JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Operational Marketing Strategy Analysis based on Online Reviews Using Machine Learning Algorithms

Divya H
M.Sc Data Science
Dept of Computing- Data Science
Coimbatore Institute of Technology
divyaharids2018@gmail.com

Nandhini S
M.Sc Data Science
Dept of Computing – Data Science
Coimbatore Institute of Technology
nandhinisiva2561@gmail.com

Dr.M. Sujithra
Assistant Professor
Dept of Computing – Data Science
Coimbatore Institute of Technology
sujisrinithi@gmail.com

Abstract:

In today’s world people exchange their thoughts by online social media platforms. At times, people give reviews and opinions on different products, brand, and their services. Those reviews towards specific products not only improve the product quality but also influence purchase decisions of the consumers. Thus, analysis of product review is a widely accepted platform where consumer can be aware about their requirements. The source dataset for the project was extracted from Kaggle. Analysis of various emotions from the reviews and ratings are done and based on that plain text review which can be categorized into eight basic emotions and two sentiments. The results show how sentiment analysis helps to identify every consumer's behaviour and overcome the risks to meet consumers' satisfaction. Machine learning algorithms are used to build ML models and predict accuracy of consumer’s behaviour. Pre–processing techniques are also used to process the data and visualizations such as word cloud which visually represents ‘mostly used’ words in the dataset. This model can be efficiently used by online retail stores to make decisions on the production of different commodities**.**

Keyword

Analysis – Text review – Emotions – Sentiments – Sentiment Analysis – Natural Language Processing – Machine learning – Consumer behaviour

Introduction:

There are numerous brands that are present in the market; selecting one will be a tough task for a consumer. The advancement of E-Commerce influences the buying routine of customers. Buyers make the desired decision centred on the reviews present in E-commerce. In present years, social networks have turned very popular; so there is a chance that because of those sites, the expansion of data can be uncontrollable in the future. The customers together with manufacturers will attain as of analysing the positive along with the negative sentiments regarding every product that can well be attained via SA. SA stands as chief tasks in NLP. By employing SA, the mood or attitude of the critic can well be determined as negative or positive. In SA, all product reviews to be summarized and sentiments are to be classified.

SA stands as a field that evaluates the people's opinions, evaluations, sentiments, attitudes, appraisals, as well as the emotion that they encompass on entities cherish products, organizations, services, and people. The links between SA and product design stay comparatively uncharted regardless of the swift advancements of SA in other fields. The primary aim of SA is to recognize the data's polarity on the Web and then to classify them. SA is text centred analysis; however, there are challenges to discover the precise polarity of the sentence. To interpret as well as understand human emotions in addition to feelings, the machines must be dependable and efficient. Most SA is grounded on supervised ML. A vital role is played via the Feature extraction (FE) in addition to classifier design of texts. Term frequency, Term Occurrence, Binary term occurrence, and Terms Frequency-Inverse Documents Frequency (TF-IDF) are the disparate methods for Feature Selections (FS). The TF-IDF usually uses the sentiment lexicon to choose the feature words as well as calculate weights that were broadly applied to traditional NLP tasks. Several methods for SA were proposed in the precedent decades, most of which are centred on computational linguistic approach and ML approach, for instance Naive Bayes (NB) and XG Boost. From several methods, it can be stated that the ML approach exhibit higher performance than the computational linguistic approach.

ALGORITHM:

1.The dataset of an online store was acquired

2.The data analysis is then carried out where the data is cleaned and is checked for any null values or duplicate values and then removed. The 3 data sets are segregated according to the ratings for an easy view.

3.The data pre-processing is then done where the ratings of customers are segregated as negative and positive ratings. A bar graph is plotted for the same.

4.The text is retrieved based on each rating

5.The most used word, the most used positive and most used negative words are collected.

6.The text pre-processing is then done where the text is cleaned by removing the punctuators, changing the upper case to lower case

7. Stop-words are removed.

8.Text lemmatization is done to analyse as a single item.

9.Fitting of the naïve bayes model is done and the accuracy is measured.

METHODOLOGY:

Data Acquisition

The dataset was acquired in 3 different JSON formats and were labelled dataset. A large amount or reviews was manually labelled and was quite impossible. Therefore, the data was pre-processed and was used Active learner to label the datasets. As amazon reviews comes in 5-star rating based generally 3-star ratings are considered as neutral reviews which means neither positive nor negative. So, any review which contains a 3-star rating will be discarded from dataset and take the other reviews and proceed to next step labelling the dataset.

Data Pre-Processing

Tokenization: It is the process of separating a sequence of strings into individuals such as words, keywords, phrases, symbols, and other elements known as tokens. Tokens can be individual words, phrases, or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens work as the input for different process like parsing and text mining.

Removing Stop Words: Stop words are those objects in a sentence which are not necessary in any sector in text mining. So generally, ignore these words to enhance the accuracy of the analysis. In different format there are different stop words depending on the country, language, etc. In English format there are several stop words.

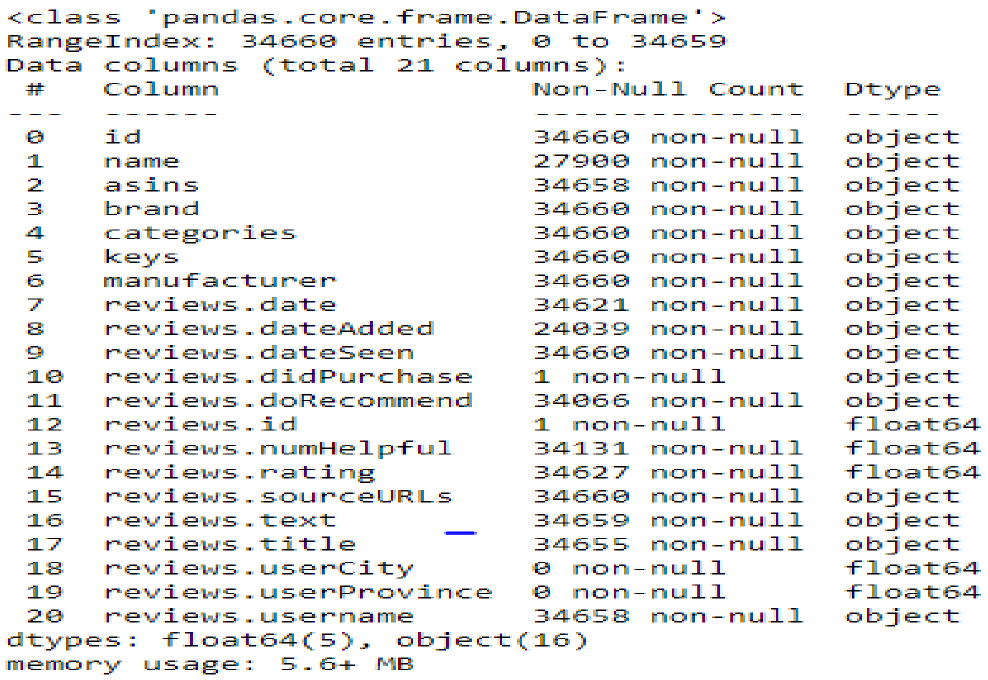
POS tagging: The process of assigning one of the parts of speech to the given word is called Parts of Speech tagging. It is generally referred to as POS tagging. Parts of speech generally contain nouns, verbs, adverbs, adjectives, pronouns, conjunction, and their sub-categories. Parts of Speech tagger or POS tagger is a program that does this job.

Dataset Description:
In this study, the Amazon review dataset is used, which comes from Stanford Network Analysis Platform (SNAP) and is publicly available at [5], [6]. Amazon reviews dataset contains about 83 million unique reviews, covering 24 main product categories, spanning from May 1996 to July 2014. Table I shows the main 24 product categories, and the total amount of reviews and the number of products covered. Each record in the dataset is related to one review for a specific product in the related category. Each record contains the following nine tags: reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, and reviewTime. Clearly, each tag expresses a certain value. The following is the description of each tag.
1) reviewerID: represents the reviewer ID in Amazon web-site.
2) asin: represents the identification numbers that identify the items.
3) reviewerName: represents the name of the reviewer.
4) helpful: represents the ratio of customers who found the review helpful, e.g [20/32]

Text Normalization:
To prepare the dataset, and to extract the reviews text with their helpfulness rating from the original dataset (JSON files), a special parser was implemented in order to extract the specific tags that we need. For each comment, only the review text and the helpfulness rates were extracted. As mentioned in figure 1, each object in the JSON files represents one review with its information by nine tags. Only the “helpful” and the “review Text” tags were considered. The remaining other tags which are (“reviewerID”, “asin”,“reviewerName”, “overall”, “summary”, “unixReviewTime”,and “reviewTime”) were ignored because they are useless forth is taskThe results showed that the review is considered as helpful if the percentage of the helpful votes on the review exceeds 60 % from all votes. In this work and due to the divergence of customers’ views, the review would be considered helpful if the percentage of the helpful votes on the review exceeds75 % from the total amount of votes. On the other hand, the review would be considered unhelpful if the percentage of the helpful votes on the review less than 35 % to ensure that the considered review is extremely unhelpful review.

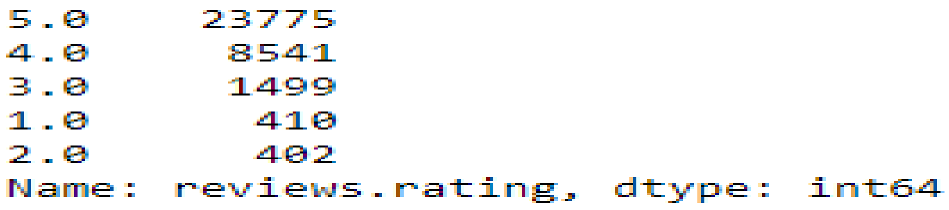
Exploratory Data Analysis:

The datasets 1,2 and 3 are pre-processed and analysed where in each data set we order them according the ratings from highest to the lowest



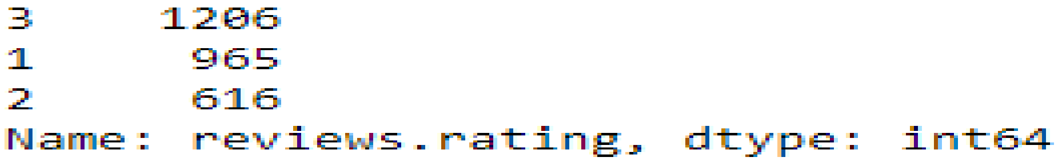
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34660 entries, 0 to 34659
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    34660 non-null  object
 1   name                  27900 non-null  object
 2   asins                 34658 non-null  object
 3   brand                 34660 non-null  object
 4   categories            34660 non-null  object
 5   keys                  34660 non-null  object
 6   manufacturer          34660 non-null  object
 7   reviews.date          34621 non-null  object
 8   reviews.dateAdded     24039 non-null  object
 9   reviews.dateSeen      34660 non-null  object
 10  reviews.didPurchase   1 non-null      object
 11  reviews.doRecommend   34066 non-null  object
 12  reviews.id            1 non-null      float64
 13  reviews.numHelpful    34131 non-null  float64
 14  reviews.rating        34627 non-null  float64
 15  reviews.sourceURLs    34660 non-null  object
 16  reviews.text          34659 non-null  object
 17  reviews.title         34655 non-null  object
 18  reviews.userCity      0 non-null      float64
 19  reviews.userProvince  0 non-null      float64
 20  reviews.username      34658 non-null  object
dtypes: float64(5), object(16)
memory usage: 5.6+ MB
```

Figure 1-Preprocessing



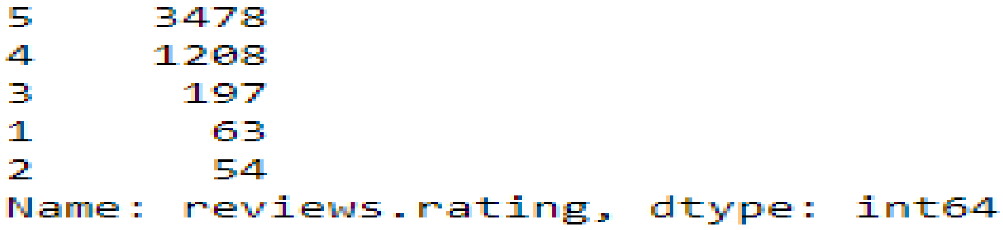
```
5.0    23775
4.0     8541
3.0     1499
1.0      410
2.0      402
Name: reviews.rating, dtype: int64
```

Figure 2- Dataset 1 rating order



```
3    1206
1     965
2     616
Name: reviews.rating, dtype: int64
```

Figure 3- Dataset 2 rating order

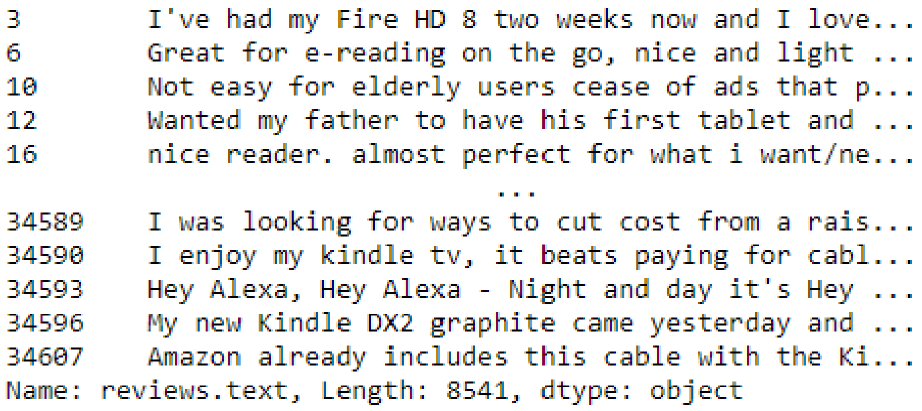


```
5    3478
4    1208
3     197
1      63
2      54
Name: reviews.rating, dtype: int64
```

Figure 4- Dataset 3 rating order

RETRIEVING PARTICULAR TEXT BASED ON RATING

Here each rating along with their respective reviews are gathered



```
3        I've had my Fire HD 8 two weeks now and I love...
6        Great for e-reading on the go, nice and light ...
10       Not easy for elderly users cease of ads that p...
12       Wanted my father to have his first tablet and ...
16       nice reader. almost perfect for what i want/ne...
                               ...
34589    I was looking for ways to cut cost from a rais...
34590    I enjoy my kindle tv, it beats paying for cabl...
34593    Hey Alexa, Hey Alexa - Night and day it's Hey ...
34596    My new Kindle DX2 graphite came yesterday and ...
34607    Amazon already includes this cable with the Ki...
Name: reviews.text, Length: 8541, dtype: object
```

MOST USED WORDS IN REVIEW

The most to least used words are collected for the datasets 1,2 and 3

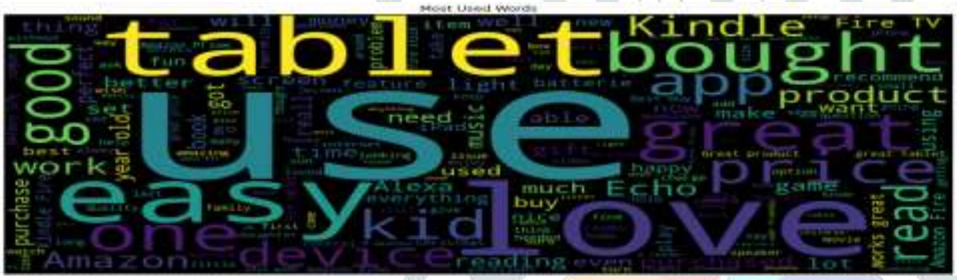


Figure 5- Most used words (1)

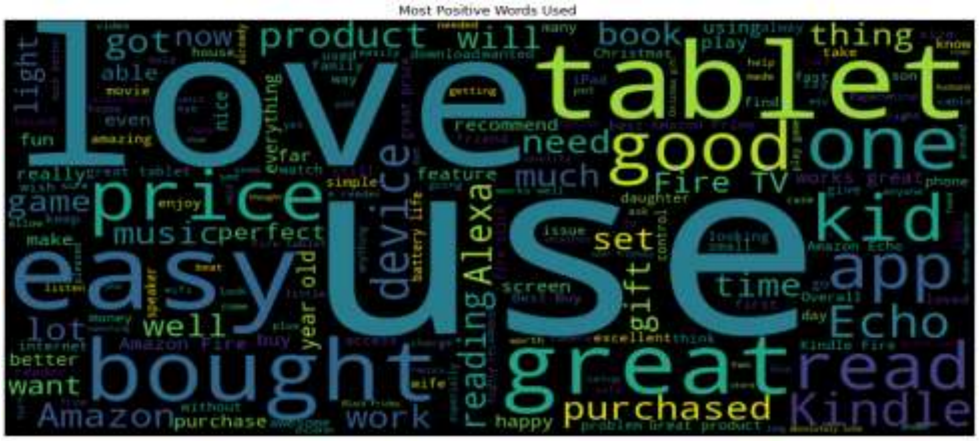


Figure 6- Most used words (2)

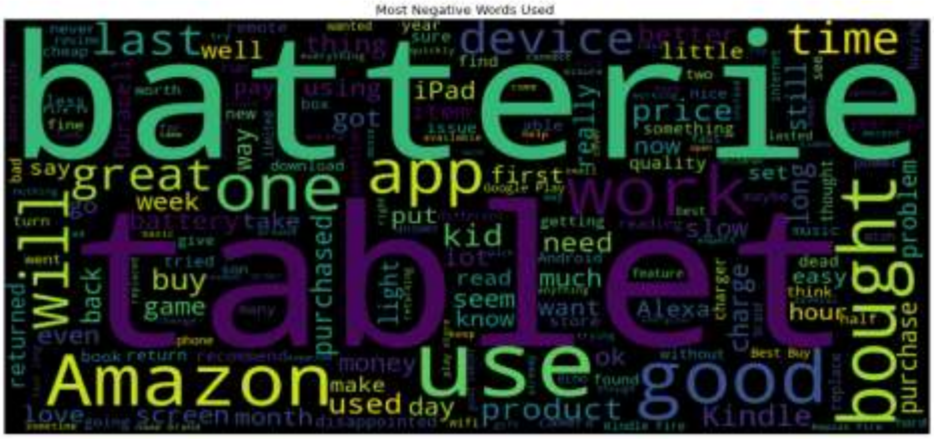


Figure 7- Most used words (3)

TEXT PREPROCESSING

- Clean text

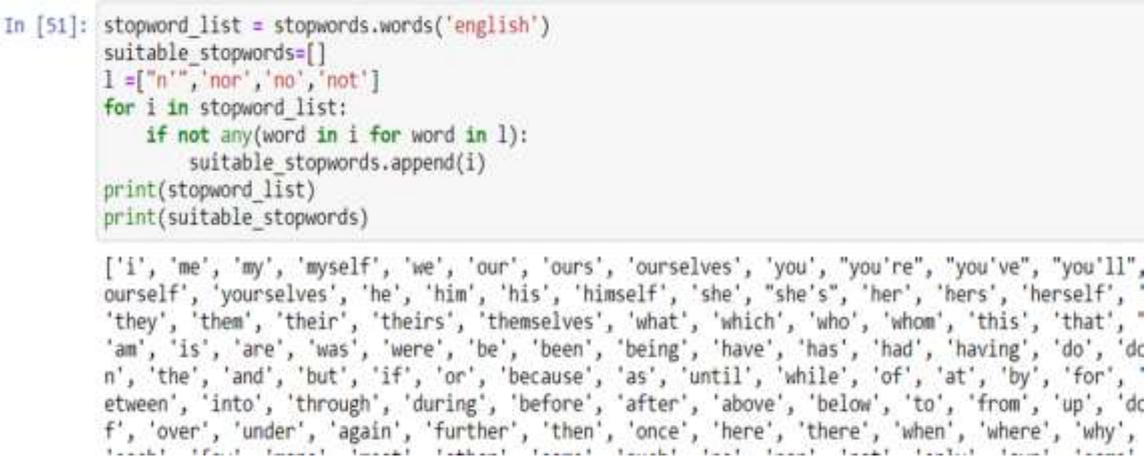
  The punctuators are removed, and the upper-case alphabets are changed to lower case.

- Remove the stop words

  Words that do not give much meaning to the sentence

- Remove stop words
- Lemmatize text

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.



```
In [51]: stopword_list = stopwords.words('english')
         suitable_stopwords=[]
         l =["n'",'nor','no','not']
         for i in stopword_list:
             if not any(word in i for word in l):
                 suitable_stopwords.append(i)
         print(stopword_list)
         print(suitable_stopwords)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll",
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', '
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'dc
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', '
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'dc
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
```

Figure 8-Text pre-processing

TRAIN AND TEST DATA SPLIT

The dataset is split into train and test

Set

```
Train Set Shape        :(30181, 700)
Test Set Shape         :(7546, 700)
```

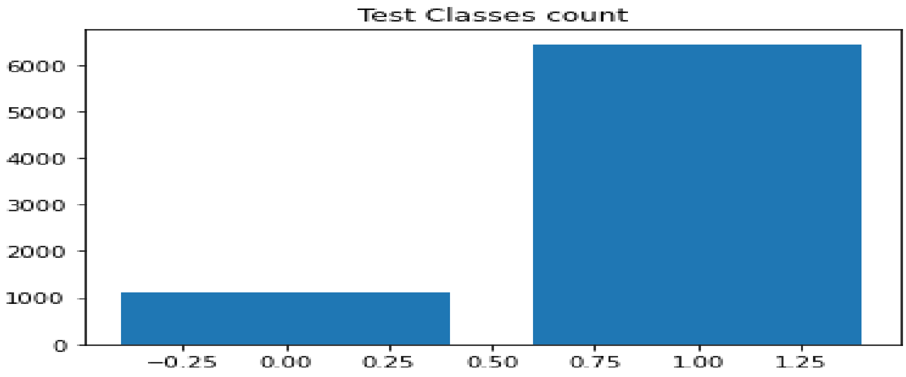


Figure 9- Test set

JETIR

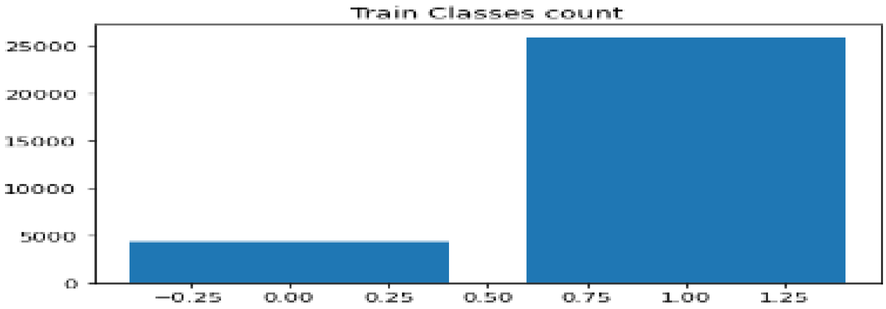


Figure 10- Train set

The Naive Bayes classification algorithm is one of the probabilistic classifiers. It is mainly based on probability models that mix strong independence assumptions. The independence assumptions mostly do not have a reality impact. The value of the probability threshold parameter is used if one of the mentioned dimensions of the cube is null/empty.

Naïve Bayes is a probabilistic technique for constructing classifiers. The characteristic assumption of the naive Bayes classifier is to consider that the value of a particular feature is independent of the value of other features, given the class variable. It is a form of supervised learning. It is considered to be supervised since naive Bayes classifiers are trained using labeled data.

RESULT:

MODEL FITTING - NAÏVE BAYES

	precision	recall	f1-score	support
0	0.48	0.82	0.60	1106
1	0.96	0.85	0.90	6440
accuracy			0.84	7546

```
macro avg      0.72     0.83      0.75      7546

weighted avg 0.89      0.84      0.86      7546

AUC  0.8337025316455697
```

INFERENCE:

The model was used for a small dataset, it can be widened to a larger dataset, and it will be efficient to be used as in large scale for useful purpose. Online shopping scams involve scammers pretending to be legitimate online sellers, either with a fake website or a fake ad on a genuine retailer site. While many online sellers are legitimate, unfortunately scammers can use the anonymous nature of the internet to rip off unsuspecting shoppers. The best shopping websites are easy to navigate, have great prices, and offer exceptional customer support. There are a lot of factors that go into a customer's decision to make a purchase from the company. The performance of the e-commerce website's shopping cart must be seamless to enable orders to go through. It must fit well with the online catalog, your customer service desk, and the payment processing gateway.

CONCLUSION:

The algorithms used like Tokenization, stop words removal, POS tagging, Text normalization are basics of Natural Language processing which helps to do teeny-tiny baby steps to build a model in future. The naïve Bayes algorithm classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. Naïve Bayes classifiers are highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. The accuracy of the model using naïve Bayes algorithms rounds up to 84% which is comparatively good at predicting output for customized inputs. The text processing is the manipulation of text, especially the transformation of text from one format to another. The initial step of text processing is Date pre-processing, and it goes into Feature extraction and finally fitting a ML model to predict the output. The appropriate text processing techniques are used before fitting the naïve Bayes ML model. Word Clouds for most frequently used words are plotted to know the intensity and frequency of the words in the dataset.

REFERENCES:

WEB:

- https://www.tandfonline.com/doi/full/10.1080/02650487.2019.1617651
- https://ideas.repec.org/a/vrs/subboe/v65y2020i1p54-66n4.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7501828/
- https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/
- https://retail-insider.com/articles/2020/09/why-product-reviews-are-important-from-a-consumers-perspective/

BOOKS:

- http://troindia.in/journal/ijcesr/vol8iss6/73-77.pdf
- https://tel.archives-ouvertes.fr/tel-02014508/document
- https://arxiv.org/pdf/1805.03687.pdf