipher knowledge. This is the process of describing the previous data and analyzing the important features. The numeric data from the analysis is pictured in the form of word clouds and charts. The preprocessing steps play a vital role in descriptive Analysis which are as follows. All the raw data is integrated into a text file. From this consolidated data, the delimiters are identified to avoid inconsistency to clean the data. Then, separators are taken into consideration to split the sentences into tokens. The frequency of each token is calculated to transform the data. The generated frequency is sorted for reduction and selection. The selected tokens and their respective frequencies are used to visualize.

The package being for performing Descriptive Analytics is Syuzhet. The methods used from this package for Analysis are get_text_as_string( )and get tokens( ). The get_text_as_string() method leads the input file which takes the path of the file as the argument. Then the get_tokens( ) method is used to split the sentences into words. The frequency of the generated tokens is calculated by as.integers() method. The tokens with the lowest frequency are eliminated. The organized data is represented as a bar graph and word cloud. This can be done theoretically by using Chi-Square Test to categorize the data based on the frequency distribution. In this method, the number of features and the number of reviews is compared with the expected values in each category to know the association. The Chi-Square Test is considered significant when there is a large difference between the observed and expected values. The difference varies depending on the size of the sample data.

Algorithm for calculating word frequency:

Step-1: importing syuzhet package

```
library(syuzhet)
```

Step-2: to get the words from text

```
 Word file <- get_text_as_string("path ")
```

Step-3: Count the repetition of words

```
wordcount <- get_tokens(wordfile, pattern = "\\W")
```

Step-4: Get the frequency of each repeated word

```
syuzhet_vector <- get_sentiment(wordcount, method="syuzhet")
```

Step-5: Extract the most frequently repeated words using sentiments

```
words.freq<-table(unlist(wordcount))
s_v <- cbind(names(words.freq),as.integer(words.freq))
```

The process of Descriptive Statistics deals with the behavior of historic data. Hence it helps us acknowledge the crucial features. For this process, the assorted reviews of lipsticks from Nykaa.com of different brands have been used. The resulting graph depicts the features that highly influence the opinion of the consumer.

Lexicon based Opinion Mining

i. Natural Language Processing:

NLP deals with the intelligence of how a computer interprets human emotions. It is used to analyze the text and extract the subjective data from the reviews. One of the major approaches of NLP is POS tagging.

ii. POS Tagging:

To perform the POS tagging, tokenization of sentence is to be carried out which means dividing the text into smaller parts called tokens. Then, the parts of speech tagger are associated with each token depending on the context and sequence. Among different POS Tagging techniques, this research concentrates on the lexical based approach of opinion mining which is an unsupervised learning method.

The view of the consumer is of two types. They can be subjective and Objective. Subjective refers to opinion in form of a sentence, whereas objective refers to the rating of the products in terms of numbers. Lexicon-based opinion Mining concentrates on the subjective part of the view. This approach is the idea of generating the sentiment based on semantics, which indicates the context of the word in the sentence. The need for the semantic orientation is to remove the ambiguity in polarizing a word. Word polarity depicts the orientation of words in a sentence whether they are positive, negative, and neutral. This results in polarity values of the sentence once all the sentences are polarized the outcome is used to analyze the product polarity.

Proposed Algorithm:

Two main packages used for this process are pandas and nltk(Natural language processing tool kit). The preprocessed file is loaded using pandas. The polarity of the sentences is generated using Sentiment Intensity Analyzer.

Step1: Importing the required nltk package and pandas package.

Step2: Create a class with the function Sentiment_Intensity_Analysis to calculate polarity.

Step2.1: Initialize all three polarity variables and count the variable to zero.

Step2.2: Excel file of a dataset is loaded and parsed into a data frame.

Step2.3: Parsed data frame is converted to list for iteration.

Step2.4: Initialize Sentiment Intensity Analyzer object.

Step2.5: Create a loop that finds polarity scores using the polarity_scores method.

Step2.6: Repeat the loop for every sentence in the dataset.

Step2.7: For each iteration the add the polar_scores to the previously initialized polar variable.

Step3: Create an instance of a class and call the function to print polarity values of each sentence and also the overall dataset polarity values.

RESULT & DISCUSSION

The outcome of this process generates the polarity (positive, negative, neutral) of different brands that are pictured in a bar graph. From this visualized data, one can analyze or depict the sentiment of customers towards various brands of lipsticks.

CONCLUSION

The increase in e-commerce websites results in generating a huge amount of unstructured data. This unstructured data is processed to perform opinion mining on the reviews of different lipstick brands from Nykaa.com. It has been observed from the Descriptive Statistics that the shade of the lipstick is given high priority compared to other features. Feature Extraction and selection help us understand the current preference of the patron.

As lexical based is a method on POS Tagging in which the context is also identified along with the parts of speech which help in determining accurate results. It can be observed from the graph that Mac has the highest positive rating and Nykaa has the most neutral rating of reviews. The Lexical based approach being unsupervised learning, it doesn't require any training data, unlike the rule-based approach. Rather it focuses on the context of the word to generate the sentiment.

REFERENCES:

Zeenia Singla, Sukhchandan Randhawa, Sushma (2017), Statistical and sentiment analysis of consumer product reviews, Jain. 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)IEEE.

Sentiment analysis-based product rating using textual reviews, C Sindhu, Dyawanapally Veda Vyas, Kommareddy Pradyoth **(2017)**. International conference of Electronics, Communication and Aerospace Technology (ICECA), IEEE,

Pankaj; Prashant Pandey; Muskan; Nitasha Soni, **(2017)**. Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

Analysis, Andréa Salinca, Business Reviews classification Using Sentiment Analysis, IEEE March **(2016)**

Aspect-Level Sentiment Analysis on E-Commerce Data, Satuluri Vanaja, Meena Belwal, International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE January **(2019)**

Sentiment Analysis Using Machine Learning Techniques on Python, Nisha Rathee, Nikita Joshi, Jaspreet Kaur,2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE June **(2018)**

Sentiment analysis on large scale Amazon product reviews, Tanjim Ul Haque, Nudrat Nawal Saber, Faisal Muhammad Shah, IEEE International Conference on Innovative Research and Development (ICIRD), IEEE June **(2018)**

An interpretation of sentiment analysis for enrichment of Business Intelligence, Bharat Singh, Nidhi Kushwaha, Om Prakash Vyas, IEEE Region 10 Conference (TENCON), IEEE February **(2017)**

Opinion mining and sentiment analysis on online customer review, K L Santhosh Kumar, Jayanti Desai, Jharna Majumdar, 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),IEEE May **(2017)**

Ecommerce sentiment analysis using graph-based approach, Monali Bardoloi, S.K Biswas, International Conference on Inventive Computing and Informatics (ICICI) 2017, IEEE May **(2018)**.

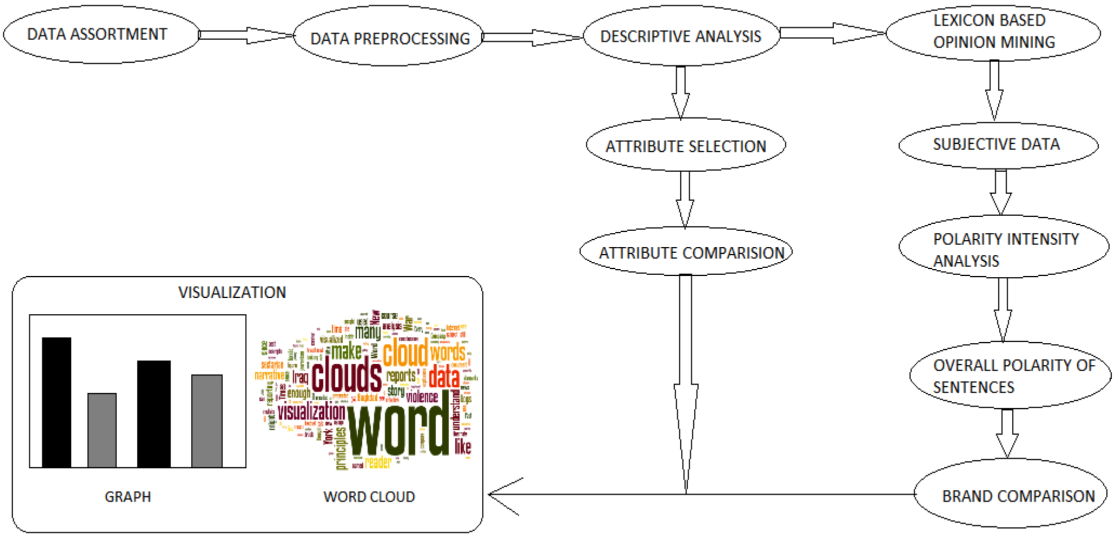


Figure 1 Flow Diagram

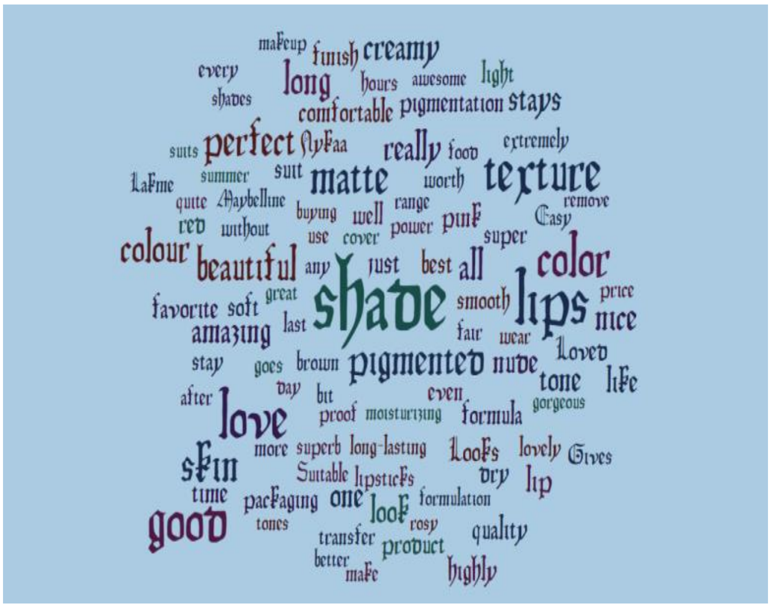
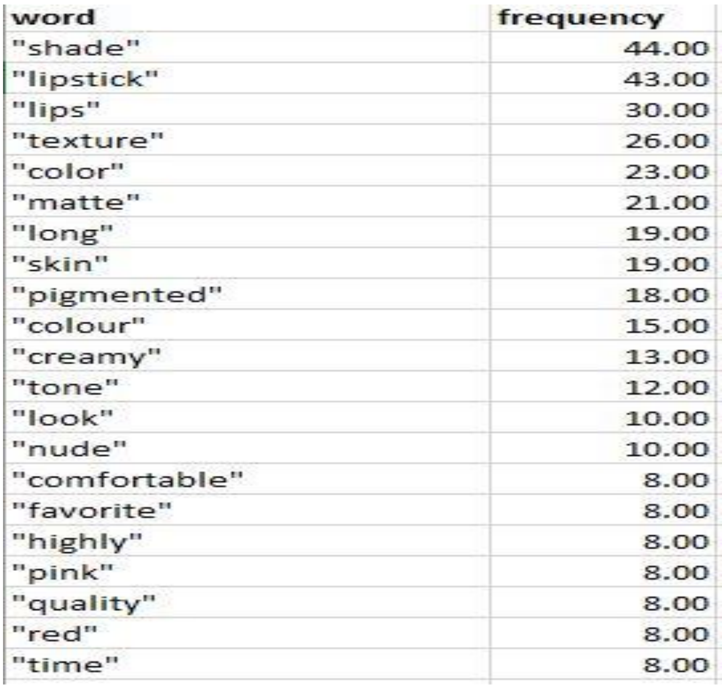


Figure 2 Word Cloud for features



word	frequency
"shade"	44.00
"lipstick"	43.00
"lips"	30.00
"texture"	26.00
"color"	23.00
"matte"	21.00
"long"	19.00
"skin"	19.00
"pigmented"	18.00
"colour"	15.00
"creamy"	13.00
"tone"	12.00
"look"	10.00
"nude"	10.00
"comfortable"	8.00
"favorite"	8.00
"highly"	8.00
"pink"	8.00
"quality"	8.00
"red"	8.00
"time"	8.00

Table 1 Frequency for words

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Figure 3 Chi-Square Test Formula

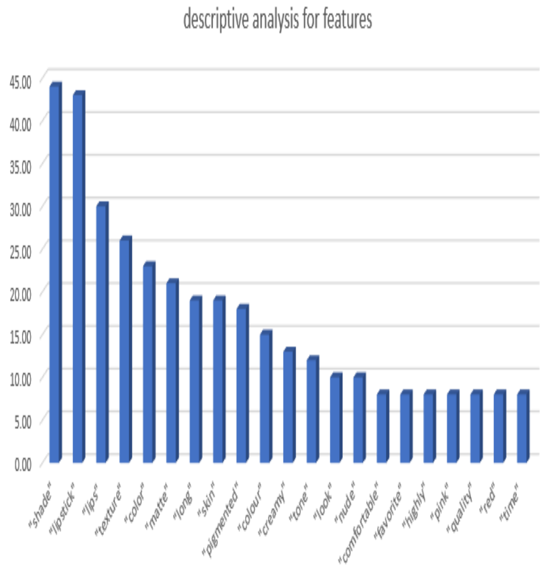
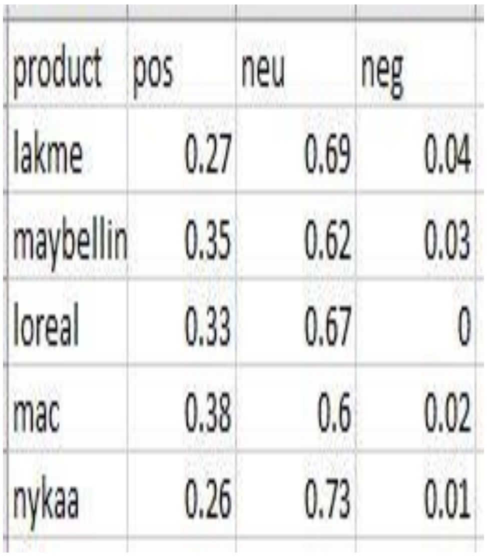


Figure 4 Number of review count by Product features



product	pos	neu	neg
lakme	0.27	0.69	0.04
maybellin	0.35	0.62	0.03
loreal	0.33	0.67	0
mac	0.38	0.6	0.02
nykaa	0.26	0.73	0.01

Table 2 Polarity for each brand

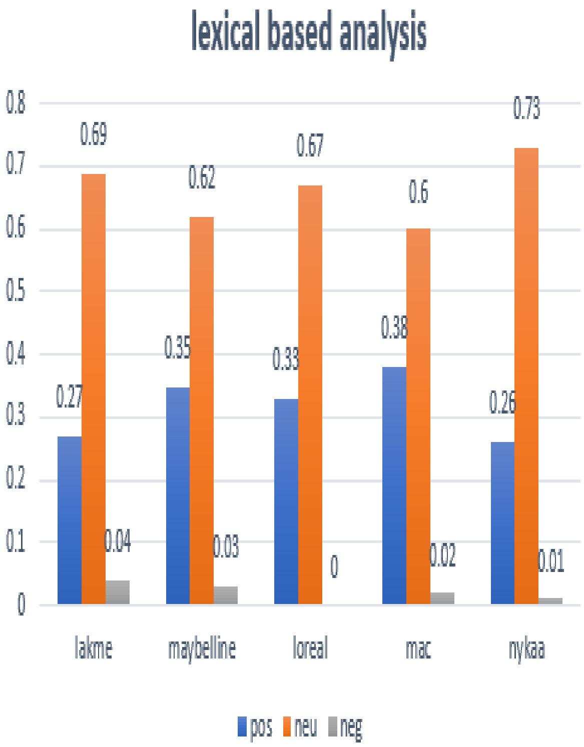


Figure 5 Brand of product and Polarity