ICIC Express Letters
Volume **16**, Number **1**, January **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **83–90**

A COMPARATIVE STUDY OF SUPERVISED LEARNING ALGORITHMS FOR ESSAY SCORING PREDICTION

Andrian Syah Putra Leong[1], Niko Pratama[1] and Derwin Suhartono[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ andrian.leong; niko.widjaja }@binus.ac.id; dsuhartono@binus.edu

Received March 2021; accepted June 2021

Abstract. *To score a written essay, the traditional method has always been the utilization of a human examiner to carefully read and assess the overall quality based on the essay's writing. Although the scores given by the human examiner are accurate and reasonable, it is still a laborious and manual task, requiring time and effort to carefully read and score the essay. Thus, automated essay scoring systems are made, which utilizes computer agents to take over the role of the human examiner. These systems utilize different algorithms in order to predict the essay score, which may vary in performance. Thus, this paper is written to compare several algorithms that can be used to predict essay scores to find out the best performing algorithm across several tasks. This paper utilizes several supervised learning algorithms, such as random forests, linear regression, and multilayer perceptron to predict the score of an essay. Our results reveal that the random forests algorithm was able to outperform all other tested algorithms in 3 out of 4 essay scoring tasks.*

Keywords: Automated essay scoring, Supervised learning, Regression task, Random forests, Linear regression, Multilayer perceptron

1. **Introduction.** Essay writing has been a commonly used method to assess a student's academic ability. To date, the examiner has been reading, checking, and assessing students' essays manually. While this is a reliable method to score the essays of different subjects and disciplines [1], it is a time consuming one. Moreover, due to the rise of Massive Open Online Courses (MOOCs), manual scoring is not a really viable solution [1,2]. Teachers, tutors, and examiners of the MOOCs will not have enough time to score each one of the essays. Therefore, a less time consuming and automated method of scoring essays must be used.

In this research, the method used to automate the essay scoring is based on steps taken by Wachsmuth and colleagues [3]. From the collected features pattern, the essay will be scored according to four different scoring dimensions: organization, thesis clarity, prompt adherence and argument strength [4-7]. For the last step, the original author used support vector machine regression from LibSVM to predict the essay score based on four aforementioned dimensions.

In the discipline of automated essay scoring, one popular essay scoring application named E-rater, has been extensively used by educational organizations and purposes, such as the Educational Testing Service (ETS). This application has been used to score the Graduate Management Admission Test Analytical Writing Assessment, and for essays submitted to ETS's writing instruction application [8].

DOI: 10.24507/icicel.16.01.83

Our approach to tackle this regression task is by comparing several different classifiers to find out which one yields the best accuracy. Our contribution in this research is to improve the accuracy of the scoring process.

2. **Related Works.** The most recent version of E-rater, (v.2.0), was developed to increase the performance of E-rater (v.1.3). E-rater (v.2.0) now utilizes multiple regression and a newer scoring method by utilizing a more optimal cutoff via signal detection [9] to maximize the agreement of scores generated by the actual human scorers and the predicted machine scorer. The newer version of E-rater (v.2.0) was able to outperform the previous version (v.1.3) in terms of agreement rates and true-score correlation between the E-rater and human scorer [10].

In a different study, different machine learning methods on argumentative segment extraction and argumentative structure prediction tasks were compared [11]. This differs from our comparative study that focuses on the scoring task. In this study, the metric used is f1 score based on the combination of numerous feature types proposed in [3], different from our metric, which uses error value.

Meanwhile, the datasets from this study are based on annotated persuasive essays and annotated Wikipedia articles. The authors did not find a classifier that showed consistent results of outperforming other classifiers in persuasive essays corpus. However, this study found that random forest yielded the best overall f1 result in Wikipedia corpus.

An AES system based on a neural network outperforms the performance of support vector regression and Bayesian Linear Ridge Regression (BLRR) by 5.6% in terms of the Quadratic Weighted Kappa (QWK) score. The experimentation utilized features such as length-based features, Parts-of-Speech (POS), word overlap with the prompt and bag of n-grams on the ASAP competition dataset [12].

2.1. **Linear regression.** One of the algorithms utilized for the essay scoring prediction task is the linear regression algorithm, one of the most well-known algorithms in the disciplines of statistics and machine learning [13]. The linear regression algorithm plots sample points with one independent variable and one dependent variable in a Cartesian coordinate system and attempts to find a linear function which will be used to predict the values of the dependent variable, where the estimated value of the dependent value can be gained via the formula

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \tag{1}$$

where y_i is the predicted value of the dependent value, b_0 is the estimate of the regression attempt, b_n is the estimate of the regression slope, and x_n is the value of the independent variable. For our experiment, we utilized the LinearRegression classifier within Weka for 100 training batches which picks the attributes that resulted in the lowest squared error.

2.2. **Random forests.** Another algorithm that was utilized for our comparative experiments was the random forests algorithm. Random forests belong to a learning method called the "ensemble learning", which utilizes multiple learning algorithms in order to obtain a high performance. In random forests, decision trees are constructed during training time and the mean of the individual trees is utilized as the output [14]. Based on an experimentation comparing the performance of random trees and other bagging techniques, the performance of random forests is generally better than utilizing only decision trees [15]. For our experiment, we utilized the RandomForest classifier in Weka utilizing the entirety of the training set size with 100 trees and 100 training batches.

2.3. **Multilayer perceptron.** We have also utilized a Multilayer Perceptron (MLP) for the task of essay scoring prediction. MLP is a class of feed-forward neural networks, which consists of layers of artificial neurons that are interconnected with each other. For every neuron connection, there exists a connection weight that is multiplied with the input

value. In all neurons, an activation function is present that maps all the weighted inputs to the output of each neuron. These input values are then multiplied with the connection weight and mapped with the activation function to produce the output, which serves as the input for the next corresponding neurons.

In MLP, the learning process involves a step called "backpropagation", which updates the connection weights between each neuron. After each iteration, an error value is calculated based on the difference between the predicted value and the actual value. The "backpropagation" process attempts to find the optimal connection weight between neurons in order to minimize the error value. Perceptrons are shown to have strong associations with discriminant analysis and regression, which is inline with the main objective of score prediction [16]. For our experiments, we utilized the MultilayerPerceptron classifier within Weka with an automated amount of hidden layers, 0.3 learning rate, and 0.2 momentum for the weight update process. The training is done for 100 epochs for 100 training batches.

2.4. **Error metrics.** In order to evaluate each algorithm's performance, we have utilized the error metrics of Mean Absolute Error (MAE) and Mean Squared Error (MSE), since the data processed is numerical. Both metrics are able to calculate the difference between the predicted value to the actual value, and the performance of each algorithm is denoted by the values yielded by these metrics. In particular, MAE calculates the absolute average in difference between the predicted value to the actual value. Since MAE is a linear score, all individual differences from the calculation are weighted equally in the average. The formula for calculating MAE is represented as

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2}$$

where y_i denotes the predicted value, x_i denotes the actual value, and n denotes the total data.

On the other hand, the MSE measures the error by calculating the square of the average difference between the predicted and actual value. Since the errors are squared and then averaged, large errors are relatively given a high weight. The formula for calculating the MSE is represented as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{3}$$

where y_i denotes the predicted value, x_i denotes the actual value, and n denotes the total data.

With the metrics of MAE and MSE, we may see just how "off" the model was when predicting the value, and thus, the smaller these error metrics are, the more accurate the model is when predicting for the essay score. These metrics have indeed been used regularly and extensively for evaluating a model's performance [17].

3. **Methodology.** This section shall discuss the details of this research's methodology, starting from the essay scoring task, data collection, feature extraction, and the evaluation metrics used to assess each model's performance.

3.1. **Task description.** Essay scoring is the process of utilizing computer systems to predict the scores of an essay based on a determined prompt. Since the yielded value of such an activity is a real-valued number, (i.e., the score given to the essay), the performance of these systems is evaluated by comparing the score generated with the score given by a human agent [18]. Thus, this task is often addressed as a supervised machine learning task, which is mostly done via regression or preference ranking [13]. For this study, we assess the score of an essay based on its argumentation quality.

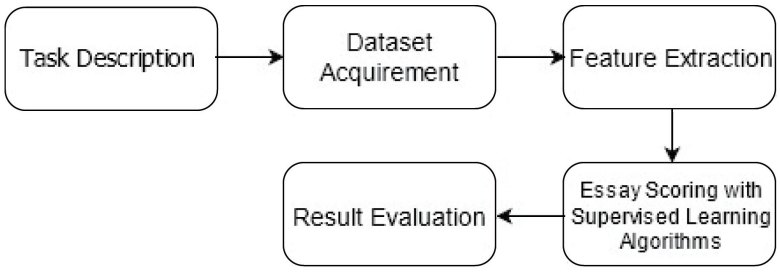

FIGURE 1. Research workflow

For the essay scoring task, the score of an essay is generated by analyzing a particular dimension of argumentation quality, such as the

1) Organization, which scores the quality of an essay's organization via its ability to introduce, argue, and conclude a topic [4];
2) Thesis Clarity, which scores the clarity of the essay via its ability to present and argue for an explanation in an essay [5];
3) Prompt Adherence, which scores the adherence of an essay via its ability to continue staying on topic of the essay's prompt [6];
4) Argument Strength, which scores the strength of the presented argument in an essay via its ability to convince the readers of the presented argument [7].

3.2. **Dataset acquirement.** For acquiring the features generated by Wachsmuth and colleagues, the Argument Annotated Essays (AAE) dataset is used, which consists of 90 persuasive student essays. For training and testing the model, the dataset used is the International Corpus of Learner English (ICLE, version 2), which consists of 6085 essays written by students [2]. The essays have 7.6 paragraphs with 33.8 sentences on average.

3.3. **Feature extraction.** For features to be fed into the classifier, we have utilized ADU features, flow features and baseline features.

Stab and Gurevych defined four different types of Argumentative Discourse Units (ADUs) that are present in essays, namely *Thesis*, *Conclusion*, *Premise*, and *None* [19]. A sentence can be described as one of these ADUs. In Wachsmuth and colleagues' approach, the ICLE corpus is mined to gain these ADU units, flow features, and standard features. The ADU units are further synthesized into three feature types:

1) *ADU flows*, which denotes the sequence of ADU types in a paragraph of an essay;
2) *ADU n-grams*, which denotes the frequencies of all ADU types in a paragraph of an essay;
3) *ADU compositions*, which denotes the frequency of occurrence of a particular ADU type in a paragraph of an essay.

Flow features were also used, which consists of

1) *Function flows*, with a basis on paragraph discourse functions for all flow features;
2) *Sentiment flows*, with a basis on a paragraph-level sentiment for all flow features;
3) *Relation flows*, with the basis of sentence-level discourse relations for all flow features [3,19].

Finally, to calculate the impact of argumentative structure within the essay itself, standard features were used, which consists of

1) *Content*, which denotes the occurrence of token 1-, 2-, and 3-gram with the minimum, average, and maximum values of prompt similarity for all sentences;
2) *POS n-grams*, which denotes occurrences of part-of-speech for 1- to 3-gram within the sentences [3].

These features are then extracted from the ICLE dataset and formed into 5 folds of training batches with Weka. These training batches are made for each essay scoring task, which totals to an amount of 25 generated Weka files. Table 1 below represents all features used for the essay scoring task.

Table 1. Features extracted

Features	Composition
ADU features	ADU flows
	ADU n-grams
	ADU compositions
Flow features	Function flows
	Sentiment flows
	Relation flows
Standard features	Content
	POS n-grams

3.4. **Essay scoring with supervised learning algorithms.** After generating the features and transforming them into training and testing batches with Weka, we utilized several different classifiers, such as linear regression, random forests, and multilayer perceptron within Weka to predict the essay score for each essay scoring task. We have utilized these classifiers due to the reasons discussed in their respective theoretical discussions above (Section 2) and they are also readily available within Weka. Some of the input parameters were also tuned to find the best results within a reasonable timeframe.

3.5. **Result evaluation.** To evaluate the results of each algorithm's performance, the metrics of MAE and MSE were utilized, as the task at hand deals with numerical data and calculation. Each algorithm has produced these metrics for all essay scoring tasks, and the results yielded were averaged for all 5 train and test folds. Since the metrics utilized were "error in prediction results" metrics, the lower the MAE and MSE values are, the better the algorithm has performed for a particular essay scoring task.

4. **Main Results.** For the training and testing purposes, the International Corpus of Learner English (ICLE, version 2) was utilized. The ICLE consists of essays written by upper intermediate and advanced English language learners. This corpus contains 6085 essays totalling to an amount of 3.7 million words written by 16 mother tongue backgrounds. The second version of the ICLE corpus is bigger in terms of words and language backgrounds compared to the first version, which consists of 2.5 million words from 11 mother tongue backgrounds.

For each essay scoring task, one distinct subset of the ICLE corpus is annotated with scores ranging from 1.0 (worst) to 4.0 (best). The argument features were also mined from each of these subsets. The final dataset contains 1003 organization, 830 thesis clarity and prompt adherence, and 1000 argument strength essays that have been scored based on each essay's quality.

After training and testing each task with each algorithm, the MAE and MSE values are produced as an indicator of their performance. In total, for each task, 5 MAE and MSE values have been produced by each algorithm for one particular task's predefined fold, which are then averaged. And thus, the performance of each algorithm denoted by their Mean Average Error (MAE) values are produced in Table 2.

For each essay scoring task, we have also evaluated each algorithm's performance on their Mean Squared Error (MSE) values. Table 3 tabulates the average MSE values for all tested algorithms for each essay scoring task.

Table 2. Average MAE

	RF	LR	MLP	Lib-SVM
Essay organization	0.338	0.388	0.580	**0.315**
Thesis clarity	**0.534**	0.540	0.570	0.542
Prompt adherence	**0.370**	0.380	0.395	0.374
Argument strength	**0.394**	0.403	0.432	0.408
Average per algorithm	**0.409**	0.427	0.494	0.410

Table 3. Average MSE

	RF	LR	MLP	Lib-SVM
Essay organization	0.182	0.248	0.191	**0.168**
Thesis clarity	**0.426**	0.433	0.507	0.486
Prompt adherence	**0.214**	0.217	0.265	0.240
Argument strength	**0.232**	0.244	0.293	0.259
Average per algorithm	**0.264**	0.285	0.314	0.288

Based on the results above, for the task of essay organization, the Lib-SVM algorithm yielded the best results, with MAE and MSE of 0.315 and 0.168. For the task of thesis clarity, the random forests algorithm yielded the best results, with MAE and MSE of 0.534 and 0.426. For the task of prompt adherence, the random forests algorithm has again yielded the best results, with MAE and MSE of 0.370 and 0.214. And finally, for the task of argument strength, the random forest has again yielded the best results, with MAE and MSE of 0.394 and 0.232.

Based on the results above, in metrics of both MAE and MSE, the random forests algorithm yielded the best results for the tasks of thesis clarity, prompt adherence, and argument strength when compared to the other tested algorithms, with average performances of 0.409 and 0.264 for MAE and MSE respectively.

The algorithm Lib-SVM utilized by Wachsmuth and colleagues was the best when tackling the task of scoring for essay organization, with an average of 0.410 and 0.288 for MAE and MSE respectively [3].

The linear regression algorithm that was employed has also performed better than the Lib-SVM algorithm in 3 out of 4 essay scoring tasks, similar to the random forests algorithm, averaging around 0.285 in terms of MSE. In terms of MAE, the linear regression algorithm has performed better than Lib-SVM in tasks of scoring for argument strength and thesis clarity.

The multilayer perceptron algorithm with our configurations was not able to yield better results than any of the other tested algorithms. Its results have averaged 0.494 and 0.314 for MAE and MSE, respectively. Perhaps a more fine-tuned configuration is able to yield better results, which can be gained by fine-tuning the parameters and having more epochs and training batches.

From the results above, we can conclude that random forests were able to yield the best results for 3 out of 4 tasks (thesis clarity, prompt adherence, and argument strength), for both metrics of MAE and MSE, when compared to the other algorithms. In terms of MSE, the linear regression algorithm was able to surpass the Lib-SVM algorithm when performing for the same 3 tasks as random forests did. The multilayer perceptron algorithm with our configurations yielded the worst results amongst all tested algorithms.

Random forest is perhaps the best algorithm amongst the tested algorithms due to its nature as a bootstrap algorithm or commonly known as bagging. The random forest is

particularly good when dealing with large amounts of data, as it is able to estimate a value based on the sample data's mean value which gives a good estimation of the true value. They are able to perform well even when dealing with a large number of features and a small number of observations. The tree-building process within random forests implicitly allows for interaction between features and high correlation between features, which may generate a predicted value close to its true value [13].

5. **Conclusions.** After gathering the results of the experiment, the conclusions and findings are then formed into this research article.

In this comparative study, we have compared machine learning models employing different supervised learning algorithms to predict the score of an essay with features generated by Wachsmuth and colleagues [3-7]. The essay scoring task shall score these essays based on its essay organization, thesis clarity, prompt adherence, and argument strength. The dataset utilized was the ICLE corpus (version 2), which contains 6085 student essays written by 16 non-English mother tongues.

Of the proposed algorithms, the random forest algorithm was able to yield the best results in terms of MAE and MSE for 3 out of 4 essay scoring tasks, namely thesis clarity, prompt adherence, and argument strength when compared to the other tested algorithms. The average MAE and MSE generated by the prediction of the random forests algorithm is 0.409 and 0.264 respectively. In terms of MSE, the linear regression algorithm performed better than Lib-SVM but worse than random forests. With our configurations, the multilayer perceptron yielded the worst results. However, by utilizing other features that can be gathered from the dataset, or by utilizing a different classifier, better results could be achieved. Our experiment's results and findings are then concluded and reported in the form of this research article.

REFERENCES

[1] P. W. Foltz, D. Laham and T. K. Landauer, Automated essay scoring: Applications to educational technology, *Proc. of ED-MEDIA 1999 – World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pp.939-944, 1999.

[2] S. Granger, F. Meunier and M. Paquot, *International Corpus of Learner English*, 2nd Edition, Presses Universitaires de Louvain, 2009.

[3] H. Wachsmuth, K. Al-Khatib and B. Stein, Using argument mining to assess the argumentation quality of essays, *Proc. of the 26th International Conference on Computational Linguistics: Technical Papers (COLING2016)*, pp.1680-1691, 2016.

[4] I. Persing, A. Davis and V. Ng, Modeling organization in student essays, *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp.229-239, 2010.

[5] I. Persing and V. Ng, Modeling thesis clarity in student essays, *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pp.260-269, 2013.

[6] I. Persing and V. Ng, Modeling prompt adherence in student essays, *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.1534-1543, 2014.

[7] I. Persing and V. Ng, Modeling argument strength in student essays, *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp.543-552, 2015.

[8] J. Burstein, M. Chodorow and C. Leacock, *Criterion*^{SM} online essay evaluation: An application for automated evaluation of student essays, *Proc. of the 15th Annual Conference on Innovative Applications of Artificial Intelligence*, vol.25, no.3, pp.27-36, 2004.

[9] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum Associates, Inc., 1996.

[10] J. Burstein and Y. Attali, Automated essay scoring with e-rater®V.2, *The Journal of Technology, Learning, and Assessment*, vol.4, no.3, 2006.

[11] A. Ahmed, A. Sliwa, Y. Ma, R. Lui, N. Borad, S. Ziyaei and M. Ghobadi, What works and what does not: Classifier and feature analysis for argument mining, *Proc. of the 4th Workshop on Argument Mining*, pp.91-96, 2017.

[12] K. Taghipour and H. T. Ng, A neural approach to automated essay scoring, *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1882-1891, 2016.

[13] C. Yu and W. Yao, Robust linear regression: A review and comparison, *Communications in Statistics – Simulation and Computation*, pp.1-22, 2016.

[14] A. Ziegler and I. R. König, Mining data with random forests: Current options for real-world applications, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp.55-63, DOI: 10.1002/widm.1114, 2014.

[15] S. M. Piryonesi and T. E. El-Diraby, Data analytics in asset management: Cost-effective prediction of the pavement condition index, *Journal of Infrastructure Systems*, vol.26, no.1, pp.36-40, 2020.

[16] B. Cheng and D. M. Titterington, Neural networks: A review from a statistical perspective, *Statistical Science*, vol.9, no.1, pp.33-35, 1994.

[17] T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?, *Copernicus Publications on Behalf of the European Geosciences Union*, pp.1525-1535, 2014.

[18] J. Wang and M. S. Brown, Automated essay scoring versus human scoring: A comparative study, *Journal of Technology, Learning, and Assessment*, vol.6, no.2, 2007.

[19] C. Stab and I. Gurevych, Annotating argument components and relations in persuasive essays, *Proc. of the 25th International Conference on Computational Linguistics: Technical Papers (COLING2014)*, pp.1501-1510, 2014.

[20] H. Wachsmuth, J. Kiesel and B. Stein, Sentiment flow – A general model of web review argumentation, *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp.601-611, 2015.