ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2021.0345

CLASS WISE LINEAR DISCRIMINANT AND REGRESSION BASED BINARIZED NEAREST LEARNING IN DIGITAL MARKETING

K.S. Narayanan[1] and S. Suganya[2]
[1]*Department of MCA, Rathnavel Subramaniam College of Arts and Science, India*
[2]*Department of Computer Science, Rathnavel Subramaniam College of Arts and Science, India*

Abstract

The employment of internet and social media has remodeled behavioral aspects of consumer or student communities and methods in which organizations or educational institutions perform their business pattern. Both social and digital marketing put forwards efficient scopes to educational institutions by way of reduced costs, enhanced brand perception and elevated sales. Nevertheless, notable disputes prevail from obstructive electronic word-of-mouth and invasive and annoying online brand existence. Nowadays, students use online promotions to know about best universities for education globally. This university choice and students' feedback observed by student-experience shared across social media platforms. Several methods have been employed for selecting the university but not providing accurate information. This paper is motivated towards applying Machine Learning for learning, analyzing and classifying the student information based on the student experience by means of tweets in twitter. The twitter data with student tweets is collected from benchmark twitter dataset and applied to the proposed method, Class-wise Linear Discriminant and Regression-based Binarized Nearest Learning (CLD-RBNL). The CLD-RBNL method is split into two sections. First, preprocessing and relevant feature selection (i.e. tweets) are acquired by employing Class-wise Linear Discriminant-based Feature Selection (CLDFS) model to obtain dimensionality reduced tweets. To this result, Regression-based Binarized Nearest Neighbor model is applied for maximum lead generation. The CLD-RBNL method is compared with other state-of-the-art methods and found to outperform in terms of sensitivity, specificity, processing time, lead generation accuracy and error rate.

Keywords:

Class-wise, Linear Discriminant, Feature Selection, Digital Marking, Educational Services, Regression, Binarized Nearest Neighbor

1. INTRODUCTION

The exponential growth of Information and Communication Technology (ICT) has resulted in an upsurge concerning magnitude and fineness in shifting course contents that enhances the learning potentiality of Digital Learning Communities (DLCs). A survey conducted by Educause Center for Applied Research (ECAR) that roughly 67% surveyed students accepted that social media play a significant role in academic performance and enhancing career aspects. This is because social media bestow significant educational e-learning potentialities to students for collaborating with academicians, course material and tutor accessing even though physical boundary exists.

A latent variance-based structural equation model was proposed in [1] for measuring the student social medial perception via collaborative learning. A Machine Learning integrated social media marketing (ML-SMM) was proposed in [2] that elucidated the ideas of social media marketing and machine learning in detail.

Recent developments in digital marketing have led to the growth of promotion and selling of various products and services via internet using new machine learning based techniques. With aid of machine learning based techniques, a classification algorithm is trained utilizing certain features in tweets, which differentiate between educational establishments and therefore contributing to lead generation. These features or tweets obtained from the twitter dataset are extracted and analysis is made for final lead generation. The existing machine learning based methods [1] though extract features with higher reliability but however the sensitivity factor and the processing time with which the reliability was obtained was not focused. Also, prediction characteristics of true negative or specificity are less focused and thousands of fake universities are mushrooming every day. Therefore, there is requirement of designing an efficient digital marketing algorithm with accurate tweets utilized in lead generation and also discarding negative tweets.

To solve above said problem, this paper presents a machine learning based CLD-RBNL method that selects dimensionality reduced tweets in a computationally efficient manner and enhances the true positive rate or the detection of true tweets in an accurate manner. This paper presents a method that can detect essential and robust tweets using the tweet information present in the twitter dataset. Proposed method extracts the tweets from page source and analyze them to detect whether the given tweet are essential or not in determining the university information.

The main objective of the research work described here. To categorize the student information depended on the student experience by using tweets on Twitter, the CLD-RBNL method is introduced.

- To reduce the dimensionality tweets with reducing the processing time, Triplet preprocessing model is applied.
- To highly sensitive and computationally efficient tweets, Class-wise Linear Discriminant-based Feature Selection (CLDFS) model is utilized.
- To accurate lead generation with higher accuracy and lesser specificity, Regression-based Binarized Nearest Neighbor is applied.

Many several research works are developed for choosing the university by using machine learning models. However, it failed to offer the exact information. But the prediction time was not reduced. In conventional methods, the sensitivity and specificity factors were not considered and processing time was not reduced. Motivated by, the Class-wise Linear Discriminant and Regression-based Binarized Nearest Learning (CLD-RBNL) is introduced using machine learning for classifying the university information in an accurate manner. CLD-RBNL is designed with the novelty of, Triplet preprocessing model, Class-wise Linear Discriminant-based Feature Selection (CLDFS) model, and

Regression-based Binarized Nearest Neighbor. Main contributions of this paper are as follows:

- We propose CLD-RBNL to effectively select dimensionality reduced tweets on user generated tweets and accelerate the convergence of the lead generation accuracy. In addition, the architecture of our method has strong scalability.
- Our method can intuitively visualize the process of obtaining highly sensitive and computationally efficient dimensionality reduced tweets from linear discriminant analysis based on feature subsets, which increases the sensitivity and specificity.
- Our Regression-based Binarized Nearest Neighbor Lead Generation algorithm has strong robustness and accurate lead generation and can be applied to educational domains in identifying and selecting universities globally via the Internet.
- In order to evaluate the performance of our CLD-RBNL method in comparison to the contemporary state-of-the-art methods, we experiment on Twitter dataset.

The remainder of paper is organized as follows. Section 2 first given presents the related work. Section 3 elaborates CLD-RBNL. Section 4 presents the experimental settings for designing CLD-RBNL method in detail. Section 5 presents the evaluation metrics and discussion. Section 6 concludes paper and presents future work.

2. RELATED WORKS

By using an aggravating pattern, the exploration of links between big data, practical marketing approach and processing data in a real time manner for business to business marketing depending on big data was proposed in [3]. In [4], the collaborative perception from numerous dominant resource persons on problems concerning digital marketing was proposed.

In [5], an in-depth investigation on numerous machine learning techniques that are utilized to improve online advertising. Some of the Digital Business Platforms (DBPs) like, eBay, Google, and Uber in the recent years have seen immeasurable development, in [6], salient features and the role played by digital marketing in DBPs was discussed. The role played by education using Artificial Intelligence (AI) was proposed in [7].

The paper in [8] concentrated on the acquirement and application of the machine learning based analytical models for three different classes, marketing agencies, media companies, and advertisers, therefore concentrating on the accuracy factor. Strategic modelling using AI for digital marketing was designed in [9].

In [10], a Self-Attentive Convolutional Neural Networks (SACNNs) that was trained on top of pre-trained word vectors for detection of robust emotion was proposed. A Self Attentive Hierarchical model for integrative enhancing summarization of text and classification of sentiment was proposed in [11]. Banks prediction customer digitalization process using machine learning approach concentrating on the factors like, accuracy and time was proposed in [12].

A systematic literature review concerning learning methodology was presented in [13]. Yet another application to digital technology on educational aspect was proposed in [14]. Certain advantages and disadvantages prevailing in online shopping behavior aspect were discussed in [15].

An ensemble algorithm was applied in [16] to improve the accuracy of purchase intention of potential customers. Machine learning methods were applied in [17] to concentrate on the accuracy with critical feature selection. In [18], a college selection procedure based on consumer decision framework was proposed. In [19], a method considering several aspects like, reputation of university reputation, branding and university brand personality were investigated to identify student loyalty in purview of higher education.Motivated by above studies, CLD-RBNL method is proposed. The CLD-RBNL method aims to offer a two-fold contribution to the existing methods on educational service digitalization. First, by employing a CLDFS, it reveals highly sensitive and computationally efficient tweets. Second, Regression-based Binarized Nearest Neighbor Lead Generation is designed to provide a more complete picture of the lead generation digitalization process.

3. METHODOLOGY

This section presents a comprehensive illustration of methodology utilized for constructing digital marketing towards educational services. To start with preprocessing technique used in our work is proposed to get rid of irrelevant data (i.e., tweets) or data of least importance. Next, an elaborate account of proposed feature selection method with a sample case is designed. Finally, Regression-based Binarized Nearest Learning classifiers classify the tweet polarity into positive, negative and neutral classes for digital market education is designed.

3.1 TRIPLET PREPROCESSING MODEL

Preprocessing necessitates the removal of irrelevant data or tweets that does not accord to the lead generation process for education services.

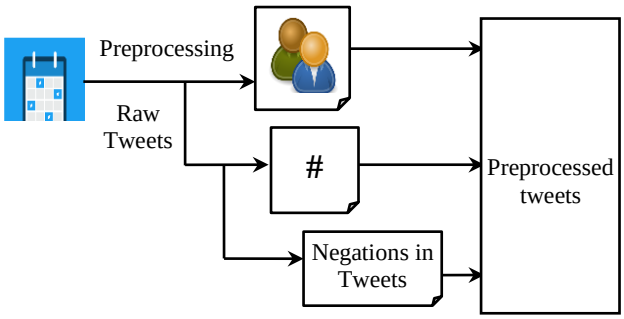

Fig.1. Block diagram of Triplet Preprocessing Model

We used three preprocessing techniques to process distance learning tweets, as shown in Fig.1. As shown in the Fig.1, the first and foremost data refers to the user ‘user (@)’ who posted the tweets. The second most essential part forms the ‘Hashtag (#)’ connected or correlated with the specific topic and indicated by the user in the tweets. Here, the symbol ‘#’ is eliminated by retaining the actual contents ‘Tweets’. During the final classification stage for estimating the lead towards educational domain negations play an essential function. For example, synchronization of negative words (‘not, don’t, dislike’) transpose

the tweet inclination into distinct polarity. Therefore, handling of negation words is utilized.

3.2 CLASS-WISE LINEAR DISCRIMINANT-BASED FEATURE SELECTION (CLDFS) MODEL

In this section, we propose a novel feature selection method based on class-wise information called CLDFS. The CLDFS model identifies linear subsets of features that separate two or more classes (i.e., users, hash and negation in our case). The resulting combination is utilized as a linear classifier for dimensionality reduction before classification for lead generation in educational services domain. The CLDFS model comprises three steps. The Fig.2 shows the block diagram of CLDFS model.

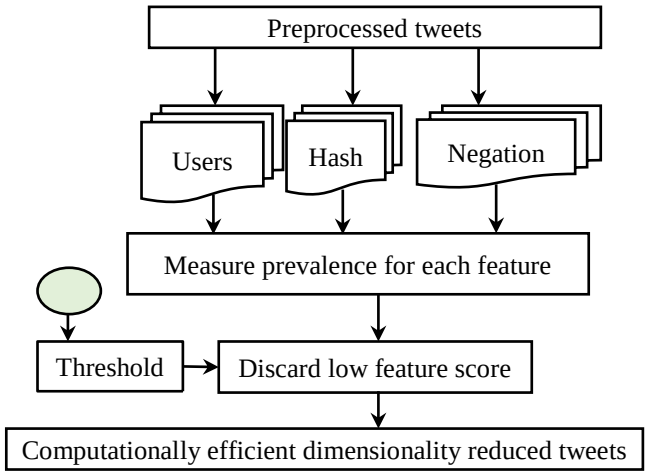

Fig.2. Block diagram of Class-wise Linear Discriminant-based Feature Selection Model

As shown in the Fig.2, to start with, the class-related short tweets are clustered. Next, the aggregate of the prevalence (i.e., repetition) of each feature equivalent to class are estimated. The acquired feature scores are arranged in descending order, where low feature scores are discarded by fixed threshold value. The value of threshold is fixed experimentally that refers to the number of features acquired from each class and repeated for each class. Finally, the chosen feature subsets from each class are aggregated to obtain overall features that are utilized for the upcoming training of classifiers.

Let us consider i number of classes and each class comprises of j number of short tweets with short tweets represented by n dimensional frequency vector. Let us further assume tweet matrix TM of size ($ij*n$) in such a manner that each row denotes a short tweet associated to class C_i and each column denotes a feature $F=\{F_1,F_2,\ldots,F_n\}$. To start with, the aggregate of the prevalence of each feature F_i analogous to class C_j is mathematically formulated as given below.

$$CP\left(C_j, F_i\right) = \sum_{i=1}^{k} P\left(TM_i, F_i\right) \quad (1)$$

From Eq.(1), $P(TM_i,F_i)$ represents the prevalent of existence of the features F_i in short tweet matrix TM_i and i denote the number of short tweets in the classes C_j respectively. Then for each class the size of the resultant $CP(C_j,F_i)$ matrix is assumed with the conditional probability density function $Prob(TM_i|F_i=0)$ and $Prob(TM_i|F_i=1)$.

Moreover, the values of $CP(C_j,F_i)$ are arranged in descending order to get the most prevalent existences or occurrences of tweets within the classes. Followed by which a feature subset of n' are obtained based on the threshold values experimentally. The dimensionality reduced selected features are $F'=\{F_1',F_2',\ldots,F_n'\}$. In a similar manner the iterations are repeated for each class. Finally, an association function is applied to feature subsets achieved from each class as given below.

$$F' = F_1' \cup F_2' \cup \ldots \cup F_n' \quad (2)$$

With the resultant values obtained from the above Eq.(2), F' are utilized for the successive training of the classifiers. Followed by which the score of polarity P and subjectivity S is estimated as given below.

$$Score\{Polarity(P)=Avg(W,P[F']),W\} \quad (3)$$

$$Score\{Subjectivity(S)=Avg(W,S[F']),W\} \quad (4)$$

From Eq.(3) and Eq.(4), the score of polarity $Score(P)$ and score of subjectivity $Score(S)$ for the respective feature subsets F' is measured on the basis of assigned weights W and the polarity P, subjectivity S of each feature.

3.2.1 Class-wise Linear Discriminant-based Feature Selection Algorithm:

Input: Dataset 'DS', Tweets '$T=T_1,T_2,\ldots,T_n$', Feature '$F=\{F_1,F_2,\ldots,F_n\}$'

Output: Highly sensitive and computationally efficient tweets

Step 1: Initialize Tweet Matrix 'TM'

Step 2: Begin

Step 3: For each Dataset 'DS' with Tweets 'T' from Feature 'F'

- **a.** Estimate prevalence of each feature as in Eq.(1)
- **b.** Apply association function to feature subsets as in Eq.(2)
- **c.** Estimate polarity as in Eq.(3)
- **d.** Estimate subjectivity as in Eq.(4)
- **e.** Return dimensionality reduced tweets (RT)

Step 4: End for

Step 5: End

Algorithm 1 explains the process of Class-wise Linear Discriminant-based Feature Selection, for each dataset DS with tweets T consisting of distinct features F provided as input, the objective remains in acquiring the dimensionality reduced correct tweets in a computationally efficient manner. To achieve this objective first, linear discriminant analysis is performed for the class-wise tweets that in turn only obtain the discriminant tweets that are said to be high of use in determining whether a subset of features is effective in predicting a specific category of class, therefore contributing to processing time. Next, with the aid of acquiring the polarity and subjectivity values for each subset of features via aggregate function not only reduces the dimensionality but also detects the tweets correctly, therefore contributing to sensitivity.

3.3 REGRESSION-BASED BINARIZED NEAREST NEIGHOR

As far as digital marketing for the educational services is considered, whenever the feature subsets are organized into

groups, intra-class correlation is used. The objective of using the intra-class correlation for providing personalized information to the users (i.e., students) is to optimize linear transformation so that a tradeoff between same and different classes of feature subsets. The intra-class correlation in our work is utilized to quantify the degree to which individuals (i.e., polarity) with a fixed degree of relatedness resemble each other in terms of a score.

Another prominent objective is the assessment of specificity of quantitative true negative measurements made by different observers in the variables within a category. Therefore, in this section with the obtained highly sensitive and computationally efficient dimensionality reduced tweets, the next step in our work remains in provided personalized information to the users (i.e., students in selecting the universities) in terms of quantitative trait based on intra-class variance.

Let us consider new dimensionality reduced tweets with its polarity value obtained is generalized and ranges from '0 to 1' with φ=1 when all nearest neighbors (i.e., tweets obtained from feature subset) are from the same class and φ=0 when all nearest neighbors are from different classes. According to this updated generalized class-wise value, the maximum intra-class correlation is defined as follows.

$$C' = \arg\max \sum_{i,j=1}^{C} \varphi_i^j \tag{5}$$

From the above Eq.(5), φ_i^j represents the updated generalized class-wise value when the data sample S assumed to be class j is added to the dimensionality reduced tweets RT, i.e. $RT_j = RT_j \cup \{S\}$. Followed by which the linear tweet transformation matrix TTM is formulated with the objective of identifying the linear tweet transformation matrix TTM that generates the largest intra-class correlation over the whole dimensionality reduced tweets sample. Mathematically, the minimization function with the intra-class correlation to its negative with respect to polarity is defined as follows:

$$TTM\left(P\right) = \arg\max - \sum_{i=1}^{C} \frac{\sum_{p \in RT_i} \sum_{p_k \in NN_0(p)} \left| TM\left(p - p_k\right) \right|}{\sum_{p \in RT_i} \sum_{p_k \in NN(p)} \left| TM\left(p - p_k\right) \right|} \tag{6}$$

From Eq.(6), the tweet transformation matrix TTM for each feature subset possessing polarity score P is estimated on the basis of the minimization function *argmin*. This function is applied for feature subset of C classes with respect to dimensionality reduced tweets RT_i and $NN_0(p)$ representing the set of k nearest neighbors of dimensionality reduced tweets RT_i (i.e., nearest tweets with respect to the actual dimensionality reduced tweets).

The existing ML-SMM [2] has no training stage, and its deterministic decision rule makes it difficult to incorporate existing prediction characteristics of true negative, i.e., disjunct loss functions. In this section, we present an alternative of the original ML classification whose training objective function is continuous. This is achieved by means of regression-based binary function. With this function, the threshold utilized for thresholding the approximated tweets are considered to be a variable represented by T_2 that may be distinct from the customized threshold T_1 being utilized for thresholding the definite tweets. Then, the variable T_2 is estimated in such a manner that the difference between the approximated tweets and the definite tweets is minimized. This is achieved using the loss function as given below.

$$Loss\left(T_2\right) = \sum_{T \in RT} \left(Fun\left(App_T, T_1\right) - Fun\left(Def_T, T_2\right) \right) \tag{7}$$

From the above Eq.(7), App_T represents the approximated tweets and Def_T stands for the definite tweets respectively. Finally, the resultant regression-based binary function is estimated as given below.

$$Fun\left(App_T, T_1\right) = \begin{cases} 1 & App_T > T_1 \\ 0 & Otherwise \end{cases} \tag{8}$$

From the above Eq.(8), the approximated and definite tweets are obtained for each class feature subsets. With this resultant value, the polarity value with minimum loss function is obtained that in turn reduces the true negative rate of lead being generated. With this minimum true negative the probability of falsely generating lead while selecting universities from a cluster is improved, therefore, causing productivity in the educational domain services.

3.3.1 Regression-based Binarized Nearest Neighbor Lead Generation Algorithm:

Input: Dataset '*DS*', Tweets '$T=T_1,T_2,\ldots,T_n$', Feature '$F=\{F_1,F_2,\ldots,F_n\}$'

Output: Minimized test negative accurate lead generation

Step 1: Initialize Tweet Matrix '*TM*', dimensionality reduced tweets (*RT*)

Step 2: Begin

Step 3: For each Dataset '*DS*' with Tweets '*T*' from Feature '*F*'

- **a.** Evaluate maximum intra-class correlation as in Eq.(5)
- **b.** Evaluate minimization function with the intra-class correlation to its negative to obtain lead based on polarity as in Eq.(6)
- **c.** Estimate the loss function as in Eq.(7)
- **d.** Estimate regression-based binary function
- **e.** If '$Fun(App_T,T_1)\leq 1$'
  - **i.** Then 'Label = positive tweets'
- **f.** End if
- **g.** If '$Fun(App_T,T_1)\leq -1$'
  - **i.** Then 'Label = negative tweets'
- **h.** End if
- **i.** If '$Fun(App_T,T_1)=1$'
  - **i.** Then 'Label = neutral tweets'
- **j.** End if
- **k.** Return (value, label)

Step 4: End for

Step 5: End

Algorithm 1 explains the process of Regression-based Binarized Nearest Neighbor Lead Generation. For each dataset *DS* with tweets *T* consisting of distinct features *F* provided as input, the objective of the algorithm remains in obtaining the lead with maximum accuracy and minimum specificity. To achieve this objective, first, intra-class correlation is improved in such a

way by employing the *argmin* function that in turn contributes to specificity.

Then, with regression-based binary function in convergence to loss function, the accuracy with which the lead generation is achieved is improved. Hence, the tweets obtained in the form of digital marketing for educational services in our work, assists in promoting higher education by selecting the universities of choice remotely.

4. EXPERIMENTAL SETTINGS

In this section, the proposed CLD-RBNL method is employed to select higher education therefore laying strong foundations for precise and fine grained education branding based on the sentiment analysis of the tweets about distance learning. First, tweets collected from distance learning dataset are preprocessed, followed by which dimensionality reduced relevant tweets for lead generation are selected in a precise fashion.

Finally, the regression-based binary functions for each user's tweets are entered into the neighbor learning system as input. Finally, the polarity of each tweet is classified as a positive, negative or neutral for all tweets with minimum loss function. Experimental evaluations are performed in Python by utilizing distance learning [20] employing three distinct csv files, i.e. raw files, processed files and sentiment files.

5. DISCUSSION

Comparative analysis of lead generation methods is performed and compared with three different methods CLD-RBNL, Latent Variance-based Structural Equation (LVSE) [1] and ML-SMM [2]. Performance analysis is made with five different parameters for the respective number of tweets and distinct tweet sizes.

5.1 PERFORMANCE ANALYSIS OF SENSITIVITY

Sensitivity is referred to as the probability of detection of dimensionality reduced tweets. In other words, sensitivity in our work estimates the ratio of positive values (i.e., correctly detected tweets) identified in a given preprocessed dataset.

$$Sen = \sum_{i=1}^{n} \frac{TP}{TP + FN} * 100 \quad (9)$$

From Eq.(9), the sensitivity *Sen* is measured based on the true positive *TP* (i.e., correctly detected tweets) and false negative *FN* (i.e., incorrectly detected unwanted tweets). Sensitivity rate is measured in percentage (%). The sample calculation of three methods are given below,

5.1.1 Sample Mathematical Calculation for Sensitivity:

- **Existing Latent Variance-based Structural Equation**: Total user counts= 500 and the number of correctly detected tweets is 110 and the incorrectly detected unwanted tweet is 40. Thus, the sensitivity is calculated as follows,

  Sen(LVSE)=110/(110+40)*100=73.33%

- **Existing ML-SMM:** The number of correctly detected tweets is 100 and the incorrectly detected unwanted tweet is 50. Thus, the sensitivity is calculated as follows,

  Sen(ML-SMM)=100/(100+50)*100=66.66%

- **Proposed CLD-RBNL:** The number of correctly detected tweets is 130 and the incorrectly detected unwanted tweet is 20. Thus, the sensitivity is calculated as follows,

  Sen(CLD-RBNL) = 130/(130+20)*100=86.66%

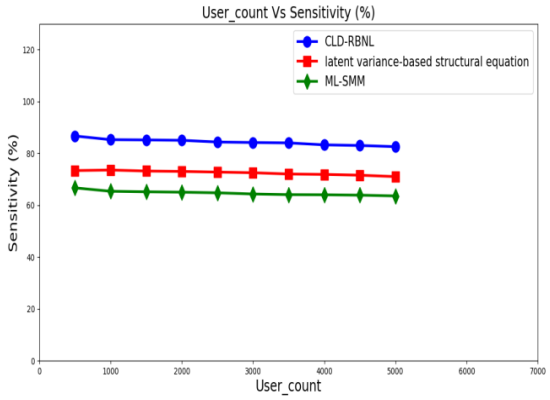

Fig.3. Sensitivity

The Fig.3 shows the sensitivity rate with 500 to 5000 user counts. It shows that the overall sensitivity rate decreases to the optimal value, the number of classes (i.e., true positive achieves the maximum value of 130) at which nearest tweets of all data feature subsets are from the same class. In the first iteration, the simulation with 500 user counts found a true positive rate of 130, 110 and 100, false negative rate of 20, 40 and 50 using the three methods. With this, the sensitivity rate was observed to be 86.55%, 73.33% and 66.66%. In the second iteration, the sensitivity rate was observed to be 85.25%, 73.55% and 65.35%. The x-axis refers to the user counts and the y-axis refers to the sensitivity of three methods. As shown in the graphical chart, there are three various colors of lines such as blue, red, and green indicate the sensitivity of three techniques such as the CLD-RBNL method, existing [1] and [2] respectively. Among the three methods, the proposed CLD-RBNL method has the ability for increasing sensitivity.

From this result, the sensitivity rate of CLD-RBNL method was better than [1] and [2] due to the application of CLDFS model based on the feature subsets. By applying this model, the resultant values of polarity and subjectivity were acquired for each subset of features via aggregate function that in turn not only minimized the dimensionality. This is aid to determine the tweets. The average of ten results shows that the sensitivity is significantly improved using CLD-RBNL method by 16% and 12% compared to [1] and [2].

5.2 PERFORMANCE ANALYSIS OF PROCESSING TIME

Processing time refers to the time consumed in obtaining the dimensionality reduced tweets. Minimum the processing time involved in obtaining dimensionality reduced tweets, maximum the related features are acquired by the students involving decision making process engaging educational services and therefore higher is the lead generation accuracy said to be.

$$PT = \sum_{i=1}^{n} UC_i * Time[DRT] \quad (10)$$

From Eq.(10), the processing time *PT* is measured on the basis of user count considered for simulation UC_i and the time involved in obtaining dimensionality reduced tweets time [*DRT*]. It is measured in milliseconds (ms). The sample calculation of three methods are given below,

5.2.1 Sample Mathematical Calculation for Processing Time:

- **Existing Latent variance-based structural equation:** Total user counts = 500 and the time involved in obtaining dimensionality reduced tweets is 0.095ms. Then, the processing time is computed as follows,

  PT(LVSE) = 500*0.095ms = 47.5ms

- **Existing ML-SMM:** The time involved in obtaining dimensionality reduced tweets is 0.105ms. Then, the processing time is computed as follows,

  PT(ML-SMM) = 500*0.105ms = 52.5ms

- **Proposed CLD-RBNL:** The time involved in obtaining dimensionality reduced tweets is 0.085ms. Then, the processing time is computed as follows,

  PT(CLD-RBNL) = 500*0.085ms = 42.5ms

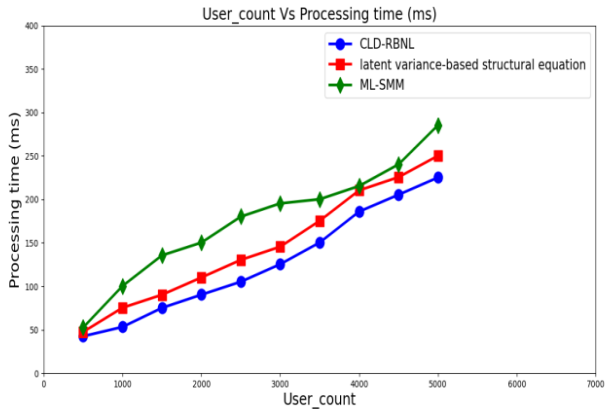

Fig.4. Processing Time

The Fig.4 shows analysis of processing time for CLD-RBNL and existing [1] and [2]. The user count is taken in the horizontal direction and the processing time is observed at the vertical axis. To explore the significance of CLD-RBNL method, we show the average processing time of CLD-RBNL, existing [1] and [2] with varying user counts. It can be seen that the processing time performance varies with the user count frequency. While this variation depends on the tweets size with respect to each user count, CLD-RBNL method shows performance improvement over [1] and [2]. In other words, though increasing the word count results in the increase of processing time, but simulations conducted in the first iteration with 500 user count show a processing time of 42.5ms using CLD-RBNL method, 47.5ms and 52.5ms using [1] and [2] respectively. In the second iteration, with consideration of 500 user counts, the processing time of existing [1] [2] is obtained as 75.35ms and 100.35ms respectively. Besides, the proposed method obtains the processing time as 53.25ms. The obtained results of the proposed method are compared to conventional methods. With this result the processing time improvement was found in CLD-RBNL method due to the application of CLDFS algorithm. By applying this algorithm, linear discriminant analysis were employed for obtaining class-wise tweets that in turn only selected only the discriminant tweets to be of highly influential in determining subset of features for predicting specific category of class. This reduced the processing time for obtaining dimensionality reduced tweets using CLD-RBNL by 15% compared to [1] and 30% compared to [2] respectively.

5.3 PERFORMANCE ANALYSIS OF SPECIFICITY

Specificity is referred to as the ratio of negatives that are correctly identified (i.e., the proportion of the users' traits who do not have the condition (i.e., negative traits who are correctly identified as not having the positive traits condition). In other words, it refers to the prediction characteristics of true negatives in feature subset within a class in a dataset.

$$Spe = \sum_{i=1}^{n} \frac{TN}{TN + FP} * 100 \qquad (11)$$

From Eq.(11), the specificity *Spe* is measured based on the true negative *TN* (i.e., correctly detected unwanted tweets) and false positive *FP* (i.e., incorrectly detected tweets). It is measured in percentage (%). The sample calculation of three methods are given below,

5.3.1 Sample Mathematical Calculation for Specificity:

- **Existing Latent variance-based structural equation:** Total user counts= 500 and the number of correctly identified tweets is 310 and the incorrectly identified unwanted tweet is 40. Thus, the specificity is evaluated as follows,

  Spe(LVSE)=310/(310+40)*100 = 88.57%

- **Existing ML-SMM:** The number of correctly identified tweets is 300 and the incorrectly identified unwanted tweet is 50. Thus, the specificity is evaluated as follows,

  Spe(ML-SMM)=300/(300+50)*100 = 85.71%

- **Proposed CLD-RBNL:** The number of correctly identified tweets is 330 and the incorrectly identified unwanted tweet is 20. Thus, the specificity is evaluated as follows,

  Spe(CLD-RBNL) = 330/(330+20)*100 = 94.28%

Table.1. Performance Values of Specificity

User count	Specificity (%)		
	CLD-RBNL	LVSE	ML-SMM
500	94.28	88.57	85.71
1000	93.25	87.55	84.35
1500	93.05	87.35	84.15
2000	93	87.15	84
2500	92.75	87	82.15
3000	92.55	86.55	82
3500	92.3	86.35	81.55
4000	92	86.1	85
4500	91.85	86	80.35
5000	91	85.55	80

The Table.1 strikes a specificity rate for CLD-RBNL method and existing [1] and ML-SMM [2]. From Table.1, specificity rates

using CLD-RBNL are higher as compared to other method. Though the specificity rate is decreasing with the increase in the user count, with simulations conducted using 500 user counts, false positive rate using the three methods are 20, 40 and 50. Therefore, the specificity rate were observed to be 94.28% using CLD-RBNL method, 88.57% using [1], and 85.71% using [2]. In the second iteration, the specificity rate was observed to be 93.25% using the CLD-RBNL method, 87.55% using [1], and 84.35% using [2].

Followed by, different performance results are observed for each method. For each method, ten different results are observed. The improvement in specificity using CLD-RBNL method was owing to application of intra-class correlation in Regression-based Binarized Nearest Neighbor Lead Generation algorithm. The maximum intra-class correlation is measured. The minimization function is estimated with the intra-class correlation to its negative based on the polarity. For each feature subset possessing polarity score *P* was evaluated based on minimization function *argmin* that improved specificity using CLD-RBNL method by 7%and 12% compared to [1] and [2].

5.4 PERFORMANCE ANALYSIS OF LEAD GENERATION ACCURACY

In this section, the performance metrics, lead generation accuracy is measured as given below.

$$LGen_{acc} = \sum \frac{TLea_{acc}}{T_{size}} * 100 \quad (12)$$

From Eq.(12), the lead generation accuracy $LGen_{acc}$ is measured based on tweet size T_{size} and tweets learned accurately $TLea_{acc}$. It is measured in percentage (%). The Table.2 show the results of lead generation accuracy for CLD-RBNL, existing [1] and [2]. The sample calculation of three methods are given below,

5.4.1 Sample Mathematical Calculation for Lead Generation Accuracy:

- **Existing Latent variance-based structural equation:** Total user counts =500 and the number of accurately learned tweets is 225 and the tweet size is 250. Thus, the lead generation accuracy is measured as follows,

$$LGen_{acc}(\text{LVSE}) = 225/250*100 = 90\%$$

- **Existing ML-SMM:** The number of accurately learned tweets is 215 and the tweet size is 250. Thus, the lead generation accuracy is measured as follows,

$$LGen_{acc}(\text{ML-SMM}) = 215/250*100 = 86\%$$

- **Proposed CLD-RBNL:** The number of accurately learned tweets is 242 and the tweet size is 250. Thus, the lead generation accuracy is measured as follows,

$$LGen_{acc}(\text{CLD-RBNL}) = 242/250*100 = 96.8\%$$

Table.2. Performance Values of Lead Generation Accuracy

Tweet size	Lead Generation Accuracy (%)		
	CLD-RBNL	LVSE	ML-SMM
250	96.8	90	86
500	96.35	89.35	85.55
750	96	89.15	85.25
1000	95.35	89	85
1250	95.15	88.55	84.15
1500	95	88.35	84.05
1750	94.65	88	84
2000	94.3	87.35	83.55
2250	94.15	87.2	83.15
2500	94	87	83

From Table.2, CLD-RBNL method has higher values of lead generation accuracy. The Table.2 demonstrates the result analysis of lead generation accuracy with respect to tweet size. To conduct experimental work, tweet size in the range of 250 to 2500 is considered. The obtained results of lead generation accuracy using the CLD-RBNL method are compared to the two existing [1] and [2]. Thus it can be observed that the CLD-RBNL method is highly accurate and provides accurate university learning model to the students requesting for their selection globally as compared to existing [1] and [2]. Let us considers 250 tweet sizes for conducting the experiments in the first iteration. By applying the CLD-RBNL method, 242 tweet sizes are correctly classified and the accuracy is 96.8% whereas the accuracy percentage of the existing [1] and [2] are 90% and 86% respectively. For each method, ten different results are observed. The performance of the proposed TPFMDANN-LSTM is compared to other existing methods. The reason behind the improvement was due to the application of regression-based binary function where approximated and definite tweets were obtained for each class feature subsets separately. The polarity of each tweet is categorized as a positive, negative or neutral for all tweets with minimum loss function. Accuracy was improved using CLD-RBNL method by 8% compared to [1] and 13% compared to [2].

5.5 PERFORMANCE ANALYSIS OF LEAD GENERATION ERROR RATE

The error rate is defined as the ratio of a tweets learned incorrectly classified to the total number of tweets. Then, the lead generation error rate calculated in terms of percentages (%) and mathematically determined using the below,

$$LGen_{err} = \sum \frac{TLea_{inacc}}{T_{size}} * 100 \quad (13)$$

From Eq.(13), the lead generation error rate $LGen_{err}$ is estimated based on tweet size T_{size} and tweets learned inaccurately $TLea_{inacc}$. Sample calculation using error rate is given below,

5.5.1 Sample Mathematical Calculation for Error Rate:

- **Existing Latent variance-based structural equation:** Total user counts =500 and the number of inaccurately learned tweets is 25. Thus, the lead generation error rate is measured as follows,

$$Sen(\text{LVSE}) = 25/250*100 = 10\%$$

- **Existing ML-SMM:** The number of inaccurately learned tweets is 35. Thus, the lead generation error rate is measured as follows,

$$Sen(\text{ML-SMM}) = 35/250*100 = 14\%$$

- **Proposed CLD-RBNL:** The number of inaccurately learned tweets is 8. Thus, the lead generation error rate is measured as follows,

$$Sen(\text{CLD-RBN}) = 8/250*100 = 3.2\%$$

Table.3. Performance values of lead generation error rate

Tweet size	Lead Generation Error Rate (%)		
	CLD-RBNL	LVSE	ML-SMM
250	3.2	10	14
500	3.65	10.65	14.45
750	4	10.85	14.75
1000	4.65	11	15
1250	4.85	11.45	15.85
1500	5	11.65	15.95
1750	5.35	12	16
2000	5.7	12.65	16.45
2250	5.85	12.8	16.85
2500	6	13	17

The Table.3 shows the performance values of error rate with respect to the tweet size taken in the range from 250 to 2500. The experiments are conducted by comparing the proposed CLD-RBNL method with the existing methods [1] and [2]. Both the proposed and existing methods successfully reduce the error rate for classifying the student information. Specifically, the proposed method attains better results on the improvement of error rate than the other two existing methods.

In the first iteration, let us consider 500 user counts, error rate of the proposed CLD-RBNL method is obtained as 3.2 % whereas, the existing [1] [2] is obtained as 10% and 14% correspondingly. In the second iteration, with 500 user counts, error rate of proposed CLD-RBNL method is obtained as 3.65%whereas, the existing [1] [2] is obtained as 10.65% and 14.45% correspondingly.

Similarly, 10 iterations are conducted with different methods. The reason for the lesser error rate is to apply the regression-based binary function to measure the difference among the approximated tweets and the definite tweets. Then, this function returns the classification results and minimizes the incorrect tweet classification. In this way, tweet are incorrectly classified into the respective classes hence it reduces the error rate. As a result, the error rate of the proposed CLD-RBNL method is reduced by 59% and 69% as compared to existing [1] and [2] respectively.

6. CONCLUSION

This work was directed towards developing a suitable learning system that obtain the student review data that is available in the form of tweets, understand the prevalence of tweet and predict the interest of students towards selecting a university globally. Eight essential features were extracted from twitter data and using this, the CLD-RBNL method was designed using machine learning based learning model. This system has been found to be instrumental in classifying the preferred choice of universities based on the subjectivity and polarity of tweets. The important characteristic of this CLD-RBNL method is that there is zero interference of factor involved when it comes to predicting the student choice of universities. The performance of this CLD-RBNL method was evaluated using four metrics namely sensitivity, specificity, processing time and lead generation accuracy and the results were compared and analyzed with other contemporary methods. These results that were discussed in the previous section show that the sensitivity, specificity and lead generation accuracy value of CLD-RBNL method is significantly higher than the existing methods that were compared. Also, the lead generation accuracy value for CLD-RBNL method was nearly 96.8% which is an explicit measure of a high degree of accuracy. Similarly, the processing time for CLD-RBNL method showed minimal time for obtaining dimensionality reduced tweets which is another robust gauge of high degree of accuracy. Thus, it can be concluded that the proposed CLD-RBNL method distinctly surpasses other contemporary methods in terms of accurate learning of student sentiment with respect to selecting a university globally.

REFERENCES

[1] Jamal Abdul Nasir Ansari, Nawab Ali Khan, "Exploring the Role of Social Media in Collaborative Learning the New Domain of Learning", *Smart Learning Environments*, Vol. 7, No. 9, pp. 1-6, 2020.

[2] B. Senthil Arasu, B.Jonath Backia Seelan and N. Thamaraiselvan, "A Machine Learning-Based Approach to Enhancing Social Media Marketing", *Computers and Electrical Engineering*, Vol. 86, pp. 1-9, 2020.

[3] Abdul Jabbara, Pervaiz Akhtarb and Samir Dania, "Real-Time Big Data Processing for Instantaneous Marketing Decisions: A Problematization Approach", *Industrial Marketing Management*, Vol. 90, pp. 558-569, 2019.

[4] Yogesh K. Dwivedi, Elvira Ismagilova, D. Laurie Hughes, Jamie Carlson and Raffaele Filieri, "Setting the Future of Digital and Social Media Marketing Research: Perspectives and Research Propositions", *International Journal of Information Management*, Vol. 59, pp. 1-37, 2020.

[5] Jin A. Choi and Kiho Lim, "Identifying Machine Learning Techniques for Classification of Target Advertising", *The Korean Institute of Communications and Information Sciences*, Vol. 6, No. 3, pp. 1-37, 2020.

[6] Arvind Rangaswamy, Nicole Moch, Claudio Felten, Gerrit van Bruggen, Jaap E. Wieringa and Jochen Wirtz, "The Role of Marketing in Digital Business Platforms", *Journal of Interactive Marketing*, Vol. 51, pp. 72-90, 2020.

[7] Ido Roll and Ruth Wylie, "Evolution and Revolution in Artificial Intelligence in Education", *International Journal of Artificial Intelligence in Education*, Vol. 26, pp. 582-599, 2016.

[8] Andrej Miklosik, Martin Kuchta, Nina Evans and Stefan Zak, "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing", *IEEE Access*, Vol. 7, pp. 85705-85718, 2019.

[9] Ming Hui Huang and Roland T. Rust, "A Strategic Framework for Artificial Intelligence in Marketing", *Journal of the Academy of Marketing Sciences*, Vol. 49, pp. 30-50, 2021.

[10] Ying Qian, Weiwei Liu, Jiangping Huang, "A Self-Attentive Convolutional Neural Networks for Emotion Classification

on User-Generated Contents", *IEEE Access*, Vol. 8, pp. 154198 - 154208, 2019.

[11] Hongli Wang and Jiangtao Ren, "A Self-Attentive Hierarchical Model for Jointly Improving Text Summarization and Sentiment Classification", *Proceedings of Asian Conference on Machine Learning*, pp. 630-645, 2018.

[12] Santiago Carbo-Valverde, Pedro Cuadros Solas, Francisco Rodrıguez Fernandez, "A Machine Learning Approach to the Digitalization of Bank Customers: Evidence from Random and Causal Forests", *PLOS ONE*, Vol. 15, No. 10, pp. 1-39, 2020.

[13] Marta Marco Gardoqui, Almudena Eizaguirre and Maria Garcian Feijoo, "The Impact of Service-Learning Methodology on Business Schools' Students Worldwide: A Systematic Literature Review", *PLOS ONE*, Vol. 15, No. 12, pp. 1-21, 2020.

[14] Iuliana Mihaela Lazar, Georgeta Panisoara Ion Ovidiu Panisoara, "Digital Technology Adoption Scale in the Blended Learning Context in Higher Education: Development, Validation and Testing of a Specific Tool", *PLOS ONE*, Vol. 15, No. 7, pp. 1-27, 2020.

[15] Cheng Ju Liu, Tien Shou Huang, Ping Tsan Ho, Jui Chan Huang and Ching-Tang Hsieh, "Machine Learning-Based E-Commerce Platform Repurchase Customer Prediction Model", *PLOS ONE*, Vol. 15, No. 12, pp. 1-15, 2020.

[16] Peng Wang and Zhengliang Xu, "A Novel Consumer Purchase Behavior Recognition Method using Ensemble Learning Algorithm", *Mathematical Problems in Engineering*, Vol. 2020, pp. 1-11, 2020.

[17] Rung Ching Chen, Christine Dewi, Su Wen Huang, and Rezzy Eko Caraka, "Selecting Critical Features for Data Classification based on Machine Learning Methods", *Journal of Big Data*, Vol. 7, No. 52, pp. 1-26,2020.

[18] Amber L. Stephenson, Alex Heckert and David B. Yerger, "College Choice and the University Brand: Exploring the Consumer Decision Framework", *Higher Education*, Vol. 71, pp. 489-503, 2016.

[19] Vikrant Kaushal and Nurmahmud Ali, "University Reputation, Brand Attachment and Brand Personality as Antecedents of Student Loyalty: A Study in Higher Education Context", *Corporate Reputation Review*, Vol. 23, pp. 254-266, 2019.

[20] Github, "Distance Learning", Available at: https://github.com/Bhasfe/distance_learning.