CONTEMPORARY EDUCATIONAL TECHNOLOGY, 2018, 9(4), 423-436
https://doi.org/10.30935/cet.471024

Home-Grown Automated Essay Scoring in the Literature Classroom: A Solution for Managing the Crowd?

Kutay Uzun
Trakya University, Turkey

Submitted: 23.01.2018 Accepted: 05.10.2018 Published: 16.10.2018

Abstract

Managing crowded classes in terms of classroom assessment is a difficult task due to the amount of time which needs to be devoted to providing feedback to student products. In this respect, the present study aimed to develop an automated essay scoring environment as a potential means to overcome this problem. Secondarily, the study aimed to test if automatically-given scores would correlate with the scores given by a human rater. A quantitative research design employing a machine learning approach was preferred to meet the aims of the study. The data set to be used for machine learning consisted of 160 scored literary analysis essays written in an English Literature course, each essay analyzing a theme in a given literary work. To train the automated scoring model, LightSide software was used. First, textual features were extracted and filtered. Then, Logistic Regression, SMO, SVO, Logistic Tree and Naïve Bayes text classification algorithms were tested by using 10-Fold Cross-Validation to reach the most accurate model. To see if the scores given by the computer correlated with the scores given by the human rater, Spearman's Rank Order Correlation Coefficient was calculated. The results showed that none of the algorithms were sufficiently accurate in terms of the scores of the essays within the data set. It was also seen that the scores given by the computer were not significantly correlated with the scores given by the human rater. The findings implied that the size of the data collected in an authentic classroom environment was too small for classification algorithms in terms of automated essay scoring for classroom assessment.

Keywords: *Automated essay scoring; Literary analysis essay; Classification algorithms; Machine learning*

Introduction

Assessing the written products of the students is one of the most time-consuming activities for teachers, especially in classes with a relatively large number of students. This is because finalizing a written product involves many processes such as providing/activating background knowledge of the students, having them write a draft essay or a paper, giving feedback for corrections/adjustments, and recollecting the essays for a final evaluation before moving on to the next assignment. In this sequence, the workload appears to be mostly on students since they are the ones to write the whole essay and correct or adjust the points suggested by the teacher. However, teacher feedback in the evaluation of writing is a crucial point which improves the skill and builds relationships between the teacher and the student (Lee, 2014; Yang, 2012).

Despite the generally acknowledged importance of individualized teacher feedback in the teaching of writing, its real-life implementation is rather a strenuous act for teachers in many institutions of various educational levels due to issues of excessive workload on behalf of teachers resulting from the high number of students they have to deal with each semester, which is still an unsolved problem (Ritten, 2012) that is also criticized by Baker (2014). The ideal conditions in terms of teacher feedback in writing are also considered to be unattainable in practical classroom conditions by Hyland and Hyland (2006).

The problem grows even deeper in literature classes where students are expected to demonstrate content knowledge through establishing connections among historical facts, social or psychological realities, and a particular author's style in terms of the way s/he handles his or her subject matter. As demanding a task as it is on behalf of students, it also puts teachers under an immense amount of workload throughout which they have to assess each student product, which is usually in the form of an academic essay in literature courses (Paran, 2006), in terms of organization, form and content.

Referring back to the problem of crowded classes in the process of feedback provision and adding to this the increased workload of essay evaluation within the context of literature courses, the present study attempts to provide a technology-oriented solution to the mentioned problem through testing the accuracy of Automated Essay Scoring (AES) by using context-bound training data and comparing the automatically acquired scores with the scores given by the course instructor.

Automated Essay Scoring

AES systems can be defined as computer systems which are able to evaluate and score written texts (Shermis & Burstein, 2003). Aiming primarily to reduce the amount of time, expense, reliability and generalizability-related problems, such systems can be designed to use both in large-scale assessment and small-scale classroom assessment (Bereiter, 2003; Page, 2003).

Although there is a variety of academic fields such as computer programming, language teaching or mathematical knowledge which use AES systems, Powers et al. (2015) note that the working principles of AES systems can be listed as follows in general:

1. Creation of a data set to be used as 'training data' (essays and their scores given by experts)
2. Extraction of the features in the training data
3. Development of a model which links the essay to its score
4. Utilizing the developed model to score the unscored essays

Paraphrasing the sequence of the machine learning process provided by Powers et al. (2015) within the context of education, it can be stated that the first thing needed for the successful evaluation of student essays by a computer is a set of essays already scored by a human. Following this, the AES system analyses the essays and tries to figure out which features student essays have. In the third step, the computer develops a model by working out the relationship between the features of student essays and the scores given by the human expert, understanding the scoring criteria. Finally, the essays which have not been evaluated yet are scored by the computer using the same criteria. Naturally, this process has its own strengths and drawbacks.

According to Dikli (2006), the main advantage of using AES systems is that they can provide scores and feedback immediately. Moreover, Yang et al. (2012) state that AES systems are cost-effective and access to these systems are fairly easy since several AES systems are available for online use. In a similar fashion, Weigle (2013) indicates that AES systems are able to reduce the amount of labor in essay scoring to a great extent, which is needed for designing essay prompts, creating scoring rubrics, scoring and providing feedback. The subjective nature of essay evaluation can also be turned into a more standardized process through AES by increasing internal consistency (Norbert & Williamson, 2013). From a functional perspective, Wilson (2018) proposes that AES systems can be used to identify learners who are at risk of failing standardized exams with a writing component. With an emphasis on teacher's heavy workload, Westera, Dascalu, Kurvers, Ruseti and Trausan-Matu (2018) argue that using AES systems substantially reduces teachers' workload without jeopardizing precision in scoring.

The process of AES, however, is not without its disadvantages. In terms of learner psychology, AES systems may get students worried about their grades, resulting in a stronger preference for human graders, objecting to the scores gives by a computer and anxiety since they doubt that a machine could understand concepts or ideas (Barker, 2011; Lai, 2010; Landauer et al., 2000). On the technical side, the creation or the gathering of an appropriate corpus as training data as suggested by Powers et al. (2015) can be problematic. Moreover, Deane (2013) argues that AES scoring criteria can be manipulated.

The validity and reliability issues in AES have also been extensively studied in the recent years. According to Attali (2013), if AES systems are to replace human raters of essays with an acceptable level of validity, they must be using the same criteria as humans and the problem with the human criteria of essay scoring is that they are not clear enough to be taught to computers. However, comparing human scores to the scores given by the computer is an approach which is generally adopted to study the validity and reliability of automatically given essay scores.

Attali and Burstein (2006) report strong correlations between human scores and computer scores. In a similar vein, Shermis and Hamner (2013) conclude that the correlation between essay scores given by a human and a computer rater is as strong as the correlation between the scores given by two humans. Imaki and Ishihara (2013) reach a similar conclusion in that they reveal a correlation coefficient between human raters and the computer which is as strong as the correlation among the essay scores given by nine human raters. On the validity and reliability of AES systems being acceptable, a striking result is put forth by Powers et al. (2015), who conclude that while automatically given scores correlate more strongly with the scores given by more expert raters, the scores correlate weakly with those given by less experienced raters. As an alternative way of ensuring validity, it should also be noted that Yamamoto, Umemura and Kawano (2018) suggest training AES models that take rubrics which have previously been validated as their basis.

Based on the relevant literature, it can be stated that AES systems can save a lot of time and effort while they have the potential to cause a certain amount of discontent among the students. In addition, according to relevant studies, AES systems can provide an acceptable level of convergent and discriminant validity along with inter-rater reliability when they are trained accurately.

Context-boundedness of Classroom Assessment

The definition of the term 'classroom assessment' is a self-explanatory one in terms of the role of context in classroom assessment. According to McMillan (2007), classroom assessment is "the collection, evaluation, and use of information to help teachers make decisions that improve student learning" (p. 8). As suggested in the definition, the aim of classroom assessment is assessing student work in order to promote learning (Gavriel, 2013), and thus, teachers' awareness of classroom assessment comes out as a necessity in the teaching and learning environment (Greenstein, 2010).

Unlike large-scale assessment, classroom assessment is a highly contextual practice in which the teacher actively participates in the learning process of the learners no matter what teacher role s/he adopts. What distinguishes classroom assessment from large-scale assessment is that teachers are already familiar with their students, their skills, abilities and potentials. Moreover, as individuals and groups, the stakeholders interact, communicate and build relationships within the context of classroom assessment, which is a quality that is non-existent in large-scale assessment. In the latter, the raters are typically independent and no relationship exists between the rater and the test-taker (Fulcher & Davidson, 2007).

The context-boundedness of classroom assessment can also be inferred from the study of Shermis and Di Vesta (2011), who point out that classroom assessment is the projected accumulation of data to investigate the effects of the processes of teaching and learning. Furthermore, feedback based on the findings are crucial in that they affect the process of decision making related to the efficiency of instruction, the changes needed, the consideration of learner differences, the observation of learner improvement and goal-setting. As seen from the suggested effects of feedback as a result of assessment by Shermis and Di Vesta (2011), assessment findings can be benefitted from to make contextual changes or taking the contextual factors into account.

According to Brookhart and Bronowicz (2003, p. 225), classroom assessment is composed of eight dimensions, which are the purpose, method, criteria and quality of assessment along with the nature of teacher feedback, teacher attitudes towards classroom assessment and the students. The last dimension suggested by Brookhart and Bronowicz (2003) is the institutional policy towards assessments. Similar to Shermis and Di Vesta's (2011) study, the dimensions proposed by Brookhart and Bronowicz (2003) also refer to the context in which classroom assessment is practiced and it can be understood that changes in one or more in these contextual factors may affect the practice of classroom assessment.

In a case study with Singaporean English Language teachers investigating the contextual factors which affect classroom assessment, Chih-Min and Li-Yi (2013) conclude that the high-stakes examinations which the students take, the management policies of the school, the physical environment surrounding the teaching and learning processes, the parents of the students, the availability of teacher training and practices in the classroom influence the way assessment in carried out in the classroom. Similar to these findings, Anderson and Ben Jafaar (2006) indicate that the assessment and educational policies influence classroom practices. Other factors that affect classroom assessment practices include teacher expertise in terms of experience (Darling-Hammond et al., 2009), the subject that is handled (Anderson and Ben Jafaar, 2006) and the methods of assessment and teacher beliefs (Young & Kim, 2010).

In short, the literature supports the notion that classroom assessment is a highly contextual practice which is affected by factors such as teachers' and students' beliefs and attitudes, physical environment and policy environment, which create the context of assessment when they come together. However, the literature appears to indicate a gap in terms of the use of AES systems for classroom-assessment purposes. In this respect, the present study aims to test the use of AES in classroom assessment and fill a gap in the literature by attempting to develop an AES system which makes strict use of contextual data in order to find out if computational algorithms can work out the context-dependent rules of scoring literary analysis essays.

Purpose of the Study

Taking the relevant literature into account, it is argued in the present study that the training data for AES systems should be as context-specific as possible so that the automated scoring of the essays can be carried out in a way that is closer to real-life conduct in terms of classroom assessment. In this respect, the present study aimed to test the accuracy of a home-grown AES system whose training data consisted of student essays which were graded by the course instructor. Secondarily, the study aimed to compare the essay scores given by the computer and the course instructor with the aim of revealing if they were correlated.

The research questions of the study are as follows:

a. Can literary analysis essay scoring criteria be adapted to an automated essay scoring environment?
b. Does using context-dependent data as training data for machine learning produce accurate results in the automated scoring of literary analysis essays in English Literature courses?
c. Is there a relationship between the literary analysis essay scores given by the instructor of English Literature course and the computer?

Methodology

Study Design

Since the automated scoring of student essays and the correlation between human-scored and machine-scored essays were sought for in the present study, a quantitative research method utilizing machine learning algorithms was preferred. According to Dornyei (2007), a major advantage of quantitative research is that reliable and generalizable results can be obtained through quantitative research methods, which also allow the researcher to look for relationships among variables. Moreover, machine learning, by nature, is a mathematical approach to problem solving and it makes heavy use of statistical and discrete analyses relying on numerical conversions and the probability theory (Stankova, Balakshiy, Petrov, Shorov & Korkhov, 2016). Considering the advantages of quantitative research methods and the nature of machine learning, the research problem was approached from a quantitative perspective in the present study.

Procedures

The study followed the following phases:

1. Data collection
2. Scoring the collected essays
3. Feature extraction
4. Feature selection
5. Testing classification algorithms / Choosing the most accurate one
6. Having the computer predict the scores for the unscored essays
7. Checking for the correlation between human-scored and computer-scored essays

Data Collection

As argued in the 'Purpose' section, the study aimed to create a context-specific database of students' literary analysis essays. For this reason, essay data was collected through the natural flow of the English Literature course, resulting in a sum of 160 essays on the following topics:

1. Madness in Hamlet by William Shakespeare (n = 16)
2. Death in Hamlet by William Shakespeare (n = 27)
3. Vengeance in Volpone by Ben Johnson (n = 16)
4. Greed in Volpone by Ben Johnson (n = 20)
5. Mastery Conflicts in Robinson Crusoe by Daniel Defoe (n = 16)
6. Love in Pride and Prejudice by Jane Austen (n = 19)
7. The significance of Journeys in Pride and Prejudice by Jane Austen (n = 15)
8. Inversion of Gender Roles in The Importance of Being Earnest by Oscar Wilde (n = 16)
9. The nature of marriage in The Importance of Being Earnest by Oscar Wilde (n = 15)

The instructions given for the essays required the participants to write literary analysis essay of 400-600 words explaining how the given theme/motif was handled in the related literary work by the author. Upon the collection of the essays, it was seen that the whole corpus of literary analysis essays included 160 essays and 53840 words, resulting in an average word count of 299 per essay.

Each essay was scored by the course instructor using the Bauer and Kohut Argumentative Writing Rubric, which helps the rater score an argumentative essay from 1 to 6, each point corresponding to a level in Bloom's taxonomy of educational objectives (Bauer, 2016). In order to ensure the reliability of the course instructor's scoring of the essays, 30% of the essays were re-scored 6 weeks after the first scoring. The scores for the first and second evaluation was investigated in terms of Cohen's Kappa to reveal the level of agreement, producing a statistically significant Kappa value, $K = .861$, $p < .001$, indicating that the scores given by the course instructor had a high level of intra-rater reliability.

Data Analysis

For the extraction of textual features within the corpus, LightSIDE (Mayfield & Rose, 2013), which is a textual feature extraction, organization, machine learning and prediction software, was used. For the extraction of the features, the 'Basic Features', 'Parse Features' and 'Stretchy Patterns' plug-ins within the software were used. As for the basic features, Unigrams, Bigrams, Trigrams, POS Bigrams, POS Trigrams, Word/POS Pairs and Line Length features were

analyzed and the features 'Count Occurrences', 'Normalize N-gram Counts', 'Include Punctuation', 'Stem N-Grams' and 'Track Feature Hit Location' functions were ticked. For the 'Parse Features' plug-in, 'Production Rules', 'Leaf Productions' and 'Dependency Relations' were preferred to include all Parts of Speech, terminal branches and dependencies among POS. In the Stretchy Patterns plug-in, pattern length and gap length were left in their defaults (2-4 and 1-2 respectively), but surface words and POS tags in patterns were also included in the analysis. Once the analysis for feature extraction was initiated, it was seen that the 1GB RAM allocated to LightSIDE by default was not sufficient for the process, for this reason, the amount of RAM allocated to the software was increased by 2GB.

Upon the completion of feature extraction, 'Filter Feature Values' function under the 'Restructure Data' tab of LightSIDE was used to remove the features which had a Kappa value below .2 since this value is the threshold value for a fair agreement among features (Viera & Garrett, 2005). Moreover, eliminating features with a Kappa below .2 was considered necessary by the researcher because the corpus of student essays included 9 essay prompts which were in the same style but different in terms of their topics. Therefore, it was aimed that the machine learning model would include only the common features across texts, which were expected to be the general features related to argumentative essays on literary topics.

The next step for the processing of the data was to find the most accurate classification algorithm for the prediction of unscored essays. According to Santos et al. (2012), Logistic Regression, Functional Tree, Random Forest, LAD Tree and Naïve Bayes algorithms are among the most accurate classification algorithms within the context of essay scoring. Similarly, Kumar and Sree (2014) count Naïve Bayes, Logistic Regression and Support Vector Machines (SMO) among the most accurate ones. Taking these studies into account, Logistic Regression, Naïve Bayes, SMO, SVM, Random Forest and Functional Tree algorithms were tested to see how accurate results they produce in the automated scoring of essays using 10-fold Cross-Validation technique, in which the algorithm uses a random 90% of the data set as training data and a random 10% of it as the test data, using a different group and testing the algorithm 10 times. The results were investigated in terms of prediction accuracy and Kappa values, the latter of which indicates an acceptable level of agreement when it is above .40 (Sakiyama et al., 2008). Confusion matrices produced for each algorithm were also scrutinized to reveal how the algorithms performed. Then, the algorithm with the best results was saved to be used in prediction.

The last phase of data analysis included the scoring of 19 new essays on 'Realization in Mrs. Dalloway by Virginia Woolf' and 'The Absurd in Waiting for Godot by Samuel Beckett'. In this phase, the essays were first scored by the course instructor. Following the scoring of the course instructor, the 'Predict Labels' function of LightSIDE was run using the trained model, which was saved beforehand, to predict the scores for the 19 essays mentioned above. In order to test the relationship between human-scored and computer-scored essays, Spearman's Rank Correlation Coefficient was calculated since the number of the essays was not sufficient to obtain normally distributed data.

Findings

Following the collection and the scoring of literary analysis essays, feature extraction, feature selection, testing algorithms and choosing the most accurate one, and score prediction for the unscored essays took place in the study.

Feature Extraction, Feature Selection and Testing Classification Algorithms

As mentioned in the methodology section of the present study, feature extraction was performed by means of 'Basic Features', 'Parse Features' and 'Stretchy Patterns' plug-ins of LightSIDE. Feature extraction resulted in the identification of 76168 textual features in the whole dataset. However, the features with Kappa values lower than .20 were removed from the feature table since they were considered to be the features pertaining to a particular essay topic. After the removal of those features, the restructured feature table was seen to have 141 features.

The restructured feature table was used to test the algorithms mentioned in the methodology section. 10-fold Cross-Validation was used to produce the prediction accuracy and Kappa values. These values are presented in Table 1.

Table 1. Comparison of the Classification Algorithms

Algorithm	Prediction Accuracy	Kappa
Logistic Regression	.55	.44
Support Vector Machines	.52	.41
SMO	.47	.34
Random Forest	.46	.32
Naïve Bayes	.44	.29
Functional Tree	.31	.15

In Table 1, the prediction accuracy and Kappa values are presented for each of the classification algorithms that were tested. According to the test results, Logistic Regression was the most accurate classification algorithm with a prediction accuracy of 55% and a Kappa value of .44. Support Vector Machines, on the other hand, was the second most accurate classification algorithm with 52% prediction accuracy and a Kappa value of .41. The least accurate algorithm, according to the results, was the Functional Tree with 31% prediction accuracy and a Kappa value of .15. The results indicated that Logistic Regression, which was the most accurate classification algorithm with the current dataset, could be used for predicting the unscored essays. For that reason, the confusion matrix for the model trained with the Logistic Regression algorithm is presented below in Table 2.

Table 2. Confusion Matrix for the Model Trained with the Logistic Regression Algorithm

A / P*	x1	x2	x3	x4	x5	x6
x1	9	1	1	4	1	0
x2	1	14	3	8	0	0
x3	1	3	18	5	6	0
x4	3	3	4	28	4	2
x5	1	0	5	5	12	0
x6	0	1	2	7	1	7

*: A – Actual, P – Predicted

In the confusion matrix in Table 2, it can be observed that x1 as the lowest score was accurately predicted in 9 essays out of 16, having been confused with x4 4 times. The second lowest score, x2, was predicted accurately in 14 essays out of 26, having x4 as the class with which it was mostly confused. The third score, x3, which had been given by the researcher 33 times, was accurately predicted in 18 essays, having been confused with x4 and x5, the latter one of which was predicted accurately 12 times by the algorithm. The highest score, x6, was rather problematic in that it was predicted accurately in 7 essays out of 18 but confused with x4 in another 7.

Score Prediction and Comparison against Human Scoring

The dataset whose scores were to be predicted contained 19 essays analyzing the themes of 'Realization' in Mrs. Dalloway by Virginia Woolf and 'The Absurd' in Waiting for Godot by Samuel Beckett. The human rater scored the essays in this data set before the prediction of the computer so as not to be biased by the predictions. The score agreement between the human rater and the computer are given below in Table 3 in the form of a matrix showing frequencies.

Table 3. The Comparison of the Scores Given by the Human Rater and the Computer

H / C*	x1	x2	x3	x4	x5	x6
x1	0	0	0	0	0	0
x2	0	2	0	0	1	0
x3	0	0	2	1	1	0
x4	0	2	0	1	0	0
x5	0	1	0	3	0	0
x6	0	1	0	3	0	1

*: H- Human, C- Computer

As seen in Table 3, there is no essay which was marked '1' in the final dataset of 19 essays. Moreover, 3 essays were assigned the score of '2' by the human rater while 6 essays were given the same score by the computer. For the score of '3', the ratio of the frequency is 4 against 2 in favor of the human rater. The third highest score, '4', was given by the human rater 3 times but 8 times by the computer. The second highest score, '5' was given by the human rater 4 times but 2 times by the computer. Similar to the confusion matrix of the training model, the score '6' was again a point of confusion, having been given by the human rater 5 times but only once by the computer.

Although it is possible to infer from the matrix in Table 3 that the scores given by the human rater and the computer are not in agreement, it is necessary to compute the correlation coefficient to reach more concrete evidence. Upon computation, Spearman's Correlation Coefficient indicated that the scores given by two parties are not significantly correlated (r_s = .246, p > .05, r^2 = .06), which showed that the scores assigned by the course instructor and the algorithm did not agree.

Discussion and Conclusion

The present study aimed to find out if general argumentative/expository essay criteria could be adapted to an automated essay scoring environment, if using entirely context-dependent data was sufficient to train the computer to produce accurate results in literary analysis essays and if the essay scores given by the computer correlated with the scores given by the course instructor.

The findings indicated that, in order to develop a generic model which can be used for multiple literary analysis essay prompts regardless of the topic, only .019% of the total amount of extracted textual features could be utilized with the dataset of the study, since the rest of the features did not seem to be common for all the essays within the dataset. However, the exclusion of features does not necessarily pose a problem because as long as the results are accurate, the low percentage of features that have an acceptable level of agreement with the whole data set is not considered to be a problem since having fewer features in the training model makes the training process less complex, contributing to the training of a more generic model (Abu-Mustafa et al., 2012), which was among the aims of the study.

As for training an accurate model, logistic regression algorithm was the most accurate one with the dataset that was used for the purposes of this study, producing an adequate Kappa value according to Sakiyama et al. (2008). Nevertheless, according to the findings, the percentage of prediction accuracy was only slightly above the half of the data set, from which finding it could be inferred that half of a given dataset would be inaccurately scored by the computer. In addition to the numerical problems that arise out of mathematical calculations, this finding also signals a pedagogical problem in that feedback should be as close as possible to the real performance of a learner so that it has a positive impact on the future performance and the motivation level of the learner (Nunan, 2010). In this respect, using a model that is only half-accurate may have a negative effect on the performance and motivation levels of the learners within a particular group.

The relatively low prediction accuracy of the trained model also posed problems related to the dataset of unscored essays gathered with the purpose of predicting their scores. According to the findings, the model could predict only 6 out of 19 essays accurately, which produced a prediction accuracy of only 32%. Related to this, correlation analysis revealed that the scores predicted by the computer were not significantly correlated with the scores given by the course instructor.

As a consequence, training an automated essay scoring model using machine learning and text classification algorithms does not seem to be a solution for managing and providing feedback for the written assignments in English Literature course due to the inaccuracy of the training model which may have unwanted effects on the performance of the learners as well as their motivation levels (Nunan, 2010). The failure of producing an accurate model may be attributed to the low number of instances in the dataset since Heilman and Madnani (2015) state that increasing the sample size of the training dataset as well as the instances per score level significantly contributes to the performance of the training model. Within the context of the study, although it could be possible to increase the number of instances to some extent, the inevitable limit is the limited number of students who take the English Literature course, thus it may not be possible to double or triple the number of essays to be submitted by the students. The algorithm that was used for the prediction task, namely logistic regression, may

also have reduced the prediction accuracy as it has been reported to be a prediction algorithm with relatively low accuracy by Jung (2017). However, it was used for the tasks in the present study since it was the most accurate algorithm among the others which were tested. Even so, data clustering methods can still be tested to improve the overall prediction accuracy of the algorithm even though clustering may risk data shrinkage and data ignorance (Trivedi, Pardos & Heffernan, 2015).

Taking the conclusions into account, the strongest limitation of the study appears to be the sample size since it may be the reason why the training model did not produce accurate results in terms of classifying student essays according to their scores. For this reason, a future research suggestion can be to double or triple the sample size and replicate the study to see if the results are more accurate. In addition, creation of larger datasets through resampling methods can also be tested to observe if resampling adds to the accuracy of the findings. At this point, however, it should also be kept in mind that each score level should have an adequate number of instances. If the increase in the sample size is combined with data clustering, it appears that more accurate results can be obtained from similar analyses. Another limitation of the study is that the psychological effects, along with the washback effect, of automated essay scoring is not scrutinized within the context of the present study. Upon creation of a working training model and utilizing it, these effects should also be measured within its own context in order to reveal if the automated scoring of the essays has a positive or negative effect on the learners.

References

Abu-Mustafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data (1st ed.)*. Seattle: AML Book.

Anderson, S.E., & Ben Jafaar, S. (2006). *Policy trends in Ontario education: 1990-2003*. (ICEC Working Paper #1). University of Toronto, Ontario Institute. Retrieved on 22 January 2018 from http://fcis.oise.utoronto.ca/~icec/policytrends.pdf

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment (JTLA), 4*(3). Retrieved on 11 March 2016 from http://ejournals.bc.edu/ojs/ index.php/jtla/article/view/1650

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York, NY: Routledge.

Baker, N. L. (2014). "Get it off my stack": Teachers' tools for grading papers. *Assessing Writing, 19*, 36-50. doi: 10.1016/j.asw.2013.11.005.

Barker, T. (2011). An automated individual feedback and marking system: An empirical study. *The Electronic Journal of E-Learning, 9*(1), 1-14.

Bauer, J. (2016). *A new approach: Closing the writing gap by using reliable assessment to guide and evaluate cross-curricular argumentative writing* (Unpublished master's thesis). The University of Wisconsin, USA.

Brookhart, S.M. & Bronowicz, D.L. (2003). "I don't like writing. It makes my fingers hurt": students talk about their classroom assessments. *Assessment in Education, 10*(2), 221-242.

Chih-Min, S. & Li-Yi, W. (2013). Factors affecting English language teachers' classroom assessment practices: A case study of Singapore secondary schools. *NIE research brief series.* Retrieved on 12 March 2016 from http://hdl.handle.net/10497/15003

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education, 53*(4), 286-302.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7-24.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1), 1-35.

Dornyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford University Press.

Duncan, C. R. & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research, 53*(1), 1-21.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* London: Routledge.

Gavriel, J. (2013). Assessment for learning: A wider (classroom-researched) perspective is important for formative assessment and self-directed learning in general practice. *Education for Primary Care, 24*(2), 93-96.

Greenstein, L. (2010). *What teachers really need to know about formative assessment?* Alexandria, VA: ASCD

Heilman, M. & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. Proceedings from NAACL HLT: *The tenth workshop on innovative use of NLP for building educational applications* (pp. 81-85). The Association for Computational Linguistics: USA. Retrieved on 13 May 2016 from http://www.cs. rochester.edu/u/tetreaul/bea10proceedings.pdf# page=270

Hyland, K. & Hyland, F. (Eds.). (2006). *Feedback in second language writing: Contexts and issues*. New York: Cambridge University Press. http://dx.doi.org/10.1017/ CBO9781139524742

Imaki, J. & Ishihara, S. (2013). Experimenting with a Japanese automated essay scoring system in the L2 Japanese environment. *Papers in Language Testing and Assessment, 2*(2), 28-47.

Jung, E. (2017). A comparison of data mining methods in analyzing educational data. In J. Park, Y. Pan, G. Yi, & V. Loia (Eds.), *Advances in computer science and ubiquitous computing CSA-CUTE2016* (pp. 173-178). Singapore: Springer.

Kumar, C. S., & Rama Sree, R. J. (2014). An attempt to improve classification accuracy through implementation of bootstrap aggregation with sequential minimal optimization during automated evaluation of descriptive answers. *Indian Journal of Science and Technology, 7*(9), 1369-1375.

Lai, Y.H. (2010). Which do students prefer to evaluate their essays: Peers or computer program? *British Journal of Educational Technology, 41*(3), 432-454.

Landauer, T. K., Laham, D. & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems & Their Applications, 15*(5), 27-31.

Lee, I. (2014). Teachers' reflection on implementation of innovative feedback approaches in EFL writing. *English Teaching, 69*(1), 23-40.

Mayfield, E. & Rose, C. P. (2013). LightSIDE: Open Source Machine Learning for Text. In M.D. Shermis& J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 124-135). New York: Psychology Press.

McMillan, J. H. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction*. Boston: Pearson.

Norbert, E. & Williamson, D. M. (2013). Assessing writing special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*(1), 1-6.

Nunan, D. (2010). Technology Supports for Second Language Learning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 204-210). Elsevier.

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). Mahwah, NJ: Lawrence Erlbaum Associates.

Paran, A. (2006). The stories of literature and language teaching. In A. Paran (Ed.), *Literature in language teaching and learning* (pp. 1-10). Alexandria, VA: TESOL.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015) Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education, 28*(2), 130-142. doi: 10.1080/08957347.2014.1002920

Ritter, K. (2012). Ladies who don't know us correct our papers: Postwar lay reader programs and twenty-first century contingent labor in first-year writing. *College Composition and Communication, 63*(3), 387-419.

Sakiyama, Y., Yuki, H., Moriya, T., Hattori, K., Suzuki, M., Shimada, K., & Honma, T. (2008) Predicting human liver microsomal stability with machine learning techniques. *J Mol Graph Model, 26*, 907-915

Santos, V. D. O., Verspoor, M., & Nerbonne, J. (2012). Identifying important factors in essay grading using machine learning. In D. Tsagari, S. Papadima-Sophocleous, & S. Ioannou-Georgiou (Eds.), *International experiences in language testing and assessment—Selected papers in memory of Pavlos Pavlou* (pp. 295–309). Frankfurt am Main, Germany: Peter Lang GmbH.

Shermis, M. D. & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Di Vesta, F. J. (2011). *Classroom assessment in action.* Maryland: Rowman and Little Field Publishers.

Shermis, M. D. & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 213–246). New York, NY: Routledge.

Stankova, E. N., Balakshiy, A. V., Petrov, D. A., Shorov, A. V., & Korkhov, V. V. (2016). Using technologies of OLAP and machine learning for validation of the numerical models of convective clouds. *Lecture Notes in Computer Science, 9788*, 463-472. doi: 10.1007/ 978-3-319-42111-7_36

Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2015). The utility of clustering in prediction tasks. *CoRR,* 1509.06163. Retrieved on 01 October 2018 from https://arxiv.org/abs/1509.06163

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360-363.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85-99.

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education, 123,* 212-224. https://doi.org/10.1016/j.compedu.2018.05. 010

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology, 68,* 19-37. https://doi.org/10.1016/j.jsp.2017.12.005

Yamamoto, M., Umemura, N., & Kawano, H. (2018). Automated essay scoring system based on rubric. In R. Lee (Ed.), *Applied computing & information technology. ACIT 2017. Studies in computational intelligence, vol. 727* (pp. 177-190). Springer, Cham. https://doi.org/10.1007/978-3-319-64051-8_11

Yang, M., Kim, M., Lee, H., & Rim, H. (2012). Assessing writing fluency of non-English speaking student for automated essay scoring – How to automatically evaluate the fluency in English essay. *Paper presented at the 4th International Conference on Computer Supported Education*. Porto, Portugal. Retrieved on 11 March 2016 from http://www.researchgate.net

Yang, W. (2012). A study of students' perceptions and attitudes towards genre-based ESP writing instruction. *The Asian ESP Journal, 8*(3), 50-73.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives, 18*(19), 1-40.

Correspondence: Kutay Uzun, Lecturer, English Language Teaching, Faculty of Education, Trakya University, Edirne, Turkey