Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 9158-9176
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.3956

Automatic assessment of short answer questions: Review

Salma Abdullbaki Mahmood[1], Marwa Ali Abdulsamad[2*]
[1,2]Department of Computer Science, Basra University, Basra, Iraq; pgs.marwaa.ali@uobasrah.edu.iq (M.A.A.).

Abstract: With the rapid development of technology, automated assessment systems have become an essential tool to facilitate the process of correcting short answers. This research aims to explore ways to improve the accuracy of automated assessment using natural language processing techniques, such as Latent Semantic Analysis (LSA) and Longest Common Sequence (LCS) algorithm, hile highlighting the challenges associated with the scarcity of Arabic language datasets.These methodologies facilitate the assessment of both lexical and semantic congruences between student submissions and benchmark answers. The overarching objective is to establish a scalable and precise grading mechanism that reduces manual evaluation's temporal and subjective dimensions. Notwithstanding significant advancements, obstacles such as the scarcity of Arabic datasets persist as a principal impediment to effective automated grading in languages other than English. This research scrutinizes contemporary strategies within the domain, highlighting the imperative for more sophisticated models and extensive datasets to bolster the precision and adaptability of automated grading frameworks, particularly concerning Arabic textual content. With the rapid development of technology, automated assessment systems have become an essential tool to facilitate the process of correcting short answers. This research aims to explore ways to improve the accuracy of automated assessment using natural language processing techniques, such as Latent Semantic Analysis (LSA) and Longest Common Sequence (LCS) algorithm, hile highlighting the challenges associated with the scarcity of Arabic language datasets.These methodologies facilitate the assessment of both lexical and semantic congruences between student submissions and benchmark answers. The overarching objective is to establish a scalable and precise grading mechanism that reduces manual evaluation's temporal and subjective dimensions. Notwithstanding significant advancements, obstacles such as the scarcity of Arabic datasets persist as a principal impediment to effective automated grading in languages other than English. This research scrutinizes contemporary strategies within the domain, highlighting the imperative for more sophisticated models and extensive datasets to bolster the precision and adaptability of automated grading frameworks, particularly concerning Arabic textual content.

Keywords: *Assessment of short answer, Cosine similarity algorithm, Longest common subsequence, Latent semantic analysis.*

1. Introduction

With the rapid technological developments, Automated Writing Assessment (AEG) systems have become essential tools to address the challenges teachers face when correcting student essays. Manual correction, despite its importance, wastes a lot of time that could be invested in more important tasks such as lesson planning and improving teaching strategies. Moreover, the increasing number of students makes it difficult for teachers to maintain consistent assessments across all students. Written essays, despite their usefulness in developing students' skills and preventing cheating, are more difficult to assess than other objective questions. Therefore, automated assessment systems can provide an effective solution by reducing the time required for correction and improving the accuracy of assessments. In this paper, we will discuss the automated assessment process starting from collecting student texts, through pre-analysis and feature extraction processes, to assessing textual similarity and providing accurate scores that reflect the quality of written essays. Teachers may then focus on other

* Correspondence: pgs.marwaa.ali@uobasrah.edu.iq

important tasks, such as creating more effective lesson plans and other forms of communication, rather than only grading students' work [21]. and We found that boredom increased as time went on, and that greater Disinterest led to worse performance in the classroom. Moreover, these undesirable consequences of boredom persisted even after an evaluation plan was provided. Concerns over the use of essays as an evaluation tool are heightened by these findings, which also cast doubt on the usefulness of such methods in general [31]. The compilation of student texts into a text corpus and their subsequent input into the AEG program is one of the several stages in the AEG system process. The AEG system first pre-processes the texts for further processing and analysis [25]. The fundamental pre-processing method comprises dividing the text sequence into chunks called tokens, deleting certain characters like punctuation, and removing any character from foreign languages. It also involves stripping the texts of white spaces. In the second step, feature extraction is usually performed. This approach seeks to transform the text sequence into a vector of measurable variables, typically used occurrences, and word frequencies [25]. Finding out how similar two short sentences are is the goal of text-similarity assessment. It is important to consider both the lexical perspective, which focuses only on character sequences, and the semantic meaning when evaluating similarities. Even if the words used to produce two messages are different, they would have identical semantic content [32]. Various natural language processing (NLP) tasks depend on assessments of semantic similarity among words, sentences, or documents, such as information retrieval [33], text summarization [34], text classification [35], essay evaluation [36], machine translation [37], question answering , and numerous others [38]. In order to address concerns with dependability, cost, and time, it is possible to use Automated Essay Grading (AEG) methods to grade student writings without human intervention. An AEG system can assess and generate a grade for a written essay without human intervention. In order to enhance and speed up the process of essay grading while avoiding human mistakes or biases, the Automatic Essay Grading (AEG) system has recently become an essential tool for universities and institutions [25].
Journal homepage: https://ijici.edu.iq

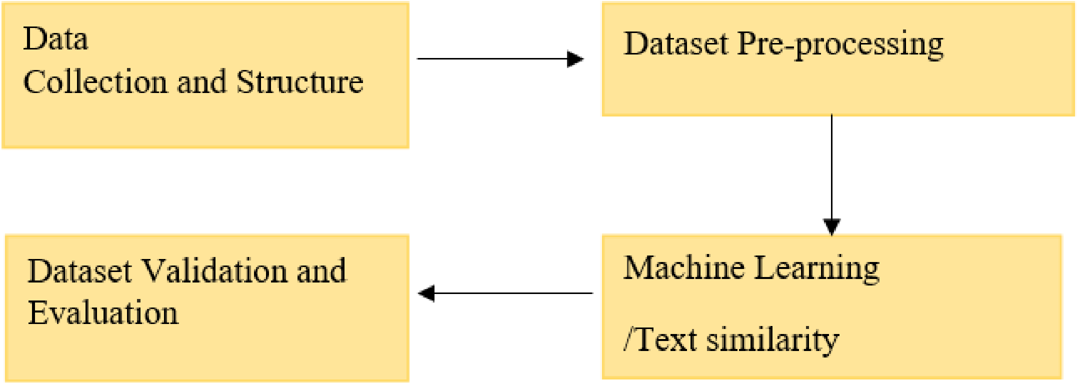

Figure 1.
Shows general structure of AEGS.

2. Related Work

Automated assessment systems for succinct responses have experienced substantial advancements in recent years, with a plethora of studies utilizing natural language processing (NLP) methodologies to systematically evaluate the caliber of students' responses. These investigations are characterized by varying approaches addressing linguistic, syntactic, and semantic evaluations, along with a range of languages and domains in which these systems are implemented.

This review seeks to furnish a thorough examination of the most significant research that has tackled the subject of automated evaluation of short answers, emphasizing the diverse methodologies employed in each investigation. We will scrutinize works that have utilized sophisticated techniques such as Latent Semantic Analysis (LSA) and the Longest Common Subsequence (LCS) algorithm in

English, as well as in Arabic, given that these systems encounter additional complexities associated with the limited availability of datasets.

We will categorize these studies according to the techniques employed, concentrating on the advantages and disadvantages of each, thereby facilitating the identification of existing gaps and potential avenues for enhancement in contemporary systems.

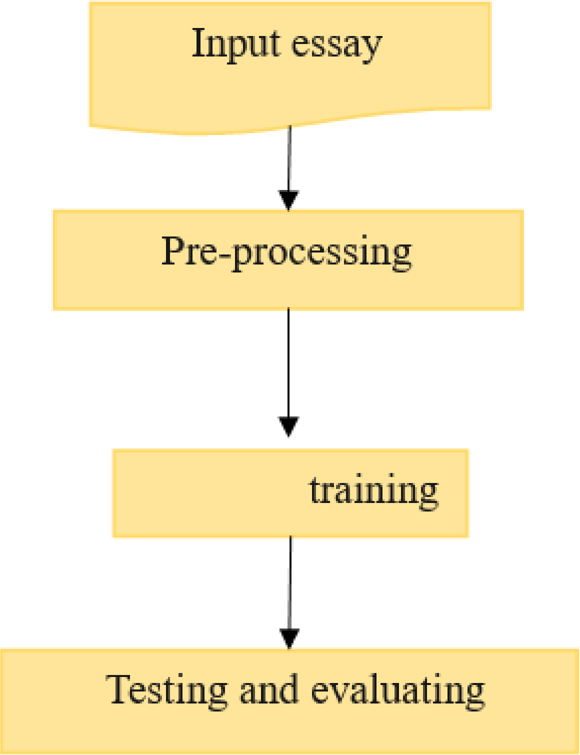

Figure 3.
Shows general structure of machine learning of AEGS.

2.1. Research Based on Deep Learning Algorithms

Hussein et al. (2020) established a trait-oriented deep learning automated essay scoring (AES) framework that delivers adaptive feedback. This framework employs a variety of deep learning architectures, including CNN, RNN, GRU, and LSTM, with the LSTM-based model demonstrating a 4.6% enhancement in scoring precision relative to the baseline. The investigation concentrated on trait-oriented assessment, enabling the system to identify particular traits within essays and provide feedback tailored to those traits. The proposed AES AUG model, assessed using the Quadratic Weighted Kappa (QWK) metric, surpassed the agreement levels of human evaluators in the 2012 ASAP dataset. This model was integrated into the I Assistant educational platform, supplying learners with estimated scores and comprehensive feedback across multiple rubric dimensions. The research concludes that trait-oriented AES contributes to improved overall scoring precision and the quality of feedback, indicating that forthcoming systems should prioritize the expansion of trait-focused evaluations to facilitate more effective feedback mechanisms in educational contexts [8].

In their 2021 study, Kumar and Boulanger investigate automated essay scoring (AES) systems, particularly those using deep learning techniques. They analyze datasets from the 2012 Kaggle Automated Student Assessment Prize, which includes manually graded essays, to assess their suitability for creating effective automated assessment tools. The research employs deep neural networks and advanced natural language processing (NLP) methods to predict rubric scores, aiming to improve the interpretability of the AES system by providing clear explanations for the predicted scores. The study

DOI: 10.55214/25768484.v8i6.3956

measures the system's effectiveness by comparing its agreement with human raters using the quadratic weighted kappa (QWK) statistic. The AES system achieved a QWK score of 0.72, surpassing the average agreement among human assessors (0.60) and previous models (0.53). Additionally, it predicted holistic essay scores with a QWK of 0.78, indicating competitiveness with other leading AES technologies. Despite its successes, the study notes some limitations, including the reliance on initial human rater agreement, which raises concerns about the consistency of evaluations. The authors also highlight that while the system provides rationales for its predictions, the deep learning models still function as "black boxes," making it challenging to fully understand their decision-making processes. The research emphasizes the importance of interpretability and accountability in AES technology and suggests future directions, such as incorporating human validation in grading practices, enhancing model clarity, and improving system performance to better mimic or exceed human evaluations [12].

In their 2021 research, Tulu, Ozkaya, and Orhan introduced an automated short answer grading (ASAG) system leveraging the Manhattan Long Short-Term Memory (MaLSTM) architecture and SemSpace sense vectors. They utilized the SemSpace methodology to generate sense vectors derived from WordNet, which enabled precise word sense disambiguation for the evaluation of student submissions. The model underwent training on two distinct datasets: the widely recognized Mohler ASAG dataset, as well as a bespoke CU-NLP dataset, which encompassed responses from students enrolled in a Natural Language Processing course. The assessment demonstrated elevated Pearson correlation coefficients, reflecting substantial concordance between the model's ratings and human assessments, with correlations frequently reaching approximately 0.95 for the Mohler dataset and 0.989 for the CU-NLP dataset. Notwithstanding these encouraging findings, the study acknowledged certain limitations, including difficulties associated with the intricate nature of word relationships within WordNet and the training methodology, which could be refined by concentrating on more concise context sets. In summary, the investigation emphasizes the potential of harnessing sophisticated machine-learning methodologies and semantic assets to improve the precision and effectiveness of automatic grading systems [5].

In their 2023 research, Wang et al. present an innovative evaluation framework termed AuPEL, aimed at assessing personalized text generation through the application of large language models (LLMs). This framework analyzes text across three fundamental dimensions: quality, relevance, and personalization. The research employed an email dataset encompassing over 279 accounts, concentrating on documents exceeding 300 characters that were produced by authors with a minimum of three documents. Participants were included if they submitted at least 10 examples, with a maximum of 100 examples per participant to mitigate bias from more prolific users. The results indicated that AuPEL surpassed conventional metrics such as BLEU and ROUGE in the assessment of personalized text generation, demonstrating enhanced capabilities to differentiate and rank models based on their level of personalization. The concordance between AuPEL and human evaluations was elevated across all dimensions, with human evaluators achieving agreement scores of 0.65 for quality, 0.61 for relevance, and 0.63 for personalization. Notwithstanding its merits, the study acknowledged that human evaluations tend to be more resource-intensive and less reliable. The assessment of personalization proved to be particularly complex and subjective, highlighting the necessity for additional test cases to derive dependable results. Although AuPEL proficiently addresses essential aspects of text quality, it does not encompass every potential dimension. The authors underline the swift progression in the assessment of text generation and advocate for future investigations to formulate more comprehensive test cases and examine the boundaries of LLM capabilities in the evaluation of personalized content. In summary, AuPEL represents the inaugural automated approach specifically designed for this objective, offering precise, consistent, and nuanced evaluations in comparison to traditional metrics and human raters [2].

In their 2023 investigation, Meccawy, Maram, and associates concentrated on the creation of an automated essay evaluation system tailored for Arabic short answer inquiries by employing text mining methodologies. They developed the Engineering Short Answer Feedback (EngSAF) dataset, which encompasses 5,800 student responses alongside reference answers and questions spanning various engineering disciplines. This dataset aspires to offer a broad spectrum of topics and response modalities.

The researchers utilized a Label-Aware Synthetic Feedback Generation (LASFG) framework, harnessing large language models (LLMs) to produce feedback that elucidates the rationale behind the awarded grades. This methodology augments the pedagogical effectiveness of the assessment process by delivering constructive critiques regarding student submissions. The system's efficacy was corroborated through its implementation in a genuine end-semester examination at the Indian Institute of Technology Bombay (IITB), illustrating its practical applicability within an academic context. The findings underscored the system's proficiency in rendering precise evaluations and substantial feedback, thereby enhancing the overall educational experience. By establishing various LLM-based zero-shot and fine-tuned baselines, the research establishes a groundwork for subsequent inquiries into automated assessment and feedback frameworks, promoting further investigation of these sophisticated technologies in educational contexts [4].

Mizumoto and Eguchi (2023) investigated the efficacy of GPT in the realm of automated essay scoring (AES) utilizing the TOEFL11 corpus, which comprises 12,100 essays authored by non-native English speakers. The essays were classified into Low, Medium, and High proficiency levels, with statistical analyses substantiating significant disparities among them. The reliability of intra-rater assessments was evaluated via Quadratic Weighted Kappa, indicating a consistent concordance between the scores produced by GPT. In an effort to enhance precision, the research compared seven distinct models, integrating a range of linguistic characteristics, including measures of lexical and syntactic complexity. The findings indicated that the amalgamation of GPT scores with linguistic features surpassed the performance of GPT scores in isolation, achieving an adjacent agreement rate of 89.15% with levels evaluated by human graders. The study concluded that GPT exhibits substantial promise for AES, particularly when synergized with linguistic features; however, additional investigation is essential to fully delineate its capabilities [7].

Cho et al. (2023) present a rubric-oriented methodology for Automated Essay Scoring (AES), where models are trained according to scoring rubrics instead of prompts. This investigation utilizes data augmentation strategies such as Prompt Swap, Grade Match, and Response Distortion to enhance the accuracy of the models. The AES model is developed using the ASAP dataset, employing the Mean Squared Error loss function, and is assessed using the Quadratic Weighted Kappa (QWK) metric. The findings indicate that the rubric-oriented models, augmented through data techniques, surpass the performance of baseline models. In particular, Grade Match markedly enhances the correspondence with rubric components, while Prompt Swap and Response Distortion facilitate the identification of irrelevant and distorted responses. The research underscores the shortcomings of prompt-oriented models and posits that rubric-centric training can improve the accuracy and resilience of AES, potentially substituting human evaluators in certain scenarios. The authors also recognize deficiencies in the present AES literature, particularly the necessity for superior performance metrics and enhanced abilities for identifying irrelevant responses. They advocate for forthcoming research to concentrate on refining these aspects to establish more dependable AES systems [9].

Waer (2023) examines the impact of Automated Writing Evaluation (AWE) on writing apprehension and grammatical proficiency among English majors at an Egyptian university. The study divided participants into two groups: an experimental group using the Cambridge "Write & Improve" platform for writing assessment and a control group receiving feedback from an instructor. Writing apprehension was measured using the English Writing Apprehension Scale (EWAS), while grammatical proficiency was assessed through a Grammar Knowledge Test (GKT) adapted from the TOEFL framework. The findings showed that the experimental group experienced a significant reduction in writing apprehension compared to the control group. Additionally, there was a modest improvement in grammatical proficiency among less apprehensive writers in the experimental group. The results highlighted a substantial positive effect of AWE on alleviating writing anxiety and enhancing performance, with a negative correlation found between writing apprehension and grammatical proficiency—indicating that as apprehension decreases, grammatical knowledge increases. The study concludes that AWE tools like Cambridge's "Write & Improve" can serve as effective resources in education, particularly for students struggling with writing apprehension. These tools can complement traditional feedback methods, helping students improve their grammatical skills while reducing writing anxiety. However, the author notes the study's limitations, including its focus on a specific group of English majors in Egypt, which may affect the generalizability of the findings. Future research is suggested to include diverse participant demographics and explore the long-term effects of AWE. Additionally, integrating AWE with other educational resources and qualitative insights could provide a more comprehensive understanding of its impact [10].

In their 2024 investigation, Sun and Wang introduce a methodology for the automated multi-dimensional evaluation of essays. The scholars leverage fine-tuning of pre-existing language models such as RoBERTa and DistilBERT to augment their efficacy in evaluating various facets of essays, including lexicon, syntax, and coherence. The research employs multiple regression methodologies to empower the models to concurrently predict scores across diverse dimensions. Furthermore, contrastive learning is utilized to refine the models' comprehension by integrating additional contextual information, such as prompts and subjects, which significantly enhances scoring precision. A composite loss function that amalgamates cross-entropy loss with mean squared error loss is also implemented to proficiently facilitate both classification and regression tasks. The researchers executed their experiments utilizing two extensive datasets, ELLIPSE and IELTS, which comprise essays evaluated on multiple metrics. The findings reveal that the proposed evaluation framework surpasses existing techniques, illustrating its efficacy in delivering nuanced feedback across various dimensions of essay quality. The study underscores the necessity of employing intricate datasets for the development of more resilient automated essay assessment systems. The authors propose that subsequent investigations could concentrate on the incorporation of more sophisticated language models and the exploration of additional evaluation dimensions, such as creativity or argumentative strength, to further enhance the overall quality of feedback [11].

Oulimani, Younes Alaoui; El Achaak, Lotfi; Bouhorma, Mohammed (2024) undertook an investigation into deep learning-driven Arabic short answer grading (ASAG) within the context of serious games. They utilized a range of models, such as LSTM, Transformer, and BERT, to assess the responses provided by sixth-grade learners. The dataset comprised 1,276 answers collected manually and evaluated by an educator utilizing a scoring framework that spanned from 0 to 2. To evaluate the efficacy of each model, four assessment metrics were applied: accuracy, precision, recall, and Cohen's Kappa. The findings revealed difficulties with grading accuracy, particularly stemming from the insufficient availability of Arabic datasets, which impedes the advancement of robust ASAG systems in comparison to their English counterparts. The study underscores the intricacy involved in automating the assessment of free-text responses and highlights the necessity for enhancements in ASAG methodologies. It proposes the augmentation of the system by integrating correction functionalities, thereby allowing it to not only assign grades but also furnish students with constructive feedback on improving their responses [1].

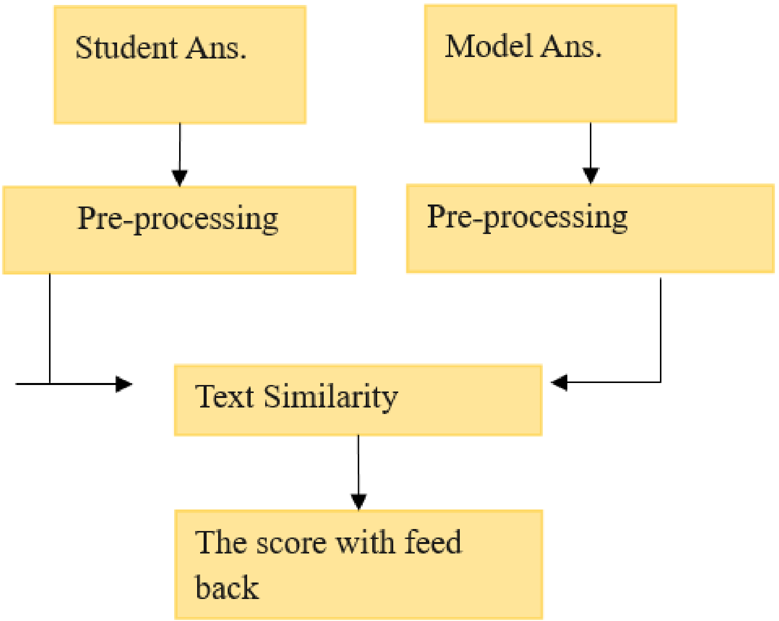

Figure 2.
Shows general structure of text similarity 0f AEGS.

2.2. Research Based on Text Similarity Algorithms

In their 2013 investigation, Omran, Ali Muftah Ben, and Ab Aziz, Mohd Juzaiddin concentrated on the establishment of an automatic essay grading (AEG) system specifically aimed at appraising short responses in the English vernacular. The principal objective of this inquiry was to examine the feasibility of constructing an AEG system that utilizes similarity matching techniques to accurately evaluate student submissions. To accomplish this, the researchers employed a subset of a dataset introduced by Mohler and Mihalcea, which encompasses 360 short responses originating from three assignments, each comprising 40 queries, along with three student replies for every question. The dataset was deliberately partitioned into two components: one segment was allocated for training the system, while the other was designated for evaluating its efficacy. The proposed system integrated several groundbreaking methodologies, initiating with an Alternative Sentence Generator Method. This approach was linked to a synonym repository, enabling the production of alternative model responses that would promote a more adaptable matching process. The assessment procedure entailed multiple stages, with considerable emphasis placed on the matching phase, wherein the system evaluated the similarity between student replies and the generated model responses. To assess the efficacy of the responses, the researchers utilized a range of techniques, including the Longest Common Subsequence (LCS), Common Words (COW), and Semantic Distance (SD). The LCS method was employed to discern common sequences between the two responses, while COW was implemented to identify corresponding key terms. Concurrently, SD offered a metric for the interrelation among individual words, thereby ensuring a comprehensive evaluation of the text. By amalgamating these techniques, the researchers aspired to develop a robust grading framework capable of proficiently aligning student responses with the model answers in English. The findings were encouraging, as the proposed system attained an 82% correlation with human grading, demonstrating its enhanced performance relative to other prevailing systems. In summary, this study tackled the complexities associated with essay evaluation for short responses, underscoring the system's capacity to adeptly manage synonyms and contribute to a more precise and efficient grading mechanism in English language assessment [3].

In their 2018 study, Ramalingam and associates investigated the realm of automated essay evaluation utilizing a machine learning algorithm, specifically implementing the e-Rater methodology in conjunction with Bayes' Theorem. Their methodology entailed the preprocessing of essays through

DOI: 10.55214/25768484.v8i6.3956

various techniques, such as stemming and the elimination of stop words, aimed at augmenting textual analysis. The essays underwent comparative analysis based on several metrics, which encompassed polarity, vocabulary utilization, and overall content richness. The dataset employed in this inquiry was procured from Kaggle and comprised essays that had been previously assessed by human evaluators, categorized into eight distinct groups according to essay type. The findings revealed an accuracy rate surpassing 80%, underscoring the efficacy of their proposed model in essay assessment. The researchers sought to model linguistic attributes such as fluency and grammatical accuracy while applying a polynomial regression model to the characterized feature space. The investigation indicated that the model could attain an average absolute error that was lower than the standard deviation of human evaluative scores across all categories. In terms of future directions, the authors pinpointed prospective avenues for further exploration, including the refinement of enhanced semantic and syntactic feature models as well as the investigation of alternative methodologies, such as neural networks, to bolster the precision of automated grading systems [6].

In the year 2018, Abdulaziz Shehab and his associates introduced an automated Arabic essay evaluation system that assesses student submissions via text similarity algorithms. This system functions on the premise of quantifying the likeness between a student's response and a reference answer through a Bag of Words (BoW) model, which scrutinizes each word in both responses to produce a conclusive score. The researchers employed an amalgamation of string-based and corpus-based algorithms to compute similarity metrics for the responses. To assess their methodology, the team implemented it on a dataset consisting of 210 brief responses from a sociology course undertaken by secondary level 3 students. Each response was evaluated by multiple annotators, with scores ranging from 0 (entirely incorrect) to 5 (entirely correct). The mean score from the annotators acted as the standard for appraising the system's grading efficacy. The investigation underscored that the character-based N-gram algorithm surpassed alternative techniques, attaining a considerable association value of 0.803 with the Damerau-Levenshtein (DL) algorithm. The application of stop-stemming, which emphasizes the root forms of words while disregarding stop words, contributed to the elevated precision of the findings. The N-gram approach exhibited numerous benefits, including its dependability in managing noisy data, such as spelling and grammatical inaccuracies, as well as its capacity to produce a substantial quantity of N-grams for comparative analysis. Moving forward, the researchers aspire to amalgamate string-based and corpus-based algorithms to further refine grading precision. They intend to investigate synonymous methodologies and content-centric strategies to diminish errors in automated grading. Furthermore, they plan to evaluate their system on diverse datasets across various disciplines to gauge the algorithms' reliability in more extensive applications [21].

In their 2019 publication, Al Awaida, Al-Shargabi, and Al-Rousan introduced an automated essay evaluation framework specifically designed for Arabic language compositions, utilizing F-score, Arabic WordNet (AWN), and text similarity metrics to enhance the precision of grading outcomes. The framework encompasses several stages: preprocessing, feature extraction via AWN, and the application of cosine similarity to assess the congruence between student submissions and model responses. The dataset employed in this investigation was developed within a MySQL environment and consists of 120 prompts, each paired with three model responses and 30 student submissions. This dataset was meticulously constructed following the guidelines established by the Hewlett Foundation for automated student evaluations. The findings demonstrated that the incorporation of Arabic WordNet markedly enhanced grading efficacy, as indicated by decreases in mean absolute error and elevated Pearson correlation coefficients when compared to systems that did not implement AWN. The research underscored the utility of AWN and F-score as potent instruments for feature extraction from both student and model responses. Future research directions are suggested to examine the utilization of machine learning and neural network architectures to further refine the accuracy of Arabic essay assessment, as well as to investigate the ramifications of word-embedding methodologies [16].

In the year 2019, scholars Aqil M. Azmi, Maram F. Al-Jouie, and Muhammad Hussain established the AAEE (Automated Arabic Essays Evaluator) system, aimed at assessing students' essays written in Arabic. The architecture of the system operates through two primary stages: training and evaluation. To construct the system, the researchers amassed a dataset consisting of 350 handwritten pre-graded essays sourced from four distinct educational institutions, encompassing a variety of topics and spanning two academic levels. Each essay underwent evaluation by two or, in some instances, three human raters to guarantee reliability and reduce bias. During the training stage, 300 of these essays were utilized, with human instructors assigning scores on a scale of 10. Should the scores from the initial two raters exhibit significant discrepancies, a third rater would assess the essay to ensure an equitable evaluation. In the evaluation phase, the remaining 50 essays were appraised by the AAEE system. The researchers gauged the system's efficacy by determining the correlation between the automated scores and those awarded by human raters. The outcomes were encouraging; the AAEE system attained a Pearson correlation of 0.756 with human assessments. This correlation is analogous to the inter-human correlation of 0.85 observed in English essay evaluations and slightly exceeds the previously reported correlation of 0.709 for Arabic essays. The system exhibited a high accuracy rate, with approximately 90% of the test essays receiving accurate scores. These results signify the competency of the AAEE system in the automated grading of Arabic essays. Moving forward, the researchers aspire to enhance the evaluation outcomes by integrating additional features into the system. They plan to incorporate Word2Vec, a widely recognized method for evaluating semantic similarity, and to broaden the dataset by gathering more essays and involving multiple educators in the grading process. This research marks a notable advancement in automated essay evaluation for the Arabic language and paves the way for future inquiries and developments in this domain [27].

In the year 2019, Saeda Esmaile Odeh Al-Awaida, under the guidance of Dr. Bassam Alshargabi, submitted a doctoral thesis concerning an automated grading system for Arabic essays. The proposed framework comprises two main components. Initially, it employs Arabic WordNet to ascertain related lexicons, selecting features via a Support Vector Machine (SVM) subsequent to a preprocessing phase. The subsequent component assesses electronic student submissions by juxtaposing them against predefined answer templates to ascertain the level of similarity through the cosine similarity algorithm. The dataset utilized for this investigation was generated in MySQL and subsequently exported as a CSV file. It consisted of 40 inquiries and three categories of responses sourced from computer science and social studies literature at Allu'lu'a Modern School, adhering to the structure outlined by the Hewlett Foundation's Automated Student Assessment Prize. The dataset featured unique identifiers for each response, scores assigned by evaluators, and response texts linked to the respective question IDs. Experimental findings revealed that the proposed system significantly enhances the performance of Arabic essay grading in comparison to human evaluation. Specifically, the mean absolute error (MAE) between the human-assigned scores and the automated scores derived from cosine similarity was markedly lower when employing Arabic WordNet, achieving an MAE of 0.117, in contrast to elevated values observed in the absence of WordNet. This advancement signifies an improvement in accuracy of 2.648%. The emphasis of the study was to devise an automated essay grading model tailored specifically for Arabic compositions, highlighting the significance of incorporating Arabic WordNet to bolster grading precision. The results suggested that the automated grading mechanism could achieve closer alignment with human assessments when utilizing WordNet. Recommended future avenues include the application of machine learning and neural network paradigms for enhanced accuracy, the use of larger datasets, and the integration of word-embedding methodologies [25].

In the year 2020, Wael Hassan Gomaa and Aly Fahmy presented the "Ans2vec" framework for the evaluation of succinct responses, notable for its straightforwardness and reliance on an unsupervised pre-trained sentence embedding model. This framework is devoid of the necessity for natural language processing procedures such as the elimination of stop words or tokenization, nor does it depend on human-annotated datasets or lexical resources like WordNet. "Ans2vec" underwent validation on three established benchmark datasets: the Texas dataset, which encompasses 80 inquiries from a Data Structures course alongside 2,273 student replies; the Cairo University dataset, which comprises 61 questions paired with 610 responses; and the SCIENTSBANK dataset, which classifies replies into five distinct categories, including "correct" and "partially correct." The system attained the highest Pearson correlation coefficient of 0.63 when compared to other analogous systems by transforming both student and model responses into semantic vectors and assessing the similarity between them. Prospective research will aim to explore alternative Sent2vec methodologies, merge their approach with conventional machine learning strategies, and extend the application of the model to datasets in additional languages, including Arabic [23].

In their 2020 scholarly article, Ouahrani and Bennouar introduced AR-ASAG, a rigorously developed Arabic dataset intended for the assessment of automated short answer grading systems. This dataset comprises 2,133 instances of both model and student responses, which are accessible in a variety of file formats, including txt, xml, and Moodle xml. The authors explored an unsupervised, corpus-based approach tailored for automatic grading specific to the Arabic language, employing the COALS (Correlated Occurrence Analogue to Lexical Semantic) algorithm to create a semantic space that reflects word distribution patterns. The study implemented a summation vector model that integrates term weighting and frequently occurring terms to evaluate the similarity between model responses and student submissions. The researchers conducted experiments to assess the impact of factors such as domain specificity, the dimensionality of the semantic space, and stemming methods on the performance of the grading model. The results indicated promising outcomes for the Arabic language, thereby laying a foundation for future research in this area. Furthermore, the authors suggest that forthcoming studies will aim to enhance grading precision by formulating a supervised model that incorporates additional educator insights and parameters that mirror real-world grading challenges. The language-agnostic nature of their approach offers avenues for exploration and implementation across various languages [14].

2.3. Path Research Using Hybrid Models (Combining Algorithms)

In 2021, Abdelrahman ElNaka and collaborators developed AraScore, an innovative framework for automating the scoring of Arabic short responses. The framework comprises four key stages: inputting Arabic responses, data preprocessing, feature extraction, and predictive modeling. The study evaluated the performance of Support Vector Machines (SVM) and Random Forest models alongside a multi-layer perceptron classifier. The researchers utilized two main corpora for their analysis. The first corpus included over 1,000 responses to two open-ended questions from the Automated Student Assessment Prize competition, originally in English and later translated into Arabic. These questions, related to science and biology, were scored using a rubric from 0 to 3. The second corpus, the AR-ASAG dataset, contained 48 questions, each with an average of 50 responses. Additionally, the team created a new dataset called AraScore-Dataset, which included manually assessed Arabic Social Studies examinations from an Egyptian educational institution, focusing on five History and Geography questions. To improve the dataset's volume and quality, they employed data augmentation techniques, particularly back-translation, which involved translating text into another language and back to Arabic to generate diverse responses. The model achieved an impressive 84% agreement between its predicted scores and those assigned by human raters. A sentence embedding approach proved particularly effective, with a feed-forward deep neural network architecture outperforming SVM and Random Forest models, achieving a classification accuracy of 95.701% and a mean squared error of 0.073 in regression analyses. The AraScore-Dataset also yielded favorable results, surpassing the performance of benchmark datasets in the field. Looking ahead, the researchers plan to explore additional data augmentation strategies, such as synonym substitution using word embeddings from AraVec or AraBERT, to further enhance the dataset's quality and size [22].

In 2021, Hikmat A. Abdeljaber introduced a framework for the automated assessment of Arabic short answers utilizing the Longest Common Subsequence (LCS) technique alongside Arabic WordNet. This framework consists of three primary elements: preprocessing, text classification, and evaluation. The preprocessing stage entails processes such as tokenization, elimination of special characters and stop words, and stemming. The text classification segment employs a modification of the LCS algorithm to assess student responses to Arabic essay prompts. Experiments were conducted using a dataset comprising 330 student submissions, resulting in a Root Mean Square Error (RMSE) of 0.81 and a Pearson correlation coefficient of 0.94. These findings suggest that the proposed methodology adeptly approximates human scoring accuracy. Nonetheless, the research possesses significant constraints. The dataset utilized is comparatively limited due to the absence of a gold standard for Arabic, and it fails to address linguistic inaccuracies such as spelling, punctuation, and grammatical errors. Furthermore, the analysis does not encompass sentence structure nor does it provide evaluative feedback. Abdeljaber's investigation underscores the promise of integrating the LCS methodology with Arabic WordNet to refine text similarity assessments. Nevertheless, the development of a more extensive Arabic WordNet could greatly enhance future research endeavors in this arena by supplying improved synonyms and minimizing lexical discrepancies between student responses and model answers [24].

In 2021, Bassam Al-Shargabi, Rawan Alzyadat, and Fadi Hamad developed the Arabic Essay Grading Dataset (AEGD) to facilitate the automation of Arabic essay grading (AEG). The dataset was carefully designed and validated, comprising questions and exemplary responses sourced from a variety of subjects delivered in Jordanian educational institutions, encompassing Islamic studies, history, geography, biology, computer science, geology, chemistry, and physics, covering grades 9 to 11. Dataset Composition and Collection: The AEGD comprises a total of 1,003 questions and 3,009 model answers, encompassing a diverse array of unique Arabic vocabulary (10,364 distinct words). The data collection process involved gathering information from teachers' manuals and soliciting feedback from educators to incorporate supplementary model answers, thereby ensuring the dataset's thoroughness and pertinence. Preprocessing and Evaluation: The dataset underwent preprocessing to identify grammatical and spelling inaccuracies, thereby enhancing its suitability for machine learning utilization. The assessment of the AEGD utilized three established machine learning algorithms: Naïve Bayes (NB), Decision Tree (J48), and a meta-classifier. The findings revealed a commendable accuracy rate of 81.29% for correctly evaluated essays, with an error rate of 18.71% in terms of incorrectly graded submissions. Future Work: The authors intend to further validate the dataset utilizing deep learning methodologies and to augment the dataset by incorporating additional subjects to enrich the Arabic vocabulary included. This endeavor represents a crucial advancement towards enhancing automated grading systems for Arabic essays by providing a robust and varied dataset for machine learning applications [26].

In the year 2021, Hossam Magdy Balaha and Mahmoud Saafan introduced an Automated Exam Correction Framework (AECF) aimed at the evaluation of multiple-choice questions (MCQs), essays, and equation matching. The framework is composed of five principal phases, commencing with the text preprocessing phase, wherein the text is refined and prepared for subsequent utilization. This is succeeded by the text tokenization phase, during which the text is segmented into smaller components referred to as tokens. The following phase encompasses the extraction of features from the text, whereby pertinent attributes are discerned for analytical purposes. The fourth phase emphasizes the calculation of semantic similarity, evaluating the extent to which students' responses align with the correct answers. Ultimately, the framework integrates various software packages to proficiently execute these processes. The researchers conducted eight experimental trials utilizing two datasets: the Quora Questions Pairs dataset and the UNT Computer Science Short Answer dataset. They attained a peak accuracy of 77.95% without the fine-tuning of the pre-trained models, particularly employing the Universal Sentence Encoder (USE). Furthermore, the proposed similarity checker algorithm (HMB-MMS-EMA) reported an accuracy of 100% when evaluated with the SymPy Python package, while an alternative method (HMB-EMD-v1) demonstrated an accuracy of 71.33%. This investigation delivers a thorough examination of the extant literature on text semantic similarity and automated correction systems. Several embedding models, encompassing Word2Vec, FastText, GloVe, and USE, were utilized in the experiments. The researchers proposed that future endeavors could expand the AECF to encompass additional question formats and enhance the HMB-MMS-EMA algorithm to address more intricate arithmetic operations. They also advocated for the exploration of alternative machine learning and deep learning methodologies to augment the framework's efficacy [17].

In 2022, Mustafa Abdul Salam and his collaborators presented a hybrid automatic evaluation framework for Arabic short answer responses. This pioneering framework enhances the Long Short-Term Memory (LSTM) deep learning architecture by employing a contemporary optimization technique termed the Grey Wolf Optimizer (GWO). The GWO improves the efficacy of the LSTM by autonomously determining the most suitable dropout and recurrent dropout rates for its hyperparameters, thus eliminating the dependence on manual tuning. This optimization not only augments the generalization capabilities of the LSTM model but also alleviates the problem of overfitting, ultimately benefiting the assessment procedure and conserving educators' time. The researchers executed eight experimental trials, each involving diverse datasets. The findings demonstrated that the hybrid LSTM-GWO model surpassed multiple other models, including conventional LSTM, Support Vector Machine (SVM), as well as variations such as SVM-GWO, N-gram, Word2Vec, Latent Semantic Analysis (LSA), Arabic WordNet, and MaLSTM (Manhattan LSTM). The efficacy of the models was evaluated using metrics such as Root Mean Square Error (RMSE), R-Squared, and the Pearson correlation coefficient. Despite the encouraging results, the investigation recognized limitations associated with an exclusive reliance on traditional deep learning and machine learning methodologies. The authors indicated their intention to delve into more sophisticated language models, like BERT and its derivatives, in subsequent research endeavors. They also aspired to broaden the utility of their grading system beyond Arabic to encompass additional languages and to amalgamate various methodologies to further enhance outcomes. The results imply that the LSTM-GWO model is not only proficient in assessing Arabic short answers but also possesses potential for wider applications within automated evaluation systems [20].

In their 2022 publication, Mijbel and Sadiq introduced an innovative methodology for evaluating short-form answers grounded in the principles of semantic networks. This technique entails the representation of both model answers and student submissions as semantic networks, which depict knowledge by interlinking concepts through semantic associations. The objective of this approach is to assess answers by determining similarities that extend beyond mere word frequency, instead focusing on the interconnections among nodes and the relationships present within the semantic networks. The methodology commences with preprocessing, which includes Parts of Speech (PoS) tagging, and progresses to the construction of semantic networks that consist of nodes (representing words) and edges (denoting relationships). The assessment of student responses is based on the Mohler dataset, a resource frequently utilized in research pertaining to automatic short answer grading (ASAG). This dataset comprises 12 assessments in the domain of computer science, covering 87 questions, with corresponding reference answers for each item. The evaluations of student responses were conducted independently by two human raters on a scale from 1 to 5, assessing their relevance and similarity to the provided reference answers. The findings demonstrated that optimal performance in answer evaluation was obtained through specific weightings assigned to nodes, relationships, and PoS similarities ($\alpha = 0.1$, $\beta = 0.6$, $\gamma = 0.3$). The authors underscored the significance of recognizing that human evaluations of answers frequently integrate experiences and contexts that transcend simple semantic or lexical comparisons. They propose that future research could broaden this methodology to encompass essay-type questions and enhance techniques for attaining more precise assessments of similarity [15].

In the year 2023, Rasha M. Badry and her research team formulated an Automatic Arabic Grading System tailored for short answer questions by employing Natural Language Processing (NLP) methodologies to evaluate student submissions. Their methodology is predicated on Latent Semantic Analysis (LSA) to assess the semantic congruence between student responses and model answers. The system encompasses several pivotal stages: data preprocessing, which entails tokenization, elimination of stop words, and lemmatization; LSA execution; representation of term weights; and the computation of semantic scores. They utilized the AR-ASAG dataset, comprising 2,133 pairs of model and student responses, for their empirical investigation. The researchers executed two experiments contrasting local and hybrid weighting frameworks. The hybrid methodology produced superior outcomes, attaining an F1-score of 82.82% and a Root Mean Square Error (RMSE) of 0.798. Notwithstanding the emphasis on Arabic—a language with scarce resources for automated evaluation—the findings underscore the model's efficacy. The manuscript concludes with aspirations for future developments, including evaluating the system across additional languages and incorporating Arabic WordNet to enhance scoring precision. Through this inquiry, they seek to propel automated grading within Arabic language processing [19].

In 2023, Nourmeen Lotfy and her research team introduced an Enhanced Automatic Arabic Essay Scoring System that leverages machine learning techniques to enhance grading precision. This system incorporates a preliminary data cleansing and natural language processing stage, succeeded by a grading mechanism that employs both string-based and corpus-based similarity algorithms, notably excelling with co-occurrence similarity metrics. The adaptive fusion engine significantly boosts grading efficacy through feature selection approaches such as Recursive Feature Elimination (RFE) and Boruta. Additionally, the system integrates a variety of machine learning methods, including Random Forest, Decision Trees, and K-Nearest Neighbor, to maximize operational efficiency. The performance assessment utilized indicators like Pearson's Correlation Coefficient and Mean Absolute Error, attaining remarkable figures of 91.30% and 94.20%, respectively, using a dataset comprising 270 student submissions from a sociology class. The manuscript emphasizes the shortcomings of existing Arabic essay assessment systems, particularly concerning accuracy and processing duration. The research accentuates the capacity of machine learning to furnish accurate and effective grading solutions. Future research will aim to broaden the dataset, explore various academic disciplines, and employ pre-trained transformer models to improve accuracy and generalizability [18] further.

In their 2023 research, Maram Meccawy and associates examined the efficacy of automatic essay scoring for Arabic short response items, with a particular emphasis on improving scoring precision attributable to the language's inherent complexities. The investigation encompassed multiple phases, beginning with the collection of student responses (SAs) and model responses (MAs), and subsequently engaged in preprocessing activities such as data cleansing, normalization, and stemming. The researchers evaluated three methodologies—WordNet, Word2Vec, and BERT—employing cosine similarity to assess the correspondence between SAs and MAs. They utilized two distinct datasets, one derived from a cybercrime examination containing 2,133 SAs, and another from a Jordanian History assessment featuring 11 queries. Among the tested methodologies, BERT exhibited superior performance, attaining a Pearson correlation coefficient of 0.84 and a root mean square error (RMSE) of 1.003. Notwithstanding the encouraging findings, the authors acknowledged the necessity for additional inquiry to further refine the accuracy of Arabic automatic essay scoring systems for effective online implementation. The study highlighted the significance of contextual word embeddings in enhancing automated evaluations in Arabic. In summary, while advancements were achieved, the authors advocated for ongoing investigation within this domain [29].

In the 2023 investigation conducted by Su-Youn Yoon, an automated assessment framework for brief responses was established utilizing a large language model (LLM) alongside one-shot prompting and a neural textual similarity evaluation mechanism. The methodology commenced with the identification of justification keys for each sub-question through one-shot prompting. Following this, similarity metrics between these keys and benchmark answers were computed employing the Sentence-BERT architecture, resulting in analytic scores that were subsequently amalgamated to produce a comprehensive score. The inquiry employed a segment of the Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) corpus, concentrating specifically on four inquiries from the realms of science and biology. The dataset was partitioned into training (80%), development (10%), and testing (10%) segments. From the training subset, 64 responses were manually annotated to enhance reference answer quality. The model attained an impressive accuracy rate of 90.3%, successfully identifying justification keys in the majority of instances. Nevertheless, certain constraints were acknowledged, including the absence of evaluation concerning justification keys and analytic scores, which will be tackled in forthcoming studies. The research emphasized that although the model demonstrated considerable performance enhancements in comparison to existing benchmarks, a disparity remains in attaining near-human scoring accuracy. Subsequent endeavors are aimed at augmenting model efficacy, assessing justification keys with greater precision, and broadening the array of correct responses within the scoring criteria. By capitalizing on LLM-based prompting, the study posits that annotations can be economically enriched, potentially enhancing the interpretability of scoring models and their overall efficacy within automated grading systems for brief responses [30].

In their 2024 publication, Konade et al. introduced an automated evaluation framework leveraging the Sentence-BERT (SBERT) architecture to augment semantic analysis within the grading paradigm. By producing vector embeddings for sentences, SBERT emerged as an essential element in facilitating a proficient and resilient assessment of descriptive responses. The system exhibited a remarkable accuracy rate of 95.1% in the evaluation of theoretical answers, suggesting its capability to surpass conventional human grading practices. Although the paper does not detail the specific dataset employed during the implementation, it underscores the efficacy of SBERT in the realm of automated assessment. The authors recognize the shortcomings of various pre-existing evaluation methodologies, accentuating the necessity for more efficient and precise automated assessment systems, particularly in light of the increasing student population in educational contexts. Moreover, notwithstanding the high level of accuracy attained, the investigation highlights the difficulties associated with sustaining uniform performance across diverse subjects, implying constraints in the system's generalizability. The paper articulates the imperative for automation in grading, ultimately endorsing the widespread adoption of such systems to foster consistent and objective assessment practices within educational settings [13].

In a research initiative spearheaded by Siddharth Jain and associates from the Department of Computer Science Engineering at Acropolis Institute of Technology and Research in Indore, India, an innovative automated grading system has been created to improve the educational experience. This system mandates that users, particularly students, register and utilize a class code to enroll in their designated classes. After obtaining the code from the instructor through email communication, students gain access to a platform that enables them to track their progress, class notes, and assignments. For submission of assignments, students are required to compile their work into a singular PDF document, which they subsequently upload to the system. The system then assesses this submitted document against the instructor's reference copy and assigns a grade based on established grading benchmarks. Instructors maintain the capacity to introduce assignments as deemed necessary. Although the paper does not specify the particular dataset utilized, it underscores the system's interface and operational capabilities. The deployment of this automated grading system employs machine learning methodologies and data integration, providing considerable benefits to the educational domain. It optimizes the grading procedure, conserves time, diminishes errors, and guarantees uniformity in grading standards. Furthermore, the system offers tailored feedback to students, thereby promoting their academic development. The proposed solution is crafted to be scalable and adaptable, catering to diverse educational institutions and curricula. Ongoing enhancements informed by user feedback are also emphasized to ensure that the system is aligned with instructional goals. In summary, the automated grading recommendation system aspires to enrich the learning experience, support educators in their assessments, and foster equitable and precise evaluations [28].

Table 1:
Summary table of research on automatic assessment of short answers.

Authors	Year	Theoretical framework	Dataset	Major findings	Limitations/Gaps
SOULIMANI Younes Alaoui et al.	2024	Deep learning (LSTM, Transformer, BERT)	Moroccan students' answers	High accuracy during training, performance drop in testing	Limited Arabic datasets
WANG Yaqing et al.	2023	AuPEL framework, LLM models	Email dataset (279 accounts)	Outperformed BLEU/ROUGE in scoring personalization	Does not cover all text quality aspects
OMRAN Ali et al.	2013	Sentence matching	Mohler dataset (360	82% correlation with human	English-only focus

		methods	answers)	grading	
MECCAWY Maram et al.	2023	Text mining, LLMs	EngSAF dataset (5.8k answers)	Effective feedback generation for engineering	Limited to engineering domains
TULU Cagatay et al.	2021	MaLSTM, SemSpace vectors	Mohler ASAG, CU-NLP datasets	Pearson correlation up to 0.95	Training time affects performance
MIZUMOTO Atsushi et al.	2023	GPT for AES	TOEFL11 dataset (12100 essays)	Close match with TOEFL11 benchmarks	Difficulty differentiating scoring levels
SUN Kun, WANG Rong	2024	Fine-tuning, multiple regression	ELLIPSE, IELTS datasets	Multi-dimensional scoring for essays	Lack of evaluation for justification keys
HUSSEIN Mohamed A. et al.	2020	Deep learning (LSTM, CNN)	ASAP dataset	Trait-based evaluation improved accuracy	Needs further feedback and analysis
BALAHA Hossam et al.	2021	Automated exam correction framework	UNT, Quora datasets	77.95% accuracy, 100% on equations	Limited to certain question types
LOTFY Nourmeen et al.	2023	Arabic AES with ML	Sociology course dataset	91.30% Pearson correlation achieved	Time efficiency is a key challenge
ABDELJABER Hikmat A.	2021	LCS, Arabic WordNet	Dataset of 330 student answers	RMSE 0.81, Pearson correlation 0.94	Small dataset, lacks semantic analysis
AL AWAIDA Saeda et al.	2019	F-score, Arabic WordNet	Jordanian school dataset	Higher accuracy with Arabic WordNet	No feedback or correction mechanism provided

RAMALINGAM V. V. et al.	2018	Bayesian methods, NLP	ASAP dataset	Over 80% accuracy	Limitations in performance metrics
GOMAA Wael Hassan et al.	2020	Ans2vec model	Texas, Cairo University, SCIENTSBANK	Achieved best Pearson correlation (0.63)	Focus on semantic vector generation
OUAHRANI Leila et al.	2020	Unsupervised corpus-based approach	AR-ASAG Arabic dataset	Promising results for Arabic grading	Dataset size and domain specificity limitations
KONADE Shashank et al.	2024	SBERT architecture	N/A	95.1% accuracy for descriptive answers	Performance varies across subjects
SHEHAB Abdulaziz et al.	2018	Text similarity (DL algorithm)	Sociology course dataset	N-gram algorithm achieved best results	Combined string and corpus algorithms in future work
MIJBEL Sumaya H. et al.	2022	Semantic networks for ASAG	Mohler dataset	Best results with (α=0.1, β=0.6, γ=0.3) weights	Focused on Arabic, no feedback or correction
AL AWAIDA Saeda et al.	2019	SVM, Cosine similarity, Arabic WordNet	Arabic essays (120 answers)	Improved accuracy using Arabic WordNet	Requires larger datasets for ML models
ABDUL SALAM Mustafa et al.	2022	Optimized LSTM (GWO)	Arabic short answers	Best accuracy with LSTM-GWO hybrid	Uses classical deep learning techniques
BADRY Rasha M. et al.	2023	LSA for semantic scoring	AR-ASAG Arabic dataset	Achieved F1-score of 82.82%	Limited focus on Arabic only
MECCAWY Maram et al.	2023	BERT, Word2Vec	AR-ASAG, Jordanian	BERT achieved best Pearson	Needs more investigation for

		for ASAG	History dataset	correlation (0.84)	accuracy
TULU Cagatay et al.	2021	ASAG with Word Sense Disambiguation (WSD)	Mohler ASAG dataset, CU-NLP dataset	High Pearson correlation, low error rates	Issues with long training times
RAMALINGAM V. V. et al.	2018	ML techniques for AES	Kaggle essay dataset	Over 80% accuracy using polynomial regression	Needs better semantic and syntactic features
HIKMAT A. ABDELJABER	2021	LCS, Arabic WordNet	Dataset of 330 student answers	High agreement with human evaluators	Small dataset, lacks advanced semantic analysis
AL AWAIDA Saeda Esmaile	2019	SVM, Cosine similarity, Arabic WordNet	Jordanian school dataset	Improved accuracy using Arabic WordNet	Needs more research on machine learning models
GOMAA Wael Hassan et al.	2020	Ans2vec model	Texas, Cairo University, SCIENTSBANK	Best Pearson correlation achieved	Need to explore other Sent2vec approaches
ABDUL SALAM Mustafa et al.	2022	Hybrid LSTM with GWO	Arabic short answers	Best results with hybrid LSTM-GWO	Focused only on Arabic
RASHA M. BADRY et al.	2023	LSA for semantic scoring	AR-ASAG dataset	F1-score 82.82%, RMSE 0.798	Primarily focused on Arabic

5. Conclusion

The escalating necessity for more effective and dependable educational instruments has rendered automated essay grading (AEG) systems indispensable within contemporary educational environments. This manuscript has examined the evolution of AEG systems, underscoring their significance in diminishing the temporal demands of manual grading and enhancing the uniformity of evaluations. The implementation of natural language processing methodologies, including Latent Semantic Analysis (LSA), Longest Common Subsequence (LCS), and various machine learning algorithms, empowers these systems to assess the semantic and lexical caliber of student submissions with precision. Despite the fact that numerous approaches, such as advanced deep learning frameworks like BERT and LSTM, have demonstrated encouraging outcomes, the obstacles associated with constrained datasets, particularly in non-English languages like Arabic, continue to pose challenges. Additional inquiry and innovation are imperative to surmount these constraints and augment the accuracy of grading as well as the efficacy of feedback mechanisms. In summary, automated grading systems not only refine the grading procedure but also assist educators by enabling them to concentrate on more vital responsibilities, such as curriculum development and fostering student engagement.

The contributions delineated in our investigation are numerous:

(a) comprehensive survey of the prevailing methodologies for assessing short text similarity; an elucidation of the principal challenges and constraints inherent in these methodologies; an exposition of the current deep learning-driven, text similarity-oriented, and hybrid approaches.

Copyright:

References

[1] SOULIMANI, Younes Alaoui; EL ACHAAK, Lotfi; BOUHORMA, Mohammed. Deep learning based Arabic short answer grading in serious games. International Journal of Electrical & Computer Engineering (2088-8708), 2024, 14.1.

[2] *WANG, Yaqing, et al. Automated evaluation of personalized text generation using large language models. arXiv preprint arXiv:2310.11593, 2023.*

[3] OMRAN, Ali Muftah Ben; AB AZIZ, Mohd Juzaiddin. Automatic essay grading system for short answers in English language. Journal of Computer Science, 2013, 9.10: 1369.

[4] MECCAWY, Maram, et al. Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques. International Journal of Advanced Computer Science and Applications, 2023, 14.6.

[5] TULU, Cagatay Neftali; OZKAYA, Ozge; ORHAN, Umut. Automatic short answer grading with SemSpace sense vectors and MaLSTM. IEEE Access, 2021, 9: 19270-19280.

[6] RAMALINGAM, V. V., et al. Automated essay grading using machine learning algorithm. In: Journal of Physics: Conference Series. IOP Publishing, 2018. p. 012030.

[7] MIZUMOTO, Atsushi; EGUCHI, Masaki. Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2023, 2.2: 100050.

[8] HUSSEIN, Mohamed A.; HASSAN, Hesham A.; NASSEF, Mohammad. A trait-based deep learning automated essay scoring system with adaptive feedback. International Journal of Advanced Computer Science and Applications, 2020, 11.5.

[9] CHO, Brian; JANG, Youngbin; YOON, Jaewoong. Rubric-Specific Approach to Automated Essay Scoring with Augmentation Training. arXiv preprint arXiv:2309.02740, 2023.

[10] WAER, Hanan. The effect of integrating automated writing evaluation on EFL writing apprehension and grammatical knowledge. Innovation in Language Learning and Teaching, 2023, 17.1: 47-71.

[11] L.K. Harper, "Computational investigation of pernicious compounds: Arsenic and high energy density materials and their relevant mechanisms," Old Dominion University, 2015.

[12] C. Devereux et al., "Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens," Journal of Chemical Theory and Computation, vol. 16, no. 7, pp. 4192-4202, 2020. Doi:https://doi.org/10.1021/acs.jctc.0c00121.

[13] M.e. Frisch, G. Trucks, H.B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, and H. Nakatsuji, Gaussian 16, Gaussian, Inc. Wallingford, CT, 2016.

[14] A. Barnawi, I. Budhiraja, K. Kumar, N. Kumar, B. Alzahrani, A. Almansour, and A. Noor, "A comprehensive review on landmine detection using deep learning techniques in 5G environment: open issues and challenges," Neural Computing and Applications, vol. 34, no. 24, pp. 21657-21676, 2022.Doi:https://doi.org/10.1007/s00521-022-07819-9

[15] E.A. Musad, R. Mohamed, B.A. Saeed, B.S. Vishwanath, and K.L. Rai, "Synthesis and evaluation of antioxidant and antibacterial activities of new substituted bis (1, 3, 4oxadiazoles), 3, 5-bis (substituted) pyrazoles and isoxazoles," *Bioorganic & medicinal chemistry letters*, vol. 21, no. 12, pp. 3536-3540, 2011.

[16] A.G. Martynov, J. Mack, A.K. May, T. Nyokong, Y.G. Gorbunova, and A.Y. Tsivadze, "Methodological survey of simplified TD-DFT methods for fast and accurate interpretation of UV–vis–NIR spectra of phthalocyanines," ACS omega, vol. 4, no. 4, pp. 7265-7284, 2019.Doi:https://doi.org/10.1021/acsomega.8b03500.

[17] J. Krupa, M. Wierzejewska, and J. Lundell, "Experimental FTIR-MI and Theoretical Studies of Isocyanic Acid Aggregates," Molecules, vol. 28, no. 3, p. 1430, 2023. Doi:https://doi.org/10.3390/molecules28031430

[18] G.A. Garcia, L. Dontot, M. Rapacioli, F. Spiegelman, P.Bréchignac, L. Nahon, and C. Joblin,"Electronic effects in the dissociative ionisation of pyrene clusters," Physical Chemistry Chemical Physics,vol.25,no.6,pp.4501-4510, 2023.Doi:https://doi.org/10.1039/d2cp05679h

[19] T.G. Mayerhöfer, S. Pahlow, V. Ivanovski, and J. Popp, "Dispersion related coupling effects in IR spectra on the example of water and Amide I bands," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 288, p. 122115, 2023. Doi:https://doi.org/10.1016/j.saa.2022.122115 M. Kim, Y. Ham, C. Koo, and T. W. Kim, "Simulating travel paths of construction site workers via deep reinforcement learning considering their spatial cognition and wayfinding behavior," *Automation in Construction*, vol. 147, p. 104715, 2023.

[20] R. Kiralj and M. Ferreira, "Basic validation procedures for regression models in QSAR and QSPR studies: theory and application," Journal of the Brazilian Chemical Society, vol. 20, pp. 770-787, 2009.

[21] PUN, Kun; WANG, Rong. Automatic Essay Multi-dimensional Scoring with Fine-tuning and Multiple Regression. arXiv preprint arXiv:2406.01198, 2024.

[22] KUMAR, Vivekanandan S.; BOULANGER, David. Automated essay scoring and the deep learning black box: How are rubric scores determined?. International Journal of Artificial Intelligence in Education, 2021, 31: 538-584.

[23] KONADE, Shashank, et al. Implementation of an Automated Answer Evaluation System. In: 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). IEEE, 2024. p. 457-462.

[24] OUAHRANI, Leila; BENNOUAR, Djamal. AR-ASAG an Arabic dataset for automatic short answer grading evaluation. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020. p. 2634-2643.

[25] MIJBEL, Sumaya H.; SADIQ, Ahmed T. Short Answers Assessment Approach based on Semantic Network. Iraqi Journal of Science, 2022, 2702-2711.

[26] AL AWAIDA, Saeda A.; AL-SHARGABI, Bassam; AL-ROUSAN, Thamer. Automated arabic essay grading system based on f-score and arabic worldnet. Jordanian Journal of Computers and Information Technology, 2019, 5.3.

[27] BALAHA, Hossam Magdy; SAAFAN, Mahmoud M. Automatic exam correction framework (aecf) for the mcqs, essays, and equations matching. IEEE Access, 2021, 9: 32368-32389.

[28] LOTFY, Nourmeen, et al. An Enhanced Automatic Arabic Essay Scoring System Based on Machine Learning Algorithms. CMC-COMPUTERS MATERIALS & CONTINUA, 2023, 77.1: 1227-1249.

[29] BADRY, Rasha M., et al. Automatic Arabic grading system for short answer questions. IEEE Access, 2023, 11: 39457-39465.

[30] ABDUL SALAM, Mustafa; EL-FATAH, Mohamed Abd; HASSAN, Naglaa Fathy. Automatic grading for Arabic short answer questions using optimized deep learning model. Plos one, 2022, 17.8: e0272269.

[31] SHEHAB, Abdulaziz; FAROUN, Mahmoud; RASHAD, Magdi. An automatic Arabic essay grading system based on text similarity Algorithms. International Journal of Advanced Computer Science and Applications, 2018, 9.3.

[32] ELNAKA, Abdelrahman, et al. AraScore: Investigating response-based Arabic short answer scoring. Procedia Computer Science, 2021, 189: 282-291.

[33] GOMAA, Wael Hassan; FAHMY, Aly Aly. Ans2vec: A scoring system for short answers. In: The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4. Springer International Publishing, 2020. p. 586-595.

[34] ABDELJABER, Hikmat A. Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet. IEEE Access, 2021, 9: 76433-76445.

[35] ODEH AL-AWAIDA, Saeda Esmaile. Automated Arabic essay grading system based on support vector machine and text similarity algorithm. Doctoral dissertation, 2019.

[36] AL-SHARGABI, Bassam; ALZYADAT, Rawan; HAMAD, Fadi. Aegd: Arabic essay grading dataset for machine learning. Journal of Theoretical and Applied Information Technology, 2021, 99.6.

[37] AZMI, Aqil M.; AL-JOUIE, Maram F.; HUSSAIN, Muhammad. AAEE–Automated evaluation of students' essays in Arabic language. Information Processing & Management, 2019, 56.5: 1736-1752.

[38] https://www.irjmets.com/uploadedfiles/paper//issue_11_november_2023/45831/final/fin_irjmets1698934910.pdf

[39] MECCAWY, Maram, et al. Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques. International Journal of Advanced Computer Science and Applications, 2023, 14.6.

[40] YOON, Su-Youn. Short answer grading using one-shot prompting and text similarity scoring model. arXiv preprint arXiv:2305.18638, 2

[41] E. B. Page, "Project Essay Grade: Peg.," 2003.

[42] Prakoso, Dimas Wibisono; Abdi, Asad; Amrit, Chintan. Short Text Similarity Measurement Methods: A Review. Soft Computing, 2021, 2

[43] Kim, Sun, Et Al. Bridging The Gap: Incorporating A Semantic Similarity Measure For Effectively Mapping Pubmed Queries To Do Informatics, 2017, 75: 122-127.

[44] Mohamed, Muhidin; Oussalah, Mourad. Srl-Esa-Textsum: A Text Summarization Approach Based On Semantic Role Labeling And Explicit Semantic Analysis. Information Processing & Management, 2019, 56.4: 1356-1372.

[45] Yoon Kim. 2014. Convolutional Neural Networks For Sentence Classification. In Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (Emnlp). 1746–1751.

[46] Janda, Harneet Kaur, Et Al. Syntactic, Semantic And Sentiment Analysis: The Joint Effect On Automated Essay Evaluation. Ieee Access, 2019, 7: 108486-108503.

[47] Zou, Will Y., Et Al. Bilingual Word Embeddings For Phrase-Based Machine Translation. In: Proceedings Of The 2013 Conference On Empirical Methods In Natural Language Processing. 2013. P. 1393-1398.

[48] Chandrasekaran, Dhivya; Mago, Vijay. Evolution Of Semantic Similarity—A Survey. Acm Computing Surveys (Csur), 2021, 54.2: 1-37.