Automatic Essay Scoring for E-Learn System

Ameer Hassan Hadi[1], Ahmed Hussein Aliwy[2]

1 ameerh.altai@student.uokufa.edu.iq, 2 ahmedh.almajidy@uokufa.edu.iq

Article Info
Page Number: 1011 – 1024
Publication Issue:
Vol. 71 No. 3 (2022)

Abstract

The e-learning system is used to support and enhance the educational process with many facilities over traditional learning. Some of these facilities are electronic exams and automatic scoring for answers of types true and false, multiple choices, and may be short answers. Therefore, many researchers take this research direction but the biggest challenge of exam scoring is the scoring of essay exams, which is an open problem. Automated essay scoring is an educational evaluation technique and part of the natural language processing (NLP) application. Numerous considerations, including cost, accountability, standards, and technology, contribute to the increased interest in automated essay scoring. In this work, a complete essay scoring system is proposed with different methodologies that are evaluated with different evaluation metrics. These methodologies are used for classification\regression tasks; (i) each one of nine classifiers/regressions is used as an independent classifier/regression; (ii) the best three classifiers/regressions algorithms are used for getting the score as the average, and (iii) using augmented combination of all the classifiers\regression for calculating the final score. Four categories of features are used; raw features, morphological features, compound features, and orthographical features with weight for each feature that reflect the feature importance. The results, on the Hewlett essay scoring dataset, showed that scoring using the average of the best three classifiers and augmented combination have the lowest error rate in most tests. Also, they are more stable than the other classifiers where there is not any huge rising in errors in all the tests.

Keywords: E-Learning, natural language processing, machine learning, and essay score.

Article History
Article Received: *12 January 2022*
Revised: *25 February 2022*
Accepted: *20 April 2022*
Publication: *09 June 2022*

1. Introduction

After the emergence of the Corona pandemic, the world began to turn towards e-learning and rely, partially or completely, on it. This makes extra burdens and efforts on the teaching staff and students in general because e-learning has many challenges and problems.Along with having more materials than the conventional classroom to facilitate learning, e-learning overcomes the time and space constraints associated with traditional instruction. Because e-learning enables learners to study autonomously, it lacks the

monitoring and enforcement mechanisms associated with traditional education [1].However, it is the best solution in times of epidemics, disasters, and abnormal conditions. Among the most important directions taken by some researchers is how to make computers score exams like humans, to facilitate work and speed up achievement.There are several examination models that are commonly used in the educational system, false or true, short answer questions, multiple-choice, and essays. The first three models facilitated the evaluation, and automated scoring can be achieved easily because they are not ambiguous. But the difficulty and the problem appearin the essay answer because we need a comprehensive evaluation to verify the accuracy of the answer. Therefore, scoring the essay is a great challenge. It is not a challenge to the system, but it challenges to teachers because they should analyze the student's answer with concentration and care and then give the grade manually.

Automated essay scoring is an educational grading technique and part of the natural language processing (NLP) application[2]. If a standard impartial training dataset is available, the automated system for essay grading will avoid these drawbacks altogether[3].On the oppositeside, Manualessay scoring is time-consuming and may be unintentionally prejudiced they are grading. For example, the same answers may result in different grading. This ensures that the grades are affected by psychology, mood, student self, and others [4].

Several researchers have used various strategies to address the challenge of automated essay scoring, also known as automatic essay assessment. The main feature of these systems is a collection of essays written by the student and manually graded by experts. Typically, these manually graded essays are referred to as training essays, whereas essays that must be scored by a machine are referred to as tested essays or automated scored essays [5].

2. Related Works

There are many researches on AES most of them use ML techniques. In this section, some of these researches will be explained briefly.

Burstein, Kukich, Wolfe & Chodorow (1998)[6] created an electronic essay scoring system that extracts features related to topical content, discourse marking, and syntactic information. Then they compared two vectors to predict, essay argument content and essay content. This system obtained 82% accuracy between human raters and argument content scores, 69% compared with essay content. Li Bin et al. (2008) [7] presented an AES using the KNN algorithm. Their system was done using different feature selection methods. The system achieved 76% accuracy on the Chinese Learner English Corpus. This system depends on the content only.

Zhen Biao Chen et al. (2010) [8] suggested an automated essay scoring system using four methodologies of Vector Space Models (VSM), including the Word-based Vector Space Model (W-VSM), the Weight Adapted Word-based Vector Space Model (WAW-VSM), the Latent Semantic-based Vector Space Model (LS-VSM) and the Sequence Latent Semantic-based Vector Space Model (SLS-VSM). The system was implemented using 970 Chinese essays with a best average correlation of 0.6123.

ManviMahana et al. (2012)[9] presented AES based on a linear regression algorithm. The system was tested using 13000 essays. Little features were used, such as Sentence Count (SC), Word Count (WC), Number of Long words (NOL), and Part Of Speech counts (POS).

The system achieved 73% accuracy score.The training and testing were done on the same dataset type (answers of same question).

Ming Qing Zhang et al. (2014) [10]presented AES by using the incremental system for Latent Semantic analysis(LSA). It was used on 15,776 essays as a training set and 1,000 test essays as the testing set. This data is written in the Chinese language.Results showed that this incremental system was more effective than (LSA) by reducing the usage of memory and time consuming without lowering the performance, where the system achieved accuracy of %88.8.

Shankar et al. (2018) [2]explained automated essay score using little features such as sentence count, word count etc.and a sequential forward feature selection algorithm to compare accuracy between different features to select the best subset of features in order to score the essays. Essays written in the English languageby 15 students were usedfor evaluation. This system succeeded with small datasets, but it was not tested on larger data to determine its validity.

Citawan et al.(2018) [11]designed an automated essay scoring system to enhance the learning process. They used latent semantic analysis (LSA) with n-gram as features. This method was combined with features of n-gram to know the order of words in sentences and to find the similarity between the student's answers and the teacher's optimal answer by knowing the patterns and relationships between the words in the matrix. they showed average accuracy ranging from 14.91%, 58.89 %,64.49%, 71.37%,78.65% by different features of a trigram, bigram, unigram + bigram + trigram, unigram + bigram, unigram, unigram.

Zhiyun Chen et al. (2019) [12]proposed AES system by a combining of ordinal regression (OR) and convolutional neural networks (CNN). They compared the results of the proposed system with alone CNN or aloneLSTM model. The used data was about 13,000 essays from an organized competition called the Automated Student Assessment Prize (ASAP). The average accuracy of the proposed model was (82.6%).

Mr P. V. Hari Prasad et al. (2020) [13] used deep learning techniques and layers like dense layers and (LSTM). They used data of eight sets of English language that represents answers to students from different levels. The extracted features include word count, sentence count, prevalence, and parts of speech count. These features consider as input to the neural network.

Almost all these works used training and test sets for answers of same question which cause unreliable system for other answers of new questions. In this work, a complete AES system is proposed that trained from answers of one question and tested on answers of other questions with different methodologies.

3. AES Approaches

All the existing AES systems are learning-based; supervised, semi-supervised, or reinforcement learning. AES tasks can be regression or classification tasks according to the specific requirements. This means all classification and regression algorithms can be used to solveAES but surely with different accuracy scores. In this work, nine algorithms for classification and regression were used, five of which were for classification, which are Multinomial Naïve base, Gaussian NB, K-Nearest Neighbors, Decision Tree, and Random

Forest classifiers, while for regression, four algorithms were used which are Decision Tree, Random Forest, Logistic Regression and Linear Regression.

4. The Proposed system

The proposed system has five distinct phases. The first phase is to produce a suitable data representation to be processed in the next stage. The second phase is features extraction and calculating feature importance to produce the weight of each feature because some features have a high effect on the classification. The third phase is the classification process and calculatingweight of each classifier.The fourth phase is the combination of these classifiers/regressions in two methodologies; (i) average score of best classifiers/regressions and (ii) augmented combination. The final phase is the evaluation of the proposed system. The next sections will discuss the component of the proposed system. Figure 1 shows a block diagram of the Proposed System stages.

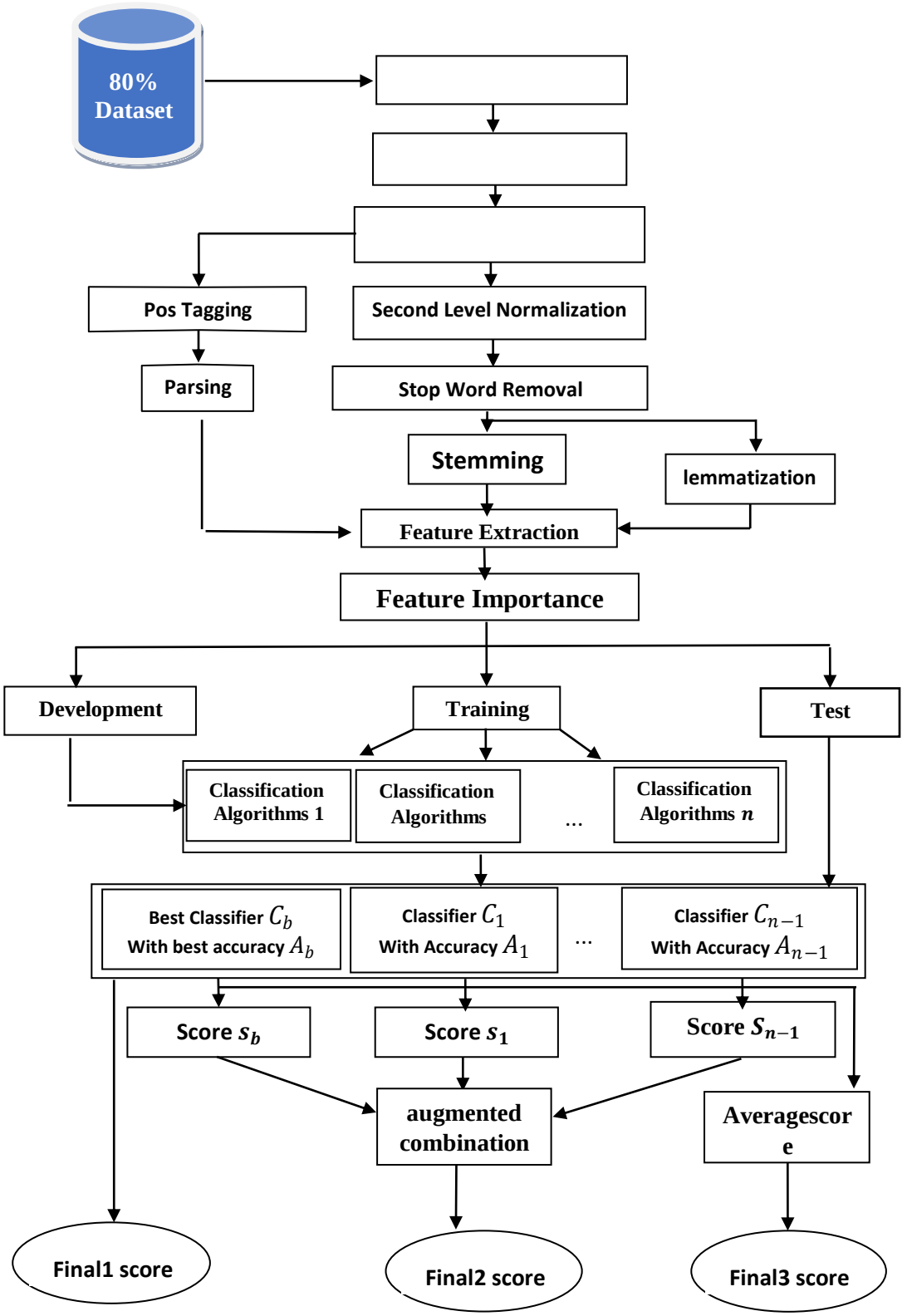

Figure 1:The Block Diagram of the first phase for the Proposed System stages.

- **Text Preprocessing (TP)**

Text pre-processing is the first stage of the proposed system. It is very important since raw data contains duplicate, noise, or unformulated data. Also, the original data may be not

proper to be used for the approaches of analysis. Almost all-natural language processing applications need the data to be initialized and pre-processed. One of these applications is text classification which needs the text data to be converted into numeric data. Classification results may be inaccurate because the raw data contains typographical errors, symbols, or abbreviations. To deal with this data, most text processing, and mining applications employ some form of pre-processing to reduce the number of features. It consists of(i) tokenization: splitting the running text into sentences and tokens, (ii) normalization: unification of letters scripts and deleting unwanted symbols), (iii) stop word removal: removing the uninformative words/tokens, (iv) stemming extracting the stem of each word, and(v) lemmatization: extracting the lemma of each word, (vi)POS tagging: extracting the part of speech for each word, and (vii)Parsing: extracting the parse tree for each sentence/phrase.

- **Features Extraction (FE) and feature importance (FI):**

Feature extraction is the second phase, and it is completed after the pre-processing stage. Four types of features are taken; Raw Features, Morphological features, Compound features, and Orthography Features. It is clear that the features are corpus-independent, where the learning process can be done for a corpus and applied to another.

In our system, many features are used with different levels of importance. For example, the number of the used stem does not equal in weight to spelling errors or grammar mistakes. Therefore, feature importance is used to calculate the importance of the features in our model. A higher score means that the feature has more effect than the other features.

In our model, each feature f_i has a weight w_i reflect the importance of this feature where $\sum_{i=1}^{feature} w_i = 1$. For this task, linear regression models are used to produce feature importance. In this work, fourteen features are used; word count, sentence counts, word greater than nine-count, word smaller than three counts, third stemmer count, 4th stemmer count, 5th stemmer count, lemmatizer counter, bigrams counter, trigrams counter, four grams counter, five grams counter, spelling errors rate, and grammar mistakes.

- **Classification using classifiers Algorithms**

The third stage of the proposed system, the classification stage, is to classify students' answers into one of a range of classes. It is done after the first twostages(pre-processing and extracting features). In our system, nine different classifiers were used (Gaussian NB, Multinomial NB, KNN Neighbors, Decision Tree Classifier, Random Forest Classifier, Random Forest Regression, Logistic Regression, Linear Regression). They were applied to three methodologies,(i) these classifiers are implemented independently,and each classifier gives its own score to test the classification and regression processes,(ii) the score is the average of the best classifiers, and (ii)the score is result of augmented combination of classifier.

4.1 Combination of theClassifiers

In this work, two methodologies of a combination of classifiers are taken; averagescore (AS) and augmented combination (AC). In the case of averagescore, the final score of a given essay is the average of the scores of the best three classifiers.

We will introduce a new methodology for using multiple classifiers where each classifier will take a weight based on its accuracy, where the sum of these weights is equal to 1.

Suppose that there are several classifiers $C_1 \ldots . C_n$; where the accuracy/score of these classifiers are$A_1 \ldots . A_n$where they were measured using development data. Surely one of them gives the highest (best) accuracy (A_b), which is C_b. If we exclude the best classifier, then:

$$A_1 + A_2 \ldots + A_{n-1} = M$$
$$A_1/M + A_2/M \ldots + A_{n-1}/M = 1$$
$$\frac{A_1}{\sum_{i=1}^{n-1} A_i} + \frac{A_2}{\sum_{i=1}^{n-1} A_i} \ldots + \frac{A_{n-1}}{\sum_{i=1}^{n-1} A_i} = 1 \quad (1)$$

Compared to $w_1 + w_2 \ldots + w_{n-1} = 1$ then.

$$w_1 = \frac{A1}{\sum_{i=1}^{n-1} Ai}, \qquad w_2 = \frac{A2}{\sum_{i=1}^{n-1} Ai} \ldots, w_{n-1} = \frac{w_{n-1}}{\sum_{i=1}^{n-1} A_i} \quad (2)$$

Where w_1 ... w_{n-1} are weights of the classifiers C_1 ... C_{n-1} respectively.

Suppose we have two variables L and S, where L represents the average of errors result from the predicted values that are greater than the correct values while S represents the average of errors results from the predicted values that smaller than the correct value.

If we want to Score an essay, it will be as input to all classifiers in addition to the best classifier. If the score value of the classifier C_i is V_i and the score value of best classifier $C_b$is V_i. The final score will be as follows:

$$\text{Vfinal } = V_b + \begin{Bmatrix} w_i\, L & V_b < V_i \,\&\, L < V_i - V_b \\ w_i\,(Vi - V_b) & V_b < V_i \,\&\, L \geq V_i - V_b \\ -(w_i\, S) & V_b >= V_i \,\&\, S < V_i - V_b \\ -(w_i (V_b - V_i)) & V_b >= V_i \,\&\, S >= V_i - V_b \end{Bmatrix}$$

In another expression for a combinationof all the classifier, the final score is:

$$Vfinal \;=\; V_b + \sum_{i \in x} w_i\, L + \sum_{i \in y} w_i (V_i - V_b) - \sum_{i \in z} (w_i\, S) - \sum_{i \in d} \big(w_i (V_b V_i)\big) \quad (3)$$

Where:

x= $\{V_b < V_i \& L < V_i - V_b\}$
y=$\{V_b < V_i \& L \geq V_i - V_b\}$
z=$\{V_b >= V_i \& S < V_i - V_b\}$
d=$\{V_b >= V_i \& S >= V_i - V_b\}$

For calculatingthe weights of the classifiers one set of n sets of data is used only. This set is split into training and test, and hence learning all the classifiers/regression from this set. The accuracy/score of each classifier will be estimated from the test part of this set. These weights will be used for the proposed method for classifying the other n-1 sets.

5. Performance metrics

Any proposed system(model)needs performance measures, whether they are regression or classification. As on the proposed system There are many metrics that are used to monitor and measure the performance of the model, which tells us how good or bad the rating is and whether the model is making progress or not.

In regression models, the outputs are continuous values, so metrics that calculate the difference between the real values and the expected values should be used in Equation (4). Some of these evaluation metrics are [14][15].

1- Mean Absolute Error (MAE) that can be defined by Equation (4):

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n} | yi - zi | \quad (4)$$

Where yi is real, and zi is predicated output.

2- Mean Absolute Percentage Error (MAPE) that can be defined by Equation (5):

$$\text{MAPE}=\frac{100}{n}\sum_{i=1}^{n} \frac{|yi-zi|}{yi} \quad (5)$$

3- Mean Squared Error (MSE) that can be defined by Equation (6):

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n} | yi - zi |^{2} \quad (6)$$

4- Root Mean Squared Error (RMSE) that can be defined by Equation (7):

$$\text{RMSE}= \sqrt{\text{MSE}}. \quad (7)$$

6. Experiment Result &Evaluation

The experiment was done using python 3.6 where many libraries were used such as Natural Language Toolkit (nltk), scikit-learn, NumPy, re, json, and xlrd that have been used for various tasks. It was implemented on laptop with 64 OS, 8 GB memory, intel core i7 processor. This section showsthe dataset, result, and evaluation.

6.1Dataset

The Hewlett essay scoring dataset was used in this work that represents answers to students from different levels. It consists of eight groups of different subjects. The levels of the students were seventh grade to the tenth. All data was stored in an excel file with six columns: (i) essay-id (the essay number), (ii) essay-set (the set number to which the essay belongs), (iii) essay(the student essay), (iv) rater1 (the first evaluators), (v) rater2 (the second evaluators), (vi) domain1 (the average or total score of the first and second correctors).All of these essays were manually evaluated by two human evaluators.

Table 1: Dataset Subjects.

# Set num.	Set Size	Average Length of Essays	Range For Score
Set1	1783	350 words	2 – 12
Set2	1800	350 words	1 – 6
Set3	1726	150 words	0 – 3

Set4	1772	150 words	0 – 3
Set5	1805	150 words	0 – 4
Set6	1800	150 words	0 – 4
Set7	1569	250 words	0 – 30
Set8	723	650Words	0– 60

7. Result and discussion

This section presents experiment test results that have been founded during the implementation of the proposed AES system.As was explained previously, three methodologies were used for classification.The first one is by using each one of the nine classifiers as an independent classifier. The second methodology is by using the best three classifiers/regression for getting the final score as the average of their scores. Finally, the third methodology is by using augmented combination whereeach classifier has weight, and the final score is evaluated according to Equation (3):

All these methodologies were tested using the mentioned dataset.The used dataset has 8 sets (set_1 …set_8)and therefore, one set is used as training, and the other set is used as tests for improving the validity of the used methods.

In this work, four metrics were used such as MSE, MAE, RMSE, MAPE.Practically the experiment was repeated eight timeswhere one set is taken as test set and the other are training but the results were huge therefore the average of these tests were written as shown in Tables 2 to 5 for MSE, MAE, RMSE, MAPE respectively.

Figures2 to 5shows the graphical representation of Tables 2 to Table 5 respectively.

Table 2:Average values of MSE of eight tests for all the used algorithms.

Data	GNB	NB	KNN	DT	RFR	RFC	DTC	LOR	LR	AC	AS
Set1	2.031	2.459	1.712	1.866	1.368	1.853	1.852	1.658	1.57	1.555	1.332
Set2	1.32	1.094	1.258	1.281	0.933	1.22	1.215	1.249	1.193	1.204	0.708
Set3	1.342	1.436	0.751	0.981	0.732	0.8	1.023	0.842	0.566	0.734	0.818
Set4	1.289	1.335	0.69	0.868	0.657	0.734	0.883	0.791	0.686	0.65	0.755
Set5	1.399	1.754	0.939	1.098	0.847	0.986	1.082	1.018	0.681	0.805	0.813
Set6	0.996	1.761	1.048	1.161	0.906	1.036	1.175	1.348	0.761	0.939	0.873
Set7	5.756	5.642	4.231	4.048	3.441	4.093	3.72	4.011	4.085	4.062	3.303
Set8	6.052	3.57	6.423	5.548	4.58	6.115	6.412	6.444	12.937	4.842	3.815

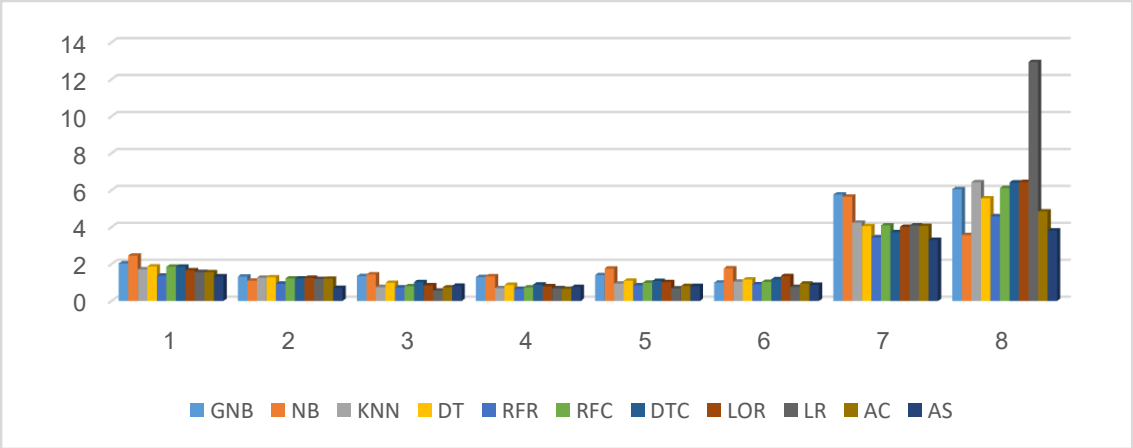

Figure 2:**Graphical representation of averagevalues of MSE of eight tests for all the used algorithms.**

Table 3: Average values of MAE of eight tests for all the used algorithms.

Data	GNB	NB	KNN	DT	RFR	RFC	DTC	LOR	LR	AC	AS
Set1	1.126	1.204	1.032	1.048	0.928	1.07	1.032	1.031	1.037	1.011	0.77
Set2	0.984	0.797	0.936	0.903	0.807	0.923	0.871	0.936	0.953	0.928	0.606
Set3	0.882	0.93	0.606	0.725	0.673	0.636	0.739	0.659	0.624	0.627	0.653
Set4	0.872	0.9	0.59	0.686	0.648	0.614	0.695	0.636	0.688	0.605	0.638
Set5	0.887	1.026	0.701	0.78	0.72	0.727	0.772	0.714	0.637	0.661	0.628
Set6	0.713	1.017	0.79	0.823	0.777	0.776	0.845	0.888	0.7	0.772	0.69
Set7	1.592	1.514	1.264	1.28	1.165	1.249	1.26	1.252	1.218	1.213	1.101
Set8	1.869	1.257	1.829	1.735	1.527	1.761	1.751	1.841	2.312	1.647	1.225

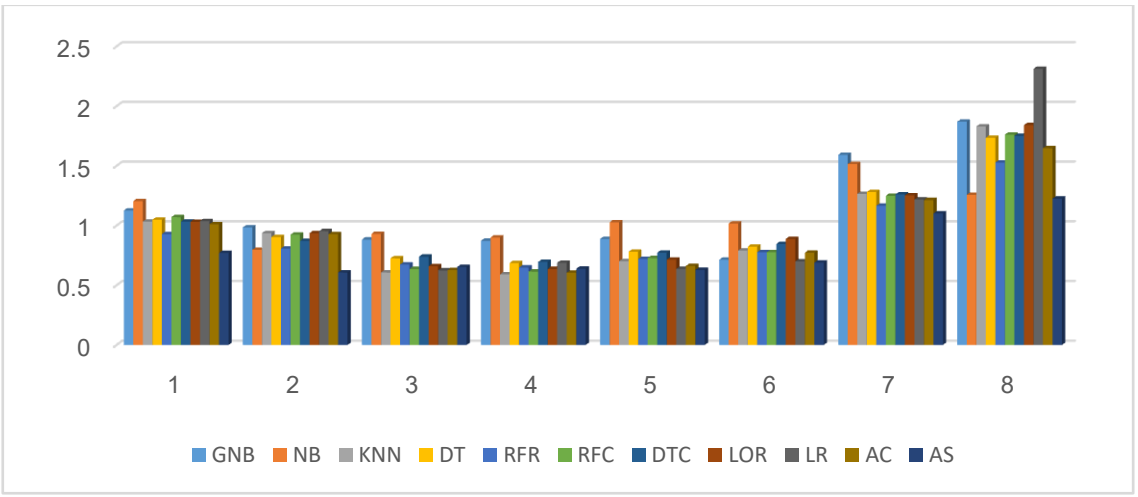

Figure 3: Graphical representation of average values of MAE of eight tests for all the used algorithms.

Table 4: Average values of RMSE of eight tests for all the used algorithms.

Data	GNB	NB	KNN	DT	RFR	RFC	DTC	LOR	LR	AC	AS
Set1	1.294	1.471	1.224	1.282	1.062	1.258	1.267	1.206	1.165	1.162	1.024
Set2	1.125	1.025	1.09	1.093	0.926	1.079	1.073	1.082	1.052	1.062	0.83
Set3	1.151	1.193	0.858	0.982	0.841	0.885	1.002	0.91	0.749	0.846	0.9
Set4	1.117	1.15	0.827	0.926	0.803	0.851	0.937	0.876	0.822	0.802	0.866
Set5	1.155	1.31	0.94	1.028	0.879	0.963	1.022	0.966	0.794	0.875	0.887
Set6	0.977	1.287	1.005	1.065	0.922	0.999	1.071	1.101	0.854	0.954	0.922
Set7	1.902	1.923	1.575	1.638	1.427	1.575	1.599	1.577	1.477	1.524	1.432
Set8	2.045	1.548	2.037	1.978	1.709	1.989	1.996	2.037	2.522	1.837	1.478

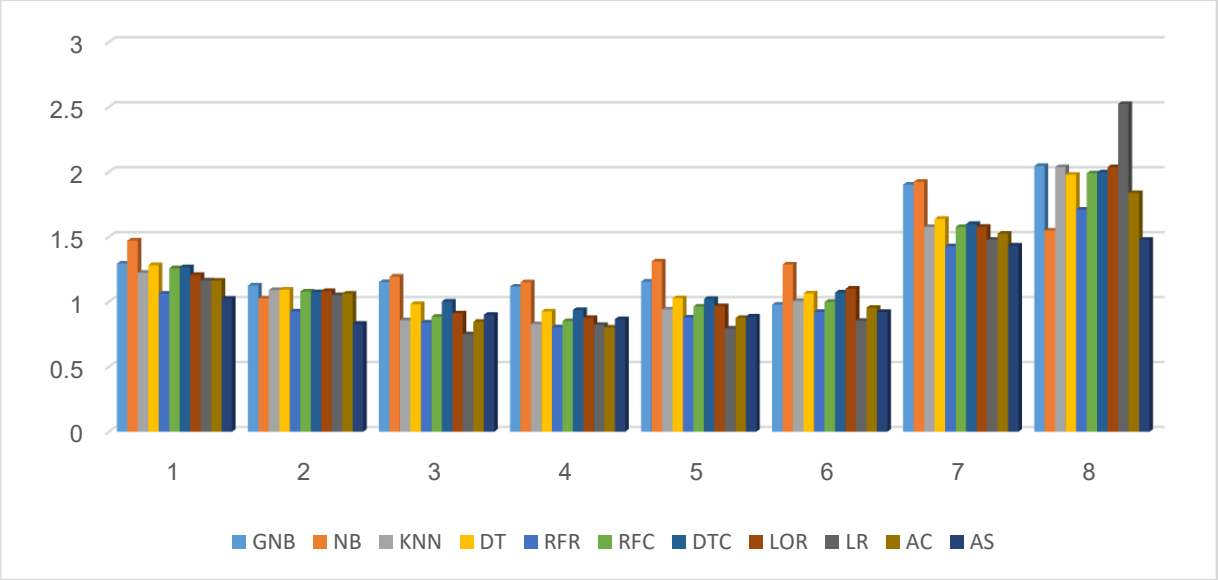

Figure 4:Graphical representation of average values of RMSE of eight tests for all the used algorithms.

Table 5: Average values of MAPE of eight tests for all the used algorithms.

Data	GNB	NB	KNN	DT	RFR	RFC	DTC	LOR	LR	AC	AS
Set1	31.2	32.2	29.5	28.9	24.7	29.4	27.8	29.9	29.8	29.1	19
Set2	44.2	34.2	43.7	41	37.2	42.7	39.7	43.9	44.6	43.4	28.7
Set3	57.1	58.9	31.9	38.1	34.4	33.1	38.6	34.4	33.3	33	34.1
Set4	49.3	42.6	26	32.2	28.7	27.7	32.9	28.5	27.3	24.9	24.2

Set5	42.8	54.6	28.3	32.5	29.1	29.4	31.9	29	28.4	27.2	28.3
Set6	35.4	47	30.2	32.1	29.6	30	32.5	33.9	29.4	30.7	28.9
Set7	43.7	45.4	33.4	36.5	31.4	33.6	37.2	33.4	32.9	31.9	30.8
Set8	48.2	31.2	46.1	45.1	37.5	45.1	40.9	46.7	49.4	44.1	26.9

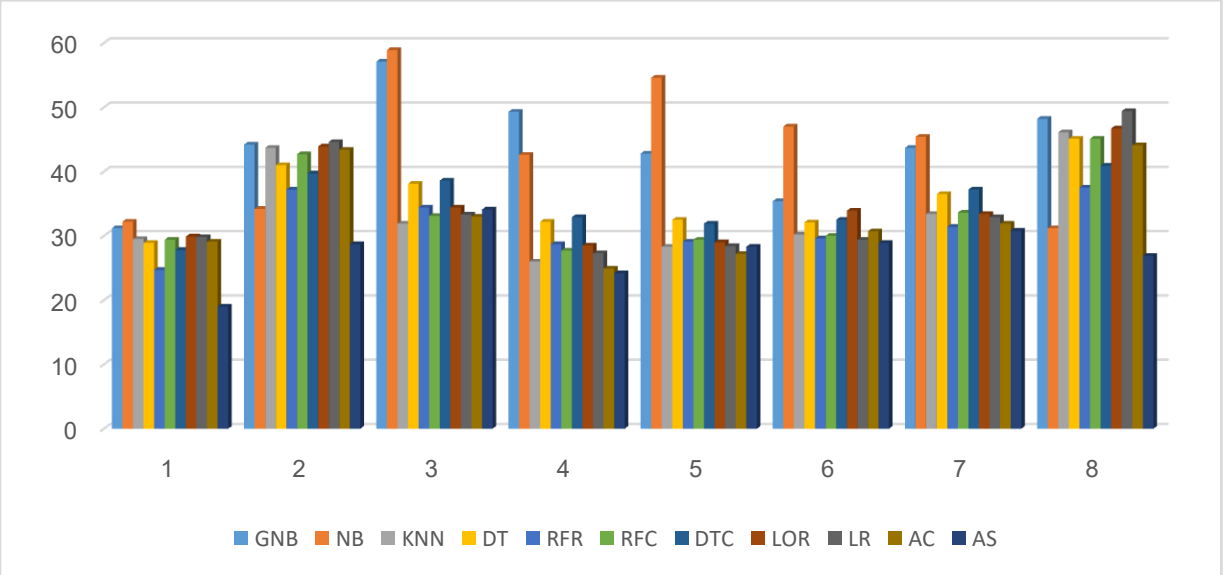

Figure 5: Graphical representation of average values of MAPE of eight tests for all the used algorithms.

As we can see from these Tables and Figures, average score and augmented combination have the lowest error rate in most tests. Also, they are more stable than the other classifiers where there is not any huge rise in errors in all the tests. There are rising in the errors rates in the augmented method results from the weak classifiers/regression.

8. Conclusions

In this work, a complete system for automatic essay scoring was implemented as a part of e-learning. It differs from the usual traditional online tests where the proposed system uses natural language processing techniques and machine learning methods that are applied in the e-learning system. three methodologies were used for classification\regression tasks; (i) each one of the nine classifiers was used as an independent classifier; (ii) the best three classifiers/regressions algorithmswere used for scoring the essay by the average of their scores, and (iii) using augmented combination of all the classifiers\regression to calculate the final score. Many metrics were used such as MSE, MAE, RMSE, MAPE for insurance the confidence of the proposed system. All the tests were taken on different sets where the test set is different in the used question and answers from the questions and answers of the training sets which give our experiment more reliability for using it in practical exam tests.

Also, the results of our test show that using the proposed features were essay-independent where they can be used for any essay scoring system with little limitations such as using mathematical equations. These features are different in their weights therefore calculating feature importance was done by assigning specific weights to each feature based on its importance. This process of choosing the features reduces the dimensions and has great importance in the good prediction of the model.

As future works, we suggest that extracting the synonyms of the exact meaning (sense) of each word in the optimal answer which make the scoring more depend on the content of the essayand in turn it gives high rank for the semantic of the essay instead of using essay-independent features only. Also, tis work can be used for other complex languages such as Arabic, Chinese and other languages.

References

[1] L. F. Motiwalla, "Mobile learning: A framework and evaluation," *Comput. Educ.*, vol. 49, no. 3, pp. 581–596, 2007, doi: 10.1016/j.compedu.2005.10.011.

[2] R. Shiva Shankar and D. Ravibabu, *Digital report grading using NLP feature selection*, vol. 758. Springer Singapore, 2018.

[3] K. Zupanc and Z. Bosnić, "Increasing accuracy of automated essay grading by grouping similar graders," *ACM Int. Conf. Proceeding Ser.*, 2018, doi: 10.1145/3227609.3227645.

[4] L. Bin and Y. Jian-Min, "Automated essay scoring using multi-classifier fusion," in *International Conference on Information and Management Engineering*, 2011, pp. 151–157.

[5] J. Kur, M. Lungu, and O. Nierstrasz, "Top-Down Parsing with Parsing Contexts A Simple Approach to Context-Sensitive Parsing," no. Iwst, 2014.

[6] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "Enriching Automated Essay Scoring Using Discourse Marking • Abstract 2 . Hybrid Feature Methodology," *Discourse*.

[7] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, "Automated essay scoring using the KNN algorithm," *Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008*, vol. 1, pp. 735–738, 2008, doi: 10.1109/CSSE.2008.623.

[8] X. Peng, D. Ke, Z. Chen, and B. Xu, "Automated chinese essay scoring using vector space models," *2010 4th Int. Univers. Commun. Symp. IUCS 2010 - Proc.*, no. October, pp. 149–153, 2010, doi: 10.1109/IUCS.2010.5666229.

[9] M. Mahana;, M. Johns;, and A. Apte;, "Automated Essay Grading using Machine Learning Algorithm," *J. Phys. Conf. Ser.*, vol. 1000, no. 1, pp. 3–7, 2018, doi: 10.1088/1742-6596/1000/1/012030.

[10] M. Zhang, S. Hao, Y. Xu, D. Ke, and H. Peng, "Automated Essay Scoring Using Incremental Latent Semantic Analysis.," *J. Softw.*, vol. 9, no. 2, pp. 429–436, 2014.

[11] R. Setiadi Citawan, V. Christanti Mawardi, and B. Mulyawan, "Automatic Essay Scoring in E-learning System Using LSA Method with N-Gram Feature for Bahasa Indonesia," *MATEC Web Conf.*, vol. 164, 2018, doi: 10.1051/matecconf/201816401037.

[12] Z. Chen and Y. Zhou, "Research on Automatic Essay Scoring of Composition Based on CNN and OR," *2019 2nd Int. Conf. Artif. Intell. Big Data, ICAIBD 2019*, pp. 13–18, 2019, doi: 10.1109/ICAIBD.2019.8837007.

[13] Mr. P. V. Hari Prasad;, G. Himaja;, Ch.Abhigna;, K.Saroja;, and K. N. Venkatesh,

"Automated Essay Grading System Using Learning," pp. 952–954, 2020.

[14] A. Botchkarev, "Evaluatingperformanceofregressionmachinelearningmodels-SSRN-id3177507.pdf." 2018.

[15] E. D. Liddy, "Natural language processing," *Syracuse University School of Information Studies - Faculty Scholarship*, vol. 19, no. 2, pp. 131–136, 2021.