Identification of Best Product from E-Commerce Reviews using Manta Ray Foraging Based Feature Selection Technique with Auto-encoder Classifier

Dr. M.A. JAMAL MOHAMED YASEEN ZUBEIR

ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE,JAMAL MOHAMED COLLEGE
(AFFLIATED TO BHARTHIDASAN UNIVERSITY), Tamil Nadu, 620020, INDIA

Abstract
Social Media has arisen as a new communication channel between consumers and enterprises to produce a huge volume of unstructured text data about products. Many web users post their opinions on several products through the blog, review sites and social networking sites based text of the attitude. Customer feedback plays a very important role in the daily movements of products. Opinions of others are also taken into account when making decisions to select the best products. Event though, it reads reviews of all the customers, it has difficulty in making decisions based on the information about whether or not to purchase the product. Keeping track of the customer's opinion, manufacturers are also finding it difficult to manage the products which lead to economic collapse. To address this issue, the research work develops the feature selection based auto-encoder based classifier to predict the opinions of the customer. Initially, the raw data is pre-processed using five different process and then, term feature extraction technique is used to extract the valuable information of review. After that the optimal features are selected by Manta Ray Foraging Optimization (MRFO) technique. Then, these features are given as an input to Auto-encoder deep neural network for final classification of product review. The experiments are carried out on e-commerce web-sites to identify the best products among others and verified the performance of projected model in terms of various metrics. The results proved that the projected model achieved 94.32% of accuracy, 96.02% of F-score, where the existing random forest technique achieved 93.27% of accuracy and 95.63% of F-score on binary classification.

Keywords:
Auto-encoder; Customer Feedback; Manta Ray Foraging Optimization Opinion Mining; Social Media; Text Mining.

1. Introduction

The development of web technology is increasing at a rapid pace today. For many people, the internet has become an indispensable source of information. There has been a growth in the use of e-commerce due to this improvement. When it comes to dealing with an ever-increasing amount of information on the internet, Real-time search play an important part in combating the problem of information overload [1]. Cooperative filtering (CF) is by far the most common recommendation system (RS) approach. People who have similar habits in the past are more likely to have similar preferences in the future, according to this strategy's main premise. Data sparseness, which is described as a lack of user evaluations with a large number of items yet has shown promising results for CF techniques, is one of the key barriers they face. As a result, recommendation systems perform less well. As e-commerce websites have grown in popularity, it has been shown that user textual reviews can be used to alleviate the problem of data sparsity, thus increasing the usefulness of RSs. Users' perspectives on a wide range of characteristics of a product are frequently included in reviews written by other users. Because these user views are so significant, they show the user's preference for products and hence affect the accuracy of RSs. Due to poor performance in rating prediction, most traditional RSs tend to ignore such user's viewpoints for rating prediction [2]–[4].

1.1. Consumer

It's important to remember that a client is someone who can afford to buy the goods and services that a business advertises, and who will do so to meet their own wants and those of their family or a certain group of people with related interests. "A consumer is the most important visitor to our premises," states Mahatma Gandhi in a well-known explanation of the term. He doesn't rely on us in any way. We rely on him for everything. He's not a stranger to our company. He is involved. When we serve him, we are not doing him any favours at all. To give us a chance to do so, he's doing us a favour" [5-6].

Manuscript received November 5, 2022
Manuscript revised November 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.11.71

1.2. Consumer Behavior

Consumer behaviour can be defined as "the physical, and social activities of possible clients as they have made their minds up to access, appraise, purchase, and notify others about any item and its qualities." Consumption is the study of how a single person's decisions about what to buy, how to use it to meet their requirements, and how to get rid of it affects other people and the entire society, whether they are individuals, groups, or organisations. For example, students at the same school or college may wear the same uniform/dressing, or groups of people working together horizontally and vertically at the same location may select whether or not to use a product collectively [7-9]. For an advertising agency/marketer, client feedback can have a significant impact on the product's market position and consumption. Services and concepts are equally important to consumer behaviour as real goods.

1.3. Internet Marketing

It is the use of the internet as a platform to evaluate the potential selling and displaying of products. E-commerce applications have been found to benefit greatly from using conventional fundamental promotion strategies [10]. Online promotion and advertising, as opposed to traditional corporate methods, have shown to be significantly more effective while posing far fewer risks. In addition to being handy for businesses, web showcasing promotes environmentally-friendly solutions around the world [11].

1.4. Data Mining and E-commerce

Building a system that takes advantage of mined information is a huge undertaking. Some uses of data mining techniques on e-commerce data were shown to be less difficult than those on other types of data, according to research. Data mining systems in Ecommerce, for example, can be developed considerably more easily than translating and correcting data for data mining purposes. Because the data was not gathered manually or through a poll but rather electronically, it is less noisy or even devoid of noise in some cases. According to [12] 1[3], the data set comprises a wide range of information.

Public information made available through e-commerce platforms has been shown to be critical to the development of regression models because of the large and diverse amount of data that can be gathered. This means that data and its inferences provided by E-commerce platforms can be used to create a trustworthy platform for clients of E-commerce services [14].

1.5. Research Objective

E-commerce is a major emphasis of this study, with the goal of creating an environment where consumers may build trust in a platform where the things they buy are authentic and the reviews they leave are authentic. To identify effective text preprocessing techniques and to determine the most important characteristics for each consumer-related aspect.. Enable machine learning to extract and classify data into predetermined categories.

1.6. Paper Structure

Currently used opinion mining approaches are discussed in Section 2. Section 3 provides a mathematical explanation of the proposed methodology. Section 4 depicts the comparison of the proposed technique to currently used techniques. Lastly, in Section 5, we come to a conclusion about the study's findings.

2. Related Works

E-commerce platform product review sentiment categorization using a large-scale and multi-domain continuous naive Bayes learning framework is presented by F. Xu [15]. Naive Bayes model parameter estimation is extended to continuous learning style while preserving the excellent computational competence of classic Naive Bayes model. Furthermore, provide strategies to fine-tune the learnt distribution based on three classes of assumptions so that it can better adapt to varied domains. Researchers have found that this approach does a better job of handling reviews from a variety of sources, including Amazon's product reviews and movie reviews, by drawing on prior domain knowledge to help drive learning in new domains.

Table 1. Samples of E-commerce comments

E-commerce Comments from Customers
I got product in given time period
They assured me to send this in a period of 2 months and they sent me

An expert in opinion mining named Uma Maheswari, S. [16] created a new method for merging social media (such as Twitter and Facebook) with blog reviews (Amazon reviews). New customers and businesses can then benefit from this information, which is mined from the reviews. Customers will be better able to make informed purchases now that they have access to this new information. As a result of analysing data gleaned from customers, businesses are able to better serve their customers by providing them with trending products. An f1-score of above 80% has been achieved by this model in sentiment prediction.

Table 2. Comments data after stemming

Comments After Stemming
Get Product give time period
Promise send throughout month

"Aspect-based opinion analysis" (reviews/comments) by Aasha, A.A. [17] is a straightforward yet crucial approach. Amazon is the primary source of data for this study, which focuses on cosmetics. Taking into account elements like product quality, packaging, and shipping results in vastly varied ratings for the same product. VADER sentiment analysis is used in the aspect-based technique. Aspect-based approach for opinion mining can be used in any online store or application, according to the findings of this study. In our technique, we categorise reviews into positive, negative and neutral polarities so that we may provide accurate and helpful feedback.

According to Truong, [18], a method to analyse Vietnamese reviews derived from e-commerce websites in Vietnam to generate product references based on the items' features/functions has been proposed by Truong. Product features cited in customer feedback and reviews are identified using a topic-based methodology suggested in this paper. VietSentiWordnet is integrated into the proposed system to calculate the position scores for each product's feature. Product suggestion databases are also built to store customers' preferences and purchases. More than 2,000 Vietnamese comments and reviews on laptop items were analysed for this project, which is intended to be useful in real-world situations.

A person's viewpoint might be favourable or negative, according to Singh, P. [19]. Movie recommendation systems and e-commerce websites can both make use of this data to gauge the quality of their offerings. It used deep learning algorithms to accurately classify people's thoughts into two groups: those who are positive and those who are negative. Data selection, data preprocessing, data tokenization and neural network creation are all part of this project's methods. For this reason, we've used the reviews dataset. Deep learning algorithms benefit from data preprocessing and data cleaning to make it easier to apply them to the data. Algorithms for deep learning are self-learning and do not require human intervention. It is the primary goal of deep learning model implementation to improve performance, accuracy, and efficiency. Using our dataset, we tested and trained three distinct neural network models to see how they fared against each other. The Recurrent Neural Model (RNN) had the least overfitting and the greatest testing and training accuracy, according to the analysis of the three models.

An e-commerce sentiment classification algorithm developed by Kim, T.Y. [20] is based on the collection and use of user reviews. SVM, SVM+, and SVM+MTL procedures were used in conjunction with a term information extraction method for this model in order to identify phrases that can boost the power and effect of information, as well as to categorise the selected terms rendering to parts of speech (POS). The evaluation of the suggested model revealed that it performed exceptionally well in terms of sentiment analysis. Improved services for customers and more e-commerce competition are predicted as a result of the suggested model's implementation.

The model proposed by Jacob, M.S. [21] can be used to detect false product reviews. Fake and authentic reviews are separated out using the model's Naive Bayes classifier and Support Vector Machine. A number of parameters are used by the model to extract features for classification, such as review length, use of personal pronouns, nature of the review, verifiable purchase status, rating of the review, and the category of product. A high classification accuracy rate is demonstrated by the model's performance in the experiments.

3. Proposed Methodology

3.1. Dataset Collection

XML X routes are used to apply X query scripts to E-commerce systems in order to collect the data set. Large data dumps necessitate the highest level of professional skill in order to retrieve the precise and contextual information needed. It is possible to mine consumer behaviour data from e-commerce platforms, which include a big amount of customer feedback.

As a result of these online platforms, new and seasoned customers and professionals alike can share their expertise with others in their field from all over the world and at any time of day or night, with no restrictions based on geographical location, linguistic barrier, or level of competence. Everyone on an E-commerce platform cannot tag their posts or comments with a certain subcategory unless they specifically request it. People who wish to learn about the thoughts of other consumers who have already used a product from the same platform will find it much easier if new comments are categorised in this way. The data set is broken down into four key categories, each of which has an associated attribute that relates to the consumer's activity. Experts in e-commerce review and manually annotate each comment in the specified data set. Table 1 provides some examples of people's remarks.

Attributes Associated with Individually Class

The phase appropriate, easy, great, best, useful, desire, effective, perfect and functionality will be in Convenience Class, where the attributes includes comfortable, excellent, recommend, described, honest, reliable, satisfaction, complains and quality will be in Time Class. The other two classes such as variety and trust, where variety class consists of forever, collection, variety, diversity, warranty, compatible, intended, specific and different attributes. Finally, the trust class have pair, come up with, successfully, time, period, received, duration and deal attributes.

3.2. Preprocessing

Prior to the extraction stage, the data that has been gathered is handled in pretreatment. There are a number of preprocessing phases in this study, including:

1) Tokenization

Tokenization is the procedure of separating a piece of text into tokens, which are phrases, words, symbols, or other meaningful pieces [22]. Tokenization is used to study the words in a given statement. The first block of text in any piece of text data is all it is. Words from the data collection are necessary for retrieval in information. As a result, we require a parser capable of handling the document tokenization process. However, punctuation and other characters such as parentheses, hyphens, etc. are still a difficulty. The most common use of tokenization is to identify terms with meaning. Standardized abbreviations and acronyms are also a source of confusion [23].

Table.3. Comparative analysis of binary class on Proposed with various existing algorithms.

Algorithm	Accuracy	Precision	Recall	F-Score
Linear Regression	87.70	99.41	85.21	91.82
Naive Bayes	92.50	99.82	90.98	95.27
KNN	92.90	99.78	91.52	95.41
DT	92.50	99.63	91.38	95.18
SVM	92.70	99.90	91.93	95.32
Random Forest	93.27	99.91	92.47	95.63
Proposed	**94.32**	**99.95**	**93.24**	**96.02**

Filtering stop words

Stop words include prepositions, articles, and pronouns, all of which are unlikely to aid with text mining. Text mining applications cannot use these phrases because they are common to all documents. This entire list of words has been omitted. For this exercise, any word group will do. It also minimises the amount of text stored on the system, which helps it run faster. "are," "I," and "you" are examples of this.

Transform cases

All lowercase letters are then used for the words that have been retrieved.

Stemming

From the words that are available in the dataset, the goal of this technique is to extract root words. To avoid calculating errors when extracting syntactic features, it is necessary to obtain a root word. A set of phrases that share a root word but differ in an affix, suffix or infix can have the same distance score from the syntactic feature if we don't stemming; however, if we use stemming we obtain a score distance of zero for this pair of sentences. If you earn a score of zero meaningful phrases, the pair is identical save for the fact that the couple has various meanings. Examples of stemming processes are shown in Table 2.

Weighting words

When it comes to finding information, numerical statistics like the TF-IDF or TFIDF indicate how essential a word is within a document or corpus. Information retrieval, text mining, and user modelling all employ this as a weighting factor. To recompense for the fact that certain words seem more often than others, TF-IDF values rise proportionally to the number of times a word is used in a document. 83% of the digital library's text-based recommendation algorithm utilises TF-IDF, one of the most widely used time-weighting techniques. The formulas for these stats can be found in the table that follows.

$$TFIDF_{i,j} = TF_{ij} \times idf_i \quad (1)$$

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (2)$$

$$IDF_i = log\frac{|D|}{|\{j:t_i \in d_j\}|} \quad (3)$$

Here n_{ij}signifies the arrival of the word i in the document d_j $\sum_k n_{kj}$ is the sum of total words which look in the document. $|D|$is the total sum of words in the corpus, and $|\{j: t_i \in d_j\}|$ is the sum of corpora which embraces the word t_i.

3.3. Term Information Extraction Technique.

Findings from opinion mining research have shown that extracting word information provides the most information about document classifying and constructing a classification model to manage challenges with sentiment classification at the document level. Sentiment classification relies heavily on the use of terms in reviews and comments that may be used to identify linguistic features and documents, and these terms can be used to determine the main sentiment in a review [24-25]. It is essential to identify the document's fundamental terms, as these terms have a substantial impact on the document's polarity.

3.3.1. Document Frequency

In this context, the phrase "document frequency" refers to the total number of documents that include a specific term. In other words, it is the percentage of documents that contain phrases that have been used at least once in a particular number of documents. This method is less complicated and involves less math than previous methods. Improves classification accuracy by deleting low-frequency phrases and removal words on the premise that low-frequency terms and words do not contribute to document accuracy [26, 27].

Table.4. Comparative analysis of multi class on Proposed with various existing algorithms.

Algorithm	Accuracy	Precision	Recall	F-score
Linear Regression	85.71	84.32	85.93	83.45
Naive Bayes	92.10	92.43	92.15	91.68
KNN	92.46	93.48	92.44	91.81
DT	89.52	90.21	89.54	89.03
SVM	94.53	96.61	92.52	92.24
Random Forest	94.16	96.17	92.32	92.10
Proposed	**95.62**	**98.32**	**94.62**	**94.53**

3.3.2. Chi-Square Statistic

Cross-tabulation analysis approaches like the chi-square statistic, which examines a correlation between categorical variables, are the most commonly utilised. If a term appears more than once in the complete document set, the correlation value is "1," and if it doesn't, it's "0." In other words, a term (t i) and a category are compared using this method to determine their relative importance (C_j).

$$x^2 = \frac{N \times (AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (4)$$

Where

$N = A + B + C + D$ (N the total sum of documents),
$i = 1,2,..n$ (n the total sum of terms used),
$j = 1,2,\ldots,m$ (m the total sum of categories).

The chi-square statistic is similar to mutual information, although the former performs better for term extraction than the latter because of its standardised value. The chi-square statistic is defined in Equation (4) (x^2) between $C_j$and $t_i$in a range of the consistent possessions applied [28-29].

3.4. Polarity Detection

Step 1:Start

Step 2: Negative keywords, Positive keywords, and Neutral keywords are all included in this list.

Step 3:Calculate numeral of Review.

Step 4:*The Polarity Reversal can be calculated (If any review is having keywords like e.g. Good, Nice, Awesome, Amazing, Excellent etc.) There is a polarity in opposition to the Polarity Reversal (if any review include keywords like e.g. Worst.*

Step 5:End

Negative keywords and Neutral keywords are also included in this algorithm's list of search terms. Calculate each sentence's positive, negative, and neutral polarity.

3.5. Feature Selection using MRFO

During this stage, the MRFO algorithm has been employed for feature selection process which assists in optimal selection of features involved in the extraction process and improves the classification accuracy. The MRFO is a bio-inspired new technique which simulates the intelligent foraging performance of manta rays (MRs) and features of its foraging performance. The model was appropriate to our current solar radiation forecast problem provided that MRs on that the MRFO is created, have 3 various foraging approaches which are utilized for searching for food that procedure the vital search methods of MRFO for optimizing the solution of our presented solar radiation forecast problem [30]. The mathematical process of chain foraging was signifying as:

$$M_m^* = \begin{cases} M_m + (M_B - M_m)(r + \sigma) & if m = 1 \\ M_m + r(M_{m-1} - M_m) + \sigma(M_B - M_m) & if m \neq 1 \end{cases} (5)$$

$$\sigma = 2r\sqrt{|\log(r)|)} \ (6)$$

In which (M_m) stands for the individual MR (m), r refers the arbitrary uniformly distributed number from the range of zero and one. M^* and MB defines the novel or optimum position of MR from the population, σ denotes the weighted co-efficient as function of all the iterations. It can be apparent in Eq. (6) that the preceding MR from the chain and spatial place of strongest plankton obviously determine the position upgrade method from the chain foraging. Cyclone foraging was separated as to 2 parts. The 1st half concentrations on improving the exploration and is upgraded as:

$$M_m^* = \begin{cases} M_R + (M_R - M_m)(r + \beta) & if m = 1 \\ M_m + r_1(M_{m-1} - M_m) + \beta(M_R - M_m) & if m \neq 1 \end{cases} (7)$$

whereas M_R signifies the individual generated arbitrarily:

$$M_R = M^{\min} + r_1(M^{\max} - M^{\min})(8)$$

The adaptive weighted co-efficient (β) was diverse as:

$$\beta = 2e^{r_2 \frac{Iter_m - Iter_m + 1}{Iter_m}} \sin(2\pi r_2)(9)$$

In which $Iter$ implies the present iteration and arbitrary uniformly distributed number, and r_2 is over of zero and one. The 2nd half concentrate on enhancing the exploitation, thus the upgrade is as per:

$$M_m^* = \begin{cases} M_B + (M_B - M_m)(r_1 + \beta) & if m = 1 \\ M_B + r_1(M_{m-1} - M_m) + \beta(M_B - M_m) & if m \neq 1 \end{cases} (10)$$

Somersault foraging: The ending foraging approach with MRs determining the food supply and exploiting backward somersaults for circling the plankton for attracting. Somersaulting is local, spontaneous, cyclical, and periodic act which MRs utilize for maximizing their food intake. The 3rd approach is where an upgrade of all individuals takes place around an optimum position:

$$M_m^* = M_m + S(r_3 M_B - r_4 M_m)(11)$$

In Eq. (11), S represents the somersault co-efficient $(S = 2)$ adjusting the domain of MRs, r_3 and r_4 are arbitrary numbers in the range of zero and one. According to an arbitrarily created number, the MRFO technique is switched amongst chain as well as cyclone foragings. Afterward, the summersault foraging gets act for updating individual's present positions utilizing an optimum solution obtainable at the time. These 3 various foraging procedures are utilized interchangeably for achieving the global optimal solution of optimized problem, so sufficient the already decided end condition.

The MRFO method made a FF for reaching higher classifier presentation. It describes a positive integer for demonstrating the best result of candidate solutions. Under this work, the minimized classification error rate is regarded as FF is given in Eq. (12). The best result is a less error rate and worst outcome reaches a higher error rate.

$$fitness(x_i) = ClassifierErrorRate(x_i) = \frac{numberofmisclassifiedsamples}{Totalnumberofsamples} * 100 \ (12)$$

3.6. Classification

Classifiers for binary and multi-class classification will use the MRFO-derived optimum features. Using an automatic encoder in this study may be a viable option for selecting the DNN's right categorization class without knowing the distribution data beforehand. Auto Encoder DNN is a pre-training technique that uses a greedy layer-by-layer approach. DNN does not have any looping functions, therefore data flow is collected directly from input to output. To put it simply, the Auto Encoder DNN Classifier has a very low chance of losing any data.

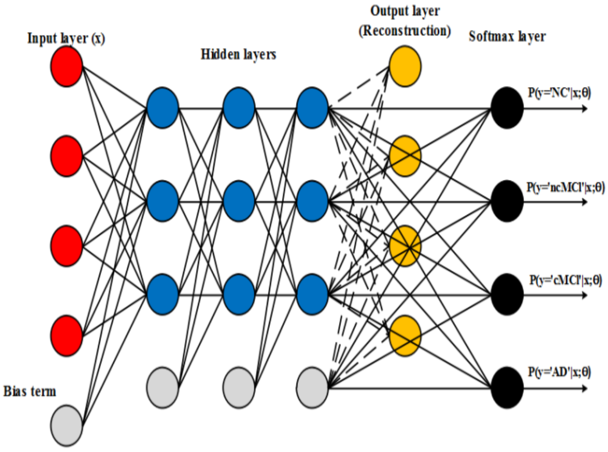

Figure.1. Structure of deep leaning.

As depicted in Figure 1, an automatic encoder consists of a multi-layer neural network. The true input value is represented by neurons that encode input. As a high-level depiction of its previous fold, each buried layer can be considered as a very minor component. Smallest representation of the inserting layer can be found on the cover of the inserting layer. Eq.(13) is used to transfer activation signals across the network. Exit rate will be reached when accrual approaches The Eq. (14) of each overlay can be used to quantify the activity of neurons.

$$\begin{cases} a_{(i)}^{(i)} = x^{(i)}, l = 1 \\ a_{(i)}^{(i)} = \sigma(W^l a + b), l = 1 \end{cases} \qquad (13)$$

$$h(W, b, x) = a^{(N)} \qquad (14)$$

Where x signifies unlabeled data $\{x^{(i)}\}_{i=1}^{m}, w = 1$,activation function can be set to hyperbolic target function to supply nonlinearity for network modelling complex relationships and input data as well as activation output layer h(W,b,x) signifies input data and activation output layer. It is necessary to employ the loss of production as an objective for optimization in order to train this unsupervised ideal (15).

$$L(W, b, x, z) = \min_{W,b} E(W, b, x, z) + \gamma||W||_2^2 + \beta K(W, b, x) \qquad (15)$$

Where, $E(W, b, x, z) + \gamma||W||_2^2$ signifies the loss demonstrated by the squared error in the experiment. A reduced weight is achieved by the second term, while the third term controls the sparsity penalty regime, which is activated at the zero point of the objective function using the KullbackLeibler deviation for all training trials.... In Eq. (16), this is explained.

$$K(W, b, x) = \sum_j^n ID_{KL}(\rho || \sum_{i=1}^m h_i(x^{(i)}; W, b)) \qquad (16)$$

This study eliminates the temporal output layer while simultaneously training the encoder's hidden layer. The previously concealed layer has been replaced with a softMax output layer at the top of the taught self-coding stack. The activation function used in the softmax layer is different from the one used in the previous layer and can be linear. Eq.(17) is the activation function for softmax.

$$h_i^l = \frac{e^{w_i^l h^{l-1+b_i^l}}}{\sum_j w_i^l h^{l-1+b_i^l}} \qquad (17)$$

Where w_i^l is i^{th} row of W^l and b_i^l is i^{th} ending layer bias term. In this study, an estimator of P(Y=i|x) can be used.

4. Results and Discussion

All tests were conducted on an Ubuntu 14.0.4 LTS with Python. Use Scikit-learn to implement proposed as well as traditional machine learning algorithms. Using GPU-enabled TensorFlow4, three DNNs were developed with a higher Keras5 framework backend. The GPU was NVidia GK110BGL Tesla K40 and the CPU was configured to run on 1 Gbps Ethernet network (32 GB RAM, 2 TB hard disk.

4.1. Performance metrics

The basis truth value is necessary in the evaluation of the various statistical measures. In the instance of binary or multi-class classification, the foundation truth consisted of several connection registers that were normal or attack. Let A and B be the sum of usual set that contains both positive and negative comments, where abnormal class contains negative comments in the test dataset and use the subsequent terms to determine the excellence of the classification model:

- ❖ True Positive (TP) - the sum of mentioned records properly categorized to the Usual class.
- ❖ False Negative (FN) - the sum of negative records incorrectly categorized to the Usual connection record.
- ❖ True Negative (TN) - the sum of mentioned records properly categorized to the abnormal class.

- False Positive (FP) - the sum of mentioned records wrongly categorized to the abnormal linking record.

Figure 2: Graphical Representation of proposed model in binary classification

The following evaluation metrics are examined based on the above given terms.

A. **Accuracy:** The ratio of the predictable connection records to the whole test dataset is estimated. If the precision is higher, then the model of ML is better. Accuracy is an appropriate metric for an experimental dataset with balanced classes.

B. **Precision:** It guesses the ratio of correctly identified attachment logs to the number of all identified attachment logs. The ML model is better if the precision is higher.

C. **F1-Score:** F1-Score is known as F1-Score, too. It is precision and recall the harmonic mean. The greater the F1-score is the better

D. **False Positive Rate (FPR):** It calculates the ratio of normal linking records to the number of standard connection records as abnormal. The lower FPR will improve the model for ML.

4.2. Performance Validation of Proposed Model for Binary Class

Here, the experiments are conducted to test whether the input is positive or negative as binary class classification. The validated analysis is shown in Table 3 and Figure 2.

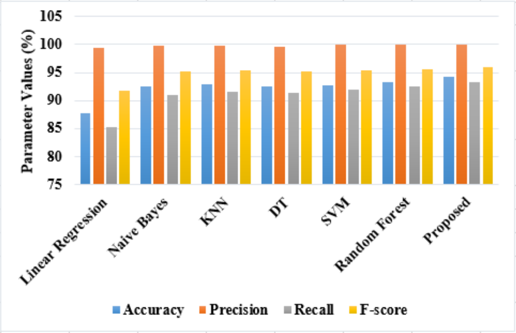

Table.3.Comparative analysis of binary class on Proposed with various existing algorithms.

In the analysis of accuracy, LR achieved 87.80%, NB, KNN, DT and SVM achieved 92%, RF achieved 93%, where the proposed model achieved 94.32%. Likewise, the KNN, DT and SVM achieved 91% of recall, RF achieved 92% of recall and proposed model achieved 93% of recall. While comparing with all techniques, LR achieved low performance, i.e. 85% of recall and 91% of F-measure. But, the proposed model achieved 96% of F-measure and 99.95% of precision. The reason for better performance is that the optimal features are selected by MRFO algorithm, before fed into classifiers. In addition, polarity calculations are done before the selection of features, which is not used in the existing techniques. The next Table 4 and Figure 3 shows the multi-class classification of various techniques in terms of different metrics.

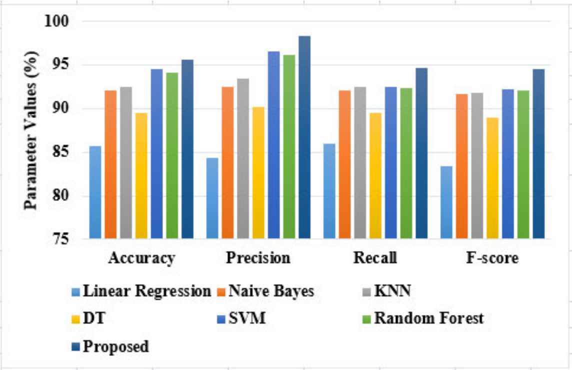

Figure 3: Graphical Representation of proposed model in multi-class classification

In the analysis of recall, the proposed model achieved 94.62%, NB, KNN, SVM, RF achieved 92% and DT achieved 89%. The existing LR achieved 85.71% of accuracy, 84.32% of precision, 85.93% of recall and 83.45% of F-score and this is because it is unable to handle large dataset for training as well as testing. In addition, the existing techniques are not used feature selection process as well pre-processing are not carried out. Hence, they achieved low performance, while comparing with proposed model. Therefore, the proposed model achieved 95.62% of accuracy, 98.32% of precision and 94.53% of F-score. When comparing with binary class, the performance of proposed model is less in multi-class classification and this is due to variations of samples in training and testing. This must be improved in the future work by introducing deep learning technique or fine-tuning the hyper-parameter of proposed model using recent techniques of meta-heuristic algorithms. In order to test the effective of feature selection technique, it is implemented with all existing models and verified in terms of accuracy, which is shown in Table 5.

When the number of feature is less, all techniques provides good performance to verify the comments. For instance, the proposed model achieved 95.89% on 11 features, 96.23% on 8 features and 97.89% on 4 features. The LR techniques provides slightly good performance in all minimal features, i.e. 89% of accuracy, 90% and 88.18% of accuracy on 11, 8 and 4 feature sets. When the techniques are tested with 8 features, NB, SVM achieved 92% of accuracy, KNN, DT achieved 93% of accuracy and RF achieved 94% of accuracy. This experimental results proves that the features plays an important role in the final classification of all models.

5. Conclusion

Text mining using feature selection and machine learning algorithms are used to review the products from customer reviews in order to discover the best products. Large product datasets from e-commerce sites that have been assessed by numerous customers that engage in online activities are the basis for our opinion mining efforts. Initially, five different process are carried out to clean the raw comments, then document frequency as well as Chi-Square is used to extract the important features. From these features, the optimal features are selected by MRFO technique and these output are fed into auto-encoder for classification process. Finally, the customer reviews on best product is analyzed as positive, negative and neutral comments. From the analysis, it is shows that the proposed model achieved 95.62% of accuracy and 98.32% of precision, where the random forest classifier achieved 94.16% of accuracy and 96.17% of precision on multi-class classification. In future work, the results of proposed model on both binary and multi-class classification must be improved by using either deep learning or fine-tuning the hyper-parameters of proposed model with recent optimization techniques.

References

[1] Rahardja, U., Hariguna, T. and Baihaqi, W.M., 2019, August. Opinion mining on e-commerce data using sentiment analysis and k-medoid clustering. In 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media) (pp. 168-170). IEEE.

[2] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," Proc. 7th ACM Conf. Recomm. Syst., pp. 165–172, 2013.

[3] G. Ling, M. R. Lyu, and I. King, "Ratings meet reviews, a combined approach to recommend," Proc. 8th ACM Conf. Recomm. Syst. - RecSys '14, pp. 105–112, 2014.

[4] Y. Tan, M. Zhang, Y. Liu, and S. Ma, "Rating-boosted latent topics: Understanding users and items with ratings and reviews," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2016-Janua, pp. 2640–2646, 2016.

[5] Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., & Bezbradica, M. (2019, September). Customer purchase behavior prediction in ecommerce: a conceptual framework and research agenda. In International Workshop on New Frontiers in Mining Complex Patterns (pp. 119-136). Springer, Cham.

[6] Raorane, A., & Kulkarni, R. V. (2011). Data mining techniques: A source for consumer behavior analysis. arXiv preprint arXiv:1109.1202.

[7] Voinea, L., & Filip, A. (2011). Analyzing the main changes in new consumer buying behavior during economic crisis. International Journal of Economic Practices and Theories, 1(1), 14-19.

[8] Nayyar, T. (2019). Analyzing Customer Buying Behavior.

[9] Saeed, R., Lodhi, R. N., Rauf, A., Rana, M. I., Mahmood, Z., & Ahmed, N. (2013). Impact of Labelling on Customer Buying Behavior in Sahiwal, Pakistan. World Applied Sciences Journal, 24(9), 1250-1254.

[10] Pahwa, B., Taruna, S., & Kasliwal, N. (2017). Role of Data mining in analyzing consumer's online buying behavior. International Journal of Business and Management Invention, 6(11), 45-51.

[11] Familmaleki, M., Aghighi, A., & Hamidi, K. (2015). Analyzing the influence of sales promotion on customer purchasing behavior. International Journal of Economics & management sciences, 4(4), 1-6.

[12] Prabhakumari, K. and Silviya, M.T., ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[13] Anggoro, M. A., & Purba, M. I. (2020). The Impact of Attractiveness of Ads and Customer Comments Against to Purchase Decision of Customer Products on the User of Online Shop Applications in the City of Medan. Jurnal Ilmiah Bina Manajemen, 3(1), 1-9.

[14] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[15] Xu, F., Pan, Z. and Xia, R., 2020. E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management*, *57*(5), p.102221.

[16] Uma Maheswari, S. and Dhenakaran, S.S., 2021. Opinion mining on integrated social networks and E-commerce blog. *IETE Journal of Research*, pp.1-9.

[17] Aasha, A.A., Hashmani, M., Baloch, F. and Ehsan, A.H., 2021. Aspect based Real-Time Opinion Mining for Online Feedback of Products at AMAZON™.

[18] Truong, Q.D., Thi Bui, T.D. and Nguyen, H.T., 2021. Product Recommendation System Using Opinion Mining on Vietnamese Reviews. In *Soft Computing: Biomedical and Related Applications* (pp. 313-325). Springer, Cham.

[19] Singh, P., Singh, Y.P., Kapil, S., Srivastava, S. and Vishwakarma, V., 2021, November. An Improved Model for Opinion Mining of Public Reviews using Recurrent Neural Network. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)* (pp. 20-25). IEEE.

[20] Kim, T.Y. and Kim, H.J., 2022. Opinion Mining-Based Term Extraction Sentiment Classification Modeling. *Mobile Information Systems*, *2022*.

[21] Jacob, M.S. and Selvi Rajendran, P., 2022. Deceptive Product Review Identification Framework Using Opinion Mining and Machine Learning. In *Mobile Radio Communications and 5G Networks* (pp. 57-72). Springer, Singapore.

[22] Deng, J., Guo, J., & Wang, Y. (2019). A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering. Knowledge-Based Systems.

[23] Yu, D., Liu, G., Guo, M., & Liu, X. An improved K-medoids algorithm based on step increasing and optimizing medoids. Expert Systems with Applications, 92, 464–473 (2018).

[24] N. F. Ibrahim and X. Wang, "A text analytics approach for online retailing service improvement: evidence from Twitter," Decision Support Systems, vol. 121, pp. 37–50, 2019.

[25] F. Hu, L. Li, X. Xu, J. Wang, and J. Zhang, "Opinion extraction by distinguishing term dependencies and digging deep text features," Neural Computing and Applications, vol. 31, no. 9, pp. 5419–5429, 2019.

[26] A. Thakkar and K. Chaudhari, "Predicting stock trend using an integrated term frequency-inverse document frequencybased feature weight matrix with neural networks," Applied Soft Computing, vol. 96, article 106684, 2020.

[27] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," IEEE Access, vol. 9, pp. 52177–52192, 2021.

[28] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," Engineering Applications of Artificial Intelligence, vol. 70, pp. 25–37, 2018.

[29] P. Sur, Y. Chen, and E. J. Candès, "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square," Probability Theory and Related Fields, vol. 175, no. 1-2, pp. 487–558, 2019.

[30] W. Zhao, Z. Zhang and L. Wang, "Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications," Engineering Applications of Artificial Intelligence, vol. 87, p. 103300, 2020.

Dr.Jamal Mohamed Yaseen Zubeir working as Assistant professor in Department of Computer Science, Jamal Mohamed College, Trichy, Tamilnadu,India. He Completed his PG and PhD from Jamal Mohamed College Affiliated by Bharathidasan University.