Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

A STUDY TO IDENTIFY THE INDIAN MARKET SENTIMENTS OF PRODUCT FEATURES

LOPAMUDRA BERA

M. Phil Scholar, Institute of Management Study. Email: lopamudra.bera26@gmail.com

Dr. MEGHDOOT GHOSH

Associate Professor, Swami Vivekananda Institute of Science and Technology.
Email: ghoshdrmeghdoot@gmail.com

Dr. RAHUL KUMAR GHOSH

Assistant Professor, Institute of Management Study. Email: rahulghosh.0101@gmail.com

Abstract

The sentiment analysis is a process which helps to categorize the people's opinion expressed in customers' reviews which determines the customers' attitude towards any service, or product is positive, neutral or negative. It is mostly used in social media monitoring to allow practitioners and researchers to gain wider public (Indian consumers in Indian market scope) opinions overview behind social networks, product review, etc. This research is interested in identifying more in expressed public sentiments (positive, negative or neutral) towards features of product rather to identify sentiments in product review texts. There are several supervised machine learning methods that have been conducted in the past to calculate product features sentiment scores and sentence-level or document-level sentiment analysis respectively. In this research work, the sentiment or polarity scores of features of product are calculated with the help of rule-based Python Text Blob library. Product feature based sentiment analysis is conducted by supervised machine learning methods. The feature extraction method, TF-IDF method is applied. Finally the result of this research shows (i) the features of product became good indicator during determination of polarity or sentiment classifications of review texts, (ii) Random Forest supervised machine learning performed well with 92% accuracy in polarity classification with the hyper parameter Information Gain.

Keywords: sentiment analysis, opinion polarity, product feature, supervised learning.

1. INTRODUCTION

Nowadays, opinion/sentiment analysis becomes a trend in research by researchers and businesses. Many industries, businesses show their interest to listen to the experience of the services they offer directly from consumers who consume their services. Along with the analysis of sales and marketing data, they have already started to accept the buying experience fact from consumers' voices. The key aspect in the B2C industry is customer relationship management (CRM), which is key to unlocking the potentiality of the business growth. CRM helps industries to keep consistent relationships and interactions with customers and potential customers. The goal of the CRM is to stay connected with customers, streamline services and improve profitability. One of the main components of four CRM is Feedback from customers. With the huge increase of internet usage through different media gadget, customers become easily reachable to the industries. So,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

collection of direct feedback from customers for industries is logging feedback by customers to industries, both becoming very quick and easy through ecommerce applications launched by industries. Also the format of feedback forms are modified by providing the opportunity to write detailed feedback instead of only the old pattern style of selecting emoji. Most of the industries, such as sales, marketing, recruiting, customer services, business development, education delivery, etc., are waiting for customers' feedback, opinion, and suggestions before they plan or decide to adopt any new long or short-term strategy.

A huge product and service-related customer review online data is available in various formats such as structured, semi-structured, or unstructured formats. The major interest has grown among many researchers and business practitioners to retrieve sentiment or opinion from those reviewing data about the products or services. Different approaches and techniques are used by researchers and practitioners for unstructured text analysis. Opinion/sentiment analysis is a well-known technique which extracts differently polarized (positive, neutral, or negative) subjective information from the source material. Further, this extracted information helps to process or make different business strategic decisions.

Using opinion/sentiment analysis, opinion or sentiment or polarity can be detected that is expressed for the product through one's posted product review on the product feature. As the usage of online shopping is increasing, and also online shopping site provides the opportunity to customers to share their opinion through reviews of the bought product and buying experiences. So every individual is keen interest to share his opinion through different mediums connected by the internet about the bought products. Other customers are utilizing these reviews while making a new purchase decision. The habit of researching any product is induced in every individual before taking a decision to buy any product either online or through direct purchase. Product reviews are becoming an essential part of any online shopping store's marketing and branding which help further to build loyalty, trust, and bonding. Sentiment analysis helps to analyze this opinioned data on different products and gives important and meaningful insight from the review text opinion data to help buyers or sellers or businesses to take decisions.

An individual needs enough information about the product before he makes the decision to buy. By providing product information to interested customers through other customers' voices, businesses can increase their sales. So, online product reviews are important to companies and businesses. Another very important factor that is revealed through customers' product reviews is the elevation and evaluation of the reputation, the standard of ecommerce stores. These are the main reasons why different brands like Puma, Samsung, Sony. JBL, etc. acquired a huge online shopping market in a very short span of time. There is a long list of available online shopping marketplace, this research work is interested in the sentiment analysis on product features specifically based upon Indian market sentiments on an online marketplace such as Flipkart.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

The main focus in sentiment analysis is to identify whether the source material texts are objective or subjective, and to detect the sentiment polarity of those subjective texts. Pang et al., Nagakawa et al. [21, 19] mentioned that the “Document-level” and “Sentence-level” classifications fall under the supervised machine learning classification. The vectorization process of Bag of Word (BoW) model is used in Document-level classification which turns document texts into fixed-length vectors by counting the frequency each word appears in documents, but this model does not maintain the order of words and where they occurred in the document. The “Document-level” classifications technique classifies vectorised words into polarity/sentiment (neutral, negative or positive) with consideration of the frequency of the words appear in the documents. Ding et al., Qu et al., Socher et al. [11, 22, and 28] mentioned that the “Document-level” classification is not providing the detail of polarity/sentiment. So the polarity/sentiment expressed in texts is further discovered by the “Sentence-level” classifications. The interest of this research work is to study sentiment classification on features of product instead of the complete product review text.

This research work is interested in identifying the polarity/sentiment scores of features of product from product review texts using the rule-based TextBlob Python library. The polarity/sentiment score is the numeric form of opinions of the features of products that are expressed in the product reviews. Analysis of polarity/sentiment is the unsupervised machine learning, detecting the polarity/sentiment scores of features of product, and involves creating word lexicon based opinion which consists of opinion polarity (positive, negative or neutral, etc.) annotated opinion words. At last, the sentiment analysis aggregates the sentiment polarities of opinion words to calculate the polarity/sentiment scores. All these above-stated tasks are embedded into the TextBlob package, hence the use of this library reduces the time of the polarity scoring technique.

The selection of the right features and feature weighting are the two most important techniques of the machine learning process. Martineau et al. [17] mentioned that there are various weighting methods have been used like “TF”, “TF-IDF”, and “delta TF-IDF” in the technique of feature weighting. Yang et al. [31] suggested various mathematical methods are suggested such as “Chi-Square”, “Information Gain”, “Mutual Information”, and term strength in the technique of feature selection, they help to improve the performance classification rate through the dimensionality reduction process. Chang et al. [7] found the result of the application of MI and IG on feature spaces are compared using the LIBSVM tool. In this research, the “TF-IDF” vectorization procedure is used in the process of feature extraction and extracted features of product is further processed in the next sentiment classification process. “Term Frequency-Inverse Document Frequency” (“TF-IDF”) performs its calculation by taking the importance of a word by considering how frequently that word appears in a review text and all reviews.

Product attributes or components can be defined as product features. A feature based sentiment analysis is the analysis where finding sentiment or polarity from the document

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

by the using of features of product. This sentiment analysis is a classification of polarities based on product features from the product reviews posted by the customers. For this work a fine grained analysis model is required which can efficiently classify and predict the polarity of given new customer's review from the existing reviews data. To find out this fine grained efficient classifier model, ten different supervised machine learning methods' classification results are compared in this research work. These ten supervised learning methods are "Logistics Regression", "Random Forest", "Bernoulli Naïve Bayes", "K-Nearest Neighbor", "Decision Tree", "Support Vector Machine", Ensemble learners: AdaBoost, AdaBoost hyper parameterized with Random Forest, XGBoost, Bagging hyper parameterized with "Logistic Regression". Each of the classifier's evaluation is done through different evaluation metrics, which includes "accuracy", "precision", "recall", "f1-score", "AUC-score" and "ROC curve". This study also extended by combining strong learners to propose a high performed stack model. To summarize, in this research work, the "Random Forest" tool is used by considering the most efficient tool with 92% accuracy in text classification rather than comparing the efficiency of different machine learning models. Also "Logistic Regression", "SVM" and AdaBoost hyper parameterized with Random Forest classifiers show the good accuracy greater than 90%.

2. RESEARCH QUESTIONS

The major researches questions arise associated with this work are as follows:

a) How to calculate the sentiment/polarity score of features of the product?

b) How to derive a specific and right feature spaces for sentiment/opinion analysis?

c) Which model can be proposed as optimized model to perform multiclass sentiment analysis?

The objectives were set to answer the above research questions:

- To calculate the sentiment scores of product features of the product reviews.
- To derive the specific feature space for sentiment analysis.
- To compare different machine learning algorithms in the classification of multiclass sentiment analysis.

3. RELATED WORK

First, the review of sentiment analysis on product features are focused, second, reviews of feature extraction and selection are focused, then reviews are concentrated on machine learning classification methods used in sentiment analysis.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

3.1 Sentiment Analysis on Product Features

Many researches and studies were conducted to extract product features and associated sentiment expressed for the product.

Blei et al. [5] proposed “Latent Dirichlet Allocation (LDA)”, a generative model for topic classification, allows observations sets to be explained by unobserved groups. Blei et al., Guo et al., Lin, Zhai et al. [5, 12, 16, 33] used this topic modelling method was used in sentiment analysis, this method can be used to group features of the product.

Zhai et al. [32] proposed a constrained semi supervised by using “Expectation Maximization (EM)” algorithm (proposed by Dempster et al. [10]) was used to group similar expressions of product features. An EM formulation was used to solve this problem with two soft constraints: (a) Some common words shared expressions of feature from the same group (like ‘battery power’ and ‘battery life”), (b) Synonyms from a dictionary from the same group (like ‘movie’ and ‘film’).

Abraham et al. [2] studied of opinion analysis based upon the aspects / features extraction through the data collection (200 Amazon reviews of 5 mobile product review datasets were collected from the well-known Kaggle website), the pre-processing (tokenization, removal of punctuation marks, removal of stop words and represented “POS Tagging” through a general Penn Treebank POS tag), the aspects extraction and opinion of words. They only extracted explicit aspects e.g. “voice quality”. Then based on the polarity identification of the opinion word (negative or positive) with the help of machine learning approaches such as “Multinomial Naïve Bayesian”, “Bernoulli Naïve Bayesian (BNB)”, “K-Nearest Neighbour (KNN)” and “SVM”. They presented precision and recall values for different mobile product reviews with respect to positive and negative sentiment models. Finally presented the accuracy values obtained for different mobile product reviews. They found “KNN” and “BNB” performed well in comparison to the other two algorithms.

3.2 Feature extraction and selection methods

As of many researchers have been used many features in sentiment analysis like “unigrams”, “bigrams”, “trigrams”, higher level of “n-grams”, “POS tags unigrams”, “dependency tree patterns”, negation-tagged token, subjective extraction patterns, and “adjectives” to detect good indicators to generalization of text classification.

A) Ngram:

Pang et al. [21] mentioned that the performance of unigrams is much better than bigram and outperformed adjectives when they used in feature space as features. Dave et al., Ng et al. [9, 20] mentioned that bigrams and trigrams, they performed better than unigrams.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

B) POS (Part-of-Speech) Tags:

Mejova et al. [18] examined the basic unit extracted from texts and got best accuracy attained by polarity classifier with adjective and verbs.

C) Dependency Tree Pattern:

Blech et al., Collins et al., Lin, Sang et al., Sha et al. [4, 8, 16, 25, 26] mentioned that the performance of the sentiment analysis became high with the use syntactic dependency trees than use of features from "Bag-of-Words".

D) Subjective Extraction Pattern:

Sivaganesan et al. [27] suggested feature-based sentiment analysis where the features are extracted from the web scrape ecommerce website reviews and corresponding polarities are predicted using the TextBlob and SVC model. They used TextBlob for data processing and Yake (Yet another Keyword Extractor) based keyword extraction technique to extract features from selected reviews. The basic SVM model offered them the best accuracy of 92.26% in comparison to TextBlob and Column Transformer models.

The selection of numerical features is also important in any sentiment analysis. The Term Frequency (TF) and presence are mostly used feature value assignment methods also can be called as feature weighting methods.

Turney et al. [29] proposed vector space models such as "term document-matrix", "word-content matrix" and "pair-pattern matrix". Based on representation of rows and columns, matrices (feature space) are different.

Bordoloi et al. [6] used a domain dedicated polarity assignment technique to assign significant and relevant polarity to the important keywords to describe and present the best underlying meaning, opinion or sentiment of the data. They proposed a six staged advanced sentiment analysis model called ESAGBA (ECommerce Sentiment Analysis using Graph Based Approach) through the comparative results for the mobile. Their model achieved 82% highest accuracy compared to the other three models built upon the supervised machine learning algorithms like "Naïve Bayes", "SVM" and "Maximum Entropy".

E) TF-IDF (term Frequency-Inverse Document Frequency):

The numerical statistic, "TF-IDF" weight, represents the importance of words to a document in a certain type of dataset and reduces the weight of the most occurring word that has less contribution to the classification.

Wang et al. [30] mentioned that for sentiment analysis, with a novel weighting method they proposed a feature-based vector model. They combined feature-based vector with three weighting methods respectively. Their proposed "High Adverb of Degree Count

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

(HADC)", the "TF" and "TF-IDF"; they used "SVM" as the classifier in their experiment. Their proposed model could classify reviews into positive class and negative class and represent the sentiment strength by "adverb" of degree. They achieved 92% accuracy in best case and 88% accuracy in worst case using "six-tuple model" and "HADC" weighting algorithm.

3.3 Machine learning classification methods

The selection of the right machine learning algorithm is also important out of all tasks in text classification using the machine learning process. "KNN (K-Nearest Neighbor)", "Naïve Bayes", "Logistic Regression", "SVM (Support Vector Machine)", "decision tree" classifier, "Random Forest", Ensemble are the different supervised machine learning algorithms used by different researchers during many years in their sentiment analysis or text analysis works.

Saifullah et al. [24] performed a study by comparing different supervised machine learning algorithms ("KNN", "Bernoulli Naïve Bayes", "Decision Tree Classifier", "Support Vector Classifier", "Random Forest", and "XG-Boost") to detect anxiety on comments of social media regarding government launched programs to handle "COVID-19 pandemic" situation. 4862 data were crawled from YouTube comments for this analysis. The supervised machine learning algorithms were fed with features extracted from count-vectorization and "TF-IDF". They found the best precision score from the "KNN" model, and best recall value from the "XG-Boost" model. Though "Random Forest" model used both count-vectorized and TF-IDF feature extracted data and showed 84.99% and 82.63% accuracy.

Jagdale et al. [14] pre-processed Amazon product review dataset (data size=13057) of different categories such as Camera, Laptops, Mobile phones, tablets, TVs, video surveillance through "tokenization", "stop word removal", "stemming", "punctuation marks removal", etc. and converted into 'Bag-of-Words". They analyzed and calculated sentiment scores for each sentence. They performed a comparative study with Naïve Bayes and SVM machine learning algorithms and found Naïve Bayes offers better accuracy of 97.18% than SVM. They compared scores of accuracy, precision and F-scores for each category of products.

Abhinand et al. [1] observed the trend in positive and negative sentiments. The Indian Budget-2022 dataset had been split into three categories with respect to the date of tweeting. The Bernoulli Naïve Bayes, SVM, Logistic Regression model has been used separately in each of these categories and the results have been observed in comparison with Bernoulli Naïve Bayes and SVM models. They presented the evaluation metrics for the models predicting negative and positive sentiments for each algorithm used in this analysis. They found the logistic regression model acquired the highest accuracy of 81.4% in comparison to the other algorithms.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

Chang et al. [7] used SVM algorithm is used in classification, regression and many other tasks.

Pang et al. [21] claimed SVM is an appropriate tool for sentiment classification and worked better than the "Naïve Bayes" and "Maximum Entropy" algorithms.

Rogati et al. [23] mentioned that SVM performed better than the KNN algorithm.

Janhavi et al. [15] mined the product which carried more percentage of product reviews from 59 Laptop product review data of Flipkart.com (data extraction happened using Flipkart product API using JSON). The ROCK algorithm was applied to cluster the review data. The CART algorithm was applied to compare the dictionary words with the 5 clustered data (output of the ROCK algorithm) and to classify the review data into positive and negative opinions.

Hung et al. [13] found the ensemble framework improves the accuracy of the system through inclusion of diversified feature selection method into the ensemble framework to select useful features and make a smaller size of the feature set.

4. METHODOLOGY

This work targets to study different opinions/sentiments expressed on features of the product using sentiment analysis. The four main steps are carried out in this study: (a) data collection, (b) system analysis and concept, (c) system design and (d) comparative study. Dataset used in this study, the secondary survey dataset identified as an option the best fit to the Indian online marketplace considered as an example of Flipkart for the business span of six years (2017 to 2022). This study has used the concept of technologies and computerization which obviously includes the use of various supervised machine learning methods to acquire the best accuracy to classify opinion polarity expressed towards the product and detect the anxiety and neutral sentiment oriented products to help the marketplace to improve their choices and productions. Ahuja et al. [3] used the result of preprocessed data is used to calculate the sentiment scores using the Python built-in natural language processing library TextBlob. The polarity scores and the scores subjectivity are calculated by using TextBlob library. "The sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [- 1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.". Only polarity scores are used in this study to evaluate three classes of sentiments (positive, negative and neutral) using simple programming decision construct. This work is developed with the application of "TF-IDF" Vectorization as feature extraction methods. "Logistics Regression", "KNN", "Random Forest", "Bernoulli Naïve Bayes", "Decision Trees", and "Support Vector Machine", Ensemble models: "AdaBoost", "AdaBoost hyper parameterized with Random Forest", "XGBoost", "Bagging hyper parameterized with Logistic Regression" and stacking models are used in machine learning models. An iterative validation process is

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

known as cross validation process where data is splitted into training data and testing data. In this study the confusion matrix, roc_auc curve and classification report are used to evaluate each method's outcome. In this study system development involves the prototyping model to detect opinion polarity or sentiment expressed by people towards the product. The prototyping model (Figure 1) includes three stages of text processing: pre-processing of data, detection of opinion polarity through sentiment classifications, and comparison study.

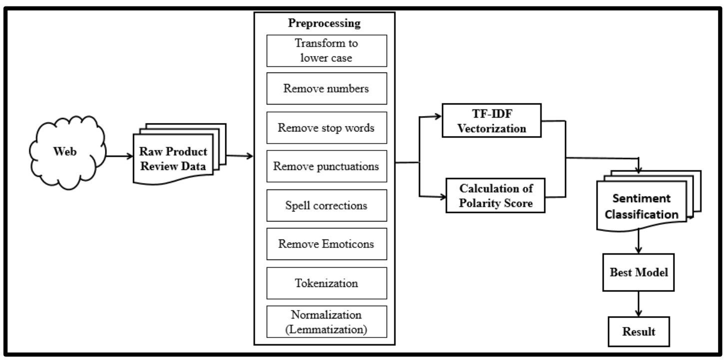

Figure 1. The prototyping model includes three stages of text processing

5. EXPERIMENTAL SETUP AND RESULT

The results and discussions of this study will be discussed in this section which includes the description of dataset, detailed procedure of the three stages of this study, the results obtained from process of polarity detection based on sentiment analysis of product reviews dataset. The validation of results are carried out through "accuracy", "precision", "recall", "f1-score", "auc-score" and through "roc-curve" visualization.

5.1: Dataset

In accordance with the research objectives, the survey dataset has been identified as an option the best fit to the Indian online marketplace considered as an example of Flipkart for the business span of six years (2017 to 2022). In this work, the Flipkart reviews dataset is used from the Kaggle site (*https://www.kaggle.com/code/naushads/flipkart-reviews-scraping/data).

Research Approach: Quantitative Research

Data Collection: Secondary Data

Sample Size: 9386

Data Collection Source: Kaggle site*

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

Population

Indian Online Marketplace: Flipkart

Product Category: Electronics

Product Line: Headset of 8 categories

Attribute Used: "review" of categorical datatype

Tool Used: Python Programming Language

Missing Value: No missing value exists

Outlier Detection: Not applicable as selected attribute is not numeric type

5.2: Data Pre-processing

The initial tasks required to perform feature-based sentiment analysis are the extraction of standard dataset considering the headset product reviews from the Kaggle website and then performing pre-processing on these dataset where transformation of lower case, removing numbers, removing stop words, removing punctuations, correction of spellings, removing emoticons, tokenization, normalization through lemmatization and calculation of opinion scores are applied. Figure 2 shows the sample output of pre-processed reviews.

	review	review_rsw
0	first of all with mi 18 watt charger,it got fu...	first mi watt charger got full charged mins am...
1	I am using this product from 2 daysLet me shar...	using product dayslet share experiencethis gen...
2	LONG BUT WORTH READING.Honest review after one...	long worth reading honest review one week go g...
3	writing review after using it for two days ..f...	writing review using two days first best batte...
4	The bass provided is a decent one and the soun...	bass provided decent one sound clarity pretty ...

Figure 2: Sample Output of pre-processed Reviews

5.3: Polarity Scores Calculation and Dataset Labelling Process

In this study Python TextBlob package calculates the polarity/sentiment scores and subjectivity scores of each review text. TextBlob module usage its own pattern library to

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

calculate polarities. The dataset is classified into three categories, namely negative, neutral and positive. The sentiment categories further are labelled from the polarity scores based on the algorithm presented in Figure 3.

```
if  polarity_score > 0 :
            sentiment= 'positive'
elif polarity_score = = 0:
            sentiment= 'neutral'
else:
            sentiment= 'negative'
```

Figure 3: Algorithm of classification of sentiment category based on polarity scores

From the Figure 5, it can be observed there are 8019 positive reviews data, 704 neutral reviews data and 663 negative reviews data exist in the dataset out of 9386 reviews data of 8 different categories of headphone products.

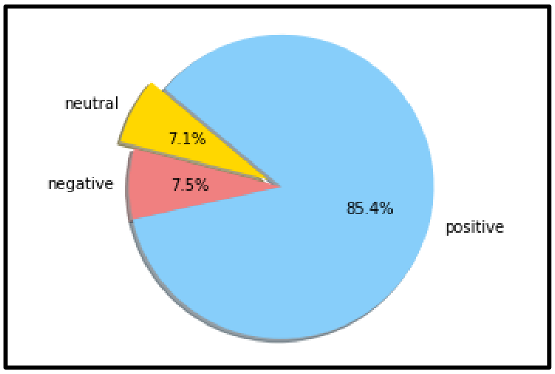

Figure 4: Percentage of three different classes

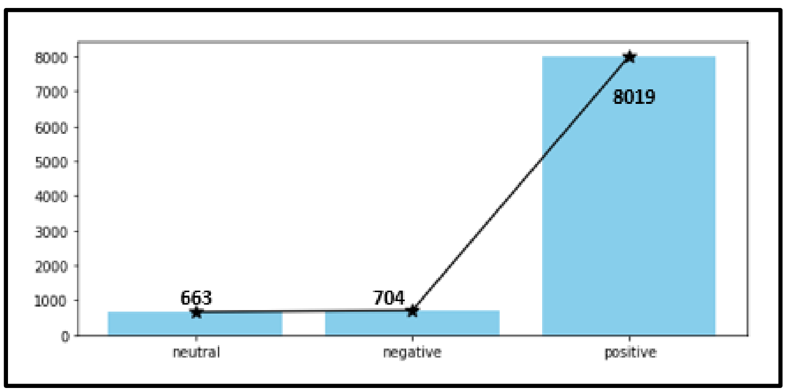

Figure 5: Count of three different classes

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

In this analysis N-gram analysis is also studied. N-grams can be defined as neighboring sequence of tokens in document. The following algorithm is used to generate n-grams in this study as given in Figure 6.

```
def generate_N_grams(text, ngram):
    temp=zip(*[text[i:] for i in range(0,ngram)])
    ans=[' '.join(ngram)  for ngram in temp]
    return ans
```

Figure 6: Algorithm for generating n-grams

5.4: Feature Extraction method

In this study, TF-IDF vectorization method is used in feature extraction process and extracted features of the product from this method is further used in the next step of sentiment classification process. “Term frequency-inverse document frequency (TF-IDF)” performs the calculation by taking the importance of a word by considering how frequently that word occurs in a document and a corpus. The TfidfVectorizer module from Python Scikit-learn package is used to extract features using TF-IDF method. TfidfVectorizer from Scikit-learn implements a different complex formulas to calculate TF-IDF. Thus, the TF-IDF values will be different than this study calculated.

5.5: Sentiment Classification using Supervised Machine Learning Models

Prior to process the data (TF-IDF vectorised data and sentiment classes) two steps of data pre-processing are performed. In the first step of pre-processing, the polarity sentiment categorical values are converted into numerical datatype by using LabelEncoder() of Python Scikit-learn package. This encoding process converts all “positive” values into 2, “neutral” values into 1 and “negative” values into 0 of the “sentiment” column of the dataset. In the next step of the pre-processing vectorised data and encoded polarity classes are splitted into train and test parts into 70:30 ratio. The “train_test_split()” of Python “Scikit-learn” package is used to perform data split operation. After splitting the train dataset consists 6570 data and test dataset consists 2816 data.

The feature extracted data from TF-IDF method and label encoded sentiment class data are further processed using supervised machine learning methods, such as, “Logistic Regression”, “Random Forest”, “Bernoulli Naïve Bayes”, “KNN”, “Decision Trees”, “Support Vector Machine”, Ensemble models: “AdaBoost”, “AdaBoost hyper parameterized with Random Forest”, “XGBoost”, “Bagging hyper parameterized with Logistic Regression” models. Each model is evaluated through “classification report”, “confusion matrix”, “auc score” and “roc curve”.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

5.6: Comparative Study

This discussion of all 5 evaluation metrics clearly provide an indication of the best model among all ten models is Random Forest which consistently provides high values of all 5 evaluation metrics. "Logistic Regression", "SVM" and "hyper parametrized AdaBoost" can be the next followers as the good polarity classifiers after the "Random Forest", the same depicted in the Figure 7. Table 1 depicts comparison of each model with respect to the each evaluation metric.

	Polarity												
	Accuracy	Positive				Neutral				Negative			
		Precision	Recall	F1-Score	AUC-Score	Precision	Recall	F1-Score	AUC-Score	Precision	Recall	F1-Score	AUC-Score
LogisticRegressionCV	91%	93%	98%	95%	69%	71%	33%	45%	35%	61%	43%	51%	29%
KNN	89%	89%	100%	94%	52%	62%	5%	10%	47%	50%	1%	1%	50%
RandomForest	92%	93%	99%	95%	70%	11%	1%	8%	51%	14%	6%	2%	48%
Bernoulli Naïve bayes	88%	86%	9%	93%	50%	55%	3%	6%	49%	89%	6%	10%	47%
Decision Trees	89%	89%	100%	94%	52%	80%	26%	40%	37%	88%	31%	46%	35%
SVM	91%	92%	99%	95%	65%	74%	35%	48%	34%	74%	44%	56%	28%
AdaBoost	89%	90%	98%	94%	52%	52%	12%	20%	45%	47%	12%	19%	44%
Hyper AdaBoost	91%	93%	98%	95%	69%	74%	34%	47%	34%	73%	42%	54%	29%
XGBoost	89%	89%	100%	94%	52%	83%	3%	5%	49%	100%	7%	13%	47%
Hyper Bagging	88%	88%	100%	94%	50%	0%	0%	0%	50%	0%	0%	0%	50%

Table 1: Comparison of each model

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

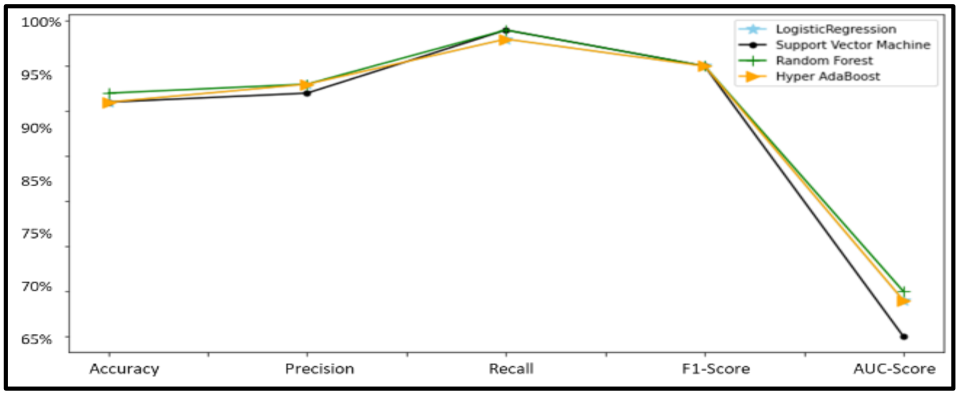

Figure 7: Comparison between four highly performed models

This discussion of all 5 evaluation metrics clearly provide an indication of the best model among all ten models is Random Forest which consistently provides high values of all 5 evaluation metrics. LogisticRegression, SVM and hyper parametrized AdaBoost can be the next followers as the good polarity classifiers after the RandomForest, the same depicted in the Figure 7. Table 1 depicts comparison of each model with respect to the each evaluation metric.

6. CONCLUSIONS

In this work, to study of different opinions expressed on product features using sentiment analysis used different supervised machine learning. We find that the features of product became good indicator during determination of polarity or sentiment classifications of review texts, and Random Forest supervised machine learning performed well with 92% accuracy in polarity classification with the hyper parameter Information Gain for studied dataset. Modern marketing concepts focuses on consumers' needs, satisfaction, sustained promotional activities, long term oriented, social selling approaches, cost-effective, digital pull marketing.

To spot opportunities in existing market and make market recommendations for launching into new markets or targeting new personas, companies require data-driven insights. Analyzing data from multiple sources to track, measure, and see the impact of changing variables is nearly impossible for marketers using traditional marketing analysis as data size is huge. Fortunately, modern analytics and modern BI tools are built and available to handle the needs of today's marketers. Sentiment analysis using machine learning methods are such kind of modern analytics. The evaluated optimized model from this study can be used to analyze to predict the users' acceptance of the product and scope of the product market. The achieved polarity will be used to provide recommendation to users. Which further helps to rectification in disturbed services or enhance the services as per customers' requirements. Which make business consistent and customer retention. Although the current research scope is determined to the specific product category and the product line, but the proposed sentiment analysis model inferenced from this study can be multiplied into other marketplaces including product and service industries as well. It is believe that fine grained pre-processing through conversion of emoticons to words, use of Vader algorithm (2019) for polarity scores calculation, modelling using BERT model of Deep Learning of the same review text data could improve the performance of the proposed model from this study.

REFERENCES:

1. Abhinand, G. Uma Maheswari, V. (2022), Corroboration of Twitter Sentiment Analysis and Event Analysis of Indian Budget 2022 on the Bitcoin Market. Proceedings of the International Journal for Research in Applied Science & Engineering Technology (IJRASET). DOI Link: https://doi.org/10.22214/ijraset.2022.43533.
2. Abraham, M. P., & KR, U. K. R. (2020). Feature based sentiment analysis of mobile product reviews using machine learning techniques. Proceeding of the International Journal, 9(2).
3. Ahuja, S. Dubey, G. (2017), Clustering and sentiment analysis on Twitter data. Proceedings of the 2nd International Conference on Telecommunication and Networks (TEL-NET), pp. 1-5, doi: 10.1109/TEL-NET.2017.8343568.
4. Blache, P., Balfourier, J.-M. (2001). Property Grammars: a Flexible Constraint-Based Approach to Parsing, in proceedings of IWPT.
5. Blei,D., Ng,A., and M. Jordan (2003). Latent dirichlet allocation, Journal of Machine Learning Research, 3(5):993–1022.
6. Bordoloi, Monali & Biswas, Saroj. (2017). E-commerce sentiment analysis using graph based approach. 570-575. 10.1109/ICICI.2017.8365197.
7. Chang, C.-C. And Lin, C.-J. (2011). LIBSVM : a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27.
8. Collins, M. (1997). Three generative, lexicalized models for statistical parsing, In Proceedings of the 35th Annual Meeting of the ACL.
9. Dave,K., Lawrence, S., and Pennock, D.M.(2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, In Proceedings of WWW, pages 519–528.
10. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.
11. Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, In Proceedings of the Conference on Web Search and Web Data Mining (WSDM).
12. Guo, H., Zhu, H., Guo, Z., Zhang, X. and Su, Z. (2009). Product feature categorization with multilevel latent semantic association, Proc. of CIKM.
13. Hung, Lai & Alfred, Rayner & Hijazi, Mohd & Ag Ibrahim, Ag Asri. (2015). A Review on the Ensemble Framework for Sentiment Analysis. Advanced Science Letters. 21. 2957-2962. 10.1166/asl.2015.6494.
14. Jagdale, Rajkumar & Shirsath, Vishal & Deshmukh, Sachin. (2019). Sentiment Analysis on Product Reviews Using Machine Learning Techniques: Proceeding of CISC 2017. 10.1007/978-981-13-0617-4_61.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 11: 2022
DOI10.17605/OSF.IO/E4XHW

15. Janhavi N. L., Santhosh, K. L. and Jharna M. (2018). Sentiment Analysis of Customer Reviews on Laptop Products for Flipkart, Proceeding of the International Research Journal of Engineering and Technology (IRJET).
16. Lin, D. (1998). Dependency-based evaluation of MINIPAR, In Workshop on Evaluation of Parsing Systems at ICLRE.
17. Martineau, J. and Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis, In Proceedings of the Third AAAI International Conference on Weblogs and Social Media, San Jose, CA.
18. Mejova, C., Srinivasan, P. (2011). Exploring Feature Definition and Selection for Sentiment Classifiers, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
19. Nakagawa, T., Inui, K. and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables, In NAACL, HLT.
20. Ng, V., Dasgupta, S. M., and Arifin, N. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews, Proceedings of the COLING/ACL on Main Conference poster sessions, p.611-618, Sydney, Australia.
21. Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques, In Proc. of EMNLP.
22. Qu, L., Ifrim, G., Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns, Proceeding in COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics.
23. Rogati, M., and Yang, Y. (2002). High-performing feature selection for text classification, Proceedings of the eleventh international conference on Information and knowledge management, November 04-09, 2002, McLean, Virginia, USA.
24. Saifullah, Shoffan & Fauziah, Yuli & Sasmito, Agus. (2021). Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data.
25. Sang, E. F. T. K. (2002). Memory-based shallow parsing. Journal of Machine Learning Research, 2:559-594.
26. Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields, In Proceedings of HLT-NAACL.
27. Sivaganesan, Aggarwal, D., Sridhar, K. Arunkumar, M. (2022). Feature based Sentiment Analysis for Product Reviews. Proceedings of the INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 06 (June 2022).
28. Socher, R., Pennington, J., Huang, E., Ng, A. and Manning, C.D. (2011). Semi Supervised Recursive Auto encoders for Predicting Sentiment Distributions, Conference on Empirical Methods in Natural Language Processing (EMNLP 2011).
29. Turney, P., and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37 (2010) 141-188.
30. Wang, H., Liu, L., Song, W., & Lu, J. (2014). Feature-based Sentiment Analysis Approach for Product Reviews. J. Softw., 9(2), 274-279.
31. Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization, In International Conference on Machine Learning, pages 412–420.

32. Zhai, Z., Liu, B., Xu, H. and Jia, P. (2010). Grouping Product Features Using Semi Supervised Learning with Soft-Constraints, to appear in Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), Beijing, China.

33. Zhai, Z., Liu, B., Xu, H. and Jia, P. (2011). Clustering Product Features for Opinion Mining, Proceedings of Fourth ACM International Conference on Web Search and Data Mining (WSDM-2011), Feb. 9-12, 2011, Hong Kong, China.