Received: October 10, 2019. Revised: November 15, 2019. 

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# A Filter Based Improved Decision Tree Sentiment Classification Model for Real-Time Amazon Product Review Data

**Maganti Syamala[1]*** **Nattanmai Jeganathan Nalini[1]**

[1]*Department of Computer Science and Engineering,*
*Annamalai University, Annamalainagar, Tamil Nadu 608002, India*
* Corresponding author's Email: shyamalamaganti54@gmaill.com

**Abstract:** E-Commerce product features and reviews are considered to be the essential factors in real-time e-commerce sites for product recommendation systems. Due to inaccuracy decision patterns, in most cases e-commerce user fails to predict the products based on the user ratings and review comments. Traditional sentiment classification models are independent of data filtering, transformation and sentiment score computing techniques which require high computing memory, time and mostly leading to false-positive rate. To overcome these issues, a novel sentiment score-based product recommendation model is proposed on the real-time product data. In this model, a new product ranking score, filtering, and hybrid decision tree classifiers are implemented. Initially, real-time amazon product review data is captured using Document Object Model (DOM) parser. The features from the review comments are extracted using lexicon Feature Dictionary (FD) and AFINN, Normalized Product Review Score (NPRS) are generated to compute the class label for product review sentiment prediction. Ranked Principal Component Analysis (RPCA) is used as a feature selection measure to overcome the problem of data sparsity. Random Tree, Hoeffding Tree, Adaboost + Random Tree, the three variants of decision tree classifiers are used for product sentiment classification. The proposed filter-based improved decision tree sentiment classification model for real-time amazon product review data recommends the product based on the user query by prediction using a new novel normalized product review sentiment score and ranked feature selection measure. The proposed product recommendation, the decision-making system maximizes sentiment classification accuracy. Experimental results are compared against the traditional decision-making classification models in terms of correctly classified instances, error rate, and PRC, F-measure, kappa statistics. The proposed model experimental results show high efficiency.

**Keywords:** Classification, Decision making, E-commerce, Features, Filtering, Sentiment score prediction.

## 1. Introduction

In recent years, the increase in the size of the online databases increased the product review comments in e-commerce applications and made the prediction process difficult. Human behaviour towards online content has become a major issue for decision making. For last two decades, the number of internet users has been increasing and according to the recent survey, every year the use of social media and messaging applications is increasing 200 percent. In the year 2019, a total of 2 billion people is using social media daily throughout the world. In the subsequent time, these statistics are expected to increase more rapidly in the coming years. Hence, there is a need for extracting knowledge from high dimensional online content using automated techniques.

According to current statistics people spend 8 hours a day on digital media. These technological advancements had changed the lifestyle of every individual person. Prior to making any purchase decision, people usually go through the web-based information present in various shopping portals, blogs, online sites and so on. Basically, most of the users check the quality of products or services before purchasing and they usually depend upon the opinions of other customers who had already purchased that product. E-Commerce sites allow the

users to share their feedback as comments about the product or shopping experience, which makes the job of the consumer easy to configure an online store for purchasing or selling. Due to the presence of huge product review data, makes it difficult to process manually. Hence, it is very much necessary to implement efficient and automated approaches to mine the required and relevant information. As different customers have different perspectives and different opinions about a particular product or service, so it is necessary to implement an efficient and effective fine-grained sentiment analysis techniques and classification algorithms.

The process of sentiment analysis plays a vital role in opinion mining. During the process of classification, the customers' opinion on the product is reviewed and positive, negative categories are mostly determined as factors for analysis. Apart of these categories there exists different levels of opinion classification models based on the type of input and they are:

1. Document-level: In the case of document level, the length of a review taken for opinion classification is a single paragraph or multiple paragraph. For example, review of a particular movie.
2. Sentence level: In the case of sentence level, the review length is restricted to a single sentence.
3. Aspect level: On the other hand, in the aspect level, the review text can be either a single word or a few words that are generally treated as aspects or features.

Nowadays, consumers are much aware of the original quality of products. Hence, the new buyers usually go through the reviews in order to take any purchase decisions. Generally, the web has become a popular medium for not only online users for making purchases but also for everyone who seeks to find relevant information on products and services before they committed to buy. To address many decision-making issues, a novel dynamic model is required to discover the hidden knowledge from the e-commerce review data.

At present product recommendation system is been used as an application in the field of E-Commerce. The product recommendation algorithm uses user associated data, such as review comments, review ratings, seller ratings, etc. A large number of classification models are used to predict the product sentiment which helps to know the changes in user behaviour. These sentiments are used for decision making by the consumer on different products. A feature is known as an attribute of a particular product.


| | |
|---|---|
| O.S | Android |
| R.A.M | 3 GB |
| Thing heaviness | 154 g |
| Creation proportions | 7.7 x 1 x 15.3cm |
| Piece replica figure | XT1643 |
| Wireless communiqué technology | Blue-tooth |
| Connectivity knowledge | G.S.M, E.D.G.E, 4G L.T.E |
| Others camera description | 5 MP |
| Screen feature | Touch screen receiver |
| Heaviness | 155Grms |
| Colour | Black |

Figure. 1 E-commerce product details

Each individual product has many features. It has a remarkable effect on customer's decision making along with the organization product development process. Therefore, the detection of essential product features plays a vital role in the process of enhancing the usability of reviews and as well as for the organizations to emphasize the advancement of product quality. Detecting these product features manually is quite impossible and impractical to date. Hence, automatic-detection techniques are to be designed to get the required features. Some of the useful information or attributes are labelled and shown in Fig. 1 E-Commerce product details.

The whole process of product sentiment analysis is divided into 3sub-phases: i) Data collection, pre-processing and feature extraction ii) Sentiment score prediction and feature selection iii) Product sentiment classification. Initially, product features are extracted along with product reviews. Later, the product sentiment classifier algorithm is implemented in order to rank the products depending on the feature frequencies and customer reviews. In general, E-Commerce sites use product sentiment classification to predict top k-products as user recommended products based on product reviews and ratings.

In the proposed work, a novel framework is implemented to predict the e-commerce product sentiment on the real-time data. In this framework, a new sentiment score is computed on the e-commerce product comments as a product feature for classification problem. Here, a new feature selection method and classification approach is implemented on the filtered data for product feature sentiment classification.

This paper is organized as follows. Section 2, describes the related works of traditional e-commerce sentiment classification models. Section 3, describes the proposed framework for e-commerce sentiment classification, Section 4, describes the experimental results and its analysis. Finally, in the section 5, conclusion and future scope is presented.

## 2. Related work

Traditional random tree classification technique is a kind of decision tree which constructs decision patterns on a set of e-commerce data using gain measure. Fig. 2. Shows the example of random forest classification, where the categorical attribute 'product reviews' defines two distinct values as Samsung and Oppo mobiles respectively.

Attribute selection, a computing measure is used for feature ranking. These sentiment classification methods select the top 'k' features based on the highest rank and eliminate those having lower feature ranks. Information gain is one of the attributes selection measure which is calculated based on the entropy value. Table 1. illustrates some of the attribute selection measures used in decision tree construction. The main limitation of these approaches is, they choose features having large distinct values and neglects the features having less distinct values.

Table 1 represents different existing attribute selection measure formulas used in decision tree construction. As these measures had a limitation to choose features having large distinct values and neglect the features having less distinct values, in the proposed model we discussed about the new attribute selection measures. Some important research concerns in the field of e-commerce on product recommendation are discussed below in brief.

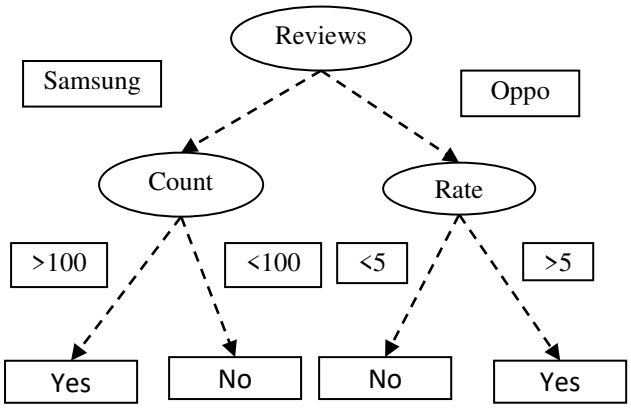


Figure. 2 Classification using random forest classifier

Table 1. Measures used in decision tree construction

| Measure Name | Formula |
|---|---|
| Entropy | $Entropy = \sum p_i \, log_{10}(p_i)$ |
| Information Gain | $Gain = \sum p_i \, log_{10}(p_i) - \sum \frac{\lvert D_V \rvert}{\lvert D \rvert} \sum p_i \, log_{10}(p_i)$<br>Where D is dataset<br>$D_v$ is subset of dataset<br>$P_i$ is the probability of the $i^{th}$ class |
| Gain Ratio | $Gain\ Ratio = Gain / \left( - \sum \frac{\lvert D_V \rvert}{\lvert D \rvert} log \frac{\lvert D_V \rvert}{\lvert D \rvert} \right)$<br>Where D is dataset<br>$D_v$ is subset of dataset<br>$P_i$ is the probability of the $i^{th}$ class |
| Pearson Correlation | $CorrelationFS = \frac{m.r_{ic}}{\sqrt{m+m(m-1)r_{ii}}}$ |
| Improved Correlation measure | $arg\ max \frac{\lvert cov(Vec(A_i,c),Vec(D)) \rvert}{\sqrt{var(Vec(A_i,c) * Vec(D))}}$ |

L. Dong et.al, presented a new unsupervised topic-sentiment joint probabilistic scheme in order to identify deceptive reviews [1]. Mostly, malicious sellers take the responsibility to provide fake reviews by hiring different buyers in order to improve their ratings and reputation. Hence, it is very much necessary to identify those deceptive reviews. Again, it is also important to mine these topics and sentiments from the reviews. So, in this model it identifies the fake reviews based on the writing style of the reviewer and for this identification, it uses an unsupervised topic-sentiment joint probabilistic model called Gibbs Sampling for topic extraction and Random Forest classifier, Support Vector Machine classifier is implemented for classification. This model makes the consumer be aware of the fake reviews and this work is limited to restaurant reviews and not applied to social media or e-commerce text classification.

M. S. Akhtar et.al, introduced an advanced feature selection framework and classifier ensemble construction method. This is basically a two-tier approach that can be implemented in Aspect Based Sentiment Analysis process [2]. In this piece of research work, they introduced an advanced cascaded framework for feature selection using particle swarm optimization algorithm and uses maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM) algorithms as classifier ensemble. They have considered different features which are identified according to the characteristics of various classifiers and domains.

R.K. Amplayo et.al, incorporated sentiment topic models on the product description to enhance the process of aspect-based sentiment analysis [3]. Usually, sentiment analysis models are implemented as unsupervised models to resolve the issues of general aspect-based sentiment analysis. The most serious issue of this approach is the substandard aspect of term extraction. This may give rise to various other problems of aspect label determination. They have emphasized on enhancing the aspect term extraction of topic models with the help of various product descriptions.

O. Araque et.al, proposed an advanced semantic similarity-based sentiment analysis on the e-commerce product data by using the lexicons [4]. It used word embedding's for finding the distributed semantics between similar words for feature extraction. The findings proved that lexicon-based computation by combining semantic distance and word embedding's for finding the similarity tends to be better when compared with the WorldNet technique. This work is mainly intended to compute the semantic distance between the words to know the similarity between words for sentiment analysis.

S. Bag et.al, emphasized his work on predicting the consumer's purchase intention on durable goods [5]. Developed an attribute level decision support prediction scheme in order to provide a better e-commerce platform to the consumers. They used the regression technique in order to predict the product sentiment. It can be considered helpful for the consumer as an effective search platform while buying any durable goods. It draws the intention of a consumer towards a product based on the attribute sensor obtained from the consumer search history.

L. Chen, et.al, focused on user perception of sentiment-integrated critiquing in recommender systems [6]. Most of the conventional approaches depend upon the static attribute values. Product reviews include various customers' sentiments which are also known as opinions. In this paper, they have presented an advanced sentiment-integrated critiquing technique in order to assist the users to decide and refine their preferences. This work defines the way how a consumer can make his decision by the sentiment towards the comment given by the customers who had already purchased the product.

G. Cosma et.al, introduced a new computational intelligence technique in order to predict the review ratings in the domain of e-commerce [7]. It is also considered as the most important research domain which has the objective to analyse people's opinions. Various e-commerce websites permit different users to share various opinions related to product or service. In the above process, the textual reviews and numerical ratings are used as training data. These opinions can affect customer purchasing decisions. In this piece of research work, an advanced computational intelligence framework is developed in order to predict customer review ratings. A small portion of data is required to construct a system in order to predict the review ratings.

R. Ireland et.al, focused on the application of data analytics for product design [8]. Advanced data analytics techniques are considered as the most important research domain in the twenty-first century. In this paper, they have emphasized sentiment analysis on online product reviews. In this piece of research work, an advanced framework is developed in order to analyse different online product reviews. The major objective of this approach is to use machine-produced data in order to determine the requirements of the consumers.

V. Jha et.al, introduced an advanced sentiment aware dictionary in order to carry out the process of multi-domain sentiment classification process [9]. Sentiment analysis is used to extract customers' opinions and then used classifiers to classify those opinions based on their polarity. However, this task is a domain-based and expensive one.

K. Kaushik et.al, concentrated on exploring reviews and review sequences on the e-commerce platform [10]. This work is intended for making the consumer know about the additional characteristics of a product other than statistically defined attributes.

X. Li, C. Wu, and F. Mai concentrated on the influence of online reviews on product sales. This technique is basically a joint sentiment-topic analysis process [11]. In this work, they have emphasized to analyse the growth in business with the influence of online reviews. Identified how the review ratings and comments are interrelated while predicting the growth of the business.

Y. Liu, J. Bi, and Z. Fan introduced a new product ranking scheme known as an intuitionistic fuzzy set theory for sentiment analysis [12]. In this model, a rank-based mechanism is used to determine the

positivity and negativity of the review which usually has a remarkable influence on consumer's purchase decision.

B. Palese and A. U sai emphasized his research on the relative importance of service quality dimensions in e-commerce experiences [13]. In this paper, they have identified the issue inadequately measuring service quality with the help of socialized data. This approach includes the basic concepts of the SERVQUAL model. Uses collaborative filtering for product ranking on static and dynamic e-commerce data in order to filter recommendations. Therefore, a wide range of e-commerce product features is not available are commended. This problem arises with large data sets where there are a greater number of customers who bought and reviewed the product.

S. Poria et.al, proposed a deep learning technique for sentiment analysis at aspect level to predict the polarity of reviews with respect to the aspects [14]. This model used a 7-layerdeep convolutional neural network to tag the words as aspects from the review comments. The aspects from neural networks are given as input to the ensemble classifier coupled with word embedding's for sentiment analysis.

B. Song, W. Yan, and T. Zhang developed a new cross-border e-commerce commodity risk assessment methodology using fuzzy-based reasoning [15]. This model aims to manage the risk in the form of safety regulations in e-commerce commodity purchases. This model uses text mining and fuzzy text rules for assessing the risk based on historical risk factors.

## 3. Proposed model

In the proposed model, a novel e-commerce product recommendation system is designed and implemented on real-time amazon data. Initially, data is collected from the real-time e-commerce site (www.amazon.com) using the user-specified product name and URL. Product comments are extracted using the Document Object Model (DOM) web parser and a set of training features are extracted from the product comments by a feature dictionary which is a lexicon-based approach and each product comment is tokenized using the Stanford Natural Language parser (NLP). Each token is matched against the feature dictionary. The positive and negative polarity of each feature is found with respect to the comment. AFINN scores and normalized product review scores are used to compute the class label of the product features that are extracted.

The product feature and its class label are used as training data for the product recommendation process. Ranked Principal Component Analysis (RPCA) algorithm is used to find the essential ranking features for the classification model. A hybrid decision tree classification model is proposed to predict the new test comment on the product for the recommendation process. The proposed framework is implemented in a three-stage architecture. In the first stage, data collection, pre-processing and feature extraction operations are performed. In the second stage, sentiment score computation and feature subset selection measures are implemented. In the third stage, a hybrid product classification model is used for sentiment polarity prediction of the test data. The basic workflow of the proposed framework is shown in the Fig. 3.

### 3.1. Data collection, pre-processing and feature extraction phase

In the data collection, pre-processing and feature extraction phase, each comment is extracted using the user defined product name and URL. These comments are collected and processed using the following pseudo code.

---

**Input:** Product name pname, URL

**Output:** Product features list FL, features dictionary FD, product comments PC [].

**Data collection**
**Step 1:** Visit URL using product name pname.
**Step 2:** if connection! =NULL
**Step 3:** then
**Step 4:** Capture all features list of the pname using the DOM xpath parser.
**Step 5:** DOM xpath parser is implemented using the web selenium driver.
**Step 6:** FD [] =Capture (URL (features_xpath))
**Step7:** Product-comments PC [] =Capture (URL (xpath_product_comments));
**Step 8:** xpath_product_comments=
/product
reviews/ref=cm_cr_arp_d_paging_btm_next_2?
Ie=UTF8&reviewerType=all_reviews&pageNumbe
r=n
**Step 9:** end if

**Pre-processing**
**Step 10:** for each product comment pc []
**Step 11:** do
**Step 12:** Tokens TC [] = tokenizer (pc[i])

---

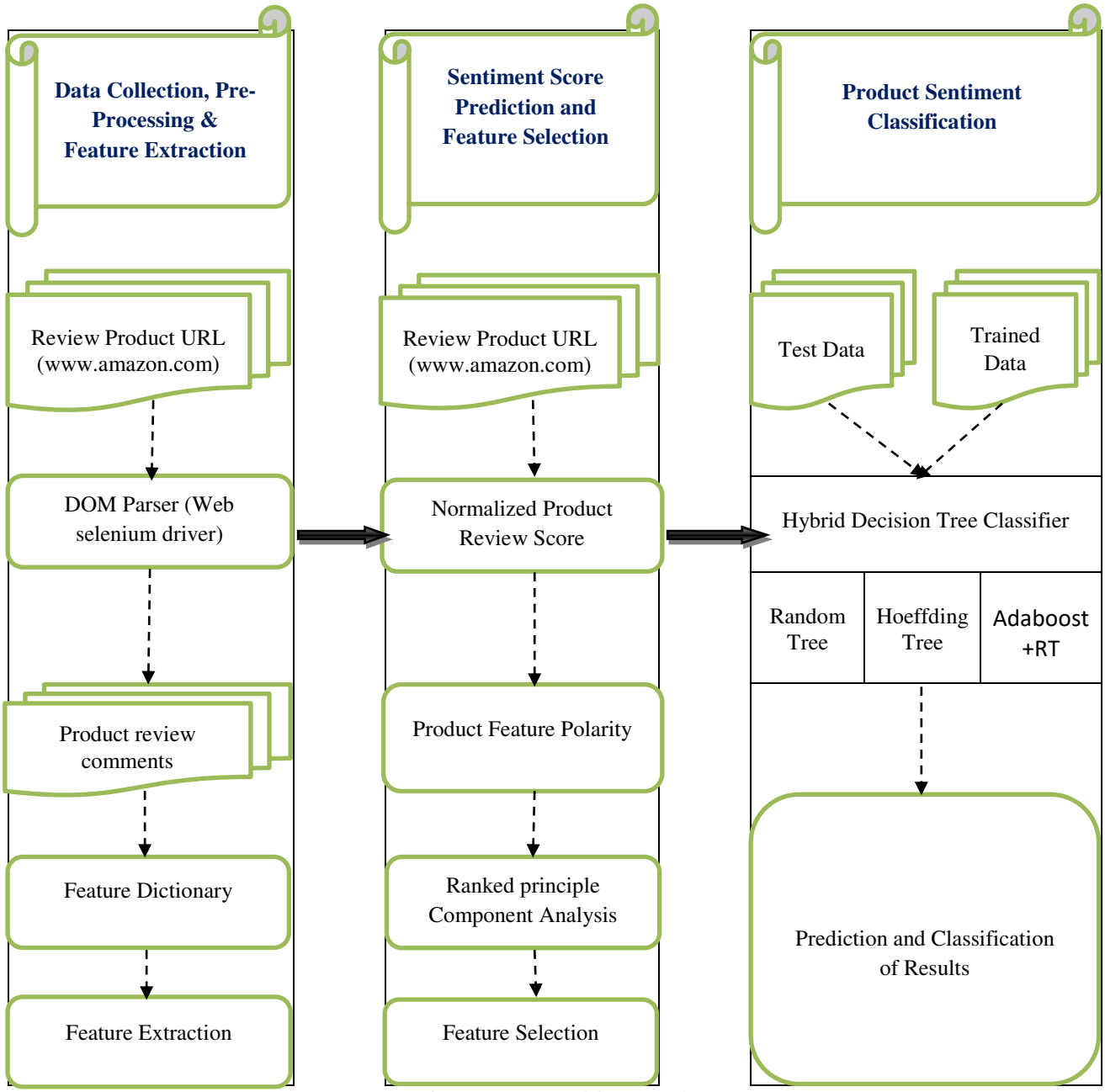


Figure. 3 Proposed framework

### 3.2. Sentiment score prediction and Feature selection

#### 3.2.1. Sentiment score prediction

In this phase, the sentimental score is computed on e-commerce product review comments using the AFINN lexicon dictionary database. AFINN lexicon database is a list of predefined training data that contains both positive and negative lexicon words with its scores. Review comments feature polarity score is computed using the trained positive and negative scores data from http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6010/zip/imm6010.zip. t

In the normalized product review score, product rating and its AFINN word score are used to compute the new product review score for the opinion classification process. The Normalized Product Review Score is computed and shown in Eq. (1). Here $AFIN\_Score$ is the product comments score, MIN is the min and MAX is the maximum values of the normalization.

$$NPRS = \frac{AFIN\_Score - MIN}{\sqrt{(Product - rating).(MAX - MIN)}} \quad (1)$$

**Step 1:** for each product dictionary FD []
**Step 2:** do
**Step 3:** if (FD[j] ==TC [])
**Step 4:** FL [] =Polarity FD[j] =" Positive"
**Step 5:** else
**Step 6:** FD [] = Polarity FD[j] =" Negative";
**Step 7:** done
**Step 8:** done

#### 3.2.2. Ranked feature subset selection

**Algorithm RPCA ():**
**Input:** Product training data.
**Output:** Ranked features.
**Step 1:** Input Data D.
**Step 2:** Normalize the Input data D using Eq. (2)

$$ND = D[i] - \mu_D$$
$$\mu_D = \sum D[i]/N \quad (2)$$

**Where** $N\ is\ the\ total\ number\ of\ records,\ \ \mu_D$ is the mean and ND is the normalized data of input data.
**Step 3:** Compute the covariance matrix between the features F using Eq. (3)
Let F= {f[0],f[1]….f[m]} be the feature space with m features.
Find the candidate features pairs CF={(f[0],f[1]),(f[0],f[2]),(f[0],f[3])……(f[m],f[0]) ….}. For m feature space we will get $\frac{m!}{(m-2)!2!}$Candidate sets.
Do
Compute covariance between features as

$$Cov(CF\{x,y\}) = \frac{\sum_{i=1}^{n}(CF[X_i]-\mu_{CF[X_i]})\left(CF[y_i]-\mu_{CF[y_i]}\right)}{(n-1)} \quad (3)$$

Where $\mu_{CF[X_i]}$ is the mean of the product features covariance matrix.
For each pair of candidate features CF
Done
**Step 4:** Compute the Eigen vector and values using the Eqs. (4) and (5)

$$Eigenvalue's\ [] = Det(\lambda I - COV(CF)) = 0 \quad (4)$$

Here $\lambda\ is\ eigen\ value,\ I$ is the identity matrix of same dimension as COV (CF). The corresponding Eigen vector is given as

$$\lambda I - COV(CF))\ v=0 \quad (5)$$

Here the optimal Eigen sum is computed as shown in Eq. (6)

$$OptimalEigenSum = (\sum Eigenvalues[i])/(Ksmallestval[\dashv], 0.5(Maxindex(Eigenalues[\dashv])) + Minindex(Eigenvalues[\dashv \quad (6)$$

**Step 5:** Selecting the highest ranked principal components using the Eigen values. Sort the Eigen values according to descending order. The highest Eigen value is the principal component of the dataset and it is more significant.

### 3.3. Product sentiment classification model

In decision tree construction, attribute selection measures play a vital role in order to prune the tree and to predict the class label with a high true positive rate. Hybrid attribute selection measures are integrated with some search strategies such as genetic, information gain, greedy, first search and brute force to find the most ranked attribute. When the number of attributes is in large number, this method requires huge computational time and memory. For instance, with 25 attributes the brute force technique needs to scan all 225 possible partitions of attributes. In the proposed work, a novel attribute selection measure is proposed to find the highest ranked attributes in the attribute set for data classification. The proposed hybrid measure minimizes the error rate and improves the true positivity and false-negative rate on large datasets. Since training data have nominal attributes and numerical attributes, the proposed measure is used to find the nominal association between the numerical and nominal attributes using the following measures of node selection. In this model, PCA ranked features are taken as input for product sentiment classification.

#### 3.3.1. Proposed ensemble classifier for review pattern detection

**Input:** Ranked PCA features RData;
Let the set of base classifiers are represented as Classifier [];
**Output:** Prediction of new product feature polarity with respect to the comment.
**Procedure:**
**Step 1:** Partition the given training dataset into 'm' classes (recommended True or False).

**Step 2:** For each partition
**Step 3:** Perform
**Step 4:** Apply the decision tree classification models on the partition.
**Step 5:** Construct the random tree, Hoeffding trees using the enhanced attribute selection measure to improve the true positive rate of the sentiment classification.

#### 3.3.2. Proposed hybrid attribute selection measures

***Hybrid Entropy Measure for Hoeffding Tree Construction***

Let Dp is the normalized product review data, the hybrid entropy for the Hoeffding tree construction is given by Eqs. (7) - (11)

$$n = \sum D_p[i].\log(D_P[i])$$
$$Ent(D_p) = \frac{(n + D_p[i].\log(\sum D_p[i])}{\sqrt[3]{(\sum D_p[i].\mu_{D_{p[i]}})}} \quad (7)$$

Where $Ent(D_p)$ is the modified entropy value. $HCondEntropy(D_p)$ *is the hybrid conditional entorpy value value* , $CramersV$ *is the cramers v rule,* $chiVal$ *is the chisquare* Value.

$$HCondEntropy(D_p) = \frac{-Math.cbrt(CramersV(D_P).total).n}{(entropyConditional(D_p) + chiVal(D_p))} \quad (8)$$

$$CramersV(D_P) = Math.sqrt(chiVal(D_p)/(n.\min(D_p)) \quad (9)$$

$$chiVal(D_p) = chisquare\ measure\ to D_p \quad (10)$$

$$EntropyConditionalent(D_p) = -n/(\log 2\ .\sum D_p[i]) \quad (11)$$

***Proposed Random forest attributes selection measure (RFASM)***

Let Dp is the normalized product review data and the proposed Random forest attributes selection measure is given by Eqs. (12) and (13).

$$Modified\ Gain = e^{-n/(\log 2.\sum D_p[i])} + Grain(D) \quad (12)$$

Hybrid Random Tree Attribute Selection Measure

$$HRTASM = \frac{-n.\sqrt[3]{entropyConditionalent(D_p)}}{\left(D_p[i].chiVal(D_p)\right)^3} \quad (13)$$

**Step 6:** Repeat until all the test data gets classified.

## 4. Experimental results

Experimental results are simulated using java programming and third-party libraries such as apache math, Jama, amazon ec2, etc. For the experimental results, amazon product review data from Kaggle [22] is tested and as well as tested against the real-time data obtained from the amazon product URL. Experimental results are compared with traditional models such as Adaboost with Random tree [16], KNN[18], Stacking Random tree[19], Bagging Random tree[20], Naïve Bayes Multinomial Text[21], RF[17] for performance analysis. The main problems identified in these models are

1. Difficult to handle nominal features with large text comments.
2. Difficult to predict the product sentiment using product features.
3. These models require high computational time and memory for data training and testing process.

Fig. 5, describes the training data features and its sequential id. These features are captured from the real-time amazon e-commerce.

| No. | Name |
|---|---|
| 1 | id |
| 2 | name |
| 3 | asins |
| 4 | brand |
| 5 | categories |
| 6 | keys |
| 7 | manufacturer |
| 8 | reviews.date |
| 9 | reviews.dateSeen |
| 10 | reviews.rating |
| 11 | reviews.sourceURLs |
| 12 | reviews.text |
| 13 | reviews.title |
| 14 | reviews.username |
| 15 | Sentiscore |
| 16 | reviews.doRecommend |

Figure. 4 Features of amazon training dataset

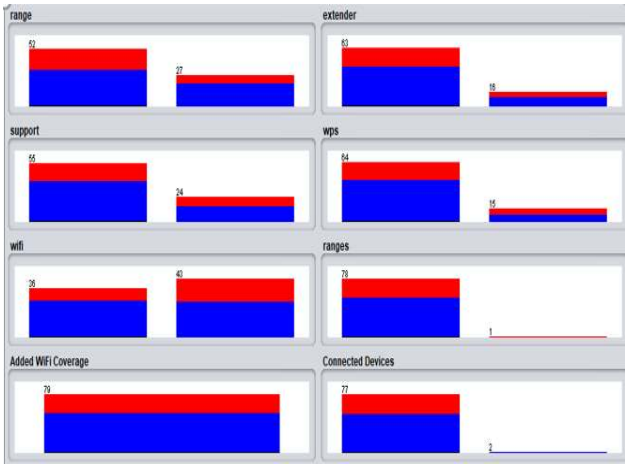


Figure. 5 Visualization of sample input features of the amazon training data for sentiment classification.

### 4.1. Proposed random tree patterns

```
Router = TRUE
| Wi-Fi Range = FALSE
| | SSID = TRUE
| | | extender = TRUE
| | | | range = FALSE: Positive (5/0)
| | | | range = TRUE
| | | | | speed = TRUE: Negative (1/0)
| | | | | speed = FALSE: Positive (1/0)
| | | extender = FALSE
| | | | signal = TRUE: Positive (1/0)
| | | | signal = FALSE: Negative (1/0)
| | SSID = FALSE
| | | strength = FALSE
| | | | ranges = FALSE
| | | | | Wi-Fi = FALSE
| | | | | | faster = FALSE
| | | | | | | speed = TRUE
| | | | | | | | support = FALSE
| | | | | | | | | signal = TRUE
| | | | | | | | | | Wi-Fi = TRUE: Positive (5/0)
| | | | | | | | | | Wi-Fi = FALSE
| | | | | | | | | | | mesh = TRUE: Negative (1/0)
| | | | | | | | | | | mesh = FALSE: Positive (1/0)
| | | | | | | | | signal = FALSE
| | | | | | | | | | extender = TRUE: Negative (2/0)
| | | | | | | | | | extender = FALSE: Positive (2/0)
| | | | | | | | | | | | button = FALSE: Positive
(2/0)
| | | | | | | | | | | | button = TRUE: Negative
(1/0)
| | | | | | | | | | signal = FALSE: Positive (1/0)
| | | | | | faster = TRUE: Negative (2/0)
| | | | | Wi-Fi = TRUE
| | | | | | mesh = TRUE: Positive (1/0)
| | | | | | mesh = FALSE
| | | | | | | button = FALSE
| | | | | | | | range = FALSE: Negative (4/0)
| | | | | | | | range = TRUE: Positive (1/0)
| | | | | | | button = TRUE
| | | | | | | | Wi-Fi = TRUE
| | | | | | | | | Bandwidth = FALSE
| | | | | | | | | | speed = TRUE: Negative (1/0)
| | | | | | | | | | speed = FALSE: Positive (1/0)
```

Fig. 6, illustrates the visualization of random tree construction on the training data. Each leaf node represents the opinion of the product labelled with class labels either true or false.

### 4.2. Performance comparison of proposed techniques with existing techniques in terms of correctly predicted and incorrectly predicted instances

Table 2, illustrates the comparison of the performance of proposed random tree technique to the traditional approaches on the e-commerce product review data. From the table, it is analysed that the proposed technique has high correctly predicted instances and low incorrectly predicted

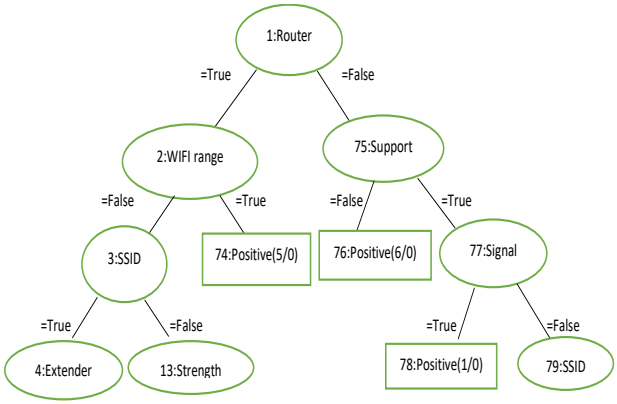


Figure. 6 Proposed random tree model on the product review data

Table 2. Performance comparison of proposed random tree to the existing approaches on e-commerce dataset

| Models | CP (in%) | IP (in %) |
|---|---|---|
| Adaboost + Random tree | 93.31 | 6.68 |
| KNN | 95.78 | 4.21 |
| Stacking + Random tree | 92.5 | 7.4 |
| Bagging + Radom tree | 92.7 | 7.2 |
| Naïve Bayes Multinomial Text | 92.52 | 7.41 |
| RT (Random Tree) | 92.92 | 7.07 |
| Proposed RT | 96.91 | 3.08 |

Table 3. Performance comparison of proposed Hoeffding tree to the existing approaches on e-commerce dataset

| Models | CP | IP |
|---|---|---|
| Adaboost + Random tree | 93.31 | 6.68 |
| KNN | 95.78 | 4.21 |
| Stacking + Random tree | 92.5 | 7.4 |
| Bagging + Radom tree | 92.7 | 7.2 |
| Naïve Bayes Multinomial Text | 92.52 | 7.41 |
| RF | 92.92 | 7.07 |
| Proposed HoeffdingTree | 97.8 | 2.13 |

Table 4. Performance comparison of proposed Adaboost + Random tree to the existing approaches on e-commerce dataset

| Models | CP | IP |
|---|---|---|
| Adaboost + Random tree | 93.31 | 6.68 |
| KNN | 95.78 | 4.21 |
| Stacking + Random tree | 92.5 | 7.4 |
| Bagging + Radom tree | 92.7 | 7.2 |
| Naïve Bayes Multinomial Text | 92.52 | 7.41 |
| RF | 92.92 | 7.07 |
| Proposed Adaboost + Random tree | 96.7 | 3.2 |

instances than the traditional method for e-commerce product sentiment classification techniques.

Table 3, illustrates the comparison of the performance of proposed Hoeffding technique to the traditional approaches on the e-commerce product review data. From the table, it is analysed that the proposed technique has high correctly predicted instances and low incorrectly predicted instances than the traditional method for e-commerce product sentiment classification techniques.

Table 4, illustrates the comparison of the performance of proposed Adaboost + Random tree technique to the traditional approaches on the e-commerce product review data. From the table, it is analysed that the proposed technique has high correctly predicted instances and low incorrectly predicted instances than the traditional method for e-commerce product sentiment classification techniques.

Table 5, illustrates the comparison of the hybrid classification models to the traditional approaches on the e-commerce product review data using F-measure. From the figure, it is analysed that the proposed models have high F-measure than the traditional method for e-commerce product sentiment classification.

Table 5. Comparison of proposed approaches to the existing approaches using F-measure

| Models | F-measure |
|---|---|
| Adaboost + Random tree | 0.909 |
| KNN | 0.956 |
| Stacking + Random tree | 0 |
| Bagging + Radom tree | 0.896 |
| Naïve Bayes Multinomial Text | 0 |
| RF | 0.909 |
| Proposed RT | 0.967 |
| Proposed Hoeffding Tree | 0.979 |
| Proposed Adaboost + Random tree | 0.964 |

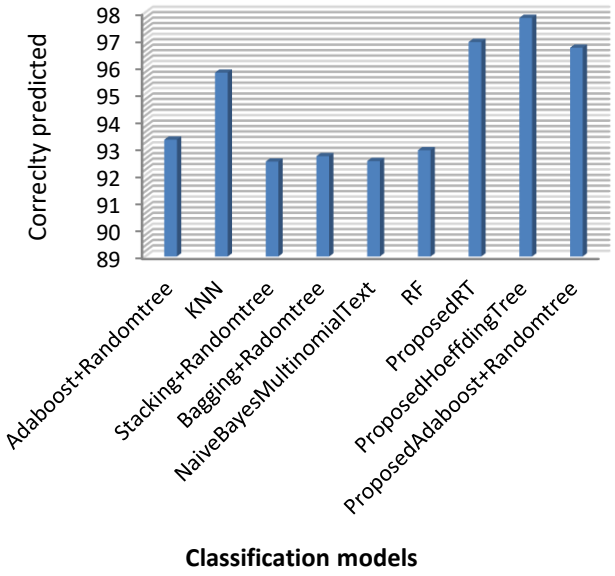


Figure. 7 Performance comparison of proposed classification models to the existing approaches on e-commerce dataset using true positivity and false negativity

### 4.3. Performance Analysis comparison of proposed techniques with existing techniques.

Fig. 7, illustrates the comparison of the proposed classification models to the traditional approaches on the e-commerce product review data. From the figure, it is analysed that the proposed models have high correctly classified instances than the traditional methods for e-commerce product sentiment classification.

Fig. 8, illustrates the comparison of the proposed classification models to the traditional approaches on the e-commerce product review data by using MCC metric. MCC is used to find the fitness of machine learning models for binary class prediction. From the

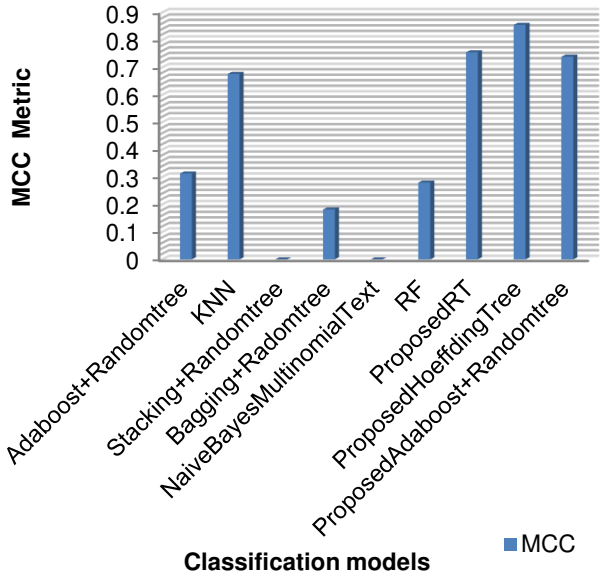


Figure. 8 Performance comparison of proposed classification models to the existing approaches on e-commerce dataset using MCC measure

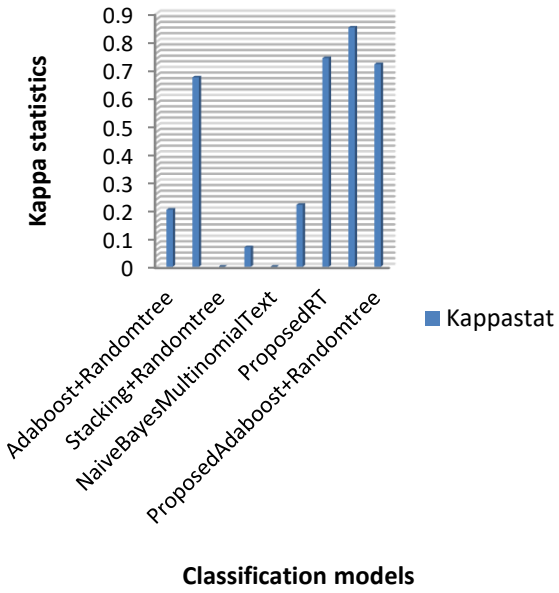


Figure. 9 Performance comparison of proposed approaches to the traditional approaches on e-commerce sentiment data using Kappa statistic measure

figure, it is analysed that the proposed models have high MCC classified instances than the traditional methods for e-commerce product sentiment classification.

Fig. 9, illustrates the comparison of the hybrid classification models to the traditional approaches on the e-commerce product review data. From the figure, it is analysed that the proposed models have high kappa statistics than the traditional method for e-commerce product sentiment classification.

## 5. Conclusion

In this paper, a novel sentiment classification model is proposed to improve the classification accuracy for amazon product reviews sentiment prediction. Traditional sentiment classification models are complicated to process all the user-generated reviews manually for decision making. The proposed model incorporates three variants of decision tree algorithms i.e. Random tree, Hoeffding tree, Adaboost + Random tree and experimental results proved to be better in terms of performance than the existing classification techniques. This model is tested with real-time amazon product review data set and as well as with the approved amazon product review data from Kaggle. In both cases, the experimental results proved to be better with high true positivity and false negativity rate. This novel model, when compared to the traditional existing models, improves the user accessibility. It will extract all the features as aspects describing the product from the review comments, which may dynamically vary from product to product. For feature extraction, it uses a trained lexicon containing various features and uses a Ranked Principal Component Analysis technique as a feature subset selection measure based on eigenvalues to remove redundant and irrelevant features. Polarity with respect to the feature is determined using AFFIN Score and later uses NPRS a normalized equation to get a normalized score for the feature extracted to determine its polarity. The proposed model for sentiment classification prediction makes the job of an individual easy to take a decision before buying anything online. From the experimental results, it is clearly observed that the proposed model has high computational efficiency for e-commerce product prediction. Also, proposed model has nearly 12% improvement over the traditional classification models on the real-time training data than the traditional models.

In the future work, this model can be extended as an improvement for detecting Sarcasm in the review text and also it is possible to design a dynamic model for filtering fake reviews.

## References

[1] L. Dong, S. Ji, C. Zhang, Q. Zhang, D. Chiu, L. Qiu, and Da. Li, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews", *Expert Systems with Applications*, Vol.114, pp. 210-223, 2018.

[2] Md. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, "Feature selection and ensemble construction: A two-step method for aspect-

based sentiment analysis", *Knowledge-Based Systems*, Vol. 125, pp.116-135, 2017.

[3] R. Amplayo, S. Lee, and M. Song, "Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis", *Information Sciences*, Vol. 454–455, pp.200-215, 2018.

[4] O. Araque, G. Zhu, and C. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis", *Knowledge-Based Systems*, Vol.165, pp.346-359, 2019.

[5] S. Bag, M. Tiwari, and F. Chan, "Predicting the consumer's purchase intention of durable goods: An attribute-level analysis", *Journal of Business Research*, Vol.94, pp.408-419, 2019.

[6] L. Chen, D. Yan, and F. Wang, "User perception of sentiment-integrated critiquing in recommender systems", *International Journal of Human-Computer Studies*, Vol.121, pp.4-20, 2019.

[7] G. Cosma and G. Acampora, "A computational intelligence approach to efficiently predicting review ratings in e-commerce", *Applied Soft Computing*, Vol. 44, pp.153-162, 2016.

[8] R. Ireland and A. Liu, "Application of data analytics for product design: Sentiment analysis of online product reviews", *CIRP Journal of Manufacturing Science and Technology*, vol.23, pp.128-144, 2018.

[9] V. Jha, R.Savitha, P.Shenoy, K. Venugopal, and A. Sangaiah, "A novel sentiment aware dictionary for multi-domain sentiment classification", *Computers and Electrical Engineering*, Vol.69, pp.585-597, 2018.

[10] K. Kaushik, R. Mishra, N. Rana, and Y. Dwivedi, "Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon.in", *Journal of Retailing and Consumer Services,* Vol.45, pp.21–32, 2018.

[11] X. Li, C. Wu, and F. Mai, "The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis", *Information & Management*, Vol.56, No.2, pp.172-184, 2019.

[12] Y. Liu, J. Bi, and Z. Fan, "Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory", *Information Fusion*, Vol.36, pp.149-161, 2017.

[13] B. Palese and A. Usai, "The relative importance of service quality dimensions in E-commerce Experiences", *International Journal of Information Management*, Vol.40, pp.132–140, 2018.

[14] S. Poria, E. Cambria, and A. Gelbukh, "Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network", *Knowledge-Based Systems*, Vol.108, pp.42-49, 2016.

[15] B. Song, W. Yan, and T. Zhang, "Cross-border e-commerce commodity risk assessment using text mining and fuzzy rule-based reasoning", *Advanced Engineering Informatics*, Vol.40, pp.69–80, 2019.

[16] H. Xing and W. Liu, "Robust AdaBoost based ensemble of one-class support vector machines", *Information Fusion*, Vol. 55, pp. 45–58, 2020.

[17] H. Liu, X. Zhang, and X. Zhang, "PwAdaBoost: Possible world based AdaBoost algorithm for classifying uncertain data", *Knowledge-Based Systems*, p. 104930, 2019.

[18] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification", *Engineering Applications of Artificial Intelligence*, Vol. 49, pp. 176–193, 2016.

[19] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin, and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection", *Applied Soft Computing*, Vol. 77, pp. 188–204, 2019.

[20] H. Gu, Y. Cui, L. Xu, M. Tu, Y. Fu, H. Fu, and Y. Zhou, "Bagging classification tree-based robust variable selection for radial basis function network modelling in metabolomics data analysis", *Chemometrics and Intelligent Laboratory Syste*ms, Vol. 174, pp. 76–84, 2018.

[21] H.-Y. Lin, "Efficient classifiers for multi-class classification problems", *Decision Support Systems*, Vol. 53, No. 3, pp. 473–481, 2012.

[22] https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products.