Evaluation of Essay using Machine Learning Techniques

[1]Bane Raman Raghunath, [2]Khandke Mahesh A.

[1]Associate Professor, [2]Assistant Professor
[1]Computer Engineering Department,
[1]Gharda Institute of Technologyr, Lavel,India.

***Abstract*: Essays are the prominent evaluation parameter of assessing the academic excellence along with linking the different ideas with the ability to recall. The process of assessing essay is notably time consuming when done manually. Manual evaluation is expensive as it takes significant amount of evaluator's time. Automated evaluation if proved effective will not only reduce the time for evaluation but makes the score realistic compared with human scores. The automated evaluation process could be useful for both educators and learners as it brings the iterative improvements in students' writings. The paper describes an automated essay evaluation system using Machine Learning Techniques such as Linear Regression, Support Vector Regression and Random Forest.**

In addition to preprocessing techniques like filling in null values, selecting valid features, normalization we applied cleaning process by removing unnecessary symbols, punctuations and stop-words from the essays in the training set. We have also added extra features like number of sentences, number of words, average word length, type of word using POS tagging, number of spelling mistakes in an essay, number of domain words in an essay, etc. We have implemented our models using 'sklearn' machine learning library. We got satisfactory results with Linear Regression, Support Vector Regression and Random Forest algorithm with Mean Squared Error values 2.88, 1.67 and 0.867 respectively.

***Index Terms*: machine learning, linear regression, support vector regression, random forest.**

I. INTRODUCTION

Essay type of exam is one of the most essential testing activities at all educational levels. It is an effective tool for assessing academic achievement, integration of ideas and ability to recall. Essays are useful for the evaluation of the learning outcomes of students since it gives student an opportunity to demonstrate their range of skills and knowledge that includes higher- order thinking. Educators choose essay questions over other forms of assessment because essay items challenge students to create a response rather than to simply select a response. Some educators use them because essays have the potential to reveal students' abilities to reason, create, analyze, synthesize, and evaluate. In short, essay items are used for the advantages they offer. Though essay questions are advantageous to student learning and assessment various challenges for the part of the teachers were noticeable. Manual grading of essays takes up a significant amount of teacher's valuable time since grading essay type exam is time consuming and tedious thing to do especially for a large population of students. Furthermore, the perception of subjectivity of the grading process can be considered since the subjective nature of essay assessment may lead to variation of different results.

When large numbers of essays are submitted at once, teachers find them self- bogged down in theirattempt to provide consistent evaluations and high quality feedback to students within a short time and takes usually a matter of days or weeks. Educational administrators are also concerned with quality and timely feedback, but in addition must manage the cost of doing this work. Automated Essay Grading is a tool that enables teacher to save their time and effort, provide more objective evaluations and refrain from being subjective. Its main goal is to predict the grade of students automatically by means of various features. In this paper, automated evaluation of an essay is introduced based on different algorithms that includes machine learning which is considered as its core component.

Our system is proposed to predict the grade of essays automatically with the help of machine learning techniques to predict the essay score. This will help to do the work fast and lower down the burden of teachers as they have to evaluate huge number of essays at a time. Students will able to get the results fast and also able to recognize their mistakes .There will be accurate grading of the essays. This system will also compare the grades of submitted essays by the students with the previously human graded score. Feedback system introduced will help to communicate well between teachers and students.

Clearly an automated system would be a highly desirable addition to the educational tool-Kit, particularly if it can provide less costly and more effective outcome. In this paper we present a description and evaluation of automated essay grading system. As we evaluate the results of our trial, a research and development direction is indicated which we believe will results in improvement over existing systems.

II. LITERATURE SURVEY

An automated evaluation of essays using computer technology has been a piece of attraction to the researchers since early sixties though it became very much useful in late 90s. IBM, the giant in computer technology focused to score essays automatically in 1938. They manufactured IBM 805 that can score an essay selecting some responses. However the earliest noticeable piece of work found in this field is the Project Essay Grade (PEG) by Ellis Page in 1960 [3]. Page believed that the information of an essay can be hypothetically divided so as the style of the essay. That's why he developed the PEG to score the essays based on the style of the candidate essays. In 1968 he published that there are two kinds of variables that are present in essays. He defined them to be proxes and trains. Proxes are the variables or indicators that a computer can recognize. Trins are on the other hand only a human rater can understand. So basically he worked with the proxes that is the surface variables like the word count, length of the essay and 30 other indicators and then weighed some regression equation to predict the score. In 1998 Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock, & Wolff [5] developed some model that is named the E-Rater which used multiple linear regression model to score essays. The second version of this model is still in use by ETS where the researchers have narrowed down the number of indicators to be used to score the essays. As both E-Rater and PEG works based on surface features that's why it got criticized by researchers that's why in 2007 Ben-Simon and Bennett worked with semantic analysis of essays to predict the scores. It mainly focused on the meaning of words sentences or paragraphs. Prior to Bennet in 1999 Foltz,Landauer and Laham [13] also worked with semantic analysis based automatic scoring which mainly focused on essay content rather than the style. In this case they developed the model that can score the candid essay based on the similarity of the essays based on the system was trained. Whenever some unique essays appeared as a candid essay this system used to mark it as anomalous essay and that essay was manually scored by a human rater. Latent semantic analysis based tools were proved much more efficient in scoring the essays than the surface indicator based system. Another approach towards automatically scoring the essays emerged later on that is based upon the text categorization technology. Williamson in 2001 used Bayesian classification techniques in text categorization and came up with an algorithm to predict the score of the essay. Larkey in 1998 developed a tool that can predict the score of an essay using text categorization with 12 surface indicators and regression model.

Limitations of existing system:

The problems of the essay writing activities are time-consuming, concerns in producing immediate result and/or feedback from teachers to students, and the teachers tend to be subjective in grading the essay activities.

Grading System in India

Automatic essay grading [AEG] has now turned into reality. Machine learning systems provide a lot to the educational community when grades have to face different kind of difficulties while rating student's writings. Analyzing student's essays in abundance within given time limit along with field back is a challenging task. But with changing times the human return [not hand written] essays are easy to evaluate with the help of automatic essay grading system. The student's writings are graded by both, human and automatic essay grading system.

III. PROBLEM STATEMENT AND OBJECTIVES

Automatic evaluation of essays would be most beneficial for today's generation. As we all know very well that an essay writing helps us to assess the creative writing ability on the parameters such as ability to recall, organize, style of writing, way of thinking etc. The proposed system would help us to detect all the things in a well-mannered way in a very short period of time as a person checking it manually will consume a lot of time and would also find it very difficult to verify each and everything thoroughly. Usually students find confusions in qualitative scoring. The proposed automated evaluation would evaluate it and give a qualitative scoring. It would consider multiple attributes of the essay to grade it.

Grammatical corrections, content of essay, organization of essay, Linear Regression to learn theessay by itself evaluation biases are the main aspects of verifying an essay which automatic essay grading system can make it in most easy and in quick manner with perfection. The problem of bribe which is very common nowadays will also be avoided.

These things will totally change the view of grading system with perfection.

IV. PROPOSED SYSTEM

We have used three machine learning algorithms for predicting the evaluation score of an essay. The techniques like linear regression with polynomial basis function, support vector regression and random forest. We applied these techniques on the data without preprocessing steps and then on the data obtained through preprocessing steps. The pre-processing steps are explained in the following paragraphs. The proposed system is represented as a block diagram in Fig. 1.

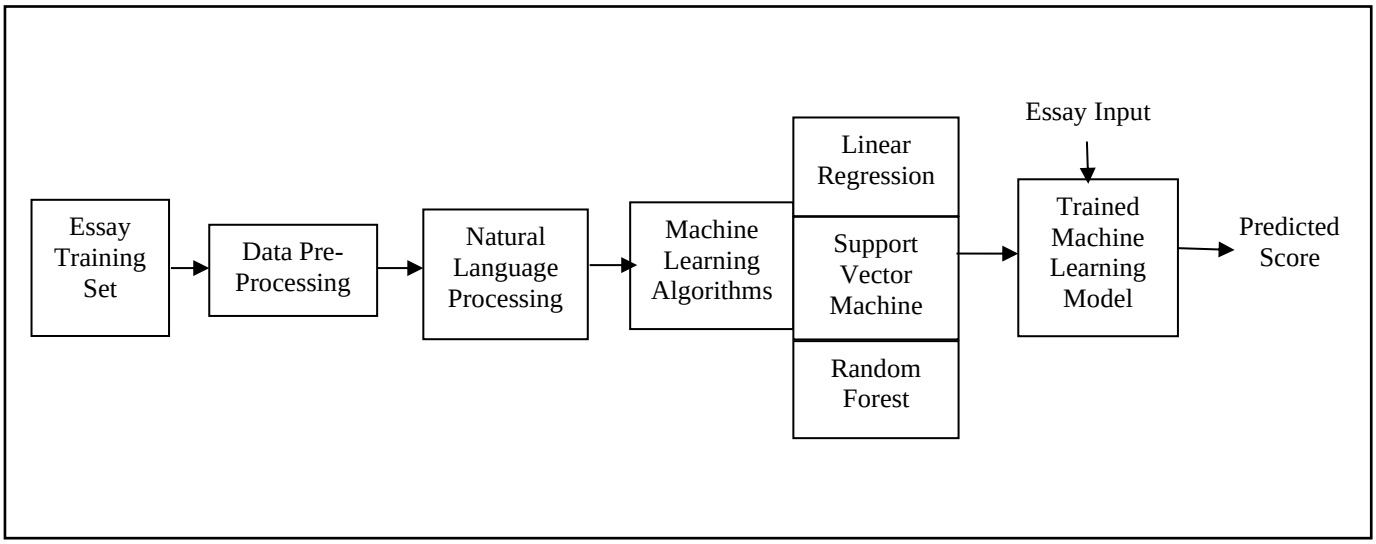

Fig 1: Proposed System Architecture

The entire process is divided into three different modules as follows:

1. Data Preprocessing

We started with some standard preprocessing steps like filling in null values. We applied correlation method to find out important features from the entire data set. Next we plotted a graph and measured the skewness of our data and applied normalization techniques to reduce this skewness. The next major step we applied is the cleaning the essays to make our training process easier for getting a better accuracy. We identified unnecessary symbols, stop words and punctuations from our essays and removed them. To increase our accuracy even more we even planned to add some extra features like the number of sentences , number of words, number of characters, average word length etc. Moreover, we even worked on techniques like getting the noun, verb, adjective and adverb counts using parts of speech tagging as well as getting the total misspellings in an essay by comparison with a corpus. We applied various machine learning algorithms on this data as explained in the next section.

We have taken the input data from Kaggle.com. We are given ~13000 essays written by school students of Grade 7, 8 and 10. These essays are divided into 8 sets - each set of essays from a different context - to ensure variability of the domain. Each set of essays was generated from a single prompt. Along with the ASCII text of each essay, we also have scores given to each essay by two human evaluators and a combined resolved score.

2. Machine Learning

One additional step is carried out to make our data ready for machine learning algorithms. Machine learning algorithms cannot be applied directly on sentences or words. It can only be used upon numeric data. The dataset contains essays that need to be converted into a numeric form first in order to train it. We used CountVectorizer to convert to numeric form. Now the CountVectorizer works by tokenizing a collection of text documents and returning an encoded vector with a length of the entire vocabulary along with an integer count for the number of times each word appeared in the document. After this step our data is finally ready for predictive modelling.

We split this data into two sets: training set and testing set, as follows:

Essay set	Total number of Essays	Number of Essays used for training	Number of Essays used for testing
1	1783	1200	583
2	1800	1200	600
3	1726	1200	526
4	1772	1200	572
5	1805	1200	605
6	1800	1200	600
7	1569	1200	369
8	723	500	223
TOTAL	12978	8900	4078

Table 1: Data Set Description

Once we have the training and testing data, we can extract features from each ofthe document and train our model. These are explained in subsequent sections.

Initially we applied machine learning algorithms like linear regression, Support Vector Regression (SVR) and Random Forest on the dataset without addition of features that were mentioned in the preprocessing section before. Our results were not really satisfactory as our mean squared error was quite high for all the above algorithms. After this initial evaluation, we added the extra features, applied CountVectorizer again on this modified dataset and applied the same three algorithms. There was a great improvement in the performance of all three algorithms especially Random forest for which the mean squared error reduced drastically.

3. Features Extraction

Machine Learning task includes feature extractions which is very important and the same with this proposal. To build effective essay scoring algorithm, we tried model attributes like language fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information etc.

At present, our model is using the following set of features extracted from the ASCII text of the essays:

i) Sentence count and Word count: These are very basic features of any text document and do influence the scoring of the document as well. We have used "NLTK" library ((http://nltk.org/) [8] in python to extract these features. Using this library, we extracted the term-document matrix for our training corpus (8900 documents). This library also provides a list of 276 common stop words in English language. Now, since these stop words are of not much importance, we skipped them while calculating the word count of each document from the term- document matrix. To get the sentence count, we simply split the document using '.' and thus, count the number of segments obtained.

ii) Part of Speech tagging (POS Tags): The number of words in various syntactic classes like nouns, adverbs, verbs, adjectives etc are another crucial set of features for evaluating an essay or any writing. These features are crucial for evaluating the quality of content in the essay. To get the counts of words in each POS (part-of-speech) class, we use the NLTK library in python (http://nltk.org/). This library gives us the POS tag for each word in an essay, and thus, we extract the number of nouns, adverbs, adjectives and verbs.

iii) Spelling Mistakes: Spelling mistakes is an important parameter while evaluating an essay. So, number of spelling mistakes in an essay is also a feature for our model. To get this number, we use the spell checker provided in python by library named 'enchant' (https://pypi.python.org/pypi/pyenchant) [11].

iv) Domain Information Content: It is perhaps the most crucial feature in our model as it tries to captures the semantics and information content of an essay. To get this feature, we first took the best essay from each set (highest scored essay) then, we pulled out nouns from that essay. These serve as keywords for the particular domain. Then, we fire these words into 'Wordnet' [9] (http://wordnet.princeton.edu/) and take out their synonyms. In this way, for each set, we get a bunch of words, most relevant to its domain. Then, we count the number of domain words in the given essay.

RESULTS

Once we are done with the task of extracting features, we use linear regression with polynomial basis functions to find out the scores of each essay. Firstly, we trained our model using all the 8900 training documents at once. Now, since, all the 8 sets had different scale for marking, we first normalized all the scores to interval [0-10].

For training, we varied the degree (d) of the polynomial basis function from d = 1 to d = 10 and looked at the testing error. The testing error in our case is the average of absolute difference between the actual and predicted scores. As the case with training, the testing is also done on essays from all the 8 sets at once.

Following plot in fig. 2 shows the behavior of test error as we vary the value of the parameter d. Actual error values are shown in Table 2.

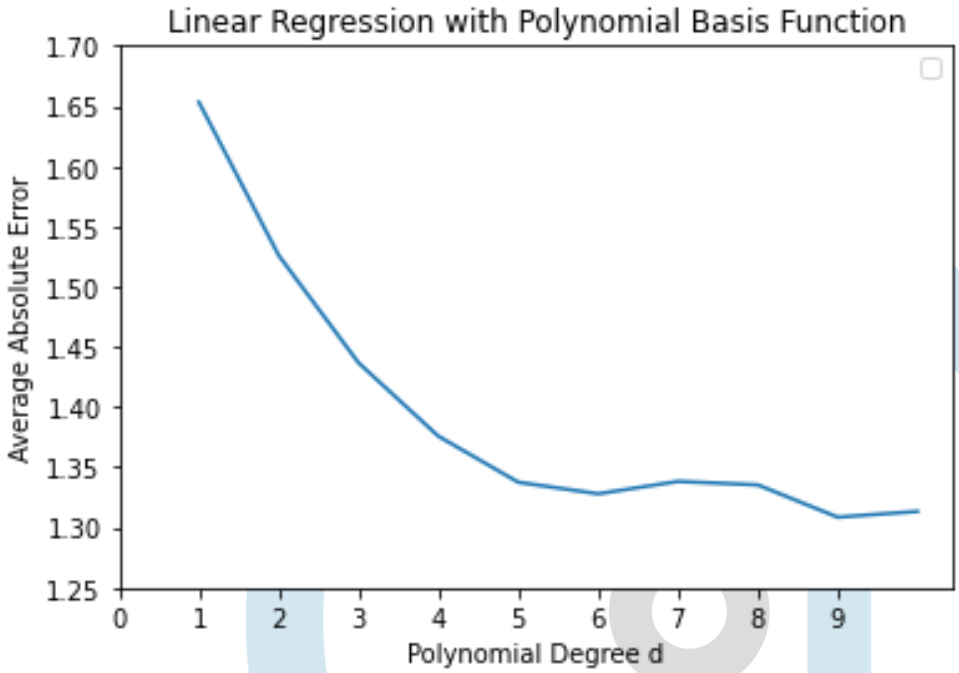

Fig. 2: Average absolute errors with different degrees of polynomials

These are the values of the test error with varying d:

Value of d	Test Error
1	1.6532
2	1.5259
3	1.4367
4	1.3758
5	1.3376
6	1.3280
7	1.3383
8	1.3354
9	1.3086
10	1.3134

Table 2: Test Errors with respect to different values of degrees of polynomials

As evident from the plot and the figures, given the current setting, the best results are obtained by fitting a polynomial of degree 9 on the data. The average discrepancy between the scores generated by our model and the actual human scores is about 1.3086 marks on a scale of [0-10].

In the above setting, we trained our model on all 8 sets in one go and then test model on essays from all sets. Thus, we could not really model the domain information well.

So, to take into account that all the 8 sets are very different, we train our model using the training data from one particular set and then test for that set. And this process is done on all of the 8 sets. So, unlike the previous case, no normalization ofgiven scores is

required. For each set, we calculate relative absolute error on the test set. The results obtained are as follows shown in Fig 3. Actual error and standard deviation values are shown in Table 3.

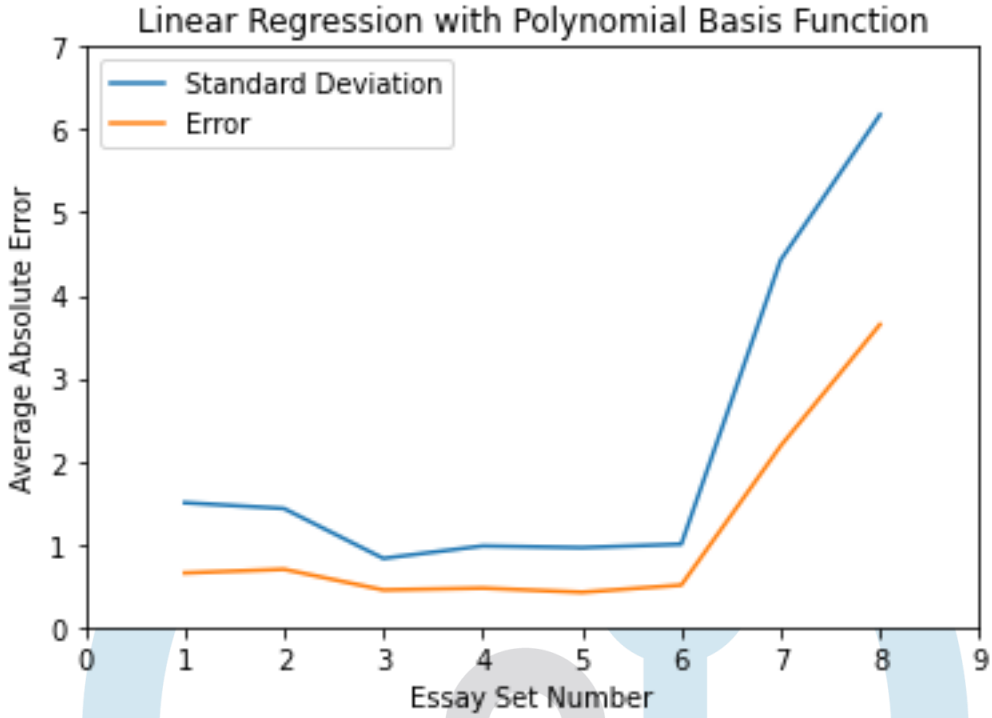

Fig.3: Average Absolute Errors with linear regression for different essay sets

This figure shows the average absolute error for the set and the standard deviationof the given scores in that test set. This was plotted to ensure that the error of the proposed model is less than the standard deviation in the given data. The followingtable gives the values of the above plot:

Essay Set	Scale of Marking	Standard Deviation	Average Absolute Error
1	2-12	1.5826	0.6423
2	2-10	1.4183	0.6856
3	0-3	0.8123	0.4299
4	0-3	0.9634	0.44598
5	0-4	0.9467	0.4043
6	0-4	0.9859	0.4989
7	0-30	4.4154	2.1734
8	0-60	6.164	3.6368

Table 3: Standard Deviation and average absolute errors for different essay sets

The relative trend of the error (and standard deviation) values is due to differentscale of marking across different data sets.

As evident from the error values, the proposed model seems to be reasonably good and reliable.

Instead of linear regression with polynomial basis function, we also tried usingSupport vector regression and Random Forest and got the following error (and standard deviation) plot shown in Fig. 4 and 5 respectively.

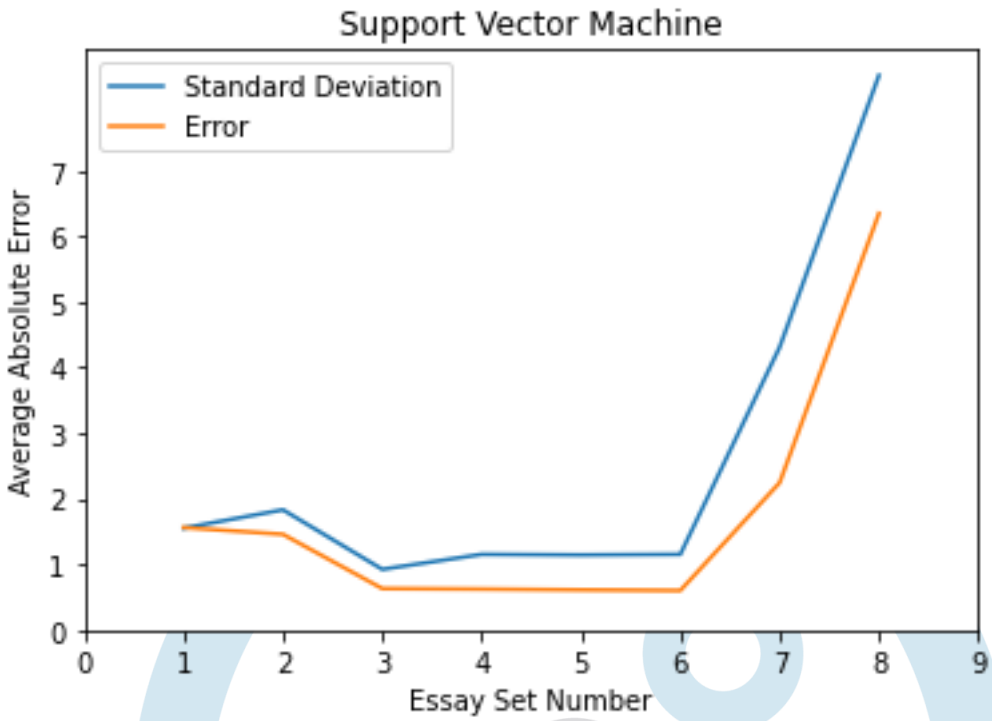

Fig.4: Average Absolute Errors with support vector regression for different essay sets

Clearly, this alternative approach doesn't work and so, we continued with linearregression with polynomial basis function.

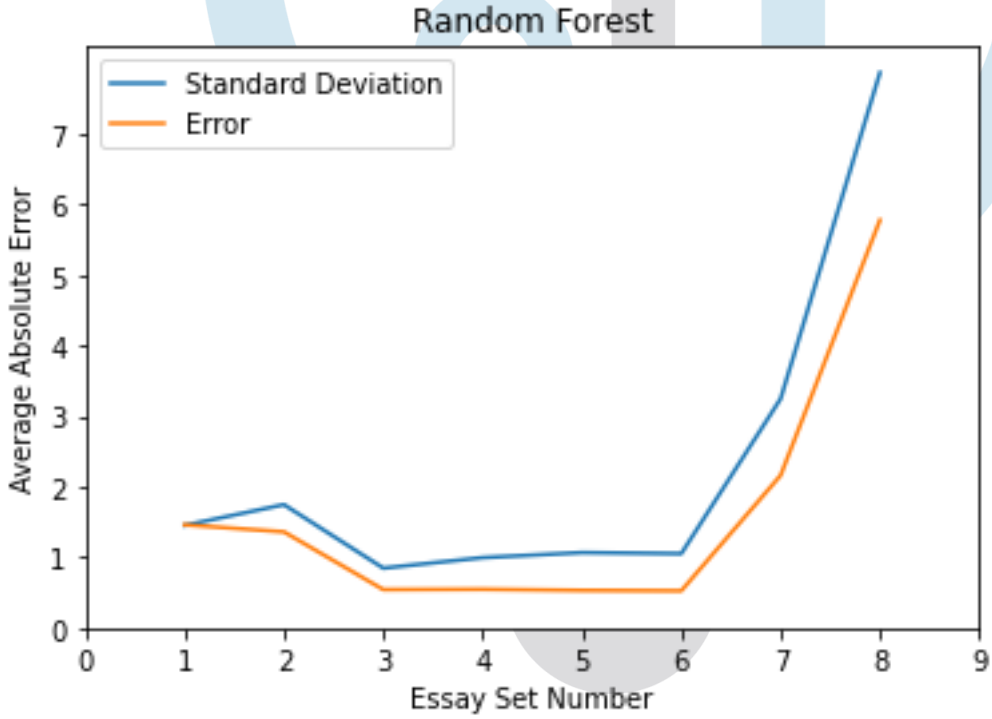

Fig.5: Average Absolute Errors with random forest for different essay sets

The mean squared error with linear regression, support vector regression and random forest before preprocessing steps is 23040.50, 3.68 and 3.68 reactively. But after applying the pre-processing steps we received adequate accuracy with mean squared values 2.88, 1.67 and 0.867 respectively.

CONCLUSION AND FUTURE SCOPE:

Automatic essay grading is a very useful machine learning application. It has been studied quite a number of times, using various techniques like latent semantic analysis etc. The current approach tries to model the language features like fluency,grammatical correctness, domain information content of the essays, and tries to fitthe best polynomial in the feature space using linear regression with polynomial basis functions.

The results obtained seem quite encouraging. We achieve average absolute error tobe significantly less than the standard deviation of the human scores. Across all domains, the proposed approach appears to work very well.

The future scope of the given problem can extend in various dimensions. One such area is to search and model good semantic and syntactic features. For this, various semantic parsers etc. can be used. Other area of focus can be to come up with a better tool than linear regression with polynomial basis functions like neural networks etc.

REFERENCES

[1] T. Miller, "Essay assessment with latent semantic analysis," Dept. of Computer Science, University of Toronto, ON M5S 3G4, Canada 2002.

[2] Alghamdi, M., M. Alkanhal, M. Al-Badrashiny, A. Al-Qabbany, A. Areshey, and A. Alharbi. (2014). "A hybrid automatic scoring system for Arabic essays," AI Communications, 27(2):103-111.

[3] Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates Publishers.

[4] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," J. of Information Technology Education, vol. 2, pp. 319-330, 2003.

[5] Jill Burstein, Susanne Wolff, Chi Lu, "Using Leexical Semantic Techniques to Clasify Free Responses", Breadth and Depth of Semantic Lexicons, Vol 10, pp 227-244.

[6] S. Drolia, P. Agarwal, S. Rupani and A. Singh, "Automated Essay Rater using Natural Language Processing," International Journal of Computer Applications (0975-8887). Vol. 163. Issue 10. April 2017.

[7] M.Young,TheTechnicalWriter'sHandbook.MillValley,CA:UniversityScience,1989.

[8] NLTK (http://nltk.org/)

[9] Wordnet (http://www.wordnet.princeton.edu)

[10] P. Reddy and G. Jambagi Automated Essay Grading", University of Illinois, Chicago, USA, 2003.

[11] enchant (https://pypi.python.org/pypi/pyenchant)

[12] textmining (https://pypi.python.org/pypi/textmining)

[13] PW Foltz, D Laham, TK Landauer Automated essay scoring: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning.