Artificial Intelligence Review (2022) 55:2495–2527
https://doi.org/10.1007/s10462-021-10068-2

An automated essay scoring systems: a systematic literature review

Dadi Ramesh[1,2] · **Suresh Kumar Sanampudi**[3]

Published online: 23 September 2021

Abstract
Assessment in the Education system plays a significant role in judging student performance. The present evaluation system is through human assessment. As the number of teachers' student ratio is gradually increasing, the manual evaluation process becomes complicated. The drawback of manual evaluation is that it is time-consuming, lacks reliability, and many more. This connection online examination system evolved as an alternative tool for pen and paper-based methods. Present Computer-based evaluation system works only for multiple-choice questions, but there is no proper evaluation system for grading essays and short answers. Many researchers are working on automated essay grading and short answer scoring for the last few decades, but assessing an essay by considering all parameters like the relevance of the content to the prompt, development of ideas, Cohesion, and Coherence is a big challenge till now. Few researchers focused on Content-based evaluation, while many of them addressed style-based assessment. This paper provides a systematic literature review on automated essay scoring systems. We studied the Artificial Intelligence and Machine Learning techniques used to evaluate automatic essay scoring and analyzed the limitations of the current studies and research trends. We observed that the essay evaluation is not done based on the relevance of the content and coherence.

Keywords Assessment · Short answer scoring · Essay grading · Natural language processing · Deep learning

✉ Dadi Ramesh
dadiramesh44@gmail.com

Suresh Kumar Sanampudi
sureshsanampudi@jntuh.ac.in

[1] School of Computer Science and Artificial Intelligence, SR University, Warangal, TS, India

[2] Research Scholar, JNTU, Hyderabad, India

[3] Department of Information Technology, JNTUH College of Engineering, Nachupally, Kondagattu, Jagtial, TS, India

1 Introduction

Due to COVID 19 outbreak, an online educational system has become inevitable. In the present scenario, almost all the educational institutions ranging from schools to colleges adapt the online education system. The assessment plays a significant role in measuring the learning ability of the student. Most automated evaluation is available for multiple-choice questions, but assessing short and essay answers remain a challenge. The education system is changing its shift to online-mode, like conducting computer-based exams and automatic evaluation. It is a crucial application related to the education domain, which uses natural language processing (NLP) and Machine Learning techniques. The evaluation of essays is impossible with simple programming languages and simple techniques like pattern matching and language processing. Here the problem is for a single question, we will get more responses from students with a different explanation. So, we need to evaluate all the answers concerning the question.

Automated essay scoring (AES) is a computer-based assessment system that automatically scores or grades the student responses by considering appropriate features. The AES research started in 1966 with the Project Essay Grader (PEG) by Ajay et al. (1973). PEG evaluates the writing characteristics such as grammar, diction, construction, etc., to grade the essay. A modified version of the PEG by Shermis et al. (2001) was released, which focuses on grammar checking with a correlation between human evaluators and the system. Foltz et al. (1999) introduced an Intelligent Essay Assessor (IEA) by evaluating content using latent semantic analysis to produce an overall score. Powers et al. (2002) proposed E-rater and Intellimetric by Rudner et al. (2006) and Bayesian Essay Test Scoring System (BESTY) by Rudner and Liang (2002), these systems use natural language processing (NLP) techniques that focus on style and content to obtain the score of an essay. The vast majority of the essay scoring systems in the 1990s followed traditional approaches like pattern matching and a statistical-based approach. Since the last decade, the essay grading systems started using regression-based and natural language processing techniques. AES systems like Dong et al. (2017) and others developed from 2014 used deep learning techniques, inducing syntactic and semantic features resulting in better results than earlier systems.

Ohio, Utah, and most US states are using AES systems in school education, like Utah compose tool, Ohio standardized test (an updated version of PEG), evaluating millions of student's responses every year. These systems work for both formative, summative assessments and give feedback to students on the essay. Utah provided basic essay evaluation rubrics (six characteristics of essay writing): Development of ideas, organization, style, word choice, sentence fluency, conventions. Educational Testing Service (ETS) has been conducting significant research on AES for more than a decade and designed an algorithm to evaluate essays on different domains and providing an opportunity for test-takers to improve their writing skills. In addition, they are current research content-based evaluation.

The evaluation of essay and short answer scoring should consider the relevance of the content to the prompt, development of ideas, Cohesion, Coherence, and domain knowledge. Proper assessment of the parameters mentioned above defines the accuracy of the evaluation system. But all these parameters cannot play an equal role in essay scoring and short answer scoring. In a short answer evaluation, domain knowledge is required, like the meaning of "cell" in physics and biology is different. And while evaluating essays, the implementation of ideas with respect to prompt is required. The system should also assess the completeness of the responses and provide feedback.

Several studies examined AES systems, from the initial to the latest AES systems. In which the following studies on AES systems are Blood (2011) provided a literature review from PEG 1984–2010. Which has covered only generalized parts of AES systems like ethical aspects, the performance of the systems. Still, they have not covered the implementation part, and it's not a comparative study and has not discussed the actual challenges of AES systems.

Burrows et al. (2015) Reviewed AES systems on six dimensions like dataset, NLP techniques, model building, grading models, evaluation, and effectiveness of the model. They have not covered feature extraction techniques and challenges in features extractions. Covered only Machine Learning models but not in detail. This system not covered the comparative analysis of AES systems like feature extraction, model building, and level of relevance, cohesion, and coherence not covered in this review.

Ke et al. (2019) provided a state of the art of AES system but covered very few papers and not listed all challenges, and no comparative study of the AES model. On the other hand, Hussein et al. in (2019) studied two categories of AES systems, four papers from handcrafted features for AES systems, and four papers from the neural networks approach, discussed few challenges, and did not cover feature extraction techniques, the performance of AES models in detail.

Klebanov et al. (2020). Reviewed 50 years of AES systems, listed and categorized all essential features that need to be extracted from essays. But not provided a comparative analysis of all work and not discussed the challenges.

This paper aims to provide a systematic literature review (SLR) on automated essay grading systems. An SLR is an Evidence-based systematic review to summarize the existing research. It critically evaluates and integrates all relevant studies' findings and addresses the research domain's specific research questions. Our research methodology uses guidelines given by Kitchenham et al. (2009) for conducting the review process; provide a well-defined approach to identify gaps in current research and to suggest further investigation.

We addressed our research method, research questions, and the selection process in Sect. 2, and the results of the research questions have discussed in Sect. 3. And the synthesis of all the research questions addressed in Sect. 4. Conclusion and possible future work discussed in Sect. 5.

2 Research method

We framed the research questions with PICOC criteria.

Population (P) Student essays and answers evaluation systems.

Intervention (I) evaluation techniques, data sets, features extraction methods.

Comparison (C) Comparison of various approaches and results.

Outcomes (O) Estimate the accuracy of AES systems,

Context (C) NA.

2.1 Research questions

To collect and provide research evidence from the available studies in the domain of automated essay grading, we framed the following research questions (RQ):

RQ1 what are the datasets available for research on automated essay grading?

The answer to the question can provide a list of the available datasets, their domain, and access to the datasets. It also provides a number of essays and corresponding prompts.

RQ2 what are the features extracted for the assessment of essays?

The answer to the question can provide an insight into various features so far extracted, and the libraries used to extract those features.

RQ3, which are the evaluation metrics available for measuring the accuracy of algorithms?

The answer will provide different evaluation metrics for accurate measurement of each Machine Learning approach and commonly used measurement technique.

RQ4 What are the Machine Learning techniques used for automatic essay grading, and how are they implemented?

It can provide insights into various Machine Learning techniques like regression models, classification models, and neural networks for implementing essay grading systems. The response to the question can give us different assessment approaches for automated essay grading systems.

RQ5 What are the challenges/limitations in the current research?

The answer to the question provides limitations of existing research approaches like cohesion, coherence, completeness, and feedback.

2.2 Search process

We conducted an automated search on well-known computer science repositories like ACL, ACM, IEEE Explore, Springer, and Science Direct for an SLR. We referred to papers published from 2010 to 2020 as much of the work during these years focused on advanced technologies like deep learning and natural language processing for automated essay grading systems. Also, the availability of free data sets like Kaggle (2012), Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE) by Yannakoudakis et al. (2011) led to research this domain.

Search Strings: We used search strings like "Automated essay grading" OR "Automated essay scoring" OR "short answer scoring systems" OR "essay scoring systems" OR "automatic essay evaluation" and searched on metadata.

2.3 Selection criteria

After collecting all relevant documents from the repositories, we prepared selection criteria for inclusion and exclusion of documents. With the inclusion and exclusion criteria, it becomes more feasible for the research to be accurate and specific.

Inclusion criteria 1 Our approach is to work with datasets comprise of essays written in English. We excluded the essays written in other languages.

Inclusion criteria 2 We included the papers implemented on the AI approach and excluded the traditional methods for the review.

Inclusion criteria 3 The study is on essay scoring systems, so we exclusively included the research carried out on only text data sets rather than other datasets like image or speech.

Exclusion criteria We removed the papers in the form of review papers, survey papers, and state of the art papers.

2.4 Quality assessment

In addition to the inclusion and exclusion criteria, we assessed each paper by quality assessment questions to ensure the article's quality. We included the documents that have clearly explained the approach they used, the result analysis and validation.

The quality checklist questions are framed based on the guidelines from Kitchenham et al. (2009). Each quality assessment question was graded as either 1 or 0. The final score of the study range from 0 to 3. A cut off score for excluding a study from the review is 2 points. Since the papers scored 2 or 3 points are included in the final evaluation. We framed the following quality assessment questions for the final study.

Quality Assessment 1: Internal validity.

Quality Assessment 2: External validity.

Quality Assessment 3: Bias.

The two reviewers review each paper to select the final list of documents. We used the Quadratic Weighted Kappa score to measure the final agreement between the two reviewers. The average resulted from the kappa score is 0.6942, a substantial agreement between the reviewers. The result of evolution criteria shown in Table 1. After Quality Assessment, the final list of papers for review is shown in Table 2. The complete selection process is shown in Fig. 1. The total number of selected papers in year wise as shown in Fig. 2.

3 Results

3.1 What are the datasets available for research on automated essay grading?

To work with problem statement especially in Machine Learning and deep learning domain, we require considerable amount of data to train the models. To answer this question, we listed all the data sets used for training and testing for automated essay grading systems. The Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE)

Table 1 Quality assessment analysis

Number of papers	Quality assessment score
50	3
12	2
59	1
23	0

Table 2 Final list of papers

Data base	Paper count
ACL	28
ACM	5
IEEE Explore	19
Springer	5
Other	5
Total	62

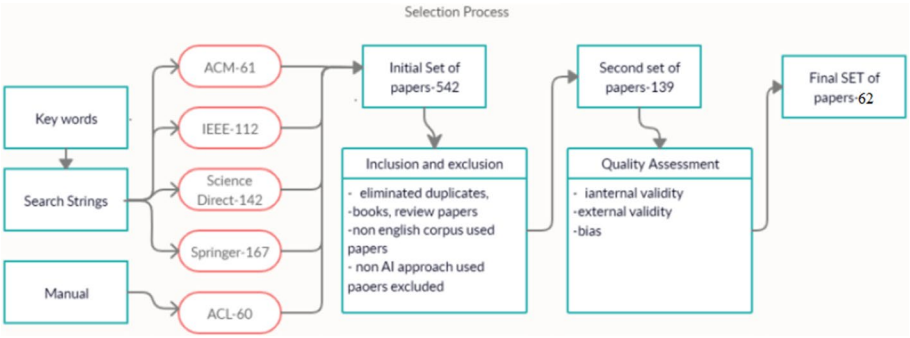

Fig. 1 Selection process

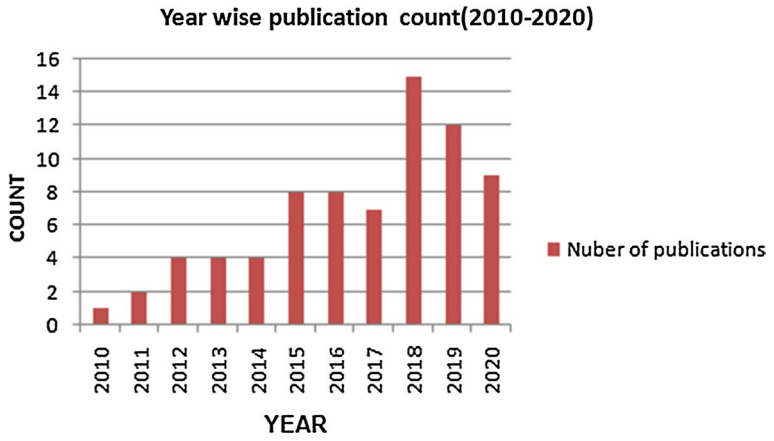

Fig. 2 Year wise publications

Yannakoudakis et al. (2011) developed corpora that contain 1244 essays and ten prompts. This corpus evaluates whether a student can write the relevant English sentences without any grammatical and spelling mistakes. This type of corpus helps to test the models built for GRE and TOFEL type of exams. It gives scores between 1 and 40.

Bailey and Meurers (2008), Created a dataset (CREE reading comprehension) for language learners and automated short answer scoring systems. The corpus consists of 566 responses from intermediate students. Mohler and Mihalcea (2009). Created a dataset for the computer science domain consists of 630 responses for data structure assignment questions. The scores are range from 0 to 5 given by two human raters.

Dzikovska et al. (2012) created a Student Response Analysis (SRA) corpus. It consists of two sub-groups: the BEETLE corpus consists of 56 questions and approximately 3000 responses from students in the electrical and electronics domain. The second one is the SCIENTSBANK(SemEval-2013) (Dzikovska et al. 2013a; b) corpus consists of 10,000 responses on 197 prompts on various science domains. The student responses ladled with "correct, partially correct incomplete, Contradictory, Irrelevant, Non-domain."

In the Kaggle (2012) competition, released total 3 types of corpuses on an Automated Student Assessment Prize (ASAP1) ("https://www.kaggle.com/c/asap-sas/") essays and

Table 3 ALL types Datasets used in Automatic scoring systems

Data Set	Language	Total responses	Number of prompts
Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE)	English	1244	
CREE	English	566	
CS	English	630	
SRA	English	3000	56
SCIENTSBANK(SemEval-2013)	English	10,000	197
ASAP-AES	English	17,450	8
ASAP-SAS	English	17,207	10
ASAP++	English	10,696	6
power grading	English	700	
TOEFL11	English	1100	8
International Corpus of Learner English (ICLE)	English	3663	

short answers. It has nearly 17,450 essays, out of which it provides up to 3000 essays for each prompt. It has eight prompts that test 7th to 10th grade US students. It gives scores between the [0–3] and [0–60] range. The limitations of these corpora are: (1) it has a different score range for other prompts. (2) It uses statistical features such as named entities extraction and lexical features of words to evaluate essays. ASAP++is one more dataset from Kaggle. It is with six prompts, and each prompt has more than 1000 responses total of 10,696 from 8th-grade students. Another corpus contains ten prompts from science, English domains and a total of 17,207 responses. Two human graders evaluated all these responses.

Correnti et al. (2013) created a Response-to-Text Assessment (RTA) dataset used to check student writing skills in all directions like style, mechanism, and organization. 4–8 grade students give the responses to RTA. Basu et al. (2013) created a power grading dataset with 700 responses for ten different prompts from US immigration exams. It contains all short answers for assessment.

The TOEFL11 corpus Blanchard et al. (2013) contains 1100 essays evenly distributed over eight prompts. It is used to test the English language skills of a candidate attending the TOFEL exam. It scores the language proficiency of a candidate as low, medium, and high.

International Corpus of Learner English (ICLE) Granger et al. (2009) built a corpus of 3663 essays covering different dimensions. It has 12 prompts with 1003 essays that test the organizational skill of essay writing, and13 prompts, each with 830 essays that examine the thesis clarity and prompt adherence.

Argument Annotated Essays (AAE) Stab and Gurevych (2014) developed a corpus that contains 102 essays with 101 prompts taken from the essayforum2 site. It tests the persuasive nature of the student essay. The SCIENTSBANK corpus used by Sakaguchi et al. (2015) available in git-hub, containing 9804 answers to 197 questions in 15 science domains. Table 3 illustrates all datasets related to AES systems.

3.2 RQ2 what are the features extracted for the assessment of essays?

Features play a major role in the neural network and other supervised Machine Learning approaches. The automatic essay grading systems scores student essays based on

Table 4 Types of features

Statistical features	Style based features	Content based features
Essay length with respect to the number of words	Sentence structure	Cohesion between sentences in a document
Essay length with respect to sentence	POS	Overlapping (prompt)
Average sentence length	Punctuation	Relevance of information
Average word length	Grammatical	Semantic role of words
N-gram	Logical operators	Correctness
	Vocabulary	Consistency
		Sentence expressing key concepts

different types of features, which play a prominent role in training the models. Based on their syntax and semantics and they are categorized into three groups. 1. statistical-based features Contreras et al. (2018); Kumar et al. (2019); Mathias and Bhattacharyya (2018a; b) 2. Style-based (Syntax) features Cummins et al. (2016); Darwish and Mohamed (2020); Ke et al. (2019). 3. Content-based features Dong et al. (2017). A good set of features appropriate models evolved better AES systems. The vast majority of the researchers are using regression models if features are statistical-based. For Neural Networks models, researches are using both style-based and content-based features. The following table shows the list of various features used in existing AES Systems. Table 4 represents all set of features used for essay grading.

We studied all the feature extracting NLP libraries as shown in Fig. 3. that are used in the papers. The NLTK is an NLP tool used to retrieve statistical features like POS, word count, sentence count, etc. With NLTK, we can miss the essay's semantic features. To find semantic features Word2Vec Mikolov et al. (2013), GloVe Jeffrey Pennington et al. (2014)

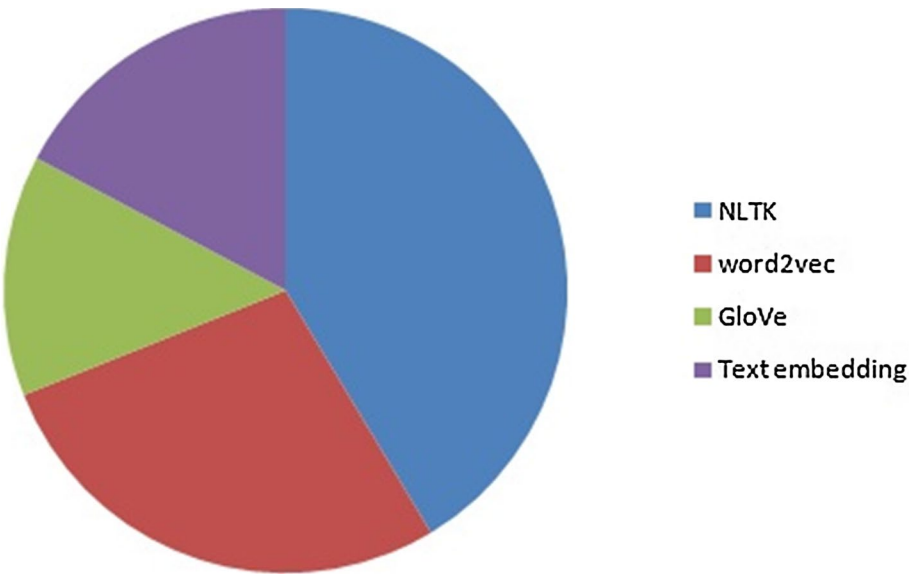

Fig. 3 Usages of tools

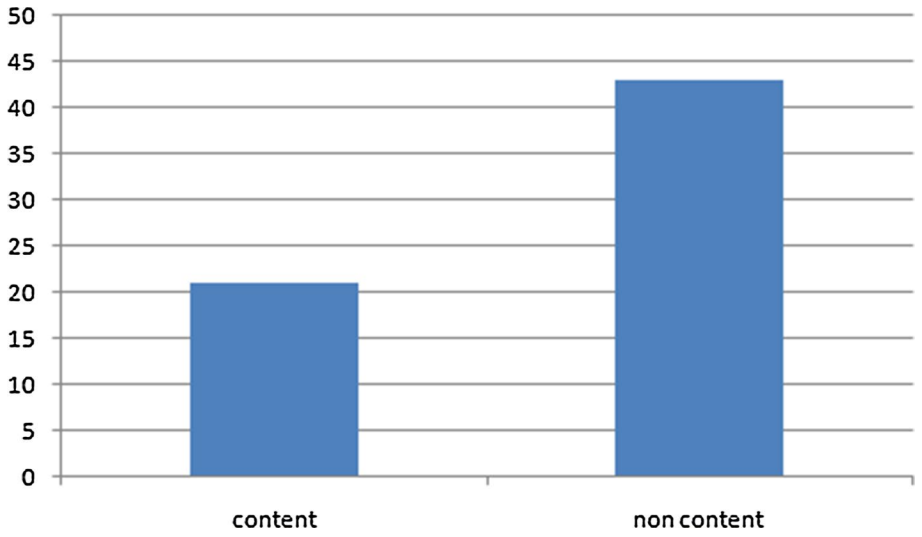

Fig. 4 Number of papers on content based features

is the most used libraries to retrieve the semantic text from the essays. And in some systems, they directly trained the model with word embeddings to find the score. From Fig. 4 as observed that non-content-based feature extraction is higher than content-based.

3.3 RQ3 which are the evaluation metrics available for measuring the accuracy of algorithms?

The majority of the AES systems are using three evaluation metrics. They are (1) quadrated weighted kappa (QWK) (2) Mean Absolute Error (MAE) (3) Pearson Correlation Coefficient (PCC) Shehab et al. (2016). The quadratic weighted kappa will find agreement between human evaluation score and system evaluation score and produces value ranging from 0 to 1. And the Mean Absolute Error is the actual difference between human-rated score to system-generated score. The mean square error (MSE) measures the average squares of the errors, i.e., the average squared difference between the human-rated and the system-generated scores. MSE will always give positive numbers only. Pearson's Correlation Coefficient (PCC) finds the correlation coefficient between two variables. It will provide three values (0, 1, −1). "0" represents human-rated and system scores that are not related. "1" represents an increase in the two scores. "−1" illustrates a negative relationship between the two scores.

3.4 RQ4 what are the Machine Learning techniques being used for automatic essay grading, and how are they implemented?

After scrutinizing all documents, we categorize the techniques used in automated essay grading systems into four baskets. 1. Regression techniques. 2. Classification model. 3. Neural networks. 4. Ontology-based approach.

All the existing AES systems developed in the last ten years employ supervised learning techniques. Researchers using supervised methods viewed the AES system as either regression or classification task. The goal of the regression task is to predict the score of an essay. The classification task is to classify the essays belonging to (low, medium, or highly) relevant to the question's topic. Since the last three years, most AES systems developed made use of the concept of the neural network.

3.4.1 Regression based models

Mohler and Mihalcea (2009). proposed text-to-text semantic similarity to assign a score to the student essays. There are two text similarity measures like Knowledge-based measures, corpus-based measures. There eight knowledge-based tests with all eight models. They found the similarity. The shortest path similarity determines based on the length, which shortest path between two contexts. Leacock & Chodorow find the similarity based on the shortest path's length between two concepts using node-counting. The Lesk similarity finds the overlap between the corresponding definitions, and Wu & Palmer algorithm finds similarities based on the depth of two given concepts in the wordnet taxonomy. Resnik, Lin, Jiang&Conrath, Hirst& St-Onge find the similarity based on different parameters like the concept, probability, normalization factor, lexical chains. In corpus-based likeness, there LSA BNC, LSA Wikipedia, and ESA Wikipedia, latent semantic analysis is trained on Wikipedia and has excellent domain knowledge. Among all similarity scores, correlation scores LSA Wikipedia scoring accuracy is more. But these similarity measure algorithms are not using NLP concepts. These models are before 2010 and basic concept models to continue the research automated essay grading with updated algorithms on neural networks with content-based features.

Adamson et al. (2014) proposed an automatic essay grading system which is a statistical-based approach in this they retrieved features like POS, Character count, Word count, Sentence count, Miss spelled words, n-gram representation of words to prepare essay vector. They formed a matrix with these all vectors in that they applied LSA to give a score to each essay. It is a statistical approach that doesn't consider the semantics of the essay. The accuracy they got when compared to the human rater score with the system is 0.532.

Cummins et al. (2016). Proposed Timed Aggregate Perceptron vector model to give ranking to all the essays, and later they converted the rank algorithm to predict the score of the essay. The model trained with features like Word unigrams, bigrams, POS, Essay length, grammatical relation, Max word length, sentence length. It is multi-task learning, gives ranking to the essays, and predicts the score for the essay. The performance evaluated through QWK is 0.69, a substantial agreement between the human rater and the system.

Sultan et al. (2016). Proposed a Ridge regression model to find short answer scoring with Question Demoting. Question Demoting is the new concept included in the essay's final assessment to eliminate duplicate words from the essay. The extracted features are Text Similarity, which is the similarity between the student response and reference answer. Question Demoting is the number of repeats in a student response. With inverse document frequency, they assigned term weight. The sentence length Ratio is the number of words in the student response, is another feature. With these features, the Ridge regression model was used, and the accuracy they got 0.887.

Contreras et al. (2018). Proposed Ontology based on text mining in this model has given a score for essays in phases. In phase-I, they generated ontologies with ontoGen and SVM to

find the concept and similarity in the essay. In phase II from ontologies, they retrieved features like essay length, word counts, correctness, vocabulary, and types of word used, domain information. After retrieving statistical data, they used a linear regression model to find the score of the essay. The accuracy score is the average of 0.5.

Darwish and Mohamed (2020) proposed the fusion of fuzzy Ontology with LSA. They retrieve two types of features, like syntax features and semantic features. In syntax features, they found Lexical Analysis with tokens, and they construct a parse tree. If the parse tree is broken, the essay is inconsistent—a separate grade assigned to the essay concerning syntax features. The semantic features are like similarity analysis, Spatial Data Analysis. Similarity analysis is to find duplicate sentences—Spatial Data Analysis for finding Euclid distance between the center and part. Later they combine syntax features and morphological features score for the final score. The accuracy they achieved with the multiple linear regression model is 0.77, mostly on statistical features.

Süzen Neslihan et al. (2020) proposed a text mining approach for short answer grading. First, their comparing model answers with student response by calculating the distance between two sentences. By comparing the model answer with student response, they find the essay's completeness and provide feedback. In this approach, model vocabulary plays a vital role in grading, and with this model vocabulary, the grade will be assigned to the student's response and provides feedback. The correlation between the student answer to model answer is 0.81.

3.4.2 Classification based Models

Persing and Ng (2013) used a support vector machine to score the essay. The features extracted are OS, N-gram, and semantic text to train the model and identified the keywords from the essay to give the final score.

Sakaguchi et al. (2015) proposed two methods: response-based and reference-based. In response-based scoring, the extracted features are response length, n-gram model, and syntactic elements to train the support vector regression model. In reference-based scoring, features such as sentence similarity using word2vec is used to find the cosine similarity of the sentences that is the final score of the response. First, the scores were discovered individually and later combined two features to find a final score. This system gave a remarkable increase in performance by combining the scores.

Mathias and Bhattacharyya (2018a; b) Proposed Automated Essay Grading Dataset with Essay Attribute Scores. The first concept features selection depends on the essay type. So the common attributes are Content, Organization, Word Choice, Sentence Fluency, Conventions. In this system, each attribute is scored individually, with the strength of each attribute identified. The model they used is a random forest classifier to assign scores to individual attributes. The accuracy they got with QWK is 0.74 for prompt 1 of the ASAS dataset (https://www.kaggle.com/c/asap-sas/).

Ke et al. (2019) used a support vector machine to find the response score. In this method, features like Agreeability, Specificity, Clarity, Relevance to prompt, Conciseness, Eloquence, Confidence, Direction of development, Justification of opinion, and Justification of importance. First, the individual parameter score obtained was later combined with all scores to give a final response score. The features are used in the neural network to find whether the sentence is relevant to the topic or not.

Salim et al. (2019) proposed an XGBoost Machine Learning classifier to assess the essays. The algorithm trained on features like word count, POS, parse tree depth, and

coherence in the articles with sentence similarity percentage; cohesion and coherence are considered for training. And they implemented K-fold cross-validation for a result the average accuracy after specific validations is 68.12.

3.4.3 Neural network models

Shehab et al. (2016) proposed a neural network method that used learning vector quantization to train human scored essays. After training, the network can provide a score to the ungraded essays. First, we should process the essay to remove Spell checking and then perform preprocessing steps like Document Tokenization, stop word removal, Stemming, and submit it to the neural network. Finally, the model will provide feedback on the essay, whether it is relevant to the topic. And the correlation coefficient between human rater and system score is 0.7665.

Kopparapu and De (2016) proposed the Automatic Ranking of Essays using Structural and Semantic Features. This approach constructed a super essay with all the responses. Next, ranking for a student essay is done based on the super-essay. The structural and semantic features derived helps to obtain the scores. In a paragraph, 15 Structural features like an average number of sentences, the average length of sentences, and the count of words, nouns, verbs, adjectives, etc., are used to obtain a syntactic score. A similarity score is used as semantic features to calculate the overall score.

Dong and Zhang (2016) proposed a hierarchical CNN model. The model builds two layers with word embedding to represents the words as the first layer. The second layer is a word convolution layer with max-pooling to find word vectors. The next layer is a sentence-level convolution layer with max-pooling to find the sentence's content and synonyms. A fully connected dense layer produces an output score for an essay. The accuracy with the hierarchical CNN model resulted in an average QWK of 0.754.

Taghipour and Ng (2016) proposed a first neural approach for essay scoring build in which convolution and recurrent neural network concepts help in scoring an essay. The network uses a lookup table with the one-hot representation of the word vector of an essay. The final efficiency of the network model with LSTM resulted in an average QWK of 0.708.

Dong et al. (2017). Proposed an Attention-based scoring system with CNN + LSTM to score an essay. For CNN, the input parameters were character embedding and word embedding, and it has attention pooling layers and used NLTK to obtain word and character embedding. The output gives a sentence vector, which provides sentence weight. After CNN, it will have an LSTM layer with an attention pooling layer, and this final layer results in the final score of the responses. The average QWK score is 0.764.

Riordan et al. (2017) proposed a neural network with CNN and LSTM layers. Word embedding, given as input to a neural network. An LSTM network layer will retrieve the window features and delivers them to the aggregation layer. The aggregation layer is a superficial layer that takes a correct window of words and gives successive layers to predict the answer's sore. The accuracy of the neural network resulted in a QWK of 0.90.

Zhao et al. (2017) proposed a new concept called Memory-Augmented Neural network with four layers, input representation layer, memory addressing layer, memory reading layer, and output layer. An input layer represents all essays in a vector form based on essay length. After converting the word vector, the memory addressing layer takes a sample of the essay and weighs all the terms. The memory reading layer takes

the input from memory addressing segment and finds the content to finalize the score. Finally, the output layer will provide the final score of the essay. The accuracy of essay scores is 0.78, which is far better than the LSTM neural network.

Mathias and Bhattacharyya (2018a; b) proposed deep learning networks using LSTM with the CNN layer and GloVe pre-trained word embeddings. For this, they retrieved features like Sentence count essays, word count per sentence, Number of OOVs in the sentence, Language model score, and the text's perplexity. The network predicted the goodness scores of each essay. The higher the goodness scores, means higher the rank and vice versa.

Nguyen and Dery (2016). Proposed Neural Networks for Automated Essay Grading. In this method, a single layer bi-directional LSTM accepting word vector as input. Glove vectors used in this method resulted in an accuracy of 90%.

Ruseti et al. (2018) proposed a recurrent neural network that is capable of memorizing the text and generate a summary of an essay. The Bi-GRU network with the max-pooling layer molded on the word embedding of each document. It will provide scoring to the essay by comparing it with a summary of the essay from another Bi-GRU network. The result obtained an accuracy of 0.55.

Wang et al. (2018a; b) proposed an automatic scoring system with the bi-LSTM recurrent neural network model and retrieved the features using the word2vec technique. This method generated word embeddings from the essay words using the skip-gram model. And later, word embedding is used to train the neural network to find the final score. The softmax layer in LSTM obtains the importance of each word. This method used a QWK score of 0.83%.

Dasgupta et al. (2018) proposed a technique for essay scoring with augmenting textual qualitative Features. It extracted three types of linguistic, cognitive, and psychological features associated with a text document. The linguistic features are Part of Speech (POS), Universal Dependency relations, Structural Well-formedness, Lexical Diversity, Sentence Cohesion, Causality, and Informativeness of the text. The psychological features derived from the Linguistic Information and Word Count (LIWC) tool. They implemented a convolution recurrent neural network that takes input as word embedding and sentence vector, retrieved from the GloVe word vector. And the second layer is the Convolution Layer to find local features. The next layer is the recurrent neural network (LSTM) to find corresponding of the text. The accuracy of this method resulted in an average QWK of 0.764.

Liang et al. (2018) proposed a symmetrical neural network AES model with Bi-LSTM. They are extracting features from sample essays and student essays and preparing an embedding layer as input. The embedding layer output is transfer to the convolution layer from that LSTM will be trained. Hear the LSRM model has self-features extraction layer, which will find the essay's coherence. The average QWK score of SBLSTMA is 0.801.

Liu et al. (2019) proposed two-stage learning. In the first stage, they are assigning a score based on semantic data from the essay. The second stage scoring is based on some handcrafted features like grammar correction, essay length, number of sentences, etc. The average score of the two stages is 0.709.

Pedro Uria Rodriguez et al. (2019) proposed a sequence-to-sequence learning model for automatic essay scoring. They used BERT (Bidirectional Encoder Representations from Transformers), which extracts the semantics from a sentence from both directions. And XLnet sequence to sequence learning model to extract features like the next sentence in an essay. With this pre-trained model, they attained coherence from the essay to give the final score. The average QWK score of the model is 75.5.

Xia et al. (2019) proposed a two-layer Bi-directional LSTM neural network for the scoring of essays. The features extracted with word2vec to train the LSTM and accuracy of the model in an average of QWK is 0.870.

Kumar et al. (2019) Proposed an AutoSAS for short answer scoring. It used pre-trained Word2Vec and Doc2Vec models trained on Google News corpus and Wikipedia dump, respectively, to retrieve the features. First, they tagged every word POS and they found weighted words from the response. It also found prompt overlap to observe how the answer is relevant to the topic, and they defined lexical overlaps like noun overlap, argument overlap, and content overlap. This method used some statistical features like word frequency, difficulty, diversity, number of unique words in each response, type-token ratio, statistics of the sentence, word length, and logical operator-based features. This method uses a random forest model to train the dataset. The data set has sample responses with their associated score. The model will retrieve the features from both responses like graded and ungraded short answers with questions. The accuracy of AutoSAS with QWK is 0.78. It will work on any topics like Science, Arts, Biology, and English.

Jiaqi Lun et al. (2020) proposed an automatic short answer scoring with BERT. In this with a reference answer comparing student responses and assigning scores. The data augmentation is done with a neural network and with one correct answer from the dataset classifying reaming responses as correct or incorrect.

Zhu and Sun (2020) proposed a multimodal Machine Learning approach for automated essay scoring. First, they count the grammar score with the spaCy library and numerical count as the number of words and sentences with the same library. With this input, they trained a single and Bi LSTM neural network for finding the final score. For the LSTM model, they prepared sentence vectors with GloVe and word embedding with NLTK. Bi-LSTM will check each sentence in both directions to find semantic from the essay. The average QWK score with multiple models is 0.70.

3.4.4 Ontology based approach

Mohler et al. (2011) proposed a graph-based method to find semantic similarity in short answer scoring. For the ranking of answers, they used the support vector regression model. The bag of words is the main feature extracted in the system.

Ramachandran et al. (2015) also proposed a graph-based approach to find lexical based semantics. Identified phrase patterns and text patterns are the features to train a random forest regression model to score the essays. The accuracy of the model in a QWK is 0.78.

Zupanc et al. (2017) proposed sentence similarity networks to find the essay's score. Ajetunmobi and Daramola (2017) recommended an ontology-based information extraction approach and domain-based ontology to find the score.

3.4.5 Speech response scoring

Automatic scoring is in two ways one is text-based scoring, other is speech-based scoring. This paper discussed text-based scoring and its challenges, and now we cover speech scoring and common points between text and speech-based scoring. Evanini and Wang (2013), Worked on speech scoring of non-native school students, extracted features with speech ratter, and trained a linear regression model, concluding that accuracy varies based on voice pitching. Loukina et al. (2015) worked on feature selection from speech data and trained

SVM. Malinin et al. (2016) used neural network models to train the data. Loukina et al. (2017). Proposed speech and text-based automatic scoring. Extracted text-based features, speech-based features and trained a deep neural network for speech-based scoring. They extracted 33 types of features based on acoustic signals. Malinin et al. (2017). Wu Xixin et al. (2020) Worked on deep neural networks for spoken language assessment. Incorporated different types of models and tested them. Ramanarayanan et al. (2017) worked on feature extraction methods and extracted punctuation, fluency, and stress and trained different Machine Learning models for scoring. Knill et al. (2018). Worked on Automatic speech recognizer and its errors how its impacts the speech assessment.

3.4.5.1 The state of the art This section provides an overview of the existing AES systems with a comparative study w. r. t models, features applied, datasets, and evaluation metrics used for building the automated essay grading systems. We divided all 62 papers into two sets of the first set of review papers in Table 5 with a comparative study of the AES systems.

3.4.6 Comparison of all approaches

In our study, we divided major AES approaches into three categories. Regression models, classification models, and neural network models. The regression models failed to find cohesion and coherence from the essay because it trained on BoW(Bag of Words) features. In processing data from input to output, the regression models are less complicated than neural networks. There are unable to find many intricate patterns from the essay and unable to find sentence connectivity. If we train the model with BoW features in the neural network approach, the model never considers the essay's coherence and coherence.

First, to train a Machine Learning algorithm with essays, all the essays are converted to vector form. We can form a vector with BoW and Word2vec, TF-IDF. The BoW and Word2vec vector representation of essays represented in Table 6. The vector representation of BoW with TF-IDF is not incorporating the essays semantic, and it's just statistical learning from a given vector. Word2vec vector comprises semantic of essay in a unidirectional way.

In BoW, the vector contains the frequency of word occurrences in the essay. The vector represents 1 and more based on the happenings of words in the essay and 0 for not present. So, in BoW, the vector does not maintain the relationship with adjacent words; it's just for single words. In word2vec, the vector represents the relationship between words with other words and sentences prompt in multiple dimensional ways. But word2vec prepares vectors in a unidirectional way, not in a bidirectional way; word2vec fails to find semantic vectors when a word has two meanings, and the meaning depends on adjacent words. Table 7 represents a comparison of Machine Learning models and features extracting methods.

In AES, cohesion and coherence will check the content of the essay concerning the essay prompt these can be extracted from essay in the vector from. Two more parameters are there to access an essay is completeness and feedback. Completeness will check whether student's response is sufficient or not though the student wrote correctly. Table 8 represents all four parameters comparison for essay grading. Table 9 illustrates comparison of all approaches based on various features like grammar, spelling, organization of essay, relevance.

Table 5 State of the art

System	Approach	Dataset	Features applied	Evaluation metric and results
Mohler and Mihalcea in (2009)	shortest path similarity, LSA regression model		Word vector	Finds the shortest path
Niraj Kumar and Lipika Dey. In (2013)	Word-Graph	ASAP Kaggle	Content and style-based features	63.81% accuracy
Alex Adamson et al. in (2014)	LSA regression model	ASAP Kaggle	Statistical features	QWK 0.532
Nguyen and Dery (2016)	LSTM (single layer bidirectional)	ASAP Kaggle	Statistical features	90% accuracy
Keisuke Sakaguchi et al. in (2015)	Classification model	ETS (educational testing services)	Statistical, Style based features	QWK is 0.69
Ramachandran et al. in (2015)	regression model	ASAP Kaggle short Answer	Statistical and style-based features	QWK 0.77
Sultan et al. in (2016)	Ridge regression model	SciEntBank answers	Statistical features	RMSE 0.887
Dong and Zhang (2016)	CNN neural network	ASAP Kaggle	Statistical features	QWK 0.734
Taghipour and Ngl in (2016)	CNN+LSTM neural network	ASAP Kaggle	Lookup table (one hot representation of word vector)	QWK 0.761
Shehab et al. in (2016)	Learning vector quantization neural network	Mansoura University student's essays	Statistical features	correlation coefficient 0.7665
Cummins et al. in (2016)	Regression model	ASAP Kaggle	Statistical features, style-based features	QWK 0.69
Kopparapu and De (2016)	Neural network	ASAP Kaggle	Statistical features, Style based	
Dong, et al. in (2017)	CNN+LSTM neural network	ASAP Kaggle	Word embedding, content based	QWK 0.764
Ajetunmobi and Daramola (2017)	WuPalmer algorithm		Statistical features	
Siyuan Zhao et al. in (2017)	LSTM (memory network)	ASAP Kaggle	Statistical features	QWK 0.78
Mathias and Bhattacharyya (2018a)	Random Forest Classifier a classification model	ASAP Kaggle	Style and Content based features	Classified which feature set is required
Brian Riordan et al. in (2017)	CNN+LSTM neural network	ASAP Kaggle short Answer	Word embeddings	QWK 0.90

Table 5 (continued)

System	Approach	Dataset	Features applied	Evaluation metric and results
Tirthankar Dasgupta et al. in (2018)	CNN -bidirectional LSTMs neural network	ASAP Kaggle	Content and physiological features	QWK 0.786
Wu and Shih (2018)	Classification model	SciEntBank answers	unigram_recall unigram_precision unigram_F_measure log_bleu_recall log_bleu_precision log_bleu_F_measure BLUE features	Squared correlation coefficient 59.568
Yucheng Wang, etc.in (2018b)	Bi-LSTM	ASAP Kaggle	Word embedding sequence	QWK 0.724
Anak Agung Putri Ratna et al. in (2018)	Winnowing ALGORITHM			86.86 accuracy
Sharma and Jayagopi (2018)	Glove, LSTM neural network	ASAP Kaggle	Hand written essay images	QWK 0.69
Jennifer O. Contreras et al. in (2018)	OntoGen (SVM) Linear Regression	University of Benghazi data set	Statistical, style-based features	
Mathias, Bhattacharyya (2018b)	GloVe,LSTM neural network	ASAP Kaggle	Statistical features, style features	Predicted Goodness score for essay
Stefan Ruseti, et al. in (2018)	BiGRU Siamese architecture	Amazon Mechanical Turk online research service. Collected summaries	Word embedding	Accuracy 55.2
Zining wang, et al. in (2018a)	LSTM (semantic) HAN (hierarchical attention network) neural network	ASAP Kaggle	Word embedding	QWK 0.83
Guoxi Liang et al. (2018)	Bi-LSTM	ASAP Kaggle	Word embedding, coherence of sentence	QWK 0.801

Table 5 (continued)

System	Approach	Dataset	Features applied	Evaluation metric and results
Ke et al. in (2019)	Classification model	ASAP Kaggle	Content based	Pearson's Correlation Coefficient (PC)-0.39 ME-0.921
Tsegaye Misikir Tashu and Horváth in (2019)	Unsupervised learning–Locality Sensitivity Hashing	ASAP Kaggle	Statistical features	root mean squared error
Kumar and Dey (2019)	Random Forest CNN, RNN neural network	ASAP Kaggle short Answer	Style and content-based features	QWK 0.82
Pedro Uria Rodriguez et al. (2019)	BERT, Xlnet	ASAP Kaggle	Error correction, sequence learning	QWK 0.755
Jiawei Liu et al. (2019)	CNN, LSTM, BERT	ASAP Kaggle	semantic data, handcrafted features like grammar correction, essay length, number of sentences, etc	QWK 0.709
Darwish and Mohamed (2020)	Multiple Linear Regression	ASAP Kaggle	Style and content-based features	QWK 0.77
Jiaqi Lun et al. (2020)	BERT	SemEval-2013	Student Answer, Reference Answer	Accuracy 0.8277 (2-way)
Süzen, Neslihan, et al. (2020)	Text mining	introductory computer science class in the University of North Texas, Student Assignments	Sentence similarity	Correlation score 0.81
Wilson Zhu and Yu Sun in (2020)	RNN (LSTM, Bi-LSTM)	ASAP Kaggle	Word embedding, grammar count, word count	QWK 0.70
Salim Yafet et al. (2019)	XGBoost machine learning classifier	ASAP Kaggle	Word count, POS, parse tree, coherence, cohesion, type token ration	Accuracy 68.12
Andrzej Cader (2020)	Deep Neural Network	University of Social Sciences in Lodz students' answers	asynchronous feature	Accuracy 0.99
Tashu TM, Horváth T (2019)	Rule based algorithm, Similarity based algorithm	ASAP Kaggle	Similarity based	Accuracy 0.68

Table 5 (continued)

System	Approach	Dataset	Features applied	Evaluation metric and results
Masaki Uto(B) and Masashi Okano (2020)	Item Response Theory Models (CNN-LSTM,BERT)	ASAP Kaggle		QWK 0.749

Table 6 Vector representation of essays

	Essay	BoW << vector >>	Word2vec << vector >>
Student 1 response	I believe that using computers will benefit us in many ways like talking and becoming friends will others through websites like facebook and mysace	<< 0.00000 0.00000 0.165746 0.280633 … 0.00000 0.280633 0.280633 0.280633 >>	<< 3.9792988e-03 − 1.9810481e-03 1.9830784e-03 9.0381579e-04 − 2.9438005e-03 2.1778699e-03 4.4950014e-03 2.9508960e -03 − 2.2331756e-03 − 3.8774475e-03 3.5967759e- 03 − 4.0194849e-03 − 3.0412588e-03 − 2.4055617e-03 4.8296354e-03 2.4813593e-03… − 2.7158875e-03 − 1.4563646e-03 1.4072991e-03 − 5.2228488e-04 − 2.3597316e-03 6.2979700e-04 − 3.0249553e-03 4.4125126e-04 2.1633594e-03 − 4.9487003e-03 9.9755758e-05 − 2.4388896e-03 >>
Student 2 response	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people	<< 0.26043 0.26043 0.153814 0.000000 … 0.26043 0.000000 0.000000 0.000000 > >	<< 3.9792988e-03 − 1.9810481e- 03 1.9830784e-03 9.0381579e-04 − 2.9438005e-03 2.1778699e-03 4.4950014e-03 2.9508960e-03 − 2.2331756e-03 − 3.8774475e-03 3.5967759e-03 − 4.0194849e-03… − 2.7158875e-03 − 1.4563646e-03 1.4072991e-03 − 5.2228488e-04 − 2.3597316e-03 6.2979700e-04 − 3.0249553e-03 4.4125126e-04 3.7868773e-03 − 4.4193151e-03 3.0735810e-03 2.5546195e-03 2.1633594e-03 − 4.9487003e-03 9.9755758e-05 − 2.4388896e-03 >>

Table 7 Comparison of models

	BoW	Word2vec
Regression models/classification models	The system implemented with Bow features and regression or classification algorithms will have low cohesion and coherence	The system implemented with Word2vec features and regression or classification algorithms will have low to medium cohesion and coherence
Neural Networks (LSTM)	The system implemented with BoW features and neural network models will have low cohesion and coherence	The system implemented with Word2vec features and neural network model (LSTM) will have medium to high cohesion and coherence

3.5 What are the challenges/limitations in the current research?

From our study and results discussed in the previous sections, many researchers worked on automated essay scoring systems with numerous techniques. We have statistical methods, classification methods, and neural network approaches to evaluate the essay automatically. The main goal of the automated essay grading system is to reduce human effort and improve consistency.

The vast majority of essay scoring systems are dealing with the efficiency of the algorithm. But there are many challenges in automated essay grading systems. One should assess the essay by following parameters like the relevance of the content to the prompt, development of ideas, Cohesion, Coherence, and domain knowledge.

No model works on the relevance of content, which means whether student response or explanation is relevant to the given prompt or not if it is relevant to how much it is appropriate, and there is no discussion about the cohesion and coherence of the essays. All researches concentrated on extracting the features using some NLP libraries, trained their models, and testing the results. But there is no explanation in the essay evaluation system about consistency and completeness, But Palma and Atkinson (2018) explained coherence-based essay evaluation. And Zupanc and Bosnic (2014) also used the word coherence to evaluate essays. And they found consistency with latent semantic analysis (LSA) for finding coherence from essays, but the dictionary meaning of coherence is "The quality of being logical and consistent."

Another limitation is there is no domain knowledge-based evaluation of essays using Machine Learning models. For example, the meaning of a cell is different from biology to physics. Many Machine Learning models extract features with WordVec and GloVec; these NLP libraries cannot convert the words into vectors when they have two or more meanings.

3.5.1 Other challenges that influence the Automated Essay Scoring Systems.

All these approaches worked to improve the QWK score of their models. But QWK will not assess the model in terms of features extraction and constructed irrelevant answers. The QWK is not evaluating models whether the model is correctly assessing the answer or not. There are many challenges concerning students' responses to the Automatic scoring system. Like in evaluating approach, no model has examined how to evaluate the constructed

Table 8 Comparison of all models with respect to cohesion, coherence, completeness, feedback

Authors	Cohesion	Coherence	Completeness	Feed Back
Mohler and Mihalcea (2009)	Low	Low	Low	Low
Mohler et al. (2011)	Medium	Low	Medium	Low
Persing and Ng (2013)	Medium	Low	Low	Low
Adamson et al. (2014)	Low	Low	Low	Low
Ramachandran et al. (2015)	Medium	Medium	Low	Low
Sakaguchi et al.. (2015),	Medium	Low	Low	Low
Cummins et al. (2016)	Low	Low	Low	Low
Sultan et al. (2016)	Medium	Medium	Low	Low
Shehab et al. (2016)	Low	Low	Low	Low
Kopparapu and De (2016)	Medium	Medium	Low	Low
Dong an Zhang (2016)	Medium	Low	Low	Low
Taghipour and Ng (2016)	Medium	Medium	Low	Low
Zupanc et al. (2017)	Medium	Medium	Low	Low
Dong et al. (2017)	Medium	Medium	Low	Low
Riordan et al. (2017)	Medium	Medium	Medium	Low
Zhao et al. (2017)	Medium	Medium	Low	Low
Contreras et al. (2018)	Medium	Low	Low	Low
Mathias and Bhattacharyya (2018a; b)	Medium	Medium	Low	Low
Mathias and Bhattacharyya (2018a; b)	Medium	Medium	Low	Low
Nguyen and Dery (2016)	Medium	Medium	Medium	Medium
Ruseti et al. (2018)	Medium	Low	Low	Low
Dasgupta et al. (2018)	Medium	Medium	Low	Low
Liu et al.(2018)	Low	Low	Low	Low
Wang et al. (2018b)	Medium	Low	Low	Low
Guoxi Liang et al. (2018)	High	High	Low	Low
Wang et al. (2018a)	Medium	Medium	Low	Low
Chen and Li (2018)	Medium	Medium	Low	Low
Li et al. (2018)	Medium	Medium	Low	Low
Alva-Manchego et al.(2019)	Low	Low	Low	Low
Jiawei Liu et al. (2019)	High	High	Medium	Low
Pedro Uria Rodriguez et al. (2019)	Medium	Medium	Medium	Low
Changzhi Cai(2019)	Low	Low	Low	Low
Xia et al. (2019)	Medium	Medium	Low	Low
Chen and Zhou (2019)	Low	Low	Low	Low
Kumar et al. (2019)	Medium	Medium	Medium	Low
Ke et al. (2019)	Medium	Low	Medium	Low
Andrzej Cader(2020)	Low	Low	Low	Low
Jiaqi Lun et al. (2020)	High	High	Low	Low
Wilson Zhu and Yu Sun (2020)	Medium	Medium	Low	Low
Süzen, Neslihan et al. (2020)	Medium	Low	Medium	Low
Salim Yafet et al. (2019)	High	Medium	Low	Low
Darwish and Mohamed (2020)	Medium	Low	Low	Low
Tashu and Horváth (2020)	Medium	Medium	Low	Medium
Tashu (2020)	Medium	Medium	Low	Low

Table 8 (continued)

Authors	Cohesion	Coherence	Completeness	Feed Back
Masaki Uto(B) and Masashi Okano(2020)	Medium	Medium	Medium	Medium
Panitan Muangkammuen and Fumiyo Fukumoto(2020)	Medium	Medium	Medium	Low

irrelevant and adversarial answers. Especially the black box type of approaches like deep learning models provides more options to the students to bluff the automated scoring systems.

The Machine Learning models that work on statistical features are very vulnerable. Based on Powers et al. (2001) and Bejar Isaac et al. (2014), the E-rater was failed on Constructed Irrelevant Responses Strategy (CIRS). From the study of Bejar et al. (2013), Higgins and Heilman (2014), observed that when student response contain irrelevant content or shell language concurring to prompt will influence the final score of essays in an automated scoring system.

In deep learning approaches, most of the models automatically read the essay's features, and some methods work on word-based embedding and other character-based embedding features. From the study of Riordan Brain et al. (2019), The character-based embedding systems do not prioritize spelling correction. However, it is influencing the final score of the essay. From the study of Horbach and Zesch (2019), Various factors are influencing AES systems. For example, there are data set size, prompt type, answer length, training set, and human scorers for content-based scoring.

Ding et al. (2020) reviewed that the automated scoring system is vulnerable when a student response contains more words from prompt, like prompt vocabulary repeated in the response. Parekh et al. (2020) and Kumar et al. (2020) tested various neural network models of AES by iteratively adding important words, deleting unimportant words, shuffle the words, and repeating sentences in an essay and found that no change in the final score of essays. These neural network models failed to recognize common sense in adversaries' essays and give more options for the students to bluff the automated systems.

Other than NLP and ML techniques for AES. From Wresch (1993) to Madnani and Cahill (2018). discussed the complexity of AES systems, standards need to be followed. Like assessment rubrics to test subject knowledge, irrelevant responses, and ethical aspects of an algorithm like measuring the fairness of student response.

Fairness is an essential factor for automated systems. For example, in AES, fairness can be measure in an agreement between human score to machine score. Besides this, From Loukina et al. (2019), the fairness standards include overall score accuracy, overall score differences, and condition score differences between human and system scores. In addition, scoring different responses in the prospect of constructive relevant and irrelevant will improve fairness.

Madnani et al. (2017a; b). Discussed the fairness of AES systems for constructed responses and presented RMS open-source tool for detecting biases in the models. With this, one can change fairness standards according to their analysis of fairness.

From Berzak et al.'s (2018) approach, behavior factors are a significant challenge in automated scoring systems. That helps to find language proficiency, word characteristics (essential words from the text), predict the critical patterns from the text, find related sentences in an essay, and give a more accurate score.

Table 9 comparison of all approaches on various features

Approaches	Grammar	Style (Word choice, sentence structure)	Mechanics (Spelling, punctuation, capitalization)	Development	BoW (tf-idf)	relevance
Mohler and Mihalcea (2009)	No	No	No	No	Yes	No
Mohler et al. (2011)	Yes	No	No	No	Yes	No
Persing and Ng (2013)	Yes	Yes	Yes	No	Yes	Yes
Adamson et al. (2014)	Yes	No	Yes	No	Yes	No
Ramachandran et al. (2015)	Yes	No	Yes	Yes	Yes	Yes
Sakaguchi et al. (2015),	No	No	Yes	Yes	Yes	Yes
Cummins et al. (2016)	Yes	No	Yes	No	Yes	No
Sultan et al. (2016)	No	No	No	No	Yes	Yes
Shehab et al. (2016)	Yes	Yes	Yes	No	Yes	No
Kopparapu and De (2016)	No	No	No	No	Yes	No
Dong and Zhang (2016)	Yes	No	Yes	No	Yes	Yes
Taghipour and Ng (2016)	Yes	No	No	No	Yes	Yes
Zupanc et al. (2017)	No	No	No	No	Yes	No
Dong et al. (2017)	No	No	No	No	No	Yes
Riordan et al. (2017)	No	No	No	No	No	Yes
Zhao et al. (2017)	No	No	No	No	No	Yes
Contreras et al. (2018)	Yes	No	No	No	Yes	Yes
Mathias and Bhattacharyya (2018a, 2018b)	No	Yes	Yes	No	No	Yes
Mathias and Bhattacharyya (2018a, 2018b)	Yes	No	Yes	No	Yes	Yes
Nguyen and Dery (2016)	No	No	No	No	Yes	Yes
Ruseti et al. (2018)	No	No	No	Yes	No	Yes
Dasgupta et al. (2018)	Yes	Yes	Yes	Yes	No	Yes
Liu et al.(2018)	Yes	Yes	No	No	Yes	No
Wang et al. (2018b)	No	No	No	No	No	Yes
Guoxi Liang et al. (2018)	No	No	No	No	No	Yes
Wang et al. (2018a)	No	No	No	No	No	Yes

Table 9 (continued)

Approaches	Grammar	Style (Word choice, sentence structure)	Mechanics (Spelling, punctuation, capitalization)	Development	BoW (tf-idf)	relevance
Chen and Li (2018)	No	No	No	No	No	Yes
Li et al. (2018)	Yes	No	No	No	No	Yes
Alva-Manchego et al. (2019)	Yes	No	No	Yes	No	Yes
Jiawei Liu et al. (2019)	Yes	No	No	Yes	No	Yes
Pedro Uria Rodriguez et al. (2019)	No	No	No	No	Yes	Yes
Changzhi Cai(2019)	No	No	No	No	No	Yes
Xia et al. (2019)	No	No	No	No	No	Yes
Chen and Zhou (2019)	No	No	No	No	No	Yes
Kumar et al. (2019)	Yes	Yes	No	Yes	Yes	Yes
Ke et al. (2019)	No	Yes	No	Yes	Yes	Yes
Andrzej Cader(2020)	No	No	No	No	No	Yes
Jiaqi Lun et al. (2020)	No	No	No	No	No	Yes
Wilson Zhu and Yu Sun (2020)	No	No	No	No	No	Yes
Süzen, Neslihan, et al. (2020)	No	No	No	No	Yes	Yes
Salim Yafet et al. (2019)	Yes	Yes	Yes	No	Yes	Yes
Darwish and Mohamed (2020)	Yes	Yes	No	No	No	Yes

Rupp (2018), has discussed the designing, evaluating, and deployment methodologies for AES systems. They provided notable characteristics of AES systems for deployment. They are like model performance, evaluation metrics for a model, threshold values, dynamically updated models, and framework.

First, we should check the model performance on different datasets and parameters for operational deployment. Selecting Evaluation metrics for AES models are like QWK, correlation coefficient, or sometimes both. Kelley and Preacher (2012) have discussed three categories of threshold values: marginal, borderline, and acceptable. The values can be varied based on data size, model performance, type of model (single scoring, multiple scoring models). Once a model is deployed and evaluates millions of responses every time for optimal responses, we need a dynamically updated model based on prompt and data. Finally, framework designing of AES model, hear a framework contains prompts where test-takers can write the responses. One can design two frameworks: a single scoring model for a single methodology and multiple scoring models for multiple concepts. When we deploy multiple scoring models, each prompt could be trained separately, or we can provide generalized models for all prompts with this accuracy may vary, and it is challenging.

4 Synthesis

Our Systematic literature review on the automated essay grading system first collected 542 papers with selected keywords from various databases. After inclusion and exclusion criteria, we left with 139 articles; on these selected papers, we applied Quality assessment criteria with two reviewers, and finally, we selected 62 writings for final review.

Our observations on automated essay grading systems from 2010 to 2020 are as followed:

- The implementation techniques of automated essay grading systems are classified into four buckets; there are 1. regression models 2. Classification models 3. Neural networks 4. Ontology-based methodology, but using neural networks, the researchers are more accurate than other techniques, and all the methods state of the art provided in Table 3.
- The majority of the regression and classification models on essay scoring used statistical features to find the final score. It means the systems or models trained on such parameters as word count, sentence count, etc. though the parameters extracted from the essay, the algorithm are not directly training on essays. The algorithms trained on some numbers obtained from the essay and hear if numbers matched the composition will get a good score; otherwise, the rating is less. In these models, the evaluation process is entirely on numbers, irrespective of the essay. So, there is a lot of chance to miss the coherence, relevance of the essay if we train our algorithm on statistical parameters.
- In the neural network approach, the models trained on Bag of Words (BoW) features. The BoW feature is missing the relationship between a word to word and the semantic meaning of the sentence. E.g., Sentence 1: John killed bob. Sentence 2: bob killed John. In these two sentences, the BoW is "John," "killed," "bob."
- In the Word2Vec library, if we are prepared a word vector from an essay in a unidirectional way, the vector will have a dependency with other words and finds the semantic relationship with other words. But if a word has two or more meanings like "Bank loan" and "River Bank," hear bank has two implications, and its adjacent words decide

the sentence meaning; in this case, Word2Vec is not finding the real meaning of the word from the sentence.

- The features extracted from essays in the essay scoring system are classified into 3 type's features like statistical features, style-based features, and content-based features, which are explained in RQ2 and Table 3. But statistical features, are playing a significant role in some systems and negligible in some systems. In Shehab et al. (2016); Cummins et al. (2016). Dong et al. (2017). Dong and Zhang (2016). Mathias and Bhattacharyya (2018a; b) Systems the assessment is entirely on statistical and style-based features they have not retrieved any content-based features. And in other systems that extract content from the essays, the role of statistical features is for only preprocessing essays but not included in the final grading.
- In AES systems, coherence is the main feature to be considered while evaluating essays. The actual meaning of coherence is to stick together. That is the logical connection of sentences (local level coherence) and paragraphs (global level coherence) in a story. Without coherence, all sentences in a paragraph are independent and meaningless. In an Essay, coherence is a significant feature that is explaining everything in a flow and its meaning. It is a powerful feature in AES system to find the semantics of essay. With coherence, one can assess whether all sentences are connected in a flow and all paragraphs are related to justify the prompt. Retrieving the coherence level from an essay is a critical task for all researchers in AES systems.
- In automatic essay grading systems, the assessment of essays concerning content is critical. That will give the actual score for the student. Most of the researches used statistical features like sentence length, word count, number of sentences, etc. But according to collected results, 32% of the systems used content-based features for the essay scoring. Example papers which are on content-based assessment are Taghipour and Ng (2016); Persing and Ng (2013); Wang et al. (2018a, 2018b); Zhao et al. (2017); Kopparapu and De (2016), Kumar et al. (2019); Mathias and Bhattacharyya (2018a; b); Mohler and Mihalcea (2009) are used content and statistical-based features. The results are shown in Fig. 3. And mainly the content-based features extracted with word2vec NLP library, but word2vec is capable of capturing the context of a word in a document, semantic and syntactic similarity, relation with other terms, but word2vec is capable of capturing the context word in a uni-direction either left or right. If a word has multiple meanings, there is a chance of missing the context in the essay. After analyzing all the papers, we found that content-based assessment is a qualitative assessment of essays.
- On the other hand, Horbach and Zesch (2019); Riordan Brain et al. (2019); Ding et al. (2020); Kumar et al. (2020) proved that neural network models are vulnerable when a student response contains constructed irrelevant, adversarial answers. And a student can easily bluff an automated scoring system by submitting different responses like repeating sentences and repeating prompt words in an essay. From Loukina et al. (2019), and Madnani et al. (2017b). The fairness of an algorithm is an essential factor to be considered in AES systems.
- While talking about speech assessment, the data set contains audios of duration up to one minute. Feature extraction techniques are entirely different from text assessment, and accuracy varies based on speaking fluency, pitching, male to female voice and boy to adult voice. But the training algorithms are the same for text and speech assessment.
- Once an AES system evaluates essays and short answers accurately in all directions, there is a massive demand for automated systems in the educational and related world. Now AES systems are deployed in GRE, TOEFL exams; other than these, we can deploy AES systems in massive open online courses like Coursera("https://coursera.

org/learn//machine-learning//exam"), NPTEL (https://swayam.gov.in/explorer), etc. still they are assessing student performance with multiple-choice questions. In another perspective, AES systems can be deployed in information retrieval systems like Quora, stack overflow, etc., to check whether the retrieved response is appropriate to the question or not and can give ranking to the retrieved answers.

5 Conclusion and future work

As per our Systematic literature review, we studied 62 papers. There exist significant challenges for researchers in implementing automated essay grading systems. Several researchers are working rigorously on building a robust AES system despite its difficulty in solving this problem. All evaluating methods are not evaluated based on coherence, relevance, completeness, feedback, and knowledge-based. And 90% of essay grading systems are used Kaggle ASAP (2012) dataset, which has general essays from students and not required any domain knowledge, so there is a need for domain-specific essay datasets to train and test. Feature extraction is with NLTK, WordVec, and GloVec NLP libraries; these libraries have many limitations while converting a sentence into vector form. Apart from feature extraction and training Machine Learning models, no system is accessing the essay's completeness. No system provides feedback to the student response and not retrieving coherence vectors from the essay—another perspective the constructive irrelevant and adversarial student responses still questioning AES systems.

Our proposed research work will go on the content-based assessment of essays with domain knowledge and find a score for the essays with internal and external consistency. And we will create a new dataset concerning one domain. And another area in which we can improve is the feature extraction techniques.

This study includes only four digital databases for study selection may miss some functional studies on the topic. However, we hope that we covered most of the significant studies as we manually collected some papers published in useful journals.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10462-021-10068-2.

Funding Not Applicable.

References

Adamson, A., Lamb, A., & December, R. M. (2014). Automated Essay Grading.

Ajay HB, Tillett PI, Page EB (1973) Analysis of essays by computer (AEC-II) (No. 8-0102). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development

Ajetunmobi SA, Daramola O (2017) Ontology-based information extraction for subject-focussed automatic essay evaluation. In: 2017 International Conference on Computing Networking and Informatics (ICCNI) p 1–6. IEEE

Alva-Manchego F, et al. (2019) EASSE: Easier Automatic Sentence Simplification Evaluation." ArXiv abs/1908.04567 (2019): n. pag

Bailey S, Meurers D (2008) Diagnosing meaning errors in short answers to reading comprehension questions. In: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (Columbus), p 107–115

Basu S, Jacobs C, Vanderwende L (2013) Powergrading: a clustering approach to amplify human effort for short answer grading. Trans Assoc Comput Linguist (TACL) 1:391–402

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. Assessing Writing, 22, 48-59.
Bejar I, et al. (2013) Length of Textual Response as a Construct-Irrelevant Response Strategy: The Case of Shell Language. Research Report. ETS RR-13-07." ETS Research Report Series (2013): n. pag
Berzak Y, et al. (2018) "Assessing Language Proficiency from Eye Movements in Reading." ArXiv abs/1804.07329 (2018): n. pag
Blanchard D, Tetreault J, Higgins D, Cahill A, Chodorow M (2013) TOEFL11: A corpus of non-native English. ETS Research Report Series, 2013(2):i–15, 2013
Blood, I. (2011). Automated essay scoring: a literature review. Studies in Applied Linguistics and TESOL, 11(2).
Burrows S, Gurevych I, Stein B (2015) The eras and trends of automatic short answer grading. Int J Artif Intell Educ 25:60–117. https://doi.org/10.1007/s40593-014-0026-8
Cader, A. (2020, July). The Potential for the Use of Deep Neural Networks in e-Learning Student Evaluation with New Data Augmentation Method. In International Conference on Artificial Intelligence in Education (pp. 37–42). Springer, Cham.
Cai C (2019) Automatic essay scoring with recurrent neural network. In: Proceedings of the 3rd International Conference on High Performance Compilation, Computing and Communications (2019): n. pag.
Chen M, Li X (2018) "Relevance-Based Automated Essay Scoring via Hierarchical Recurrent Model. In: 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, p 378–383, doi: https://doi.org/10.1109/IALP.2018.8629256
Chen Z, Zhou Y (2019) "Research on Automatic Essay Scoring of Composition Based on CNN and OR. In: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, p 13–18, doi: https://doi.org/10.1109/ICAIBD.2019.8837007
Contreras JO, Hilles SM, Abubakar ZB (2018) Automated essay scoring with ontology based on text mining and NLTK tools. In: 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 1-6
Correnti R, Matsumura LC, Hamilton L, Wang E (2013) Assessing students' skills at writing analytically in response to texts. Elem Sch J 114(2):142–177
Cummins, R., Zhang, M., & Briscoe, E. (2016, August). Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.
Darwish SM, Mohamed SK (2020) Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In: Hassanien A, Azar A, Gaber T, Bhatnagar RF, Tolba M (eds) The International Conference on Advanced Machine Learning Technologies and Applications
Dasgupta T, Naskar A, Dey L, Saha R (2018) Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications p 93–102
Ding Y, et al. (2020) "Don't take "nswvtnvakgxpm" for an answer–The surprising vulnerability of automatic content scoring systems to adversarial input." In: Proceedings of the 28th International Conference on Computational Linguistics
Dong F, Zhang Y (2016) Automatic features for essay scoring–an empirical study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing p 1072–1077
Dong F, Zhang Y, Yang J (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) p 153–162
Dzikovska M, Nielsen R, Brew C, Leacock C, Gi ampiccolo D, Bentivogli L, Clark P, Dagan I, Dang HT (2013a) Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge
Dzikovska MO, Nielsen R, Brew C, Leacock C, Giampiccolo D, Bentivogli L, Clark P, Dagan I, Trang Dang H (2013b) SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. *SEM 2013: The First Joint Conference on Lexical and Computational Semantics
Educational Testing Service (2008) CriterionSM online writing evaluation service. Retrieved from http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf .
Evanini, K., & Wang, X. (2013, August). Automated speech scoring for non-native middle school students with multiple task types. In INTERSPEECH (pp. 2435–2439).
Foltz PW, Laham D, Landauer TK (1999) The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1, 2, http://imej.wfu.edu/articles/1999/2/04/ index.asp

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). International corpus of learner English. Louvain-la-Neuve: Presses universitaires de Louvain.
Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. Educational Measurement: Issues and Practice, 33(3), 36–46.
Horbach A, Zesch T (2019) The influence of variance in learner answers on automatic content scoring. Front Educ 4:28. https://doi.org/10.3389/feduc.2019.00028
https://www.coursera.org/learn/machine-learning/exam/7pytE/linear-regression-with-multiple-variables/attempt
Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5, e208.
Ke Z, Ng V (2019) "Automated essay scoring: a survey of the state of the art." IJCAI
Ke, Z., Inamdar, H., Lin, H., & Ng, V. (2019, July). Give me more feedback II: Annotating thesis strength and related attributes in student essays. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3994-4004).
Kelley K, Preacher KJ (2012) On effect size. Psychol Methods 17(2):137–152
Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering–a systematic literature review. Inf Softw Technol 51(1):7–15
Klebanov, B. B., & Madnani, N. (2020, July). Automated evaluation of writing–50 years and counting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7796–7810).
Knill K, Gales M, Kyriakopoulos K, et al. (4 more authors) (2018) Impact of ASR performance on free speaking language assessment. In: Interspeech 2018.02–06 Sep 2018, Hyderabad, India. International Speech Communication Association (ISCA)
Kopparapu SK, De A (2016) Automatic ranking of essays using structural and semantic features. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), p 519–523
Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019, July). Get it scored using autosas—an automated system for scoring short answers. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9662–9669).
Kumar Y, et al. (2020) "Calling out bluff: attacking the robustness of automatic scoring systems with simple adversarial testing." ArXiv abs/2007.06796
Li X, Chen M, Nie J, Liu Z, Feng Z, Cai Y (2018) Coherence-Based Automated Essay Scoring Using Self-attention. In: Sun M, Liu T, Wang X, Liu Z, Liu Y (eds) Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. CCL 2018, NLP-NABD 2018. Lecture Notes in Computer Science, vol 11221. Springer, Cham. https://doi.org/10.1007/978-3-030-01716-3_32
Liang G, On B, Jeong D, Kim H, Choi G (2018) Automated essay scoring: a siamese bidirectional LSTM neural network architecture. Symmetry 10:682
Liua, H., Yeb, Y., & Wu, M. (2018, April). Ensemble Learning on Scoring Student Essay. In 2018 International Conference on Management and Education, Humanities and Social Sciences (MEHSS 2018). Atlantis Press.
Liu J, Xu Y, Zhao L (2019) Automated Essay Scoring based on Two-Stage Learning. ArXiv, abs/1901.07744
Loukina A, et al. (2015) Feature selection for automated speech scoring." BEA@NAACL-HLT
Loukina A, et al. (2017) "Speech- and Text-driven Features for Automated Scoring of English-Speaking Tasks." SCNLP@EMNLP 2017
Loukina A, et al. (2019) The many dimensions of algorithmic fairness in educational applications. BEA@ACL
Lun J, Zhu J, Tang Y, Yang M (2020) Multiple data augmentation strategies for improving performance on automatic short answer scoring. In: Proceedings of the AAAI Conference on Artificial Intelligence, 34(09): 13389-13396
Madnani, N., & Cahill, A. (2018, August). Automated scoring: Beyond natural language processing. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1099–1109).
Madnani N, et al. (2017b) "Building better open-source tools to support fairness in automated scoring." EthNLP@EACL
Malinin A, et al. (2016) "Off-topic response detection for spontaneous spoken english assessment." ACL
Malinin A, et al. (2017) "Incorporating uncertainty into deep learning for spoken language assessment." ACL

Mathias S, Bhattacharyya P (2018a) Thank "Goodness"! A Way to Measure Style in Student Essays. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications p 35–41

Mathias S, Bhattacharyya P (2018b) ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Mikolov T, et al. (2013) "Efficient Estimation of Word Representations in Vector Space." ICLR

Mohler M, Mihalcea R (2009) Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009) p 567–575

Mohler M, Bunescu R, Mihalcea R (2011) Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies p 752–762

Muangkammuen P, Fukumoto F (2020) Multi-task Learning for Automated Essay Scoring with Sentiment Analysis. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop p 116–123

Nguyen, H., & Dery, L. (2016). Neural networks for automated essay grading. CS224d Stanford Reports, 1–11.

Palma D, Atkinson J (2018) Coherence-based automatic essay assessment. IEEE Intell Syst 33(5):26–36

Parekh S, et al (2020) My Teacher Thinks the World Is Flat! Interpreting Automatic Essay Scoring Mechanism." ArXiv abs/2012.13872 (2020): n. pag

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532–1543).

Persing I, Ng V (2013) Modeling thesis clarity in student essays. In:Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) p 260–269

Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K (2001) Stumping E-Rater: challenging the validity of automated essay scoring. ETS Res Rep Ser 2001(1):i–44

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: challenging the validity of automated essay scoring. Computers in Human Behavior, 18(2), 103–134.

Ramachandran L, Cheng J, Foltz P (2015) Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications p 97–106

Ramanarayanan V, et al. (2017) "Human and Automated Scoring of Fluency, Pronunciation and Intonation During Human-Machine Spoken Dialog Interactions." INTERSPEECH

Riordan B, Horbach A, Cahill A, Zesch T, Lee C (2017) Investigating neural architectures for short answer scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications p 159–168

Riordan B, Flor M, Pugh R (2019) "How to account for misspellings: Quantifying the benefit of character representations in neural content scoring models."In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications

Rodriguez P, Jafari A, Ormerod CM (2019) Language models and Automated Essay Scoring. ArXiv, abs/1909.09482

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1(2).

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. The Journal of Technology, Learning and Assessment, 4(4).

Rupp A (2018) Designing, evaluating, and deploying automated scoring systems with validity in mind: methodological design decisions. Appl Meas Educ 31:191–214

Ruseti S, Dascalu M, Johnson AM, McNamara DS, Balyan R, McCarthy KS, Trausan-Matu S (2018) Scoring summaries using recurrent neural networks. In: International Conference on Intelligent Tutoring Systems p 191–201. Springer, Cham

Sakaguchi K, Heilman M, Madnani N (2015) Effective feature integration for automated short answer scoring. In: Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies p 1049–1054

Salim, Y., Stevanus, V., Barlian, E., Sari, A. C., & Suhartono, D. (2019, December). Automated English Digital Essay Grader Using Machine Learning. In 2019 IEEE International Conference on Engineering, Technology and Education (TALE) (pp. 1–6). IEEE.

Shehab A, Elhoseny M, Hassanien AE (2016) A hybrid scheme for Automated Essay Grading based on LVQ and NLP techniques. In: 12th International Computer Engineering Conference (ICENCO), Cairo, 2016, p 65-70

Shermis MD, Mzumara HR, Olson J, Harrington S (2001) On-line grading of student essays: PEG goes on the World Wide Web. Assess Eval High Educ 26(3):247–259

Stab C, Gurevych I (2014) Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) p 46–56

Sultan MA, Salazar C, Sumner T (2016) Fast and easy short answer grading with high accuracy. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies p 1070–1075

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. Procedia Computer Science, 169, 726–743.

Taghipour K, Ng HT (2016) A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing p 1882–1891

Tashu TM (2020) "Off-Topic Essay Detection Using C-BGRU Siamese. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, p 221–225, doi: https://doi.org/10.1109/ICSC.2020.00046

Tashu TM, Horváth T (2019) A layered approach to automatic essay evaluation using word-embedding. In: McLaren B, Reilly R, Zvacek S, Uhomoibhi J (eds) Computer Supported Education. CSEDU 2018. Communications in Computer and Information Science, vol 1022. Springer, Cham

Tashu TM, Horváth T (2020) Semantic-Based Feedback Recommendation for Automatic Essay Evaluation. In: Bi Y, Bhatia R, Kapoor S (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1038. Springer, Cham

Uto M, Okano M (2020) Robust Neural Automated Essay Scoring Using Item Response Theory. In: Bittencourt I, Cukurova M, Muldner K, Luckin R, Millán E (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science, vol 12163. Springer, Cham

Wang Z, Liu J, Dong R (2018a) Intelligent Auto-grading System. In: 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS) p 430–435. IEEE.

Wang Y, et al. (2018b) "Automatic Essay Scoring Incorporating Rating Schema via Reinforcement Learning." EMNLP

Zhu W, Sun Y (2020) Automated essay scoring system using multi-model Machine Learning, david c. wyld et al. (eds): mlnlp, bdiot, itccma, csity, dtmn, aifz, sigpro

Wresch W (1993) The Imminence of Grading Essays by Computer-25 Years Later. Comput Compos 10:45–58

Wu, X., Knill, K., Gales, M., & Malinin, A. (2020). Ensemble approaches for uncertainty in spoken language assessment.

Xia L, Liu J, Zhang Z (2019) Automatic Essay Scoring Model Based on Two-Layer Bi-directional Long-Short Term Memory Network. In: Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence p 133–137

Yannakoudakis H, Briscoe T, Medlock B (2011) A new dataset and method for automatically grading ESOL texts. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies p 180–189

Zhao S, Zhang Y, Xiong X, Botelho A, Heffernan N (2017) A memory-augmented neural model for automated grading. In: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale p 189–192

Zupanc K, Bosnic Z (2014) Automated essay evaluation augmented with semantic coherence measures. In: 2014 IEEE International Conference on Data Mining p 1133–1138. IEEE.

Zupanc K, Savić M, Bosnić Z, Ivanović M (2017) Evaluating coherence of essays using sentence-similarity networks. In: Proceedings of the 18th International Conference on Computer Systems and Technologies p 65–72

Dzikovska, M. O., Nielsen, R., & Brew, C. (2012, June). Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 200-210).

Kumar, N., & Dey, L. (2013, November). Automatic Quality Assessment of documents with application to essay grading. In 2013 12th Mexican International Conference on Artificial Intelligence (pp. 216–222). IEEE.

Wu, S. H., & Shih, W. F. (2018, July). A short answer grading system in chinese by support vector approach. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (pp. 125-129).

Agung Putri Ratna, A., Lalita Luhurkinanti, D., Ibrahim I., Husna D., Dewi Purnamasari P. (2018). Automatic Essay Grading System for Japanese Language Examination Using Winnowing Algorithm, 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 565–569. https://doi.org/10.1109/ISEMANTIC.2018.8549789.

Sharma A., & Jayagopi D. B. (2018). Automated Grading of Handwritten Essays 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp 279–284. https://doi.org/10.1109/ICFHR-2018.2018.00056