Asian Journal of Distance Education

Unveiling the Landscape: Studies on Automated Short Answer Evaluation

Abdulkadir Kara, Eda Saka Şimşek, Serkan Yildirim

Abstract: Evaluation is an essential component of the learning process when discerning learning situations. Assessing natural language responses, like short answers, takes time and effort. Artificial intelligence and natural language processing advancements have led to more studies on automatically grading short answers. In this review, we systematically analyze short-answer evaluation studies. We present the development of the field in terms of scientific production features, datasets, and automatic evaluation features. The field has developed with pioneering studies in the US. Researchers generally conduct applications with English datasets. There has been a significant increase in research in recent years with large language models that support many different languages. These models have applications that achieve accuracy close to that of human evaluators. In addition, deep learning models do not require traditional approaches' detailed preprocessing and feature engineering processes. The dataset size trend is 1000 and above regarding the number of responses. It was observed that metrics such as accuracy, precision, and F1 score were used in performance determination. It is seen that the majority of the studies focus on scoring or rating. In this context, there needs to be more literature on the context of evaluation system applications that can provide descriptive and formative feedback. In addition, the developed assessment systems must be actively used in learning environments.

Keywords: natural language responses, artificial intelligence, short answers, automatic evaluation, ASAG trends

Highlights

What is already known about this topic:

- Assessing natural language responses, such as short answers, requires a significant amount of time and effort.
- Developments in natural language processing and artificial intelligence have made automatic evaluation necessary.

What this paper contributes:

- By stressing the crucial characteristics of the developed apps, effective application features are highlighted.
- Researchers can understand the conditions for a good application in this field.

Implications for theory, practice and/or policy:

- Highly effective applications, nearly as accurate as humans, are possible now. In recent years, deep learning approaches have come to the fore in this context.
- The development of artificial intelligence and big language models may accelerate the spread of the field and the acceptance of its active use in learning environments.

Published by Asian Society for Open and Distance Education (ASODE), Japan
ISSN 1347-9008 http://www.asianjde.com/

Introduction

Change is being embraced more rapidly than in previous times. Emerging innovations that yield advantages are rapidly becoming the standard. The academic community actively discovers and effectively applies these emerging phenomena in relevant domains. Indeed, the advancement and utility of humanity are contingent upon the pursuit and perpetuation of these disciplines. The academic realm exhibits a notable inclination for novelty in education. Currently, there is a significant level of popularity surrounding research in the realm of educational and instructional technologies. The purpose is clear: How can we simplify learning, teaching, and process management?

So, where do we put our finger on this field with this research? Lately, there has been a rise in research focusing on the automated evaluation of natural language answers (NLRs). Of course, breakthroughs in the domains of natural language processing (NLP), machine, and deep learning have substantially impacted this increase. With the pandemic, most educational disciplines and levels have undergone a mandatory transformation to open and distance learning. In this process, the popularity of e-learning applications has increased (Gocmez & Okur, 2023). The change in learning environments, a movement from summative evaluations to formative assessments, has been noticed (Bozkurt et al., 2020). The use of NLR in homework and exam applications became significant during this period.

Evaluation activities are carried out in learning environments to determine students' achievements. Evaluation is essential to the learning process when discerning learning situations (Pribadi et al., 2016; Yan, 2020; Putnikovic & Jovanovic, 2023). From past to present, it has been observed that the superficial multiple-choice grading method is commonly used to represent learning success (Hasanah et al., 2016; Garg et al., 2022). Experts prefer NLRs such as fill-in-the-blank, short answers, and essays less than choice-based responses (Benli & Ismailova, 2018). NLRs are less preferred due to assessment and evaluation's laborious and time-consuming nature (Gombert et al., 2023; Wilianto & Girsang, 2023; Westera et al., 2018). Additionally, NLR assessments require human intervention. Human intervention can lead to subjective judgments and reduce assessment reliability.

We require a fairer way to collect sufficient proof of students' knowledge and abilities (Noyes et al., 2020). Learners' NLRs are crucial evidence for a comprehensive learning process evaluation (Westera et al., 2018). Furthermore, this evidence can aid in assessing learning materials and teaching techniques to enhance teaching scenarios (Noyes et al., 2020). NLR can evaluate learners' critical thinking and self-expression skills (Uto & Uchida, 2020).

Today, studies on using artificial intelligence (AI) in learning environments have increased. Song and Wang (2020) expect that learning environments will actively use AI technologies in the future. AI is becoming widespread in personalized, expert, and intelligent learning systems (Goksel & Bozkurt., 2019). AI components also produce opportunities to help teachers' assessment processes (Botelho et al., 2023). Automatic evaluation of the NLR is one of them. With the rise of ODL and lifelong learning applications, the tendency for research in this field has increased (Jadidinejad & Mahmoudi, 2014). Automatic assessment of NLR has become essential due to its ability to offer impartial grading, reduce human workload, and save time via a speedy evaluation procedure (Abdul Salam et al., 2022).

NLR automatic grading studies date back to the 1960s (Page, 1966). However, advanced systems have attracted attention in the recent past. Especially since 2010, numerous studies have been on the rating of NLR (Ghavidel et al., 2020). Automatic short answer gradings (ASAG) are one of the prominent researches in this field. ASAG is a system that compares and scores the learner's answer with one or more correct reference answers (Mohler & Mihalcea, 2009). ASAG plays a significant role in resolving the challenges arising from the nature of NLR. ASAG has positively influenced the use of brief answers in educational contexts instead of multiple-choice appraisals, widely viewed as practical (Noyes et al., 2020). ASAG systems offer various benefits, including objective scoring, prompt feedback, reduced

teacher workload in managing teaching situations, and enhanced monitoring of learners' performance (Liu et al., 2016).

Literature

Short Answers

Short answers consist of a few words or sentences (Nath et al., 2023) and are included in NLR with fill-in-the-blank and essay responses. Burrows et al. (2015) outline three dimensions: length, focus, and clarity to differentiate response types. Their characteristics include objectivity and closed-ended, with a focus on content. Short answers should be concise and require recall of external knowledge (Burrows et al., 2015).

In evaluations of short answers, the semantic evaluation of content is paramount. This situation presents some challenges. Variations in NLR from students complicate the automatic evaluation process (Benli & İsmailova, 2018; Gomaa et al., 2023). Additionally, the results generated by these systems, which incorporate scores from human evaluators, have faced criticisms regarding their validity (Attali, 2015). Researchers are using the capabilities of natural language processing (NLP) and artificial intelligence (AI) technologies to overcome current challenges. Recently, there has been a focus on using artificial neural networks (ANN) and machine learning techniques to enhance evaluations in the domain of meaning. As new approaches emerged, evaluators have followed different paths in the evaluation process. After examining the literature, researchers have made various classifications for these approaches by focusing on different dimensions.

Automatic Short Answer Grading Approaches and Models

When examining automatic short answer evaluation applications, it becomes clear that varying preferences exist to achieve more effective results according to developments in related fields. In the literature, we identified classifications regarding approaches.

Burrows et al. (2015) periodically analyzed short-answer automatic assessment approaches under four headings. The approaches are concept mapping, knowledge extraction, corpus-based, and machine learning. Another study classifies approaches as similarity-oriented. Abdul Salam et al. (2022) provide a breakdown of the approaches, which include string-based, semantic-based, hybrid-based, and machine & deep learning-based techniques.

String-based similarity is a technique employed to determine the similarity of a student's response to the correct answer. The method evaluates the similarity between two sentences independently of their meaning, expressing it via a score based on word-level similarity (Gomaa & Fahmy, 2012). The C-rater short answer scoring system is one of this approach's most classic and successful applications (Leacock & Chodorow, 2003).

The semantic-based similarity approach involves computing the semantic similarity between a question and an answer (Corley & Mihalcea, 2005). The vector space model (VSM) and latent semantic analyzis (LSA) are frequently applied techniques in this methodology (Wang & Dong, 2020). VSM measures how important words are to the document to calculate intertextual similarity (Rodrigues & Araújo, 2012). LSA changes words into vectors by studying significant amounts of text (Ratna et al., 2019).

The hybrid approach typically combines statistical methods and NLP techniques. This approach combines the benefits of different methods (Prakoso et al., 2021). Statistical methods incorporate measures, including Jaccard similarity, Levenshtein distance, and cosine similarity. NLP techniques

encompass word embedding models and deep learning techniques, such as recurrent neural networks (RNN). The hybrid approach employing various similarity measures can lead to more precise results.

The machine learning method typically comprises extracting features, selecting features, and applying classification algorithms. The student's responses are analyzed based on their attributes and juxtaposed with the correct answer. Conversely, deep learning often employs neural networks such as Convolutional Neural Networks (CNN), RNN, and Long short-term memory (LSTM). LSTM classifies students' answers as correct or incorrect by considering the order of each word or phrase, especially when their answers have a series of words or sentences. LSTM works exceptionally well with complex and long sentences.

Related Work

Burrows et al. (2015) reviewed the field's historical development over time using preferred methodological trends in ASAG systems. The study examines method-based approaches such as concept mapping, knowledge extraction, corpus-based, and machine learning for automatic evaluation. This study presents 35 ASAG systems based on their approaches and basic features. The components of the systems are then analyzed in terms of datasets, NLP, model development, rating models, evaluation, and effectiveness. The study revealed a low number of shared datasets for confidentiality reasons. The systems used 17 distinct techniques in the context of NLP. Recent studies have shown a preference for machine learning in these systems. Evaluation results are typically represented by accuracy, kappa, and Pearson. The researchers who conducted this study analyzed efficiency and concluded that the concept mapping approach was more effective than others.

Hasanah et al. (2016) adopted a specialized approach and examined 10 ASAG systems using only the knowledge extraction method. The research details the development process of an ASAG system, employing the knowledge extraction method with subsections outlining dataset creation and preprocessing, application of knowledge extraction techniques, and model evaluation. They analyze the preferred information extraction techniques, datasets, and evaluation results of the systems. It was found that syntactic pattern matching is the most prominent information extraction technique. The datasets are generally selections oriented towards science and biology. During the analyzis of 10 ASAG systems suitable for the approach, it is worth noting that the most recent study was conducted in 2012. The popularity of the knowledge extraction approach over the years explains this situation (Hasanah et al., 2016). Another factor is that researchers can combine the techniques used in knowledge extraction with other approaches.

Galhardi and Brancher (2018) focus on the machine learning approach Burrows et al. (2015), which has recently been favored in ASAG. The study analyzes 44 ASAG systems using a machine-learning approach and highlights the importance of automatic evaluation in training. The themes identified for the systematic review are the structure of the preferred datasets, the NLP and machine learning techniques applied, the features selected, and the evaluation results. It was found that 28 different datasets were used in the 44 systems analyzed. The datasets were primarily scientific (57%) and mainly in English (75%). When analyzing the NLP techniques applied to the datasets in the systems, it was found that more than ten techniques were identified. It is noted that the n-gram method, which is mainly below the lexical dimension, is preferred in feature selection.

Putnikovic and Jovanovic (2023) focused on embeddings, which have recently been widely used in ASAG systems. In their study, they conducted a systematic literature review covering the period between 2016 and 2021 in 7 different databases. In this research, in which they used the PRISMA model, 17 full-text studies were analyzed. In this study, embedding model types are analyzed in detail. The use cases of embedding models, component preferences, and comparison of their performance contributions with non-embedded systems are investigated. The types of embedding are explained

under four headings: word, context, sentence, and sense. The analyzed studies found that the embedding technique was often used with cosine similarity or a neural network. It is found that systems with embedding need to make a clear positive contribution to performance compared to systems without embedding.

The four review studies described above focus on ASAG systems based on traditional AI components. Burrows et al. (2015) analyzed the identified ASAG approaches holistically. The other three studies focused on one of the approaches or models used.

Purpose of the Research

Our study aims to reveal the current status of ASAG studies in the literature with a systematic review of research. Our study focuses on ASAG studies classified by similarity-based approaches based on traditional approaches and AI technologies. We analyze the developed applications under three themes: scientific production (1), the data set (2), and automatic evaluation (3). In this context, the research questions that we seek answers to are as follows.

What is the trend of scientific production characteristics of the studies?

1. What is the distribution of scientific production in the field according to states?
2. What is the distribution of years of scientific production in the field?

What are the general characteristics of the datasets used?

1. What is the distribution of languages in the datasets?
2. What are the dimensional tendencies of the datasets?
3. What are the scoring formats in the datasets?

What is the tendency of the features used in the automatic scoring model?

1. What are the tendencies of the scoring techniques used?
2. What is the tendency of the scoring analyzes used?
3. What is the trend in the use of evaluation results?

Importance of the Study

In our study, we wanted to portray the development of the field with the findings obtained from the research we analyzed. There are significant differences that distinguish our study from other studies. First, we conducted our study to cover all approaches in the field of ASAG. In addition, we based our study on similarity-oriented approaches. This study is the first comprehensive study conducted on similarity. Under the theme of scientific production characteristics, information on the states and years in which ASAG studies were conducted was also included in our study for the first time. This information is valuable in seeing how widespread the field has become. Another important differentiating factor is that our study discusses the dataset characteristics in detail. The language, dimension, and scoring type characteristics were carefully extracted. We also attach great importance to including our study's latest approaches and models. Interest in the ASAG field has increased, especially in recent years, with the developments in the field. In this context, while presenting the past studies in the field in detail, we also offer the trends of current studies.

As a result of all this research, we wanted to illustrate the current trends in the field and discuss the promising features and existing gaps. The study's results will identify the characteristics of current, effective, and efficient applications in the literature and guide researchers working in the field or who will work in the future.

Methodology

Research Method and Design

In this study, the aim is to uncover the present state of ASAG applications regarding research queries. In line with the purpose of the study, it was appropriate to conduct a systematic review of the literature. The systematic literature review examines and evaluates the information previously produced on a topic in a focused way. Using a systematic, reliable, and verifiable approach, researchers can identify the current state of the field of study in a systematic literature review (Galhardi & Brancher, 2018). This method aims to reach the most appropriate studies to answer the research questions (Liberati et al., 2009). In this study, the PRISMA 2020 statement method developed by Page et al. (2021) was followed.

Data Collection

The study focuses on the use of ASAG systems. The study group was defined as studies using ASAG systems published in generally accepted indexed journals. We analyzed the studies by searching the Web of Science (WoS) database. We analyzed articles on ASAG applications, considering years, states, evaluation approaches, dataset characteristics, model characteristics, and evaluation results obtained.

First, a preliminary search was carried out for commonly used terms related to the subject of the study. The most commonly used related terms in the literature were identified. The database was filtered under three headings during the screening process, as shown in Table 1.

Table 1. Database filtering keywords

Automated process	automatic, automated
Short answer	short, text, response, answer, question
Evaluation	assessment, scoring, marking, grading

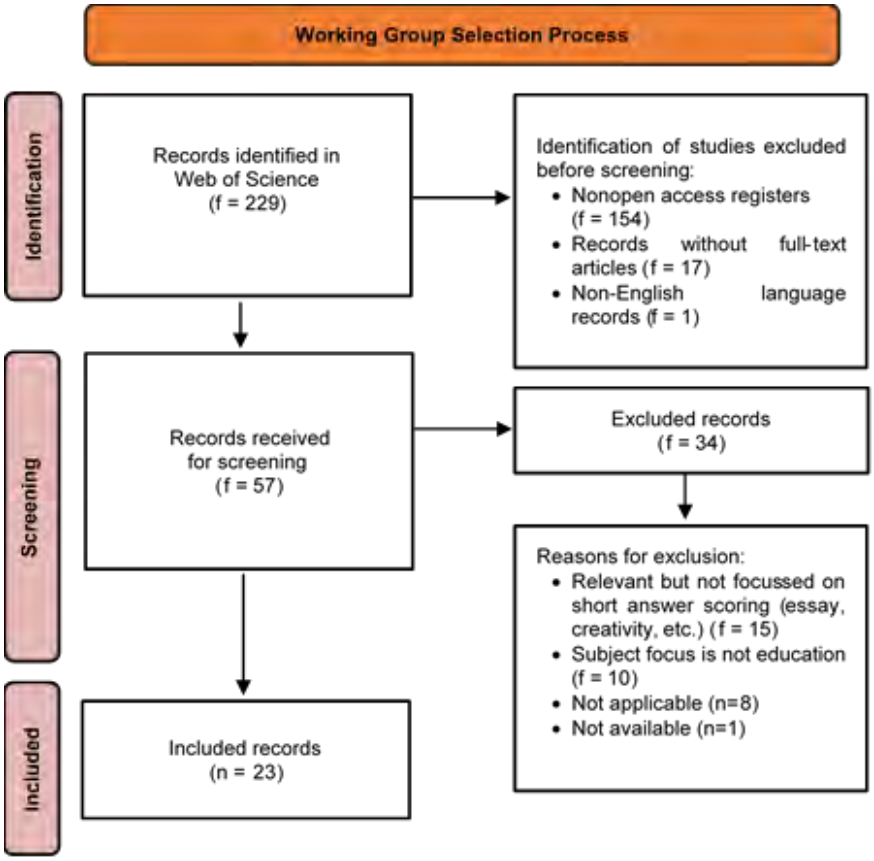
In the literature, artificial intelligence or computer-based concepts are generally defined as automatic processes, and the main concepts are "automatic and automated." In our study, we found that when specifying the type of answer, terms such as "short, text, response, answer" were used, which are the focus of our study. When we examined the concept of assessment, we found that this subject area used terms such as "assessment, scoring, marking, and grading." During the screening process in this context, we organized these terms directly related to the topic into keywords and synonyms. Following the method used by Galhardi and Branche (2018) in their systematic review, a similar database filtering sequence was used to identify ASAG-focused studies;

ALL=(((((("automatic assessment" OR "automatic scoring" OR "automatic marking" OR "automatic grading" OR "automated grading" OR "automated scoring" OR "automated marking") AND (short OR "short answer" OR text) AND (response OR question OR answer)).

The screening process used the selection criteria set by the researchers. The selection criteria were as follows:

1. Being scanned in the WoS database,
2. Inclusion of keywords,
3. The research topic is related to education and training,
4. Full-text publication in article type,
5. Written in English or Turkish,
6. Have open access,
7. An application-oriented study.

Figure 1. Working Group Selection Process



In the screening process, the PRISMA method developed by Liberati et al. (2009) was followed for the reliability of the systematic review. Figure 1 shows the data collection selection process. The keyword search yielded 229 studies. The filtering process identified 75 open-access studies and 57 studies that met the language requirement in the article type. We selected 23 studies that met the inclusion criteria for review. The researchers analyzed the included studies in detail to answer the research questions. We summarized the obtained data, visualized and presented the findings concerning the research questions, and presented the current findings.

Limitations

The limitations of our study are listed below:

- Our analyzis is limited to the relevant studies in the WoS database.
- Studies that are not open access could not be included in the study.
- The scope is limited to the set of keywords used in the study.
- We could not include studies unsuitable for the research questions and whose results were uncertain.

Findings and Discussions

In the WoS database, 23 applied studies on ASAG that met the review criteria were analyzed under three themes concerning the research questions identified by the researchers. First, (1) the publications' state and year were analyzed for scientific production characteristics. Then, (2) dataset characteristics were analyzed regarding language, subject, number of responses, and scoring. Finally, (3) the characteristics of the automated evaluation were analyzed in terms of preferred approaches and models, evaluation metrics, and system performance. The findings obtained at the end of the review process were quantified with values such as frequency (f) and percentage (%) and visualized and presented in

tables and figures. The conclusions drawn from the findings are also presented under the relevant topics.

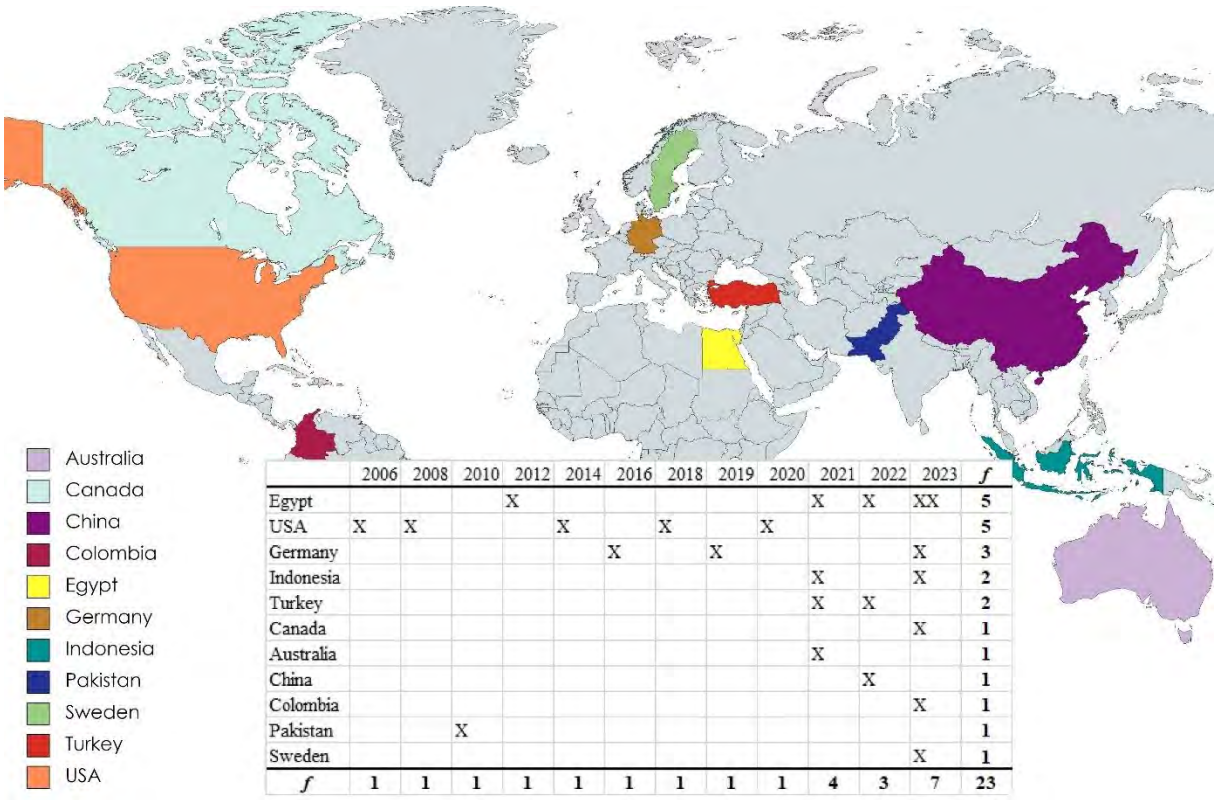
Scientific production characteristics

This topic aims to identify the main scientific production characteristics of ASAG studies. In the expectation that this will contribute to the study's primary purpose, the scientific production characteristics are limited to states, years of publication, and keyword selections. The aim is to identify the states and years in the field. To this end, the studies' results are visualized, presented, and explained under the relevant headings.

First, we superficially extracted the scientific production characteristics of ASAG studies by looking at the state and year information. Our aim here is to reveal which state researchers are prominent in the orientation of the field and the development of the field over the years. Notably, these data have yet to be addressed in previous studies in the field.

For this purpose, the states and years of the publications were analyzed. The findings and results obtained are discussed together. To obtain more meaningful results, the distribution of state and year of publication information is presented in Figure 2. Figure 2 also shows the geographical distribution of the studies.

Figure 2. Distribution of ASAG studies by state and year



	2006	2008	2010	2012	2014	2016	2018	2019	2020	2021	2022	2023	*f*
Egypt				X						X	X	XX	5
USA	X	X			X		X		X				5
Germany						X		X				X	3
Indonesia										X		X	2
Turkey										X	X		2
Canada												X	1
Australia										X			1
China											X		1
Colombia												X	1
Pakistan			X										1
Sweden												X	1
f	1	1	1	1	1	1	1	1	1	4	3	7	23

The USA (f=5), Egypt (f=5), and Germany (f=3) stand out in the number of publications in the studies conducted in the field. However, there are also studies in Indonesia (f=2), Turkey (f=2), Australia (f=1), Canada (f=1), China (f=1), Colombia (f=1), Pakistan (f=1) and Sweden (f=1). The first study included in the study was published in 2006. Since the study limit was 2023(November), it was determined that the last studies were conducted in 2023. It is noteworthy that the studies are concentrated in 2021 (f=4), 2022 (f=3) and 2023 (f=7).

Two states, the USA and Egypt, stand out in the orientation and development of the field over the years. The USA has laid the foundations of ASAG studies, with pioneering studies centered mainly in the USA. When Figure 1 is examined, it is seen that the USA is at the center of the first studies in the field. Especially in the year 2021, a dramatic increase in the number of studies was observed. In recent years, studies conducted in different states have stood out.

In scientific production, the US dominance has been broken in recent years with research conducted in other states. Pre-learned large language models such as GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and T5 (Text-To-Text Transfer Transformer) have played a significant role in the growth of native language studies. In recent years, it is clear that researchers from different states have carried out studies on native languages.

The distribution of ASAG applications by years was analyzed in relation to the developments from the past to the present. The field of ASAG attracted attention in the 2000s and intensified in the second half of the 2010s and after 2020. Technological and methodological developments directly affected the frequency of publications. In the last three years, it has been observed that ASAG applications have become widespread, and there is a severe research trend.

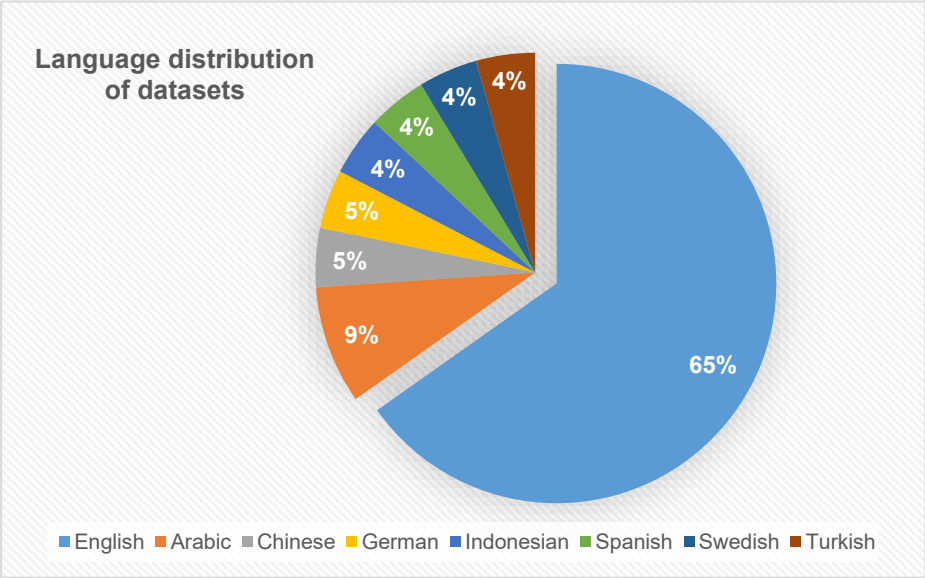
Dataset Characteristics

Datasets are essential because they reflect the characteristics, such as the language in which the system is trained, the amount of data it is trained with, and how it scores the responses. Therefore, in this study, the characteristics of the datasets used in ASAG systems were analyzed by classifying them in terms of (1) language, (2) the number of responses, and (3) scoring methods. The findings and discussion are presented under the relevant headings.

Language distribution

The language distribution of the datasets used is visualized and presented in Figure 3. It was found that the use of datasets in English (%65) was common in the studies analyzed. It is followed by Arabic (%9). However, it was also observed that there are datasets using different languages (Chinese, German, Indonesian, Spanish, Swedish, and Turkish).

Figure 3. Language distribution of datasets

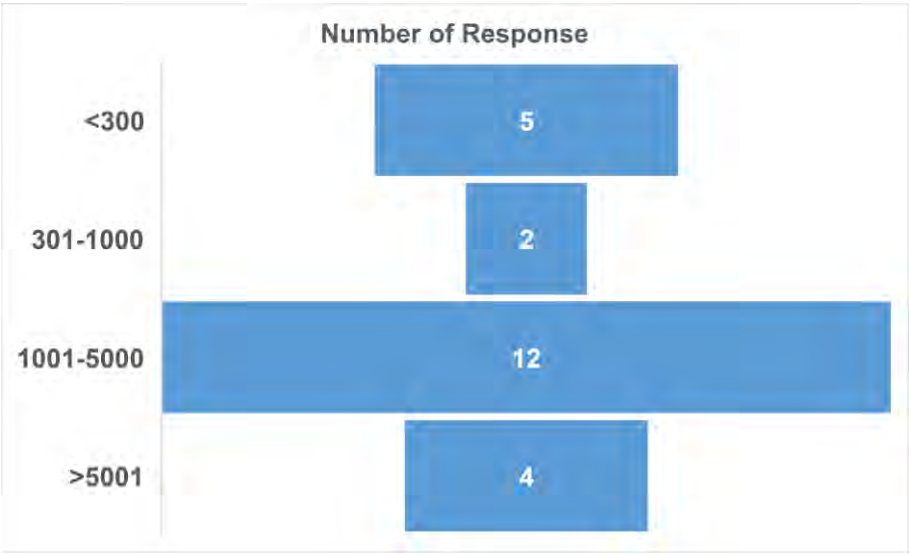


ASAG involves the active operation of NLP components. The analyzed studies observed that English was mainly preferred as the language in the datasets. Using traditional ASAG approaches in previous studies is essential when choosing an English data set. Because NLP research on English is much older (Jones, 1994), in this context, English data sets are more frequently preferred in ASAG. The use of English datasets is widespread in the field. NLP has recently made progress in various languages. Developing effective ASAG systems for different languages is now easier than ever. This is evident in the literature, with more and more work recently being done in different languages. With the expansion of this dataset language, ASAG can reach a more widespread and effective use in the future.

Dataset size

The dataset size (number of responses) is a prominent feature of the datasets used to train systems in ASAG studies. This study analyzed the number of responses in the datasets. To draw more meaningful conclusions, the distribution of the number of responses is analyzed according to five categories defined by the researchers. The aim is to determine the typical range of response numbers preferred for ASAG in the field. The results are visualized and presented in Figure 4.

Figure 4. Proportional distribution of the number of responses by category



The number of responses was concentrated in the range of 1001-5000 (f=12). This was followed by the <300 range (f=5). It was also observed that there were studies (f=4) that used 5001 and above number of responses. The least preferred response number range was observed in the 301-1000 range (f=2).

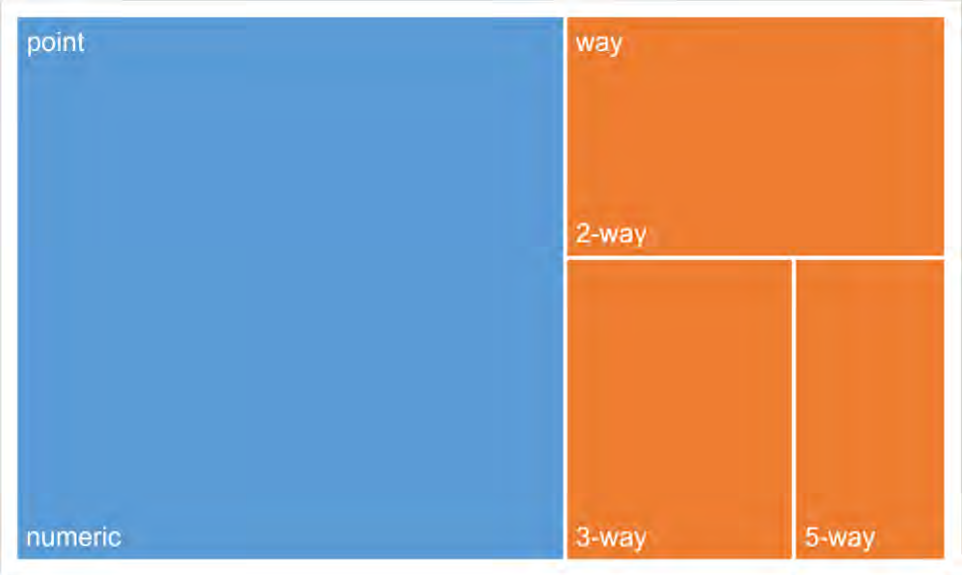
The number of responses in the datasets significantly influences the accuracy of system results. The studies concentrated on datasets of 1000 or more responses. The findings may indicate that this range (1001-5000) is the optimal data set size for system effectiveness. Nevertheless, other components for system effectiveness should be addressed.

Scoring method

The scoring methods were analyzed to determine how the developed ASAG systems scored the responses. The scoring methods were categorized as (1) point and (2) way scoring. Way scoring includes Likert-type classification-based scoring. Point scoring includes numerical scoring based on

proportions. The results of the preferred scoring methods in the responses are visualized and presented in Figure 5.

Figure 5. Distribution of scoring methods of datasets



As can be seen in Figure 5, two methods are used for scoring student responses in the data sets: point and way method. Point scoring is used relatively more frequently in the studies. It is more advantageous for summative assessments to differentiate student achievement. The way method, on the other hand, facilitates the use of the ASAG system by categorizing response accuracy more flexibly. In this context, the path method is also frequently preferred in studies. The way method structures used are categorized as 2-way, 3-way, and 5-way according to the accuracy level of the responses.

Automatic Evaluation Features

In this study, automatic evaluation features are analyzed in terms of (1) approaches and models, (2) evaluation methods and metrics, and (3) system performance. Thus, it will be possible to reveal how the automatic evaluation processes of the systems included in the study are carried out. The findings and discussion are presented under the relevant headings.

Approaches and models

The approaches preferred in the ASAG systems were analyzed based on the similarity of the automatic evaluation approaches in which the systems were developed. The analyzed studies were assigned to the appropriate ones from (1) machine and deep learning, (2) semantic-based, (3) string-based, and (4) hybrid-based approaches described in Abdul Salam et al. (2022). Thus, the approaches and models in the studies in the field will be identified, and it is expected to reveal the approach and model trends. In Table 2, the approaches preferred in the studies are visualized and presented.

Table 2. ASAG Approaches

ASAG Approach	Frequency	Percentage
Machine & Deep Learning	11	48%
Hybrid Based	6	26%
Semantic Based	4	17%
String Based	2	9%

As seen in Table 2, machine and deep learning (f=11) approaches are prominent in the studies. It is followed by hybrid-based (f=6) and semantic-based (f=4) approaches. In Table 6, the hybrid-based approaches stand out in the numerical context due to the combination of semantic, machine, and deep learning approaches. Only two studies were conducted with a string-based approach. The prominent models and techniques in the context of the approaches used in the analyzed studies are shown in Figure 6.

Figure 6. Approaches and Models Used

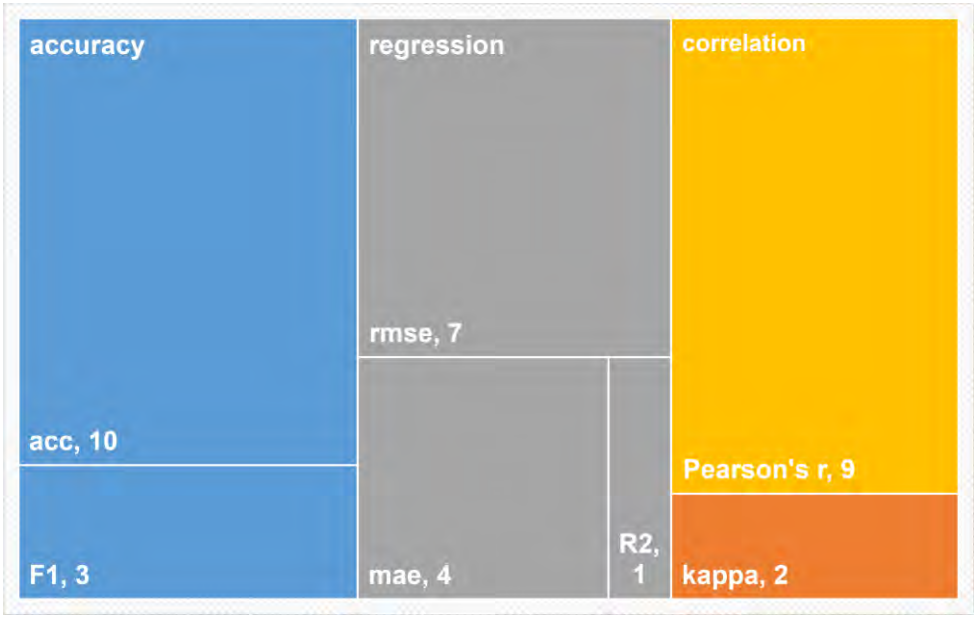
As seen in Figure 6, semantic-based and machine & deep learning-based approaches are widely used in ASAG. The most prominent modeling techniques in the field are LSA, LSTM, and BERT. Focusing on the meaning of the content in short answers has an essential share in the preference of these techniques. In addition, successful prediction performances, especially with BERT and LSTM modeling, make these models frequently preferred in studies

Evaluation metrics

The evaluation metrics and methods used in the studies are analysed to determine how the performance of the ASAG systems is determined. This is expected to provide a clearer understanding of the performance results of the systems. The metrics are (1) accuracy, (2) agreement, (3) regression, and

(4) correlation. Figure 7 shows the results of the evaluation metrics. In addition, the calculation techniques used in applying the metrics in the studies are presented.

Figure 7. Evaluation metrics types distribution



Evaluation metrics are essential in revealing the performance of the system. The analyzed studies showed that accuracy (f=13) was the most preferred metric for determining system performances. The prominent evaluation metrics are acc (f=10) and f1 score (f=3). This is followed by regression (f=10). In regression metrics, rsme (f=7), mae (f=4) and R2 (f=1) metrics were used. Pearson's r metric (f=9), one of the correlation metrics, was also frequently used in the studies.

System performance

It can be said that system performance is the most fundamental indicator of how effective the developed ASAG system is. The results obtained can be categorized according to specific criteria, and evaluations such as poor, medium, and sound can be made on system effectiveness. However, such a categorization may be considered rigid (Burrows et al., 2015). For this reason, the results of the system performance in the analyzed studies are presented in terms of the best and lowest values. The findings of the performance results obtained in this context are presented in Table 3.

Table 3. System performance

Metric Type	Metric	Highest	Lowest
Accuracy	acc	%98.66	%63.5
	F1	0.828	0.72
Agreement	kappa	0.503	0.955
Regression	mae	0.02	0.738
	rmse	0.04	0.807
	R^2	0.826	0.826
Correlation	Pearson's r	0.989	0.708

After analyzing the preferred evaluation performance results in the studies, we found that the best accuracy metric is 98.66%. For this metric type, 63.5% stands out as the lowest ACC value observed. It can be said that the best MAE (0.02) and RMSE (0.04) values were achieved in the study using the

regression metric. The lowest and highest kappa values observed are 0.503-0.955. In studies where the correlation metric was preferred, the best Pearson's r value was 0.989. The lowest Pearson's r value observed for this metric was 0.504.

System performance is the most fundamental indicator of the effectiveness of the developed ASAG system. The results obtained can be categorized according to specific criteria, and evaluations such as poor, medium, and sound can be made on system effectiveness. Burrows et al. (2015) consider such a categorization to be rigid. Therefore, the results of system performance in the studies analyzed are given directly. Nevertheless, the type of system performance can be highlighted in the context of the approach. This is because the approaches and models preferred in the system developed are practical in system performance. From this point of view, the studies analyzed generally achieved high results, especially in the case of deep learning models. Although other models can also show high results, the variability of the results is higher. Researchers observed that different studies utilize the Texas SA dataset in this study. The study, which used the deep learning approach, showed a very high success compared to other studies.

Discussion

We discuss the characteristics of the ASAG applications that we examined in our study and the increasing interest in the field in the context of these characteristics. In recent years, it has been observed that ASAG applications have become widespread, and there is a severe research trend. Developments in NLP and AI technologies have triggered the rise of the field and the spread of applications. The COVID-19 pandemic period has caused educational environments to shift to digital environments (Gabriel et al., 2022). In recent years, the rise of digital learning environments has increased interest in automated assessment (Zhang et al., 2020; Nath et al., 2023). Using ASAG for detailed assessment processes can be an essential opportunity (Putnikovic & Jovanovic, 2023). Many researchers from different states have addressed this opportunity.

When examining the dataset features, our first topic is the language of the dataset used. With English, it is possible to develop effective systems with less difficulty (Hasanah & Hartato, 2020). Badry et al. (2023) also emphasize that most researchers conducted fieldwork using English datasets. The dominance of English datasets can be traced back to previous approaches that required detailed feature engineering for preprocessing. Working in different languages with approaches that require detailed feature engineering is challenging. Each preprocessing step can accumulate errors and cause problems in predictive modeling (Zhu et al., 2022). In this context, past approaches' limitations have been discussed regarding reliability and validity (Attali, 2015). The available ASAG datasets in English are reasonably sufficient regarding topic and dimension distribution. Especially after 2010, competitions with financial awards like ASAP and SemEVAL contributed to the deepening of the field. They paved the way for the diversification of data sets and the development of different ASAG approaches.

Although the use of English datasets is prevalent in our findings, the use of datasets in different languages is increasing. NLP has recently made progress in various languages. There are dictionaries for different languages, such as WordNet and FastText. These dictionaries are enormous. Researchers have also created pre-learning models for NLP, such as BERT. These models can be used in more than one language. This is noticeable in the current work's diversity of languages and choice of approaches. With current approaches such as BERT, GPT, and T5, the dominance of English has been broken. Since it is based on deep learning, BERT models can be said to adapt much more strongly to different language features (Sayeed & Gupta, 2022). Arabic-oriented studies have increased in Egypt. It can be said that ASAG applications have been developed in various languages, including German (Zehner et al., 2016), Chinese (Li et al., 2022), Turkish (Uysal & Dogan, 2021), Swedish (Weegar & Idestam-Almquist, 2023), Spanish (Mardini et al., 2023) and Indonesian (Lubis et al., 2021), in addition to Arabic.

ASAG ultimately focuses on predicting student responses as close to accuracy as possible. The accuracy of the system results is significantly affected by the number of responses in the datasets. Studies have shown that high-dimensional responses lead to high accuracy and low error rates (Abdul Salam et al., 2022). Studies have focused on datasets of 1000 or more responses. Thus, the negative situations that may occur in multiple scoring formats are minimized. Because as the size of the data set increases, its homogeneity also increases (Andersen et al., 2023).

Various evaluation metrics have been applied to determine the effectiveness of ASAG models (Zesch et al., 2023). Metrics such as accuracy, fit, and correlation are commonly used in research (Burrows et al., 2015). Accuracy metrics use deeper measures such as F1 and precision to achieve more precise results, making them more suitable for evaluating natural language studies (Burrows et al., 2015). These metrics are often preferred for evaluating natural language studies (Hou et al., 2010). Correlation metrics, such as Pearson's r correlation measure, evaluate datasets with pointwise assessments and allow for proportional comparisons. This metric evaluates datasets with point-by-point assessments and allows for proportional comparisons (Burrows et al., 2015). For this reason, these metrics are generally used in the studies. It should also be noted that the metrics used are chosen depending on the approaches taken into focus.

More stable systems have been developed with AI components in the ASAG field. LSA was preferred because it allows for independent language use and accessible domain-specific corpora. Furthermore, similarity inference lets us evaluate without labeled datasets (Zhang et al., 2022). Recently, machine learning and deep learning approaches came to the fore (Galhardi & Brancher, 2018; Mardini et al., 2023). The studies we examined showed that the highest system performance was obtained with the LSTM deep learning model (Tulu et al., 2021). It has been observed that deep learning models give more accurate results on the same dataset (Ramesh & Sanampudi, 2023). Successful results make deep learning models preferred in current studies in NLP (Baburoglu et al., 2019). Unlike LSA, machine and deep learning methods use labeled data and factual/false statements. This function is an essential factor in obtaining more successful results. In traditional approaches, false statements are usually ignored.

Since it is based on deep learning, BERT models can be said to adapt much more strongly to different language features (Sayeed & Gupta, 2022) because effective results can be obtained with basic skills without going into detailed feature engineering processes (Lottridge et al., 2023). Sawatzki et al. (2021) achieved high performance in English and German using the BERT model. In the field of NLP, models such as GPT, BERT, and T5, referred to as large language models (LLM) in the literature, show that a critical threshold has been reached (Bozkurt & Sharma, 2023).

Conclusion

ASAG research, which started under the USA's guidance, is being explored in numerous states today. It is easily observed that research has risen in recent years. With the necessary transition to ODL practices throughout the pandemic phase, challenges experienced in assessment activities, and improvements in the field of NLP, ASAG has garnered attention in educational technologies. The interest in short-answer assignments and tests has gained momentum with the pandemic. Automatic assessment, where students may view their learning status quickly, has been seen as a chance to solve the limitation of engagement due to distance.

It is seen that English is dominant in the datasets used in ASAG applications, but recent studies have increased research on different languages. This is a significant development in expanding and accepting ASAG use in learning environments. Datasets containing 1000 or more replies are typically preferred in studies. This data size selection may be ideal for a practical ASAG application. Numerical scoring is

prominent for student assessment. However, scoring ranges are often limited. Developing such a mindset in research may be vital for the success of automated assessment performance.

Preferences for assessment approaches have varied over time. Semantic approaches have included more complicated replies in the automatic evaluation process within a semantic framework. In recent years, machine & deep learning approaches have dominated the area. Successful results obtained with this technique can be considered a justification for preference in approach selection. It was observed that classification, regression, and correlation metrics were distributed evenly in determining application performance. The automatic evaluation technique directly influences the choice of metric. The measurements used while working on publicly available data sets in different studies should be comparable. Thus, the degree of the produced application may be better understood. High performance has been reported in applications developed with the combination of LSTM and BERT. They have become quite popular in recent studies. BERT and other pre-learning models can be employed to generate an effective and efficient application. Finally, artificial intelligence technologies used in ASAG are generally based on traditional approaches. However, the potential of generative artificial intelligence has recently attracted attention in the field of ASAG. We predict that there will be a trend towards these technologies in ASAG applications. In summary, in this review, we have evaluated the scientific production, data set, and automatic evaluation elements of empirical ASAG research. We have comprehensively presented the field's current state with the features we have examined. We have stressed the practical and acceptable aspects of the system through improvements in the area. We have completed our study with suggestions for future research based on our experiences in the research process.

Looking Forward

ASAG stands out in the literature as a relatively new educational technology. The increasing number of studies in recent years has drawn researchers' attention to this field for its practical use. For researchers, using new technologies to improve the quality of education and training is an encouraging factor.

Today, applications of open and distance learning (ODL) are increasing. ASAG systems can be advantageous for conducting in-depth evaluations on large groups in natural language. In this context, studies can be carried out on developing ASAG systems that can be integrated into ODL applications.

It is seen that the majority of the studies focus on scoring or rating. In this context, more literature on ASAG applications must be available to provide descriptive and formative feedback. Immediate feedback is an essential component in online environments. The development of ASAG systems that can provide real-time assessments can enable students to analyze their performance quickly. This allows students to take early action to improve their performance. In ODL environments, generative artificial intelligence technologies such as ChatGPT can be included in ASAG systems developed with deep learning models such as LSTM and BERT. Thus, it will be possible to personalize formative feedback. Practical guidance with ChatGPT positively affects students' thinking processes (Cronje, 2023). These activities, where students test their knowledge, offer formative and descriptive feedback that facilitates permanent learning.

It was observed that English datasets were preferred in the majority of the studies analyzed. It is clear that research on different languages has yet to reach the desired maturity. Studies to be conducted in this focus are essential regarding the widespread use of the field, its acceptance in learning environments, and its practical use.

The subjects of the datasets were not examined in detail in this study. However, the preferred subjects are predominantly scientific. Future studies could focus on developing a single ASAG system for different subjects. Providing subject diversity with multiple datasets will be possible. This may increase

the usability of the developed system and contribute to a more accurate interpretation of its effectiveness.

The consideration of ASAG systems in different dimensions in educational environments may be necessary for their acceptance and widespread use. Feedback from educational administrators, teachers, and students can be highlighted. It can focus on how it can be used in education and training and improve student performance. The approach trend in recent studies is centered on deep learning. Deep learning is gaining prominence in the field with successful results and less technical skills required. In this context, it is important for researchers to focus more on this approach and investigate its effectiveness on many different datasets.

How ASAG systems work is essential. In particular, research can be conducted on the explainability of the assessment beyond the scoring of NLR responses. This makes the systems easier to understand. The accuracy of the assessments can be analyzed in detail. Such studies contribute to the acceptance of the systems.

Finally, of course, it is essential to carry out studies for ASAG systems to make evaluations more effective and accurate. The development of new approaches and techniques can pave the way for sharper evaluations. For example, it may be possible to evaluate different types of content and these contents in different dimensions.

References

Abdul Salam, M., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using an optimized deep learning model. *Plos one*, *17*(8), e0272269. https://doi.org/10.1371/journal.pone.0272269

Andersen, N., Zehner, F., & Goldhammer, F. (2023). Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning, 39*(3), 841-854. https://doi.org/10.1111/jcal.12717

Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied psychological measurement, 39(4),* 303–313. https://doi.org/10.1177/0146621614561630

Baburoglu, B., Tekerek, A., & Tekerek, M. (2019). Türkçe için derin öğrenme tabanlı doğal dil işleme modeli geliştirilmesi. In *13th International Computer and Instructional Technology Symposium* (pp. 671-679).

Badry, R. M., Ali, M., Rslan, E., & Kaseb, M. R. (2023). Automatic Arabic Grading System for Short Answer Questions. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3267407

Benli, I., & Ismailova, R. (2018). Use of open-ended questions in measurement and evaluation methods in distance education. *International Technology and Education Journal, 2*(1), 1-8. https://eric.ed.gov/?id=EJ1301381

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning.* https://doi.org/10.1111/jcal.12793

Bozkurt, A., Jung, I., Xiao, J., Vladimirschi, V., Schuwer, R., Egorov, G., ... & Paskevicius, M. (2020). A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education*, *15*(1), 1-126. https://doi.org/10.5281/zenodo.3878572

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, *18*(2), i-vii. https://doi.org/10.5281/zenodo.8174941

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, pp. 25,* 60–117. https://doi.org/10.1007/s40593-014-0026-8

Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essays and short answers. *Proceedings of the 5th International Computer Assisted Assessment Conference (CAA 01),* Loughborough University. https://hdl.handle.net/2134/1790

Corley, C. D., & Mihalcea, R. (2005, June). Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modelling of semantic equivalence and entailment* (pp. 13-18). https://aclanthology.org/W05-1203.pdf

Cronje, J. (2023). Exploring the Role of ChatGPT as a Peer Coach for Developing Research Proposals: Feedback Quality, Prompts, and Student Reflection. *Electronic Journal of e-Learning*. https://doi.org/10.34190/ejel.21.5.3114

Filighera, A., Ochs, S., Steuer, T., & Tregel, T. (2023). Cheating Automatic Short Answer Grading with the Adversarial Usage of Adjectives and Adverbs. *International Journal of Artificial Intelligence in Education, 1-31*. https://doi.org/10.1007/s40593-023-00361-2

Gabriel, F., Marrone, R., Van Sebille, Y., Kovanovic, V., & de Laat, M. (2022). Digital education strategies around the world: practices and policies. *Irish Educational Studies, 41(1),* 85-106. https://doi.org/10.1080/03323315.2021.2022513

Galhardi, L. B., & Brancher, J. D. (2018). Machine learning approach for automatic short answer grading: A systematic review. In *Advances in Artificial Intelligence-IBERAMIA 2018: 16th Ibero-American Conference on AI, Trujillo, Peru, November 13-16, 2018, Proceedings 16* (pp. 380-391). Springer International Publishing. https://doi.org/10.1007/978-3-030-03928-8_31

Garg, J., Papreja, J., Apurva, K., & Jain, G. (2022, June). Domain-specific hybrid bert based system for automatic short answer grading. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/CONIT55038.2022.9847754

Ghavidel, H. A., Zouaq, A., & Desmarais, M. C. (2020). Using BERT and XLNET for the Automatic Short Answer Grading Task. In *CSEDU (1)* (pp. 58-67). https://doi.org/10.5220/0009422400580067

Goksel, N., & Bozkurt, A. (2019). Artificial intelligence in education: Current insights and future perspectives. *In Handbook of Research on Learning in the Age of Transhumanism* (pp. 224-236). IGI Global. https://doi.org/10.4018/978-1-5225-8431-5.ch014

Gocmez, L. & Okur, M.R. (2023). Artificial intelligence applications in open and distance education: a systematic review of the articles (2007-2021). *Asian Journal of Distance Education, 18*(1), 1-32. https://doi.org/10.5281/zenodo.7514874

Gomaa, W. H., & Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications (IJACSA), 3*(11). https://doi.org/10.14569/IJACSA.2012.031119

Gomaa, W. H., Nagib, A. E., Saeed, M. M., Algarni, A., & Nabil, E. (2023). Empowering Short Answer Grading: Integrating Transformer-Based Embeddings and BI-LSTM Network. *Big Data and Cognitive Computing, 7*(3), 122. https://doi.org/10.3390/bdcc7030122

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., ... & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning, 39*(3), 767-786. https://doi.org/10.1111/jcal.12767

Hasanah, U., & Hartato, B. P. (2020, October). Assessing Short Answers in Indonesian Using Semantic Text Similarity Method and Dynamic Corpus. In *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 312-316). IEEE. https://doi.org/0.1109/ICITEE49829.2020.9271696

Hasanah, U., Permanasari, A. E., Kusumawardani, S. S., & Pribadi, F. S. (2016, August). A review of an information extraction technique approach for automatic short answer grading. In *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 192-196). IEEE. https://doi.org/10.1109/ICITISEE.2016.7803072

Hou, W. J., Tsao, J. H., Li, S. Y., & Chen, L. (2010). Automatic ssessment of students' free-text answers with support vector machines. In *Trends in Applied Intelligent Systems: 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems,*

IEA/AIE 2010, Cordoba, Spain, June 1-4, 2010, Proceedings, Part I 23 (pp. 235-243). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13022-9_24

Jadidinejad, A. H., & Mahmoudi, F. (2014). Unsupervised Short Answer Grading Using Spreading Activation over an Associative Network of Concepts/La notation sans surveillance des réponses courtes en utilisant la diffusion d'activation dans un réseau associatif de concepts. *Canadian Journal of Information and Library Science*, *38*(4), 287-303. https://doi.org/10.1353/ils.2014.0018

Jones, K. S. (1994). Natural language processing: a historical review. *Current issues in computational linguistics: in honor of Don Walker,* pp. 3–16. https://doi.org/10.1007/978-0-585-35958-8_1

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37,* pp. 389-405. https://doi.org/10.1023/A:1025779619903

Li, X., Li, X., Chen, S., Ma, S., & Xie, F. (2022). Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science*, *34*(1), 1638-1653. https://doi.org/10.1080/09540091.2022.2078279

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), e1-e34. https://doi.org/10.7326/0003-4819-151-4-200908180-00136

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), pp. 215-233. https://doi.org/10.1002/tea.21299

Lottridge, S., Woolf, S., Young, M., Jafari, A., & Ormerod, C. (2023). The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning, 39*(3), pp. 787-803.

Lubis, F. F., Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. *International Journal of Technology*, *12*(3), pp. 571-581. https://doi.org/10.14716/ijtech.v12i3.4651

Mardini G, I. D., Quintero M, C. G., Viloria N, C. A., Percybrooks B, W. S., Robles N, H. S., & Villalba R, K. (2023). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies,* pp. 1-26. https://doi.org/10.1007/s10639-023-11890-7

Mohler, M., & Mihalcea, R. (2009, March). 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)* (pp. 567-575). https://aclanthology.org/E09-1065.pdf

Nath, S., Parsaeifard, B., & Werlen, E. (2023, August). Automated Short Answer Grading using BERT on German datasets. Presentation at *the 20th biennial EARLI Conference (EARLI 2023),* Thessaloniki.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing Computer Resources to Automate Analysis of Students' Explanations of London Dispersion Forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 48, 238-243. http://www.jstor.org/stable/20371545

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery, 88,* 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Prakoso, D. W., Abdi, A., & Amrit, C. (2021). Short text similarity measurement methods: a review. *Soft Computing, 25*, 4699-4723. https://doi.org/10.1007/s00500-020-05479-2

Pribadi, F. S., Adji, T. B., & Permanasari, A. E. (2016, October). Automated short answer scoring using weighted cosine coefficient. In *2016 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)* (pp. 70-74). IEEE. https://doi.org/10.1109/IC3e.2016.8009042

Putnikovic, M., & Jovanovic, J. (2023). Embeddings for Automatic Short Answer Grading: A Scoping Review. *IEEE Transactions on Learning Technologies.* https://doi.org/10.1109/TLT.2023.3253071

Ramesh, D., & Sanampudi, S. K. (2023). Semantic and Linguistic Based Short Answer Scoring System. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(3), 246-251. https://www.ijisae.org/index.php/IJISAE/article/view/3164

Ratna, A. A. P., Santiar, L., Ibrahim, I., Purnamasari, P. D., Luhurkinanti, D. L., & Larasati, A. (2019, October). Latent semantic analysis and winnowing algorithm based automatic Japanese short essay answer grading system comparative performance. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST),* pp. 1-7. https://doi.org/0.1109/ICAwST.2019.8923226

Rodrigues, F., & Araújo, L. (2012, April). Automatic assessment of short free text answers. *In International Conference on Computer Supported Education*, 2, pp. 50-57. https://doi.org/10.5220/0003920800500057

Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021, July). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In *International Conference on Artificial Intelligence in Education Technology,* pp. 65-75. https://doi.org/10.1007/978-981-16-7527-0_5

Sayeed, M. A., & Gupta, D. (2022, December). Automate Descriptive Answer Grading using Reference based Models. In *2022 OITS International Conference on Information Technology (OCIT)*, pp. 262-267. https://doi.org/10.1109/OCIT56763.2022.00057

Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review, 21,* pp. 473-486. https://doi.org/10.1007/s12564-020-09640-2

Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic short answer grading with semspace sense vectors and malstm. *IEEE Access, 9*, 19270-19280. https://doi.org/10.1109/ACCESS.2021.3054346

Uto, M., & Uchida, Y. (2020). Automated short-answer grading using deep neural networks and item response theory. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21* (pp. 334-339). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_61

Uysal, I., & Doğan, N. (2021). Automated essay scoring effect on test equating errors in mixed-format test. *International Journal of Assessment Tools in Education*, *8*(2), 222-238. https://doi.org/10.21449/ijate.815961

Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information, 11*(9), 421. https://doi.org/10.3390/info11090421

Weegar, R., & Idestam-Almquist, P. (2023). Reducing Workload in Short Answer Grading Using Machine Learning. *International Journal of Artificial Intelligence in Education*, pp. 1-27. https://doi.org/10.1080/09540091.2022.2078279

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers and Education, 123,* pp. 212–224. https://doi.org/10.1016/j.compedu.2018.05.010

Wilianto, D., & Girsang, A. S. (2023). Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods. *TEM Journal, 12*(1). https://doi.org/10.18421/TEM121-37

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education, 45*(2), pp. 224-238. https://doi.org/10.1080/02602938.2019.1629390

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and psychological measurement*, *76*(2), pp. 280-303. https://doi.org/10.1177/0013164415590022

Zesch, T., Horbach, A., & Zehner, F. (2023). To Score or Not to Score: Factors Influencing Performance and Feasibility of Automatic Content Scoring of Text Responses. *Educational Measurement: Issues and Practice, 42*(1), pp. 44-58. https://doi.org/10.1111/emip.12544

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic short-answer grading model for semiopen-ended questions. *Interactive learning environments, 30*(1), pp. 177–190. https://doi.org/10.1080/10494820.2019.1648300

Zhang, Y., Lin, C., & Chi, M. (2020). Going deeper: Automatic short-answer grading by combining student and question models. *User modelling and user-adapted interaction, pp. 30,* pp. 51–80. https://doi.org/10.1007/s11257-019-09251-6

Zhu, X., Wu, H., & Zhang, L. (2022). Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. *IEEE Transactions on Learning Technologies*, *15*(3), pp. 364-375.

About the Author(s)

- Abdulkadir Kara (Corresponding author); abdulkadirkara@bayburt.edu.tr; Bayburt University, Turkey, ORCID ID: https://orcid.org/0000-0003-3255-1408
- Eda Saka Şimşek; edasaka@bayburt.edu.tr, Bayburt University, Turkey, ORCID ID: https://orcid.org/0000-0001-9210-5034
- Serkan Yıldırım; serkanyildirim@atauni.edu.tr, Atatürk University, Turkey, ORCID ID: https://orcid.org/0000-0002-8277-5963

Author's Contributions (CRediT)

Abdulkadir Kara: Conceptualization, Methodology, Formal Analysis, Visualization, Writing – original draft, Writing – review & editing; Eda Saka Şimşek: Methodology, Visualization, Data curation, Formal Analysis, Writing – review & editing; Serkan Yıldırım: Conceptualization, Methodology, Writing – review & editing.

Acknowledgements

We would like to express our gratitude to Atatürk University for their assistance in our investigation of automated assessment techniques for brief responses, as part of the Scientific Research Project.

Funding

Not applicable.

Ethics Statement

This research is not applicable for ethical review as it involves a review of existing studies.

Conflict of Interest

The authors do not declare any conflict of interest.

Data Availability Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Article History

Submitted: December 17, 2023 / Accepted: March 13, 2024

Suggested citation:
Kara, A., Saka Şimşek, E., &Yildirim, S. (2024). Unveiling the Landscape: Studies on Automated Short Answer Evaluation. *Asian Journal of Distance Education, 19*(1), 178-199.
https://doi.org/10.5281/zenodo.10829027

Appendix

Table 4. Articles included in the study group

ID	References	Doi
A1	[Srihari et al., 2006]	https://doi.org/10.1007/11669487_7
A2	[Srihari et al., 2008]	https://doi.org/10.1016/j.artint.2007.06.005
A3	[Siddiqi, Harrison & Siddiqi, 2010]	https://doi.org/10.1109/TLT.2010.4
A4	[Gomaa & Fahmy, 2012]	https://doi.org/10.14569/IJACSA.2012.031119
A5	[Saunders et al., 2014]	https://doi.org/10.1371/journal.pone.0093251
A6	[Zehner, Salzer & Goldhammer, 2016]	https://doi.org/10.1177/0013164415590022
A7	[Zimmerman et al., 2018]	https://doi.org/10.1080/10691898.2018.1443047
A8	[Horbach & Zesch, 2019]	https://doi.org/10.3389/feduc.2019.00028
A9	[Lavoie et al.,2020]	https://doi.org/10.1177/0013164419860575
A10	[Lubis et al., 2021]	https://doi.org/10.14716/ijtech.v12i3.4651
A11	[Somers, Samuel & Boles, 2021]	https://doi.org/10.14742/ajet.7121
A12	[Uysal & Dogan, 2021]	https://doi.org/10.21449/ijate.815961
A13	[Balaha & Saafan, 2021]	https://doi.org/10.1109/ACCESS.2021.3060940
A14	[Tulu, Ozkaya & Orhan, 2021]	https://doi.org/10.1109/ACCESS.2021.3054346
A15	[Li et al., 2022]	https://doi.org/10.1080/09540091.2022.2078279
A16	[Abdul Salam, El-Fatah & Hassan, 2022]	https://doi.org/10.1371/journal.pone.0272269
A17	[Gomaa et al., 2023]	https://doi.org/10.3390/bdcc7030122
A18	[Filighera et al., 2023]	https://doi.org/10.1007/s40593-023-00361-2
A19	[Mardini et al., 2023]	https://doi.org/10.1007/s10639-023-11890-7
A20	[Weegar & Idestam-Almquist, 2023]	https://doi.org/10.1007/s40593-022-00322-1
A21	[Wilianto & Girsang, 2023]	https://doi.org/10.18421/TEM121-37
A22	[Badry et al., 2023]	https://doi.org/10.1109/ACCESS.2023.3267407
A23	[Bernard et al., 2023]	https://doi.org/10.1002/ase.2305