A Hybrid Automated Essay Scoring Using NLP and Random Forest Regression

Muhammad Zaim Azri Bin Azahar(✉) and Khairil Imran Bin Ghauth

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia
1181101108@student.mmu.edu.my, khairil-imran@mmu.edu.my

Abstract. Assessing the performance of students through subjective assessments namely essays is critical in measuring their achievement during the learning process in an educational system. The essay test will evaluate the student's ability to remember and express their ideas or opinions toward certain topics. A teaching staff is usually required to assess and grade the students' essays. This paper presents a hybrid Automated Scoring System based on Natural Language Processing (NLP) and Random Forest Regression. The model focused on regression task where the predicted score is in a continuous value. Natural Language Processing (NLP) has also been applied in this work to extract features from essays. Finally, all the proposed model is compared to Linear Regression and Deep Learning and are then evaluated to compare the performance of the models by using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Keywords: Automated Essay Scoring · Linear Regression · Random Forest · Deep Learning · Natural Language Processing

1 Introduction

Nowadays, teaching staffs are overwhelmed to assess a lengthy essay. Sometimes the teaching staff might overlook on certain keywords or important points in the given answer. In this research, we are motivated to overcome the said problem by proposing an automated essay scoring system using a hybrid approach, which is the combination between NLP and Random Forest Regression (HRFR). The proposed method is then benchmarked against several well-known techniques in machine learning such as Linear Regression and Deep Learning. The focus of this comparison is to identify which model produces the best performance result by using evaluation metrics such as MAE, MSE, and RMSE. The difference between this research from others is that this study will focus only on the content of the essay by applying only the Bag-of-Words methods as the feature extraction. Statical features are not applied in this study because we believe that feature extraction such as the average length of the word, characters, and sentence will not positively contribute to the final score.

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 448–457, 2022.
https://doi.org/10.2991/978-94-6463-094-7_35

1.1 Motivation

The motivation for conducting this research on Automated Essay Scoring is because of looking at the inconsistency of teachers when marking essay papers. Inconsistent marking of essay papers has a negative effect on student grades. As a result, the establishment of AES has the potential to alleviate the problem.

This paper is divided into the following sections: Sect. 1 is an Introduction. Section 2 discusses the Literature Review. Section 3 will briefly explain the dataset used for this research. Section 4 discusses the proposed methodology such as framework and architecture. Section 5 discusses the conclusion and followed by Sect. 6 which suggests future works.

2 Literature Review

2.1 Essay

An essay is a type of written work that combines facts and opinions. The assessment of an essay is complex compared to a non-essay which is normally in the form of multiple-choice questions that just have only true or false answers. The essay, which is analytical, speculative, and interpretive, is more subjective from the researcher's point of view. It includes narratives that can take the form of criticism, arguments, literature based on observations of everyday life, and researcher reflections.

The purpose of writing an essay is to convince the reader to believe or accept the researcher's viewpoint on a particular situation. It assesses an individual's ability to express their understanding of a certain topic. An essay can be a source of information about a point of view or research that has been conducted. Readers can learn to gain more information from the writing.

2.2 Automated Essay Scoring

Automated Essay Scoring is part of modern computer technology and it is an education-based system that can examine essays, directly display the scores for the essay, and improve consistency. For example, teachers may no longer spend so much time marking students' essay papers while students also will be able to view their scores immediately after submission of the essay assessments.

2.3 Linear Regression

Linear Regression (LR) is a statistical technique used to model the connection between a scalar dependent variable y and one or more explanatory factors (or independent variables) indicated by x [1]. Linear Regression is one of the most straightforward and often used machine learning methods [2]. As a result, many people who are unfamiliar with machine learning will utilise this model to introduce themselves to the concept. The equation of Linear Regression is shown in Eq. (1).

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

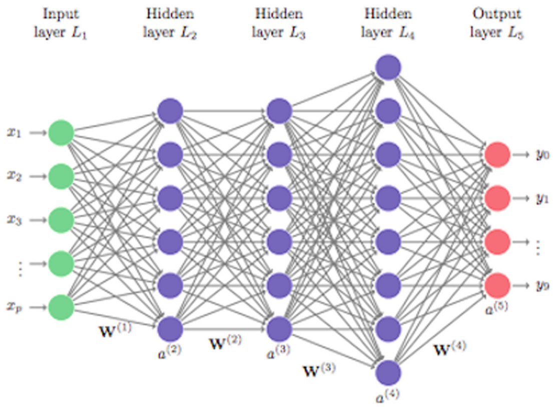

Fig. 1. Deep Learning Network [8].

The formula was cited from a paper that was also studied in the same research field [4]. $x_1 \ldots x_n$ denotes the features of an essay that, when combined with specific weights which is $b_0 \ldots b_n$. y provide a continuous value for the essay score.

2.4 Random Forest Regression

Random Forest Regression (RFR) is a form of supervised learning in which an ensemble of decision trees is constructed to perform regression or classification. The training data is subdivided into random subsets, and a decision tree is constructed from each subset of the training data [4].

2.5 Deep Learning

Deep Learning (DL) is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the brain. These algorithms are referred to as artificial neural networks [7]. Deep Learning is basically a term that refers to the use of a Neural Network with multiple layers of nodes between the input and output. It has multiple hidden layers which is more than a normal Neural Network that consists of only one hidden layer in it. Figure 1 is an example of a Deep Learning Network where it consists of one input layer, three hidden layers, and one output layer [8].

2.6 Natural Language Processing

Natural Language Processing (NLP) is a branch of computer science, Artificial Intelligence (AI), and Linguistics that is concerned with the interaction of computers and human natural language like the English language. The purpose of the existence of NLP is to perform the process of creating a computational model of language so that humans and computers can interact through the medium of natural language. Computational models are useful for both scientific purposes, such as studying the properties of a natural language, and for everyday use [9].

Table 1. Dataset's Column Details.

No.	Column Name	Description
1	essay_id	An essay id for student's essay.
2	essay_set	Essay set.
3	essay	Written essay in ASCII form.
4	rater1_domain1	All sets have a score from this rater1_domain1.
5	rater2_domain1	All sets have a score from this rater2_domain1.
6	rater3_domain1	Only a few essays in set 8 have a score from rater3_domain1.
7	Domain1_score	Finalized score that has been agreed upon by the raters.
8	Rater1_domain2	Only essays in set 2 have this score from Rater 1.
9	Rater2_domain2	Only essays in set 2 have this domain 2 score, which comes from rater 2.
10	Domain_score	Only essays in set 2 have a resolved score between the raters; the other sets do not.
11	rater1_trait1 score—rater3_trait6 score	Available only for set 7 and 8.

3 Dataset

The data for this project is taken from the Kaggle.com website Hewlett Foundation has organized a competition called Automated Essay Scoring (ASAP) which was organized in 2010 which was about 10 years ago [3].

The contestant was given a total of 8 essay sets to be used in order to compete in this competition. According to the Kaggle.com website, for each set of essays, a single prompt was used to generate the essays, which were then collected and analysed. In order to pick the writings, they look for those with an average word count ranging from 150 to 550 words each essay. The Hewlett Foundation has made a training dataset available for use in developing an AES. The training dataset was provided in three different file formats which are Table Separated Value (TSV), Microsoft Excel 2010 spreadsheet, and Microsoft Excel 2003 Spreadsheet. The training dataset details that consisted of all 8 essay sets are shown in the Table 1.

4 Methodology

4.1 Architecture

Figure 2 shows the Architecture diagram for this Automated Essay Scoring (AES) system. Based on the diagram, there will be two main entities; the Kaggle Dataset which

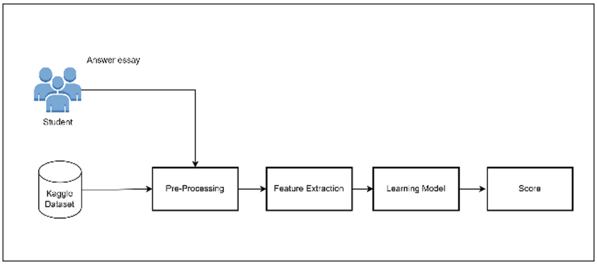

Fig. 2. Proposed Architecture.

Table 2. Essay Sets Details.

Set	Essay Type	Training Set Size
1	Persuasive/Narrative/Expository	1785 essays
2	Persuasive/Narrative/Expository	1800 essays
3	Source Dependent Responses	1726 essays
4	Source Dependent Responses	1805 essays

will be used to prepare the training materials for the learning algorithms, and also the student who will answer the essay question. However, only 4 essay datasets will be used for training the models. This is due to the time constraint which the system for this Automated Essay Scoring also needs to be completed within the allocated time. The details of the essay sets are shown in Table 2.

As illustrated in Fig. 2, the Kaggle dataset and essay answers by a student will undergo 3 processes. The process consists of *pre-processing, feature extraction,* and a *learning model*. The pre-processing process encompasses a variety of processes such as *case folding*, *symbol removal, punctuation removal*, and *stopwords removal*. After the *pre-processing* process, the data will proceed to the second process which is the *feature extraction* process. This process will apply the NLP technique and the feature that will be extracted in this process is called Bag-of-Words (BOW). Bag-of-Words (BOW) is a method that enables words to be represented in a numerical form. The reason for applying this technique is because learning models only accept numerical as the input. After extracting the feature from the data, the feature will be fitted into the learning algorithm. The learning algorithm will be trained based on a regression task. Hence, it will predict a numerical value rather than classification.

4.2 Framework

The proposed framework for this Automated Essay Scoring system is depicted in Fig. 3. The steps of the methodology proposed are as follows; To begin, the data pre-processing will be done towards the dataset. The pre-processing will include several methods like *Case Folding, Symbol Removal, Punctuation Removal,* and *Stopwords Removal.*

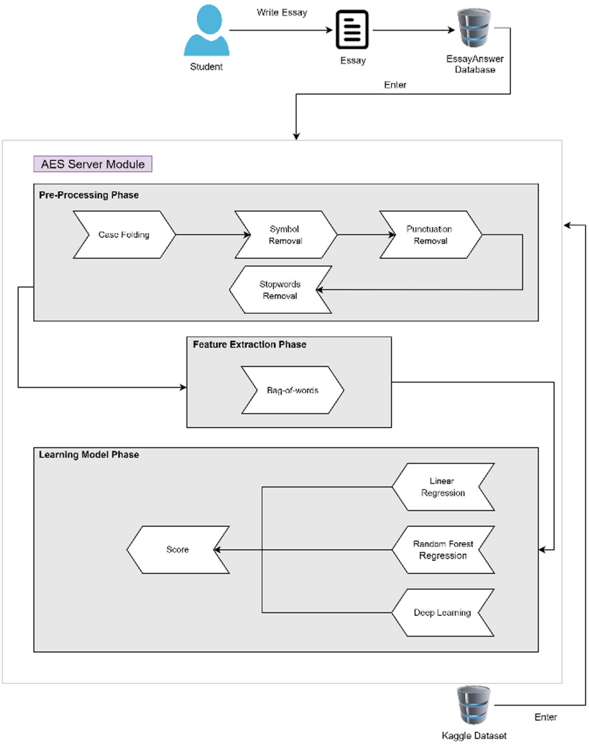

Fig. 3. Proposed Framework.

After the pre-processing, the essays will then undergo the Feature Extraction process. The feature that will be extracted is words from the essay. The word will be extracted using an NLP method called text modelling. The text modelling chosen for this project is Bag-of-Words (BOW).

Lastly, the feature will be fitted into learning models. Learning models that will be used to evaluate the essay are Linear Regression (LR), Hybrid Random Forest Regression (HRFR), and Deep Learning (DL). All of these models are intended to produce a score in regression-based.

5 Result and Analysis

We evaluate the performance using 4 sets of essays from the Kaggle dataset. Evaluation metrics for regression models are used to evaluate and compare the performance of every model. The evaluation metrics are taken based on the most frequently used by researchers when evaluating the regression models. Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are the most popular metrics for regression models [6].

MAE is calculated by averaging the absolute error values. Absolute or abs() is a mathematical function that converts a negative integer to a positive integer. Thus, the

Table 3. Evaluation Metrics result.

COMPARISON OF MODELS PERFORMANCE RESULT									
Essay Set	MAE			MSE			RMSE		
	LR	HRFR	DL	LR	HRFR	DL	LR	HRFR	DL
1	1.08	**0.87**	1.03	1.93	**1.27**	1.77	1.39	**1.13**	1.33
2	1.17	0.93	1.15	2.27	1.32	2.2	1.51	1.15	1.48
3	2.59	1.56	1.62	11.87	4.1	4.33	3.45	2.03	2.08
4	1.48	1.33	1.42	3.57	2.72	3.32	1.89	1.65	1.82

difference between an expected and forecasted value might be either positive or negative, but must be positive when calculating the MAE [5].

The mean or average of the squared discrepancies between predicted and expected target values in a dataset is used to determine the MSE and RMSE is calculated by the root square of the value of MSE [5].

A model is expected to show great performance results when each metric is close to 0 value. The closer the result to the 0 value, the less the model will produce an error. Table 3 summarises the result of the performance comparison of models for each set.

As illustrated in Table 3. The HRFR model had the best performance for MAE, MSE, or RMSE in every set. The best result was produced in the Essay Set 1 with the value of 0.87 for MAE, 1.27 for MSE, and 1.13 for RMSE. The DL model had the second-best performance result and it was followed by the LR model.

We plot bar chart graphs for each essay set to make the performance results of models more readable. The result for Essay Sets 1, 2, 3, and 4 are depicted in Figs. 4, 5, 6, and 7. As can be seen in the all figures, the HRFR model outperforms the other 3 models.

The finding of the experiments shows that Hybrid Random Forest Regression (HRFR) outperform the other 2 models and produced the best results for MAE, MSE, and RMSE. The graph comparison between the three models can be seen in Table 3. Supposedly, the Deep Learning model should produce a better result compared to the other models. However, it did not perform a great performance when combined with Bag-of-Words (BOW). As explained in the objectives of this project. The objective of this project is not to improve the performance of the model but to compare which model produces the best performance.

6 Conclusion

In conclusion, Automatic Essay Scoring allows users to receive an instantaneous score on their essays. It could also assist teachers in reducing the amount of time they spend when grading essays. This research makes use of three models to score the essay. The models are evaluated by using the evaluation metrics called Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The result revealed that the hybrid method combining an NLP technique and Random Forest Regression model beat the performance shown by Linear Regression and Deep Learning model.

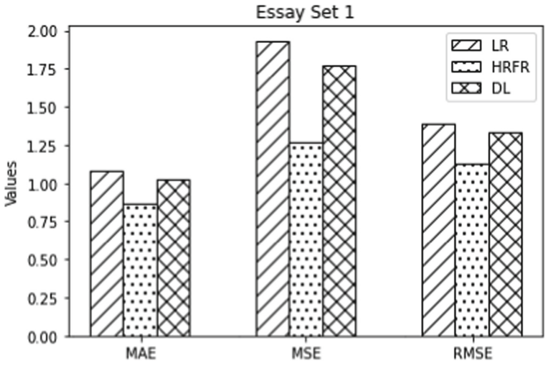

Fig. 4. Performance models of Essay Set 1.

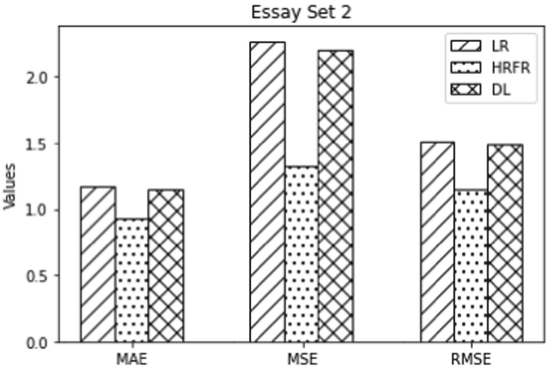

Fig. 5. Performance models of Essay Set 2.

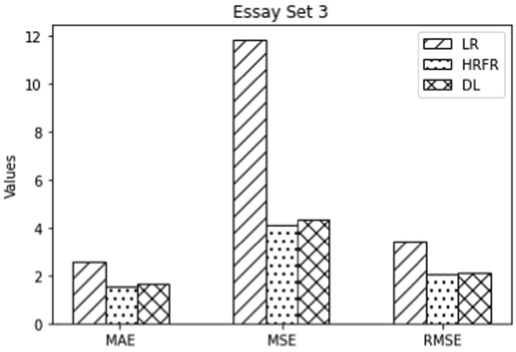

Fig. 6. Performance models of Essay Set 3.

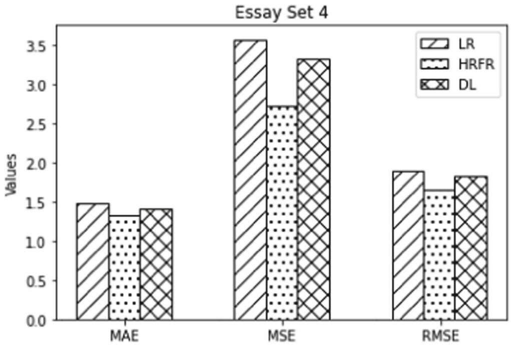

Fig. 7. Performance models of Essay Set 4.

7 Future Work

In future work, we would like to implement the different type of Word Embedding that is available in the context of Natural Language Processing. Additionally, an improvement of the performance of learning models could also be done. Besides that, we would like to implement additional learning models that are capable to perform regression tasks such as Support Vector Regression and observe the performance of the model compared to other models accessible in the context of Automated Essay Scoring. Another future work is to do live training with fewer dataset. This is because we believe that teachers will not have enough answer resources to be fitted into the learning models as done in this research.

References

1. S. Drolia, S. Rupani, P. Agarwal, and A. Singh, "Automated Essay Rater using Natural Language Processing," *International Journal of Computer Applications*, vol. 163, no. 10, pp. 44–46, Apr. 2017, Accessed: Apr. 12, 2022. [Online]. Available: https://www.ijcaonline.org/archives/volume163/number10/27435-2017913766
2. D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020. https://doi.org/10.38094/jastt1457
3. Kaggle, "The Hewlett Foundation: Automated Essay Scoring | Kaggle," Kaggle.com, 2012. https://www.kaggle.com/c/asap-aes
4. I. Arya, "A Comparative Study of Methods for Automated Essay Grading," 2021. Accessed: Apr. 12, 2022. [Online]. Available: https://ishaanarya.com/A%20Comparative%20Study%20of%20Approaches%20to%20Automated%20Essay%20Grading.pdf
5. J. Brownlee, "Regression Metrics for Machine Learning," Machine Learning Mastery, Jan. 19, 2021. https://machinelearningmastery.com/regression-metrics-for-machine-learning/#:~:text=We%20cannot%20calculate%20accuracy%20for (accessed Apr. 13, 2022)
6. A. Botchkarev, "A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045–076, 2019. https://doi.org/10.28945/4184

7. J. Brownlee, “What is Deep Learning?,” Machine Learning Mastery, Aug. 15, 2019. https://machinelearningmastery.com/what-is-deep-learning/#:~:text=Deep%20learning%20allows%20computational%20models
8. Sunpark, “It’s Deep Learning Times: A New Frontier of Data,” Medium, Dec. 14, 2019. https://towardsdatascience.com/its-deep-learning-times-a-new-frontier-of-data-a1e9ef9fe9a8
9. D. Khurana, A. Koli, K. Khatter, and S. Singh, “(PDF) Natural Language Processing: State of The Art, Current Trends and Challenges,” ResearchGate, 2017. https://www.researchgate.net/publication/319164243_Natural_Language_Processing_State_of_The_Art_Current_Trends_and_Challenges