

# Automatic Exam Correction System Involving XAI for Admission to Public Higher Education Institutions: Literature Review

Joaquim João Nsaku Ventura<sup>1,2</sup><sup>a</sup>, Cleyton Mário de Oliveira Rodrigues<sup>1</sup><sup>b</sup>  
and Ngombo Armando<sup>2</sup><sup>c</sup>

<sup>1</sup>Postgraduate Program in Computer Engineering Polytechnic School of Pernambuco, POLI/UPE, Recife, Brazil

<sup>2</sup>Department of Computer Engineering, Polytechnic Institute, Kimpa Vita University IP-UNIKIVI, Uíge, Angola

**Keywords:** Automatic Evaluation of Short Answers, Automated Evaluation, Natural Language Processing, Explainable Intelligence, Machine Learning.

**Abstract:** The process of correcting entrance exams is an essential procedure for assessing the academic performance of student candidates and ensuring fairness and accuracy in the awarding of marks for their future selection. Most lecturers at Angolan higher education institutions carry out the corrections manually, especially subjective corrections. Due to the high number of students, ensuring a high-quality correction process while meeting institutional deadlines becomes challenging. In this context, this article aims to find the techniques and metrics that are used for the automated correction process of assessments with discursive questions, involving Explainable Artificial Intelligence (XAI). This literature review follows the PRISMA 2020 methodology and includes studies from three bibliographic databases: ACM Digital Library, IEEEExplore and Science Direct. The results obtained show that the use of a combination of similarity measures and Natural Language Processing (NLP) provides greater efficiency for the automated correction of discursive questions.

## 1 INTRODUCTION


### 1.1 Generality


In Angolan Public Higher Education Institutions (IPES), the entrance exam is mandatory for undergraduate student candidates (Trujillo, 2024). Access to IPES courses is limited by the number of vacancies available each year (Angola, 2019). An admission process to IPES includes a set of steps and criteria used to select students who will be admitted to undergraduate courses offered by these institutions (Dez, 2022). The admission process consists of registration, preparation, an entrance exam, candidate selection, and enrollment.


Overall, in this process, many assessment criteria are available to candidates, such as reproduction of content taught in previous classes and general knowledge questions. In this approach, the aim is to also carry out the assessment using subjective criteria

(failure and success) (Botega et al., 2005). Assessments can be presented in a discursive or objective manner, with the former presenting a unique challenge compared to the latter. In fact, in the latter case, the subjectivity inherent in the answers requires a more thorough and comprehensive correction process, which aims to ensure impartiality, reliability and fairness in the assessment of students' knowledge.

The purpose of the entrance exam for the IPES network in Angola is to assess the knowledge and skills of candidates to ensure that those who enter the institution have the necessary foundations to have the necessary foundation to succeed in their courses and graduate. On the other hand, this exam serves to select the best prepared students, based on academic criteria established by the institution, thus ensuring that candidates have the level of knowledge necessary to meet the requirements of the course. In short, such a process aims to select the best candidates, standardize

<sup>a</sup> <https://orcid.org/0000-0001-8328-9724>

<sup>b</sup> <https://orcid.org/0000-0003-3816-656X>

<sup>c</sup> <https://orcid.org/0000-0001-7493-4365>

the admission process, and guarantee the quality of downstream education (Netvistas, 2024).

In education, evaluation is a broad area that results from a systematic effort to define criteria, based on which precise information is collected to judge the value of each alternative presented. To evaluate is, therefore, to issue a value judgment on a characteristic in focus, and this value may be based, partially but not exclusively, on quantitative data (Vianna, 2014).

Luckesi (2022) defines that exams or tests have the purpose, in school learning, of verifying the student's level of performance in a given content and classifying it in terms of approval or failure. According to Luckesi (2011) evaluating involves analyzing the current context through investigation, research, and diagnosis, with the aim of proposing solutions. In contrast, examining is a punctual and selective process focused on the final product and the past, which adopts a classificatory and exclusive approach—labeling it as “Approved” or “Rejected.”

Automated grading of essay questions and automated grading refer to the use of computational tools to assist in the analysis and grading of written responses by students (Weegar & Idestam-Almquist, 2024). This area of research has gained increasing attention in recent years due to its potential to streamline the grading process, reduce teacher workload, and provide faster and more personalized feedback to interested parties.

In the field of automatic evaluation of short discursive questions, there are two main lines of research, namely: one corpus and similarity between texts and another using similarity metrics between networks of concepts extracted from the texts of the answers using Machine Learning (ML) and NLP techniques (Sirotheau et al., 2019).

Automated exam correction tools, known as ASAG (Automated Short Answer Grading), are systems based on NLP and Artificial Intelligence (AI) designed to automatically correct short answers in exams (Tornqvist et al., 2023).

## 1.2 Main Contributions

Knowing that assessment tools and techniques, as well as automated exam correction, is a widely explored topic in the English language, and is currently gaining considerable visibility in Portuguese, it reduces human effort applied in the correction of a large number of assessments in a short space of time (Lima et al., 2023). A key contribution is the elimination of human bias, as automated tools ensure consistent and impartial grading, reducing

errors and favoritism. Furthermore, this review will allow us to identify the main techniques and metrics used in the process of automating exam correction.

## 1.3 Manuscript Organization

The sequence of this article is structured as follows: section II presents the method of how the research was carried out using the PRISMA criteria and checklist, highlighting the study design and ethical aspects, the research questions, the search and eligibility criteria, the selection of articles and the exclusion criteria in the databases used to constitute the research.

In section III, the results obtained in the reviewed literature are presented and compared with other works used as reference based on the seven research questions.

In Section IV – Discussion, we explore this section by explaining what the results represent and answering the questions. The last section, Conclusion, is where we describe the main considerations of the review.

## 2 METHOD

### 2.1 Study Design and Ethical Aspects

The main objective of this study is to review the literature based on the PRISMA guidelines (Page et al., 2023). This review was carried out by three researchers from May to July 2024. It is limited exclusively to primary and public sources and was not submitted to an ethics committee.

#### 2.1.1 Research Question

The automatic assessment of the correction of short questions in Portuguese is related to several issues such as spelling correction and automatic punctuation of answers. The following research question was then developed:

How can Angolan IPES streamline the process of correcting entrance exams with discursive questions, ensuring transparency, security and economically viable costs, through modern computing techniques?

#### 2.1.2 Eligibility Criteria

To conduct this research, eligibility criteria were established regarding the research question, which was divided into several sub-questions, namely:

- Q1. What techniques and metrics are used to perform automated assessment correction?

- Q2. What are the explainability techniques and metrics used to automatically correct discursive questions?
- Q3. What are the techniques and metrics used to explain the evaluations made by an automated system?
- Q4. What tools are available for the automatic correction process of discursive or essay questions?
- Q5. What are the opportunities in this area, in terms of research, engineering and business?
- Q6. What are the challenges described in the study?
- Q7. What are the most used databases for the study?

## 2.2 Search Strategy

As shown in Table 1, the PICOC strategy (acronym for P: population/patients; I: intervention; C: comparison/control; O: outcome; C: context) was used to help determine what the research question should specify (Santos & Galvão, 2014).

To develop this literature review, the following key terms were used;

“Questões discursivas”, “correção automatizada”, “correção de exames”, “Inteligência explicável”, “Pontuação automatizada respostas curtas”, “Processamento de Linguagem Natural”. Since the search was done in English, we translated the key terms in question, completing the search with Boolean operators “AND” and “OR” it would be “Discursive questions”, “Essay Scoring”, “Automated Grading”, “Automated Essay Scoring”, “Artificial Intelligence Explainability”, “Artificial Intelligence Explainability” and “AI- based essay grading”, “Natural Language Processing”. Searches were performed in the Science Direct, IEEE (Institute Electric Electronic Engineer), ACM Digital Library, Redalyc and Web of Science.

Some databases were excluded for several reasons, i) Redalyc returns 5 articles, but its interface does not have the resources to export the results obtained. ii) given the ergonomic problems of the interface of the portal <http://www.isiknowledge.com> and redirects you to <https://www.webofscience.com/wos/author/search>, we decided to discard this database.

Additionally, we included studies that met different search criteria but were fundamental to understanding automated exam correction using Explainable Artificial Intelligence (XAI).

We conducted pilot searches to iteratively refine the search term string. Keywords whose inclusion did not return additional articles in the automated searches were excluded (see Table 1).

Table 1: Search strategies for the research.

Population	Teachers, Managers of Public Higher Education Institutions, Candidates for admission to public institutions of higher education
Intervention	Essay Scoring, Automated Grading, Automated Essay Scoring, AI- based essay grading, machine learning in text grading, linguistics analysis, text analysis, Artificial Intelligence Explainability
Comparison	It was not applied
Outcome	machine learning in text grading, natural language processing assessment
Context	in the Higher Education environment
Research Base	Search String
Science Direct	("AI- based essay grading " OR " Automated Essay Automated " OR " Scoring Grading " OR " Essay Scoring "OR" machine learning in text grading ") AND ("Artificial Intelligence Linguistic "OR" explainability analysis "OR" natural language processing " OR " text analysis ")  (((Automatic Short Answer Grading) OR (automatically assessing short) OR (C-rater)) AND ((machine learning)))
ACM Digital	
IEEEExplore	

## 2.3 Inclusion and Exclusion Process

In this inclusion and exclusion stage, the criteria mentioned in Table 2 were considered.

After identifying the literature in the databases, the manual selection stage followed, with the reading of the titles and abstracts of the studies returned from the search stage. The objective here was to evaluate the articles in general terms regarding the importance of their application for the mapping performed. The articles that met the inclusion criteria and the articles that did not present sufficient information for exclusion went on to the next stage of the selection process. In this stage, the authors read the introduction and final considerations of the articles, with the aim of including or excluding the articles based on the selection criteria, see Table 2.

Table 2: Inclusion and Exclusion Criteria.

#	TYPE	DESCRIPTION
1.	Inclusion	Studies dealing with automated correction of exams or short-answer questions, which were published between 2022 and 2024 and published in English, Spanish or Portuguese.
2.	Inclusion	Studies that focus on NLP to correct exams or essays.
3.	Exclusion	Duplicate studies, same studies that were published on different databases.
4.	Exclusion	Secondary or tertiary studies. They include literature reviews or knowledge maps, for example.
	Exclusion	Grey literature. Articles that are unavailable in the sources defined in this research, or require payment to obtain access;
5.	Exclusion	Out-of-scope literature. All articles that do not fit the pre-defined criteria in the automated exam correction process;
6.	Exclusion	Short Papers. Studies with less than 4 pages in total

After reading the full articles, the following data were extracted:

Table 3: Data Extraction from Articles.

#	TYPE	DESCRIPTION
1.	Authors	Authors of the article
2.	Publication Date	Article publication date
3.	Country of Publication	Country of publication of the article
4.	Correction Type	Type of correction (Essay or Discursive question)
5.	Methods Used	Method used to provide solution
6.	ML Techniques (Q1)	Machine Learning techniques applied in the study
7.	XAI Techniques (Q2)	Explainable intelligence techniques
8.	XAI Metrics (Q2)	Explainable Intelligence Metrics
9.	ML Metrics (Q1)	Metrics to evaluate ML or NLP models
10.	Available Database (Q7)	Database Information Used
11.	Link	Full article link

### 3 RESULTS

To assist in searching for studies in the databases and pre-filtering, the Parsifal tool was used, designed so that researchers can collaboratively build systematic reviews or mappings (Silva et al., 2024). All search strings were executed in this application, which returned 2396. Given certain limitations, such as articles not available in the database and limited access, it was necessary to execute the same search strategy individually in each selected database.

The Conducting stage is the phase in which the bases, selection and screening of articles are defined. Thanks to this tool, it was possible to detect duplicate articles. It helped the authors in the selection process in several stages, such as:

- Removal of duplicate articles, as the tool detects potentially duplicate articles and then it is possible to confirm or rectify this information manually;
- Identification of selection criteria by reading abstracts and titles;
- Application of selection criteria by reading the Introduction and Final Considerations.

In the first phase, the tool detected 277 articles found in the three databases, distributed as follows: ACM Digital Library (n = 127), IEEEExplore (n = 36) and Science Direct (n = 114). Among these results, 36 were duplicate studies, after confirmation by the authors, they were selected and removed, moving on to the next phase. In the second phase, after reading the titles and abstracts, the inclusion and exclusion criteria were applied to select the articles for the subsequent phase, to all articles that focus on the area of NLP and studies that talk about correction of automated exams in Portuguese, English and Spanish. In the last phase, the authors searched for the articles through the title, DOI or name of the authors using other tools such as Publish or Perish is a free desktop software that extracts data from Google Scholar, Scopus, PubMed, Web of Science and other databases to help authors analyze various statistics about research impact, including total number of citations, average number of citations per article, average number of citations per year, H-index and related parameters, and an analysis of the number of authors per article (ABCD-USP, 2016), for those with incomplete metadata.

Research carried out using other methods was applied using the IEEE database (n = 34) as it presented better results, since many databases presented results outside of CE5 and complementary studies through Google Scholar (n = 4).

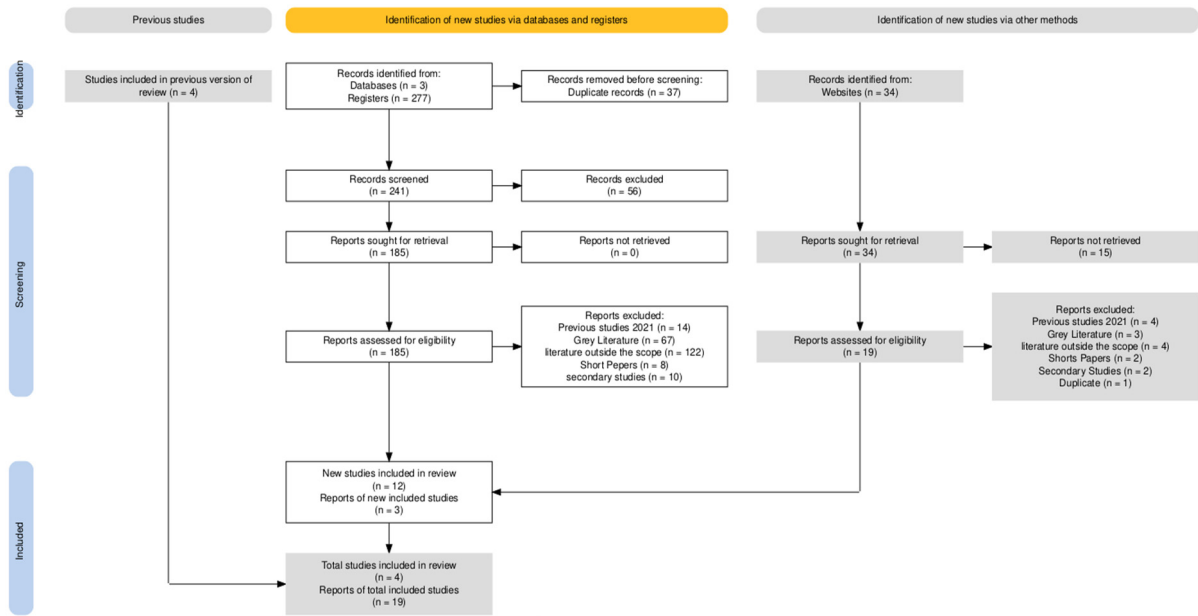


Figure 1: Figure 4: Study selection flowchart according to PRISMA criteria.

For summary purposes, the flowchart in Figure 1 shows details of the results obtained in the three phases of the PRISMA recommendations.

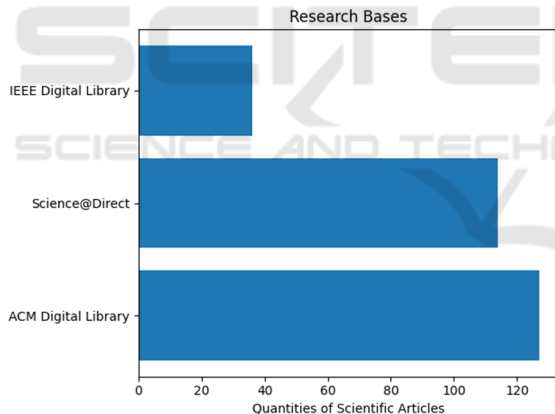


Figure 2: Research source. Source: The Authors (2024).

### 3.1 Study Selection

The selected studies were published between 2020 and 2024, and the country of publication included China, Croatia, Egypt, United States of America, India, Indonesia and Thailand.

In total, 19 articles were included, as the research's main focus is to identify studies related to tools that perform automated correction of exams with discursive questions using Machine Learning (ML) or, NLP involving XAI, so we sought to understand in more detail how it is done, what are the limitations

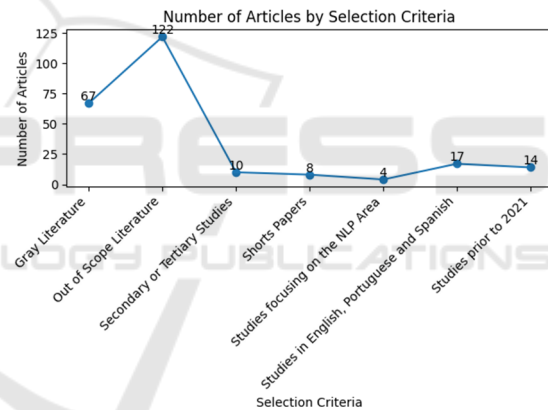


Figure 3: Selection Criteria. Source: The Authors (2024).

the techniques and metrics that are used to evaluate the models, and what are the challenges. In which 4 previous studies and 3 studies identified by other methods are included in the final analysis.

### 3.2 Key Techniques and Metrics (Q1)

Deep Learning (DL) techniques are the algorithms and methods used to build models. This question does not simply analyze the ML and DL techniques used, but also the XAI techniques. The selected studies, the techniques are directly linked to the type of approach, it depends a lot on the type of correction, whether it is discursive or dissertative questions. The authors Tambe and Kulkarni (2022) use Embeddings, Tokenization, Encoder and Attention to perform



Table 4: Selected Works with Databases.

ID	AUTHOR	DATABASE
A1	Also and Kulkarni (2022)	ASAP
A2	Nugroho et al. (2022)	ASAP and CELA1
A3	Song et al. (2024)	Mounted
A4	Suriyasat et al. (2023)	DifferSheet
A5	Petricioli et al. (2023)	Kaggle
A6	Noiyoo and Thutkawornpin (2023)	Mounted
A7	Chamidah et al. (2023)	Kaggle
A8	Ruseti et al. (2024)	Mounted
A9	Meccawy (2023)	Kaggle
A10	Saeed and Gomaa (2022)	Texas Dataset
A11	Badry et al. (2023)	AR-ASAG
A12	Chakraborty and Mishra (2023)	Kaggle
A13	Weegar and Idestam-Almquist (2024)	Dataverse
A14	Wilianto and Girsang (2023)	Mounted
A15	Zhang et al. (2022)	Kaggle
A16	Galhardi et al. (2020)	Kaggle
A17	Oliveira et al. (2020)	Mounted
A18	Abdalkareem; and Min-Allah (2024)	Mounted
A19	Silva (2023)	Kaggle

Table 5: Most commonly used techniques.

ML TECHNIQUES	QTY	%
Word- Embedding	9	47%
Tokenization	5	26%
Cosine Distance	3	16%
Stop Word	2	11%
TF-IDF	2	11%
Semantic-based similarity	2	11%
Morphological and lexical analysis, Attention, CBOW, Cosine Similarity, CWV, Encoder, Ensemble, Information Extraction, FastText, Lemmatization, Data Mining, Ngrams, Normalization, Parsing, Pattern recognition, embedding-based similarity, String - based similarity, Lexical Similarity, Skip-Gram, Word Embeddings Similarity, Word2Vec, LI, LIN, WPATH and JCN	1	5%

automated scoring, while Nugroho et al. (2022) String- based Similarity, Semantic-based Similarity, in addition to Embedding to perform essay correction.

Suriyasat et al. (2023) kNN, SVM, Random Forest, Gradient models. Boosting, XGBoost, LSTM, LSTM and LSTM + CNN; while Petricioli et al.

(2023) perform clustering for short answer classification using Machine Learning K- Means and Cosine Distance algorithms; Noiyoo and Thutkawornpin (2023) use transformers and neural networks to score Thai essays using LSTM, CNN, BERT techniques; and Chakraborty and Mishra (2023) Deep Learning -based technique for evaluating text-based answers or for short answer classification. Researchers Wilianto and Girsang (2023) use Cosine Similarity in his article entitled "Automatic classification of short answers in high school using semantic similarity methods". As you can see in Table 6.

Table 6: Most used models.

MODEL	QTY	%
BERT	12	63%
Embedding	6	32%
Long Short- Term Memory (LSTM)	4	21%
Random Forest (RF)	4	21%
Cosine Distance	3	16%
Decision Tree (DT)	3	16%
K- Means	3	16%
Logistic Regression (LR), RoBERTa, Support Vector Machine (SVM)	3	16%
Convolutional neural network (CNN)	2	11%
eXtreme Gradient Boosting (XGBoost)	2	11%
GloVe, Naive Bayes (NB), Cosene Similarity, Transformer, WangchanBERTa, Word2Vec	2	11%
ANN, CBOW, CNB, Information Extraction, GB, GBM, kNN, LSA, MiniLM-L6, GMM, MPNET, Siamese	1	5%
Manhattan LSTM, Soft-6, Stacking, Word Mover's Distance, XGB, MaLSTM, SBERT, WordNet, SVD		

One of the main objectives regarding the techniques was to determine the main, most used and efficient techniques for automated exam correction. Deep Learning models like GPT and BERT are widely used to assess essay quality by evaluating semantic coherence and argumentation, assigning scores based on trained examples.

To evaluate models for automatic correction of short-answer exams and automated scoring of essays, the metrics used are like those used to evaluate other text classification and prediction models, in addition to specific metrics for comparison with human evaluations.

<sup>1</sup> CELA - Chinese EFL Learners' Argumentation

Table 7: Metrics for evaluating the Models.

METRICS	TOTAL	%
QWK	8	42.11%
Accuracy	7	36.84%
Cohen's Kappa	5	26.32%
F1-Score	5	26.32%
Precision	5	26.32%
Pearson Correlation	4	21.05%
Recall	4	21.05%
Confusion Matrix	3	15.79%
Quadratic Kappa Scores	3	15.79%
RSME	3	15.79%
MOTHER	2	10.53%
AUC	1	5.26%
Cluster purity	1	5.26%
NMAE	1	5.26%
Normalized Mutual Information	1	5.26%
R2	1	5.26%
Silhouette Scores	1	5.26%
The Rand Index	1	5.26%
MSE	1	5.26%

The QWK (Quadratic Weighted Kappa) is a statistical metric widely used to measure the degree of agreement between two ordinal classifications, such as in essay scoring systems or automated assessments, representing a total of 42.11% of the selected studies use it (Noiyoo & Thutkawornpin, 2023; Nugroho et al., 2022; Song et al., 2024; Suriyasat et al., 2023; Tambe & Kulkarni, 2022). Followed by accuracy, which measures the percentage of correct predictions made by the model compared to the total predictions. And one of the metrics that is used was Cohen's Kappa: it measures the level of agreement between two evaluators (in this case, the automated model and humans), considering the possibility of hits by chance. Root Mean Square Error (RMSE), measures the average error between the model's scores and the human scores. The lower it is, the better the performance (Badry et al., 2023; Meccawy, 2023). Confusion Matrix: Tool to visualize the model's performance, showing the hits and errors in each class (Song et al., 2024).

### 3.3 Key XAI Techniques and Metrics (Q2 and Q3)

One of the fundamental objectives of this study is to identify XAI techniques and metrics for the automated grading process of exams with discursive questions. For the automated grading of discursive

questions, where the goal is to evaluate short texts or essays, XAI can play a crucial role, especially in providing transparency in the grading process. Explainability makes the system more reliable in justifying how it arrived at a given grade or evaluation.

The authors Abdalkareem; and Min-Allah (2024) conduct a study with the main objective of creating multiple Machine Learning predictive models and selecting the most effective one to predict the academic paths of students in Saudi secondary schools. This study extends the research by applying XAI to interpret the higher model to better understand and increase the transparency of the predictive model, therefore, the prediction model was easier to understand and more interpretable when the SHAP value technique was applied. This was a selected work that uses XAI to explain the model.

### 3.4 Automated Correction Tools for Essay Questions (Q4)

It can be said that yes, there are automated correction tools, which use artificial intelligence (AI) NLP algorithms to analyse the grammar, coherence, content and relevance of the answers and assign an automatic score. The authors Ruseti et al. (2024) "ReaderBench" platform that was designed primarily for professionals and researchers who are not experts in machine learning. ReaderBench supports languages such as English, French, German, Romanian, Portuguese, Italian, Dutch and Russian. Tobler Tobler (2024) presents GenAI Smart Grading a tool allows generative AI models to evaluate the answers provided (by students) to their questions.

### 3.5 Study Opportunities (Q5)

The research area in automated correction tools for discursive exams offers several interesting and innovative opportunities, especially with advances in AI, ML and NLP. Concerning this issue. The opportunities are related to: (i) Development of more accurate correction algorithms, which will be able to understand the semantics, algorithms that can understand the meaning of the text; (ii) Multidimensional correction of essays, a significant opportunity is to develop systems that can evaluate originality, creativity and argumentation in discursive texts; (iii) Application in standardized exams and large-scale assessments, the development of algorithms that ensure fairer and more impartial correction, minimizing human bias and ensuring uniformity in assessments in different contexts; (iv)

Correction of tests in innovative contexts, creating tools that not only evaluate, but also help students practice their writing skills, offering suggestions, revisions and even examples of essays.

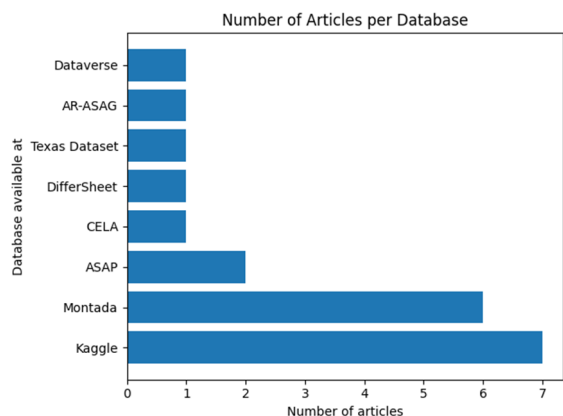


Figure 4: Database. Source: The Authors (2024).

### 3.6 Challenges (Q6)

In Q6, we sought to highlight some challenges that researchers describe. Computational Resources, automated correction, especially when involving deep learning models, can require large computational power, Song et al. (2024) encounter this limitation when using LLMs and DL models. Generalization Problems, a model trained on a specific set of data may not generalize well to new data. Specialized Subject Domain, automated correction tools may have difficulties when evaluating answers from disciplines that involve complex and specialized concepts, such as exact sciences or humanities.

Lack of Representative Data, the development of automated correction tools depends on large amounts of data for training, usually in the form of essays or answers already evaluated by humans.

Wilianto and Girsang (2023) use semantic similarity to classify short answer scores, because in the absence of contextualization, in some cases, the system may not understand the context or reasoning behind an answer, leading to unfair grading. So, these challenges highlight the complexity of developing effective and fair tools for grading discursive exams, but they also open up opportunities for research seeking innovative solutions.

### 3.7 Database (Q7)

The lack of quality data can compromise the model's performance, regarding Q7, many authors do not use

private or assembled databases and some are available in research centers and Kaggle and ASAP. The authors Suriyasat et al. (2023) use a set of 31,175 essays written by fourth to sixth grade students and provided by DifferSheet, in addition Song et al. (2024) use data collected by a writing campaign called Ability Assessment of Expressive Ability, the assessment was applied to 2,870 Chinese primary school students in the 3rd grade to evaluate the performance of their proposed models.

Ruseti et al. (2024) in their research, they evaluate the platform on three publicly available "ASAP" datasets: two focused on quality reflected in scoring tasks in English and Portuguese, and one focused on validity, considering a fake news detection task in French. Yang (2023) makes a comparison of several algorithms such as Linear Regression, Ridge Regression, Gradient Boosting Regression, Random Forest Regression And Xgboost Regression, applied to a total of 3000 essays, with parameter adjustments using the Xgboost algorithm Regression showed better results.

Chakraborty and Mishra (2023); Nugroho et al. (2022); Sethi and Singh (2022); Tambe and Kulkarni (2022) use an ASAP dataset available on Kaggle to carry out their respective studies.

## 4 DISCUSSIONS

The use of LSA in automated grading of short answer questions is a natural language processing-based approach that aims to evaluate textual responses in a more intelligent and flexible way, and is able to identify the correct meaning of words with multiple meanings based on the context, which is essential for assessing the understanding of the answers of candidates or students, the authors Badry et al. (2023) determine the similarity of the student's answer with that of the model (teacher's reference answer), aiming to build an automatic classification model of Arabic short answers using semantic similarity approaches.

The use of similarity approaches semantics in the correction automated question answering short allows bigger precision and flexibility in comparison with methods traditional, providing one assessment more fair and scalable in several scenarios educational, many studies selected present best results, such as Wilianto and Girsang (2023) use three models pre-trained ROBERTA, MPNET and MiniLM-L6, to classify automatically answers short put through similarity techniques semantics. Also authors Saeed and Gomaa (2022) propose a scoring system automatic for responses short, which



calculates an efficient and quick grade for the student's response based on different text similarity techniques.

The cosine similarity measure is often used in NLP and ML to compare the similarity between two vectors. The Meccawy (2023) uses the cosine similarity measure to provide scores ideals of student responses, he he does one comparison of three approaches, specifically you WordNet, Word2vec and BERT models, and arrives at a conclusion that, the models advanced contextual embedding features provide best results. Nugroho et al. (2022) use semantic textual similarity, exploiting the cosine similarity method, to make correction of writing. This approach offers one manner automated compare student responses with a ideal reference, facilitating the correction and assignment of grades in an objective and consistent manner, and in your majority use Embeddings, TF-IDF and Bag of Words (BoW) techniques.

Despite the presence of ML in the application of automated exam correction, NLP presents excellent results using Transformer-Based Techniques, as can be seen in Table 6, Bidirectional Encoder Representations from Transformers (BERT) represents 63% of the selected studies use this model, as authors Sethi and Singh (2022) do this exploration.

In semantic similarity assessment, the goal is to measure how similar the meaning of two texts is, regardless of the exact words used. One of the main metrics used is cosine similarity. According to the selected articles, QWK ranks first, representing a total of 42.11%, accuracy, Cohen's Kappa, Precision, Recall, F1-Score, confusion matrix, and others, as we can see in Table 7

The central idea of XAI is that the AI model not only provides the answer or score but also explains in an understandable way how and why it reached that conclusion. Of the selected articles, 94.73% do not use any technique to explain the results obtained. Using it in automated correction would offer a powerful and transparent tool that could enhance learning and improve the quality and acceptance of automatic assessment systems.

The lack of a database with student responses and teacher scores is one of the barriers faced in many languages, not only in Portuguese, but also in Indonesian, Thai, Arabic, Hindi, Croatian, and others. Many authors translate from English into their own language to train the models. For this reason, many researchers create their own databases to meet this need.

Automated short answer scoring is not only a technical issue but also involves pedagogical and

cultural considerations that vary from language to language. Our dissertation will investigate how ASAG models trained in English (or with translated data) may not capture specific nuances of the language and culture of Portuguese or other languages. This may include variations in writing style, different forms of argumentation, and pedagogical expectations. However, we can make a significant contribution with our study by exploring one or several of these gaps in the literature, focusing on the lack of teacher-scored databases and the specific challenges faced by less represented languages, such as Portuguese.

## 5 CONCLUSIONS

This work presents the state of the art of automated correction of discursive questions, where different types of techniques and metrics were presented to provide solutions to the research area. During its development, it was found that the use of NLP presents satisfactory results compared to other ML techniques, in particular transformers.

XAI is not yet a common practice in automated exam grading. Classical literature on ASAG generally focuses on the accuracy of predictive models, with little emphasis on explainability. The transition to XAI in ASAG is a recent field, with many challenges to be faced, such as creating useful and accessible explanations for teachers and students.

As for metrics, most of the selected works use Pearson Correlation, QWK, RMSE, Accuracy, F1-Score, Recall and Precision to calculate the performance of the models, to ensure the effectiveness of the models compared to others.

Many studies identified in this research show that most use proprietary databases, which in a certain way greatly hinder the progress of this field. Most of the available databases are in English, making other languages such as Thai, Mandarin, Indonesian, Arabic, Polish, Portuguese, Japanese and others difficult to understand.

## REFERENCES

- ABCD-USP, A. d. B. e. C. D. d. U. d. S. P. (2016, 09/12/2016). Publish or Perish. ABCD-USP. <https://www.abcd.usp.br/apoio-pesquisador/indicadore-s-pesquisa/publish-or-perish/>
- Abdalkareem, M., & Min-Allah, N. (2024). Explainable Models for Predicting Academic Pathways for High

- School Students in Saudi Arabia. IEEE Access. <https://ieeexplore.ieee.org/abstract/document/10444523/>
- Regulamento Geral de Acesso ao Ensino Superior; Decreto Presidencial nº 5/19, de 8 de Janeiro, DP nº 59/2019 31 (2019).
- Badry, R. M., Ali, M., Rslan, E., & Kaseb, M. R. (2023). Automatic Arabic Grading System for Short Answer Questions. IEEE Access, 11, 39457-39465. <https://doi.org/10.1109/ACCESS.2023.3267407>
- Botega, M. B., Franco, J. L., & Lemes, S. d. S. (2005). Desenvolvimento de Sistema para Avaliação On-Line com Tecnologia Java/Servlets. <https://doi.org/10.25061/2527-2675/rebram/2006.V9i2.277>
- Chakraborty, U. K., & Mishra, A. (2023, 29-30 July 2023). Automatic Short Answer Grading Using a LSTM Based Approach. 2023 IEEE World Conference on Applied Intelligence and Computing (AIC),
- Dez, A. N. (2022). O que é Ingresso universitário. Aula Nota Dez. Retrieved 21/03/2024 from <https://aulanotadez.com.br/glossario/o-que-e-ingresso-universitario/>
- Lima, T. d., Silva, I. d., & ... (2023). Avaliação automática de redação: Uma revisão sistemática. Revista Brasileira de .... <https://journals-sol.sbc.org.br/index.php/rbie/article/view/2869>
- Luckesi, C. C. (2011). Avaliação Da Aprendizagem Escolar: Estudos e proposições (Cortez, Ed.). Cortez.
- Luckesi, C. C. (2022). Avaliação da aprendizagem escolar (19 ed.). Cortez Editora.
- Meccawy, M. (2023). Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques. International Journal of Advanced Computer Science and Applications, 14(6), 768-775. <https://doi.org/10.14569/ijacsa.2023.0140682>
- Netvistas. (2024). O que é: Prova de admissão na instituição de ensino. Netvistas Soluções em Turismo Ltda. <https://netvistas.com.br/glossario/o-que-e-prova-de-admissao-na-instituicao-de-ensino/>
- Noiyoo, N., & Thutkawkornpin, J. (2023, 28 June-1 July 2023). A Comparison of Machine Learning and Neural Network Algorithms for An Automated Thai Essay Quality Checking. 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE),
- Nugroho, A. H., Hidayah, I., & Kusumawardani, S. S. (2022, 8-9 Dec. 2022). Transformer Model Fine-Tuning for Indonesian Automated Essay Scoring with Semantic Textual Similarity. 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),
- Page, M., McKenzie, J., Bossuyt, P., & ... (2023). A declaração PRISMA 2020: diretriz atualizada para relatar revisões sistemáticas. SciELO Public Health. <https://www.scielo.org/article/rpsp/2022.v46/e112/pt/>
- Petricioli, L., Skračić, K., Petrović, J., & Pale, P. (2023, 22-26 May 2023). Exploring Pre-scoring Clustering for Short Answer Grading. 2023 46th MIPRO ICT and Electronics Convention (MIPRO),
- Ruseti, S., Paraschiv, I., Dascalu, M., & ... (2024). Automated Pipeline for Multi-lingual Automated Essay Scoring with ReaderBench. International Journal of .... <https://doi.org/10.1007/s40593-024-00402-4>
- Saeed, M. M., & Gomaa, W. H. (2022, 8-9 May 2022). An Ensemble-Based Model to Improve the Accuracy of Automatic Short Answer Grading. 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC),
- Sethi, A., & Singh, K. (2022, 29-31 March 2022). Natural Language Processing based Automated Essay Scoring with Parameter-Efficient Transformer Approach. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC),
- Silva, T. d., Vidotto, K., Tarouco, L., & ... (2024). Inteligência artificial generativa no ensino de programação: um mapeamento sistemático da literatura. ... Novas Tecnologias na .... <https://seer.ufrgs.br/renote/article/view/141553>
- Sirotheau, S., Santos, J., Favero, E., & ... (2019). Avaliação Automática de respostas discursivas curtas baseado em três dimensões linguísticas. ... on Computers in .... <http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/8888>
- Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models. IEEE Transactions on .... <https://ieeexplore.ieee.org/abstract/document/10520824/>
- Suriyasat, S., Chanyachatchawan, S., & ... (2023). A Comparison of Machine Learning and Neural Network Algorithms for an Automated Thai Essay Scoring. ... Joint Conference on .... <https://ieeexplore.ieee.org/abstract/document/10201964/>
- Tambe, A. A., & Kulkarni, M. (2022, 10-11 Dec. 2022). Automated Essay Scoring System with Grammar Score Analysis. 2022 Smart Technologies, Communication and Robotics (STCR),
- Tobler, S. (2024). Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. Elsevier. <https://www.science-direct.com/science/article/pii/S2215016123005277>
- Tomqvist, M., Mahamud, M., Guzman, E., & ... (2023). ExASAG: Explainable framework for automatic short answer grading. Proceedings of the .... <https://aclanthology.org/2023.bea-1.29/>
- Trujillo, D. S. (2024). Mineração de dados aplicada ao Exame Nacional de Desempenho dos Estudantes (ENADE). bdm.unb.br. <https://bdm.unb.br/handle/10483/38462>
- Vianna, H. (2014). A prática da avaliação educacional: algumas colocações metodológicas. educa.fcc.org.br. <http://educa.fcc.org.br/pdf/eae/v25n60/1984-932X-eae-25-60-00178.pdf>
- Weegar, R., & Idestam-Almquist, P. (2024). Reducing workload in short answer grading using machine learning. Springer. <https://doi.org/10.1007/s40593-022-00322-1>
- Wilianto, D., & Girsang, A. (2023). Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods. TEM Journal. <https://www.cceol.com/search/article-detail?id=1103595>
- Yang, S. (2023). Automated English Essay Scoring Based on Machine Learning Algorithms. 2023 2nd International Conference on Data Analytics .... <https://ieeexplore.ieee.org/abstract/document/10361235/>