International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
***Article Received:** 28 May 2023 **Revised:** 22 July 2023 **Accepted:** 10 August 2023*

An Implementation of Computerized Valuation of Descriptive Answers: A Machine Learning Approach

Monika Dandotiya[1*], Dr. Devesh Kumar Bandil[2], Dr. Kriti Sankhla[3], Dr. Priyanka Yadav[4]
Monika Kumari[5].

[1]*Assistant Professor Poornima University, Jaipur, Rajasthan Email- dandotiyamonika@gmail.com
[2]Professor Poornima University, Jaipur, Rajasthan Email- mcagwalior@gmail.com
[3]Associate Professor Poornima University, Jaipur, Rajasthan E-mail- kriti.sankhla@gmail.com
[4]Assistant ProfessorPoornima University, Jaipur, Rajasthan E-mail- priyadav1990@gmail.com
[5]Assistant professor Poornima University, Jaipur E-mail- monika.kumari@poornima.edu.in

***Corresponding author:-** Monika Dandotiya
*Assistant Professor Poornima University,Jaipur,RajasthanEmail-dandotiyamonika@gmail.com

Abstract
Evaluation is an essential part of the education and it is carried out by the system of examinations. When evaluating a big number of pupils, a significant amount of physical labor is needed. In addition to being a labor-intensive process, manual valuation varies in quality depending on the examiner's disposition. Many of the aforementioned issues would be resolved in the modern world if this could be machine controlled. Thus, utilizing computers to assess responses is one way to find a solution. However, computers still have a difficult time evaluating descriptive responses. Therefore, it is crucial to investigate and implement techniques for the automated assessment of declarative responses.

This study proposes a machine learning strategy based on classifiers for evaluating descriptive responses. We conduct an experiment in our academic institution to construct the necessary

1. INTRODUCTION

Online examinations are growing exponentially. In recent scenario al-most all online examinations are objective type (contains multiple choice questions). Descriptive test usually evaluated by human evaluator. Replacing a human evaluator with Computers is beneficial in terms of saving time, money and improves accuracy. Therefore a systems is required that offers capabilities of auto-evaluation of descriptive answers. Classification methods Naive Bayes, j48 and logistic regression are explored for the evaluation of descriptive answers. Selection of these three classifiers is totally based on their text classification strength [1][2][3]. Brief description of each method is given here.

a) Naïve bayse: Built on the foundation of Bayes' theorem, the Naive Bayesian classifier utilizes every feature included in the data and treats each one as if it were independently significant and of equal importance. The posterior probability may be computed using the Bayes theorem. The naive bayes classifier makes the assumption that a predictor's (x) value has an independent impact on a given class (c) regardless of the values of other predictors. We refer to this presumption as class conditional independence.

$$K\left(\frac{c}{x}\right) = \frac{K\left(\frac{x}{c}\right)K(c)}{K(x)}$$

K(c|x) represents the class (target) posterior probability given a predictor (attribute). The class prior probability is denoted by K(c). K(x|c) represents the likelihood, or the probability of a predictor for a particular class. K(x) represents the predictor's previous probability.In [4]

b) J48: A decision tree is a predictive machine-learning model that uses different attribute values from the available data to determine the goal value (dependent variable) of a new sample. It must first build a decision tree using the attribute values of the available training data before it can

***Article Received:** 28 May 2023 **Revised:** 22 July 2023 **Accepted:** 10 August 2023*

categorize a new object. J48 uses labeled training data to construct decision trees. By dividing the data into smaller subsets, it makes use of the notion that any data attribute may be utilized to inform a choice. The information gain or difference in entropy that arises from dividing the data after selecting an attribute is examined by the J48 classifier. The selection is made using the highest normalized information gain of the characteristics, and the process then repeats on the smaller subsets.

c) Logistic regression:

The Multinomial Logistic Regression is a supervised learning algorithm which can be applied in numerous glitches including text classification. It is a regression model which generalizes logistic regression to classification problems whenever the output can take more than two possible values. Multinomial logistic regression is employed when the dependent variable in question is nominal

1.2 Related works:

Mohan et al. [5] Recommended Feature Clustering Process for Descriptive Type Inspection Assessment. Their method makes use of pre-specified clusters made up of parts of speech components such as adjectives, verbs, adverbs, pronouns, and nouns. SVM classifier is applied to assess test cases. The authors assert that this approach works well for essay-only tests and is ineffective for problems using formulas or mathematics.

Kaur et al. [6] suggests an algorithm for assessing replies that are descriptive in a single sentence. Based on a whole or partial string match, the similarity between student responses and the standard solution is measured. There are not enough instances in the work to support the system's validation.

C. Sunil Kumar et al.[7] provided a noteworthy study that evaluates descriptive responses using a bagging classifier. When evaluated across 5 datasets using ten fold cross validation with Decision Stump, Random Forests,Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Decision Trees, the authors claim that an average of 76% accuracy is attained. Nevertheless, there seem to be two problems with this work. Student essays make up the applied dataset; however, it would be better if author provided some specific questions and their responses to train classifiers, as the evaluation of essays differs greatly from that of descriptive answers. Moreover, an unseen test dataset should be provided to test system in place of a 10-fold validation.

Mamčenko et al. [8] suggested a descriptive model to use data mining techniques to find hidden trends in students' responses. The purpose of clustering techniques is to organize comparable things into groups. The results show how much time was spent overall, how long it took to provide an inaccurate response, and how long it took to provide a correct response. However, the suggested study has nothing to do with how descriptive replies are assessed.

2. Experimental details, methods, materials

2.1 Dataset collection

To form mandatory data set a descriptive test is performed consisting of 10 questions given to 34 students (PG level students) in two sessions. In the first session questions are directly asked to the students and after the first session 30 minute time is given to the students in which student can search the answers using internet. In second session same 10 questions are asked. First session is designed to expect wrong answers from students and second session is designed to get correct answers[9-12]. An online interface is designed to get answers using Google forms. We have received more than 650 answers. 389 answers are selected after the required preprocessing (preprocessing includes, redundancy removal, long answer removal, noise reduction, unification and normalization). Each answer is evaluated manually under the scale of 0 (wrong) to 2(best) by the human expert. After the evaluation, out of 389 answers 80% of it i.e. 311 answers are selected as training samples and 20% that is 78 answers are selected as testing samples (selection is totally Random). Following questions are asked in the test.

Q 1. How Does the Thermometer Work?
Q 2. What is Ladli Laxmi Yojana in M.P ?
Q.3. What do you mean by severe tropical cyclone? Who was the recent severe cyclone hit
the Indian coast.
Q 4. What is meant by Statutory Liquidity Ratio?
Q 5. What is Ebola fever?
Q 6. List some main features of Tejas fighter jet.
Q.7 Write Short note on "Mangalyaan mission”.
Q 8. What is Android ?
Q 9. Who is Magnus Carlsen ?
Q 10. What do you mean by "break-even point (BEP)”?
Table 1 shows the summary of training and testing datasets.

Table 1: Question wise Training and Test samples summary

Questions	Training Samples	Test Samples
1	28	7
2	29	7
3	41	10
4	27	7
5	32	8
6	38	9
7	35	9
8	34	9
9	24	6
10	23	6
Total	**311**	**78**

On an average each question is trained by 31 respective answers and about 8 unseen answers for each question are given to the classifier as testing samples. Classifier evaluation consequences are compared to manual evaluation results with the objective that classifier evaluation and manual evaluation will products the comparable outcomes. Dataset is available and can be downloaded from Institute website link. http://www.vns.ac.in/vnsitpdf/dataset.pdf .

2.2 Methods and Experiment:

2.2.1 Experiment Set-up: The PC used for the experiment is running Ubuntu LTS Linux v14.04 with a 3 GB RAM and a 1.3 GHz Intel i3 CPU. The university of Waikato's Weka V 3.6.11 machine learning workbench is used to classify descriptive responses.

2.2.2 Experiment steps: Figure 1 show the experimental steps performed.

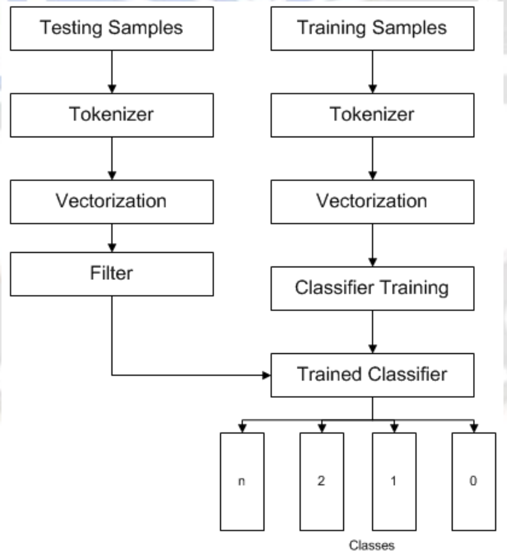

Figure 1: Experimental steps

2.2.2.1 Training and Test Samples: As was previously noted, 78 responses were chosen as test samples and 311 replies as training samples. Classifiers are independently trained and evaluated using sets of responses to each unique question (that is, each question's collection of answers is trained separately). Ten training and testing files in ARFF format are therefore produced.

2.2.2.2 Tokenizer: It separates the response string that is entered into a stream of phrases, or tokens. The string is split up into terms whenever it comes across punctuation or whitespace using a basic tokenizer.

2.2.2.3 Vectorization: After being processed, tokens that were retrieved from the tokenizer are converted into a column vector. The structure below represents each vector row.

Vec = [Vec$_1$, Vec$_2$, Vec$_3$….Vec$_n$, Class]

Example: The tokens that were identified in the previous example are converted into a column vector. The marks (scale 0-2) provided by the manual evaluator are represented by the Classattribute in this instance.

$$\begin{array}{lcccccccc} & \text{Great} & \text{work} & \text{excellent} & \text{worst} & \text{ever} & \text{no} & \text{comment} & \text{Class} \\ \text{Vec1} = [& 1, & 1, & 0, & 0, & 0, & 0, & 0, & 2] \\ \text{Vec2} = [& 0, & 1, & 1, & 0, & 0, & 0, & 0, & 2] \\ \text{Vec3} = [& 0, & 1, & 0, & 1, & 1, & 0, & 0, & 0] \\ \text{Vec4} = [& 0, & 0, & 0, & 0, & 0, & 1, & 1, & 1] \end{array}$$

2.2.2.4 Training: Naive Bayes, j48 and logistic regression classifiers are trained on the basis of known contents. 311 manually evaluated answers are given to the classifiers for training.

2.2.2.5 Filter: Both the training and test files must have the same name, type, and number of attributes (column vectors) for the classification to be effective. Nonetheless, the training and test samples in this work have uneven column vectors. Therefore, in order to achieve vector dimension compatibility, test samples are preprocessed by the arbitrary filter. This is necessary to make them compatible. The filter's structure is entirely derived from the training set, and test cases are processed by the filter without undergoing any structural modifications.

2.2.2.6 Classification: As previously indicated, three algorithms are used to assess test samples on a 0–2 scale and get a result. It is anticipated that trained classifiers would assess test samples in the same way that humans do.

3. Observations, Results and Discussions:

The data set is useful to the 3 classifiers and pragmatic result is dignified by following factors. [9]

Table 2: Matrix (Confusion):

		Detected	
		Positive	**Negative**
Actual	**Positive**	A: True Positive	B: False Negative
	Negative	C: False Positive	D: True Negative

Recall = A / A + B
FP Rate = C / C + D
Precision = A / A + C

$$F = 2.\frac{Precision * Recall}{Precision + Recall}$$

ROC Curve Plotting excellent, good, and useless test ROC (Receiver Operating Characteristic) curves on the same graph. How thoroughly the test isolates the group being examined determines how accurate the test is. A test with an area of 1 is considered ideal, whereas one with an area of 0.5 is considered useless.

Table 3: RESULTS OF NAIVE BAYES CLASSIFIER FOR INDIVIDUAL QUESTIONS

Dataset	Correctly Evacuated	incorrectly Evacuated		FP Rate	Precision	Recall	F-Measure	ROC Area	kappa Statics	Classification %
1	6	1		0.357	0.738	0.857	0.972	0.905	0.6111	85.7143
2	5	2		0.214	0.529	0.714	0.603	1	0.4815	71.4286
3	6	4		0.193	0.483	0.6	0.535	0.708	0.3443	60
4	5	2		0.048	0.905	0.714	0.757	0.976	0.3636	71.4286
5	7	1		0.075	0.906	0.875	0.871	0.975	0.8095	87.5
6	5	4		0.264	0.556	0.556	0.556	0.854	0.1429	55.5556
7	5	4		0.472	0.347	0.556	0.427	0.699	0.0526	55.5556
8	7	2		0.178	0.778	0.778	0.778	0.8	0.625	77.7778
9	6	0		0	1	1	1	1	1	100
10	6	0		0	1	1	1	1	1	100
Total.	**58**	**20**	**Avg.**	**0.76**	**0.1801**	**0.724**	**0.7499**	**0.8917**	**0.54305**	**76.4961**

Table 4: RESULTS OF J48 CLASSIFIER FOR INDIVIDUAL QUESTIONS

Dataset	Correctly Evacuated	incorrectly Evacuated		FP Rate	Precision	Recall	F-Measure	ROC Area	Kappa Static	Classification%
1	6	1		0.024	0.929	0.857	0.873	0.952	0.7308	85.7143
2	6	1		0.19	0.743	0.857	0.794	0.857	0.72	85.7143
3	6	4		0.193	0.483	0.6	0.535	0.768	0.3443	60
4	5	2		0.048	0.905	0.714	0.757	0.833	0.3636	71.4286
5	7	1		0.075	0.906	0.875	0.871	0.863	0.8095	87.5
6	7	2		0.444	0.722	0.778	0.72	0.674	0.4194	77.7778
7	6	3		0.333	0.563	0.667	0.596	0.667	0.325	66.6667
8	6	3		0.267	0.674	0.667	0.661	0.7	0.4375	66.6667
9	6	0		0	1	1	1	1	1	100
10	5	1		0.033	0.917	0.833	0.852	0.9	0.5714	83.3333
Total	**60**	**18**	**Avg.**	**0.1607**	**0.7842**	**0.7848**	**0.7659**	**0.8214**	**0.57215**	**78.48017**

Table 5: RESULTS OF LOGISTIC REGRESSION (LR) CLASSIFIER FOR INDIVIDUAL QUESTIONS

Dataset	Correctly Evacuated	incorrectly Evacuated		FP Rate	Precision	Recall	F-Measure	ROC Area	Kappa Static	Classification%
1	5	2		0.048	0.905	0.714	0.75	0.905	0.5333	71.4286

***Article Received:** 28 May 2023 **Revised:** 22 July 2023 **Accepted:** 10 August 2023*

2	5	2		0.048	0.619	0.714	0.643	0.971	0.5333	71.4286
3	7	3		0.168	0.667	0.7	0.675	0.88	0.5	70
4	4	3		0.024	0.929	0.571	0.667	1	0.2759	57.1429
5	7	1		0.075	0.905	0.875	0.871	0.913	0.8095	87.5
6	6	3		0.25	0.778	0.667	0.704	0.872	0.3415	66.4476
7	6	3		0.333	0.563	0.667	0.596	0.565	0.325	66.6667
8	7	2		0.178	0.778	0.778	0.778	0.867	0.625	77.7778
9	6	0		0	1	1	1	1	1	100
10	2	4		0.133	0.867	0.333	0.333	0.983	0.0769	33.3333
Total	**55**	**23**	**Avg.**	**0.1257**	**0.8011**	**0.7019**	**0.7017**	**0.8956**	**0.50204**	**70.17255**

Table 6: Average of Classification %

Methods	Classification%
NAIVE BAYES	76.49605
J48	78.48017
LR	70.17255
Avg.	**75.04959**

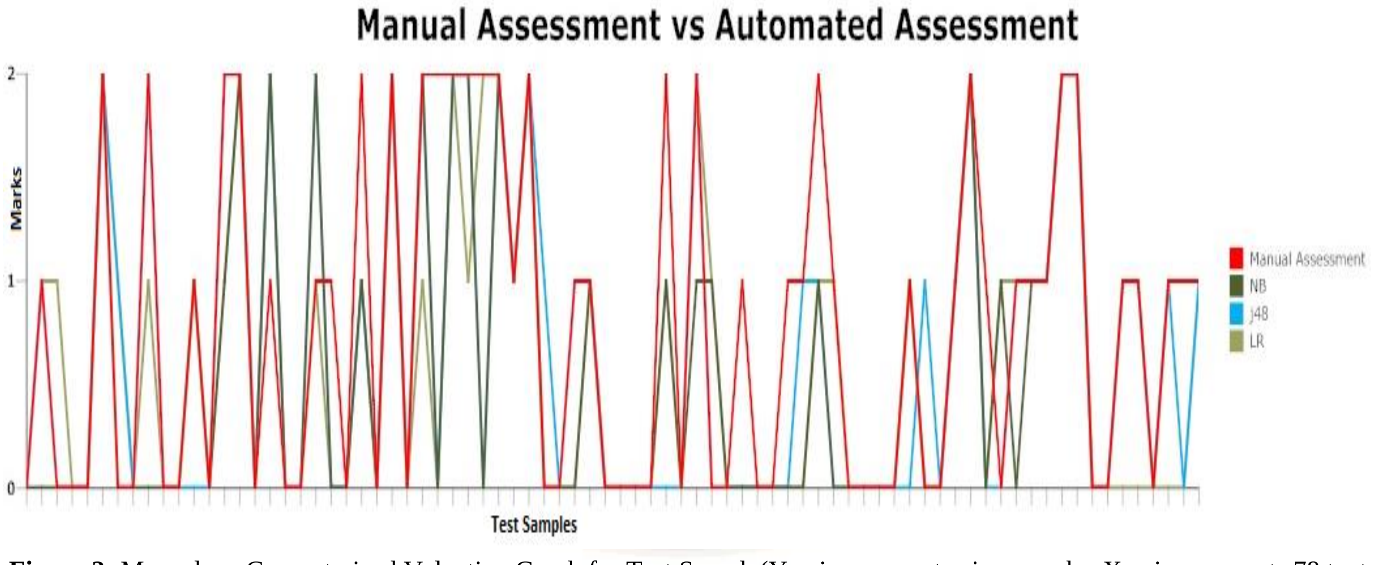

Figure 2: Manual vs. Computerized Valuation Graph for Test Sample(Y axis represents given marks, X axis represents 78 text samples)

It is observed that J48 is producing the best evaluation among the all used methods. Automated evaluation is producing on an average 75.049% correct evaluation and producing less than 25% incorrect evaluation compared to the manual evaluation results.

It is also observed that despite of high classification of Naive Bayes, logistic regression is producing less variance (Here variance is calculated by subtracting mean of automated evaluated marks with manual evaluated marks mean), as mentioned bellow:

Article Received: *28 May 2023* ***Revised:*** *22 July 2023* ***Accepted:*** *10 August 2023*

Table 7: Mean and Variance of Classification Methods

	Manual Evaluation	Naive Bayes	J48	Logistic Regression
Mean	0.7820	0.5769	0.6410	0.6025
Variance		0.2051	0.1410	0.1794

The variance of Naïve Bays is high because in some cases it evaluating the best (2 marks) answer to the wrong (0 marks) answer with overlooking the medium (i. e. 1 marks). Same as, despite of high classification of J48, logistic regression is producing high precision rate. It is also observed that ROC area of logistic regression is higher than the remaining two classifiers. This reflects better ability of correct valuation.

4. Conclusion and Future Work

Experimental results show that the proposed method is useful for the automated valuation of descriptive answers. However still research focus is required in this area, where natural language processing and semantic analysis can be explored. To the best of my knowledge there is no significant research has been taken place and still no such effective method introduced that practically replaces manual evaluation of descriptive answers with automated evaluation. The suggested method, however, may also be used as "Computer assisted Manual valuation of descriptive answers," in which each response is first assessed by computers before manual valuation begins. If the difference between the marks assigned manually and those automatically evaluated exceeds a predetermined threshold, an alert is sent to the human assessor, requesting that they revise and reevaluate the response.

References:

1. .L. Ting, W.H. Ip, Albert H.C. Tsang "Is Naïve Bayes a Good Classifier for Document Classification", International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011, PP: 37-46.
2. Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal Of Computer Science And Applications, Vol. 6, No.2, Apr 2013, PP: 256-261
3. Vasilis Vryniotis, "Machine Learning Tutorial: The Multinomial Logistic Regression" weblink:http://blog.datumbox.com/machine-learning-tutorial-the-multinomial-logistic-regression-softmax-regression/.
4. Naive Bayesian, web. link: http://www.saedsayad.com/naive_bayesian.htm.
5. A Krishna Mohan, MHM Krishna Prasad, "A Novel Feature Clustering Algorithm for Evaluation of Descriptive Type Examination", International Journal of Computer Applications, Volume 98– No.9, July 2014, PP: 35-41.
6. Amarjeet Kaur, M Sasikumar, Shikha Nema, Sanjay Pawar, "Algorithm for Automatic Evaluation of Single Sentence Descriptive Answer", International Journal of Inventive Engineering and Sciences (IJIES), Volume-1, Issue-9, August 2013, PP:6-9
7. C. Sunil Kumar, R. J. Rama Sree," An Attempt to Improve Classification Accuracy through Implementation of Bootstrap Aggregation with Sequential Minimal Optimization during Automated Evaluation of Descriptive Answers", Indian Journal of Science and Technology, Vol 7(9) , September 2014,PP: 1370-1375.
8. J. Mamčenko, I. Šileikienė, J. Lieponienė and R. Kulvietienė, "Evaluating the Data of an e-Examination System using a Descriptive Model in Order to Identify hidden Patterns in Students Answers", The Online Journal on Computer Science and Information Technology (OJCSIT), Vol. (1) – No. (2), PP: 45-49.
9. Performance Measures, http://www.seas.gwu.edu/~bell/csci243/lectures/performance.pdf
10. Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Burlington.
11. Bryson, A.E. and Ho, Y.C. (1969) Applied Optimal Control: Optimization, Estimation, and Control. Blaisdell Publishing, Waltham.
12. Russell, S.J. and Norvig, P. (2010) Introduction. In: Russell, S. and Norvig, P., Eds., Artificial Intelligence: A Modern Approach, 3rd Edition, Prentice Hall, Upper Saddle River, 1-33.
13. Harrington, P. (2012) Splitting Datasets One Feature at a Time: Decision Trees. In: Harrington, P., Ed., Machine Learning in Action, Manning Publications, Shelter Island, 37-60
14. LeMoyne, R. and Mastroianni, T. (2021) Global Algorithm Development. In: LeMoyne, R. and Mastroianni, T., Eds., Applied Software Development

Article Received: *28 May 2023* ***Revised:*** *22 July 2023* ***Accepted:*** *10 August 2023*

with Python & Machine Learning by Wearable & Wireless Systems for Movement Disorder Treatment via Deep Brain Stimulation, World Scientific Publishing, Singapore, 63-86

15. LeMoyne, R. and Mastroianni, T. (2021) Automation of Feature Set Extraction Using Python. In: LeMoyne, R. and Mastroianni, T., Eds., Applied Software Development with Python & Machine Learning by Wearable & Wireless Systems for Movement Disorder Treatment via Deep Brain Stimulation, World Scientific Publishing, Singapore, 107-135.